Sociedad de Estadística e Investigación Operativa
Test (2000) *Vol.* **9**, *No.* 1, pp. 1–96

Contributions of empirical and quantile processes to the asymptotic theory of goodness-of-fit tests

Eustasio del Barrio
Departamento de Estadística e Investigación Operativa
Universidad de Valladolid, Spain

Juan A. Cuesta-Albertos
Departamento de Matemáticas, Estadística y Computación
Universidad de Cantabria, Spain

Carlos Matrán[*]
Departamento de Estadística e Investigación Operativa
Universidad de Valladolid, Spain

Abstract

This paper analyzes the evolution of the asymptotic theory of goodness-of-fit tests. We emphasize the parallel development of this theory and the theory of empirical and quantile processes. Our study includes the analysis of the main tests of fit based on the empirical distribution function, that is, tests of the Cramér-von Mises or Kolmogorov-Smirnov type. We pay special attention to the problem of testing fit to a location scale family. We provide a new approach, based on the Wasserstein distance, to correlation and regression tests, outlining some of their properties and explaining their limitations.

Key Words: Goodness-of-fit, correlation tests, Cramér-von Mises, empirical and quantile processes, Kolmogorov-Smirnov, Shapiro-Wilk, strong approximations, Wasserstein distance.
AMS subject classification: Primary 62F05, 62E20; secondary 60F25.

1 Introduction

This year has been declared by UNESCO as World Year of Mathematics. This year of mathematical celebrations also commemorates the centenary of some landmarks in the history of Mathematics. In the 2nd International Congress of Mathematics, held in Paris in 1900, David Hilbert postulated

[*] Correspondence to: Carlos Matrán Bea, Departamento de Estadística, Facultad de Ciencias, Universidad de Valladolid, 47005 Valladolid, Spain.

Dedicated to Miguel Martín Díaz, whose scientific criticism has greatly inspirated our research by years.

Research partially supported by DGICYT, grants PB98-0369-C02-01 and 02. E. del Barrio and C. Matrán have also been supported by PAPIJCL grant VA08/97.

his celebrated 23 problems as the main challenges to which the mathematical community should pay attention, without any reference to Probability or Statistics. The development of statistical methods became, though, a major source of motivation for the mathematical research in this century.

It was also in 1900 that Karl Pearson proposed the first test of goodness-of-fit: the χ^2 test. The subsequent research devoted to enhancements of this elementary goodness-of-fit procedure became a major source of motivation for the development of key areas in Probability and Statistics, such as the theory of weak convergence in general spaces and the asymptotic theory of empirical processes. Commemorating this centennial we wish to analyze, with absolute subjectivity, some aspects which arise from the development of the asymptotic theory of goodness-of-fit tests through this century.

We will pay special attention to stressing the parallel evolution of the theory of empirical processes and the asymptotic theory of goodness-of-fit tests. Doubtless, this evolution is a good indicator of the vast transformation that Probability and Statistics experienced during this century. Certainly, the names that contributed to the theory are the main guarantee for this assertion: Pearson, Fisher, Cramér, von Mises, Kolmogorov, Smirnov, Feller, et al. laid the foundations of the theory. In some cases, the mathematical derivation of the asymptotic distribution of goodness-of-fit tests in that period had the added merit that, in a certain sense, the limit law was blindly pursued. In Mathematics the main difficulty in showing convergence consists of obtaining a convincing candidate for the limit. Thus, proofs in that period could be considered as major pieces of precision and inventiveness.

A systematic method of handling adequate candidates for the limit law begins in 1950 with the heuristic work by Doob (1949), revised by Donsker through the Invariance Principle. The subsequent construction of adequate metric spaces and the development of the corresponding weak convergence theory as the right probabilistic setup for the study of asymptotic distributions had a wide and rapid diffusion, with notable advances due to Prohorov and Skorohod among others. The contribution of Billingsley's book (Billingsley 1968) to this diffusion must also be pointed out.

The study of Probability in Banach spaces has been another source of useful results for the goodness-of-fit theory. The names of Varadhan, Dudley, Araujo, Giné, Zinn, Ledoux, Talagrand, et al. are necessary references to anyone interested in asymptotics in Statistics. For example, the Central Limit Theorem (C.L.T.) in Hilbert spaces played a main role in the obtaining of the asymptotic behaviour of Cramér-von Mises-type statistics.

Lastly, we must indicate the significance of the "Hungarian school", developing the strong approximation techniques initiated by Skorohod with his "embedding". Breiman's book (Breiman 1968) had the merit of initially spreading Skorohod's embedding. Today, the strong approximations due to Komlós, Major, Tusnády, M. and S. Csörgő, Révész, Deheuvels, Horváth, Mason, et al. are an invaluable tool in the study of asymptotic in Statistics, as we will point out in the last section of this paper.

This paper is organized in two sections concerning, respectively, tests of fit to a fixed distribution, Section 2, and tests of fit to a parametric family of distributions, Section 3. A major goal in our approach consists of providing an adequate setup for the analysis of regression and correlation tests of fit, associated with the well-known probability plots. Subsection 3.2 is devoted to these tests. The analysis of correlation tests in the setup provided by the Wasserstein distance, Subsection 3.3, will give additional justification to the good behaviour of the most popular representatives of this class of tests, and will also explain their poor performance in testing fit to a family of heavy-tailed distributions. The asymptotic for tests of fit based on Wasserstein distance will be obtained through the use of strong approximations. Of course, we do not pretend to cover the wide range of existing tests of fit.

The notation to be employed in this paper is as follows. All the random variables will be defined on the same, rich enough, probability space (Ω, σ, P). Given $X_1, \ldots, X_n$ independent and identically distributed (i.i.d.) real valued random variables, $\overline{X}_n$ and S_n^2 will, respectively, denote their sample mean and variance and F_n will denote the associated empirical distribution function, i.e., $F_n(x) = n^{-1} \sum_{1 \leq i \leq n} I_{\{X_i \leq x\}}, x \in \mathbb{R}$. $\{U_n\}$ will represent a sequence of i.i.d. random variables uniformly distributed on the interval $(0,1)$ and G_n will denote its associated empirical distribution function. As usual, the uniform empirical process is defined by

$$\alpha_n(t) := \sqrt{n}\,(G_n(t) - t)\,,\ t \in (0,1).$$

The so-called Brownian bridge, $\{B(t) : 0 < t < 1\}$, is a Gaussian process with continuous trajectories and covariance operator $\operatorname{Cov}(B(s), B(t)) = s \wedge t - st$.

Two metric spaces will appear very frequently. The space $\mathcal{C}[0,1]$ which consists of all real, continuous functions on the interval $[0,1]$, endowed with the supremum norm $\|x\|_\infty := \sup_{0<t<1} |x(t)|$, $x \in \mathcal{C}[0,1]$; and the space $\mathcal{D}[0,1]$ (respectively $\mathcal{D}[-\infty,\infty]$) of all real functions on $[0,1]$ (resp. on $[-\infty,\infty]$) which are right-continuous and have left limits, càdlàg (from the French "continue à droit avec limits à gauche"), endowed with the Skorohod distance (see Skorohod 1956).

Convergence in distribution and in probability will be respectively denoted by $\xrightarrow{w}$ and $\xrightarrow{P}$. Given a random variable X its probability distribution will be represented by $\mathcal{L}(X)$. Finally, Φ and ϕ will, respectively, denote the distribution function and density function of a standard normal random variable.

2 Testing fit to a fixed distribution

The simplest goodness-of-fit problem consists of testing fit to a single fixed distribution, namely, given a random sample of real random variables $X_1, X_2, \ldots, X_n$ with common distribution function F, testing the null hypothesis $F = F_0$ for a fixed distribution function F_0. While this procedure is usually of limited interest in applications, the solutions proposed for this problem provided the main idea in subsequent generalizations designed for testing fit to composite null hypotheses.

Pearson's chi-squared test can be considered as the first approach to the problem of testing fit to a fixed distribution. The solution proposed by Pearson consisted of dividing the real line into k disjoint categories or "cells" $C_1, \ldots, C_k$ into which data would fall, under the null hypothesis, with probabilities $p_1, \ldots, p_k$. That is, if $F = F_0$ were true, then $P(X_1 \in C_i) = p_i$, $i = 1, \ldots, k$. If O_i is the number of observations in cell i, then O_i has a binomial distribution with parameters n and p_i; hence, the de Moivre-Laplace C.L.T. states that $(np_i(1-p_i))^{-1/2}(O_i - np_i) \xrightarrow{w} N(0,1)$.

The multivariate C.L.T. shows that, if $l \leq k$, then $B_l = n^{-1/2}(O_1 - np_1, \ldots, O_l - np_l)^T$ has a limit distribution which is centered Gaussian and has covariance matrix Σ_l whose (i,j) element, $\sigma_{i,j}$, satisfies $\sigma_{i,j} = -p_i p_j$, for $i \neq j$, and $\sigma_{i,i} = p_i(1-p_i)$. On the other hand, if $p_i > 0, i = 1, \ldots, k$, Σ_{k-1} is nondegenerate and Σ_{k-1}^{-1} has element (i,j), $\nu_{i,j}$, satisfying $\nu_{i,j} = p_k^{-1}$, for $i \neq j$, and $\nu_{i,i} = p_i^{-1} + p_k^{-1}$. Simple matrix algebra shows that $B_{k-1}^T \Sigma_{k-1}^{-1} B_{k-1}$ converges in law to a χ_{k-1}^2 distribution. Further, straightforward computations show that

$$\chi^2 := \sum_{j=1}^{k} \frac{(O_j - np_j)^2}{np_j} = B_{k-1}^T \Sigma_{k-1}^{-1} B_{k-1}$$

providing, therefore, a well-known result in the asymptotic theory of tests of fit:

Theorem 2.1. *Under H_0, χ^2 has asymptotic distribution χ^2_{k-1}.*

Theorem 2.1 reduces the problem of testing fit to a fixed distribution to analysing a multinomial distribution, thus providing a widely applicable and easy-to-use method for testing fit which immediately carries over to the multivariate setup. Moreover, this test also allows some freedom in choosing the number, the location or the size of the cells $C_1, \ldots, C_k$. This point will be discussed in the next section.

However, as pointed out by many authors (see, e.g., Moore 1986), considering only the cell frequencies when F is continuous produces a loss of information that results in lack of power (the χ^2 statistic will not distinguish two different distributions sharing the same cell probabilities). Therefore, in order to improve our method for testing fit, we should try to make use of the complete information provided by the data. However, the multivariate C.L.T. and elementary matrix algebra were the only tools needed in the derivation of the asymptotic distribution in Theorem 2.1. This will not be the case when handling more complicated statistics.

One way to improve Pearson's statistic consists of employing a functional distance to measure the discrepancy between the hypothesized distribution function F_0 and the empirical distribution function F_n. The first representatives of this method were proposed in the late 20's and in the 30's. Cramér (1928) and, in a more general form, von Mises (1931) proposed

$$\omega_n^2 = n \int_{-\infty}^{\infty} (F_n(x) - F_0(x))^2 \rho(x)\, dx,$$

for some suitable weight function ρ as an adequate measure of discrepancy. Kolmogorov (1933) studied

$$D_n = \sqrt{n} \sup_{-\infty < x < \infty} |F_n(x) - F_0(x)|$$

and Smirnov (1939, 1941) the closely related statistics

$$\begin{aligned} D_n^+ &= \sqrt{n} \sup_{-\infty<x<\infty} \left(F_n(x) - F_0(x)\right), \\ D_n^- &= \sqrt{n} \sup_{-\infty<x<\infty} \left(F_0(x) - F_n(x)\right), \end{aligned}$$

which are more adequate for tests against one-sided alternatives. The statistics D_n, D_n^+ or D_n^- are known as Kolmogorov-Smirnov statistics and present the advantage of being distribution-free: for any continuous distribution function F_0, D_n has, under the null hypothesis, the same distribution as $\sup_{0<t<1} |\alpha_n(t)|$. Similar statements hold for D_n^+ and D_n^-. Thus, the same p-values can be used to obtain the significance level when testing fit to any continuous distribution. This desirable property is not satisfied by ω_n^2, but it also holds for the following modification:

$$W_n^2(\Psi) = n \int_{-\infty}^{\infty} \Psi(F_0(x))(F_n(x) - F_0(x))^2 \, dF_0(x),$$

which was proposed by Smirnov (1936, 1937). All the statistics which can be obtained by varying Ψ are usually referred to as statistics of Cramér-von Mises type. Consideration of different weight functions Ψ allows the statistician to put special emphasis on the detection of particular sets of alternatives. For this reason, some weighted versions of Kolmogorov's statistics have also been proposed, namely,

$$K_n(\Psi) = \sqrt{n} \sup_{-\infty<x<\infty} \frac{|F_n(x) - F_0(x)|}{\Psi(F_0(x))}.$$

The convenience of employing $W_n^2(\Psi)$ instead of D_n^2 as a test statistic can be understood when taking into account that D_n^2 accounts only for the largest deviation between $F_n(t)$ and $F(t)$, while $W_n^2(\Psi)$ is a weighted average of all the deviations between $F_n(t)$ and $F(t)$. Thus, as observed by Stephens (1986a), $W_n^2(\Psi)$ should have more chance of detecting alternatives that are not very far from F at any point t, but are moderately far from F for a large range of points t (think of location alternatives). These heuristic considerations are confirmed by simulation studies (see, for reference, Stephens 1986a).

Two particular statistics have received special attention in the literature. When $\Psi = 1$,

$$W_n^2 = n \int_{-\infty}^{\infty} (F_n(x) - F_0(x))^2 \, dF_0(x)$$

is called the Cramér-von Mises statistic; when $\Psi(t) = (t(1-t))^{-1}$ then

$$A_n^2 = n \int_{-\infty}^{\infty} \frac{(F_n(x) - F_0(x))^2}{F_0(t)(1 - F_0(t))} \, dF_0(x)$$

is called the Anderson-Darling statistic. A_n^2 has the additional appeal of weighting the deviations according to their expected value, and this results in a more powerful statistic for testing fit to a fixed distribution, see Stephens (1986a).

To be able to use any of these appealing statistics in practice we should be able to obtain the corresponding significance levels. Smirnov (1941), using combinatorial techniques, obtained an explicit expression for the exact distribution of D_n^+. Kolmogorov (1933) also gave an expression that enabled the tabulation of the distribution of D_n. Further difficulties were found when dealing with the exact distributions of statistics of Cramér-von Mises type. But even in those cases where a formula allowed the computation of the exact p-values, the interest in obtaining the asymptotic distribution of the test statistics was clear, for it would greatly decrease the computational effort needed to obtain the (approximate) p-values (and this was of crucial importance by the time these tests were proposed). The celebrated first asymptotic results regarding D_n and D_n^+ are summarized in the following theorem:

Theorem 2.2. *For every $x > 0$:*

i) (Kolmogorov 1933)

$$\lim_n P\left(D_n \leq x\right) = \sum_{j=-\infty}^{\infty} (-1)^j e^{-2j^2x^2}$$

ii) (Smirnov 1941)

$$\lim_n P\left(D_n^+ > x\right) = e^{-2x^2}.$$

Kolmogorov's proof of *i)* was based on the consideration of a limiting diffusion equation. Smirnov used the exact expression of $P\left(D_n^+ > x\right)$ to show *ii)*. Also, Smirnov (1936) derived the asymptotic distribution of the Cramér-von Mises statistic, W_n^2.

Feller (1948) claimed that Kolmogorov's and Smirnov's proofs were "very intricate" and were "based on completely different methods" and presented his paper as an attempt to give "unified proof" of those theorems (which could provide a systematic method of deriving the asymptotic distribution of other test statistics expressible as a functional of the empirical distribution function) It seemed unnatural that, since D_n, D_n^+ and W_n^2 are measures of the discrepancy between F_n and F_0 based on the same object, namely, the empirical process, a particular technique had to be used in the derivation of the asymptotic distribution of each statistic. Thus, Feller's paper is a remarkable step in the development of a unified asymptotic theory for tests of fit based on the empirical process. Nevertheless, a study of the empirical process itself and of its asymptotic distribution (a concept which would have to be made precise) was not considered and, as claimed in Doob (1949), all these proofs (including Feller's) "conceal to some extent ... the naturalness of the results (qualitatively at least) and their mutual relations".

It was Doob (1949) who, considering the finite dimensional distributions, conjectured the convergence of the uniform empirical process to the Brownian bridge. A useful consequence of this fact would be that, under some (non explicit) hypotheses, the derivation of the asymptotic distribution of a functional of the uniform empirical process could be reduced to the derivation of the distribution of the same functional of the Brownian bridge. Doob proved that

$$P\left(\left(\sup_{0\leq t\leq 1}|B(t)|\right)\leq x\right)=\sum_{j=-\infty}^{\infty}(-1)^j e^{-2j^2x^2} \tag{2.1}$$

and

$$P\left(\left(\sup_{0\leq t\leq 1}B(t)\right)>x\right)=e^{-2x^2}.$$

Thus, justification of Doob's conjecture would provide a new, simpler proof of the results of Kolmogorov and Smirnov.

This justification was given by Donsker through his invariance principle in Donsker (1951, 1952). His results showed that the distribution of a continuous functional of the partial sum process (obtained from a sequence of i.i.d. random variables with finite second moment) converges to the distribution of the corresponding functional of a Brownian motion, and that the distribution of a continuous functional of the uniform empirical process converges to the distribution of the corresponding functional of a Brownian bridge.

The development of the theory of weak convergence in metric spaces by, among others, Kolmogorov, Prohorov and Skorohod in the fifties (see Prohorov 1953; Kolmogorov and Prohorov 1949; Prohorov 1956; and Skorohod 1956) allowed a better understanding of this invariance principle, as presented in Billingsley (1968). The space $\mathcal{C}[0,1]$ was one of the first metric spaces for which this theory was developed, through the work of Prohorov (1956). The scheme consisting of proving the convergence of the finite dimensional distributions plus a tightness condition allowed the obtaining of distributional limit theorems for slight modifications of the partial sum and the uniform empirical processes, because both processes could be approximated by "equivalent" processes obtained from them by linear interpolation, so that all the random objects considered in the limit theorems remained in $\mathcal{C}[0,1]$.

This last approximation is somewhat artificial. In order to avoid it, a wider space had to be considered. A proper study of the weak convergence of the uniform empirical process could be attempted in the space $\mathcal{D}[0,1]$. The fact that the empirical process is not measurable when the uniform norm is considered led to the introduction of a more involved topology, namely the Skorohod topology that turned $\mathcal{D}[0,1]$ into a separable and complete metric space in which the empirical process was measurable. In this setup the weak convergence of the empirical process could be properly stated (see, e.g., Billingsley 1968, pp. 141)

Theorem 2.3. *If we consider α_n and B as random elements taking values in $\mathcal{D}[0,1]$, then*

$$\alpha_n \xrightarrow{w} B.$$

Theorem 2.3 enables us to rederive Theorem 2.2 in a very natural way. Note that $D_n = ||\alpha_n||_\infty$ and that the map $x \mapsto ||x||_\infty$ is continuous for the Skorohod topology outside a set of B-measure zero. Thus, we can conclude that $D_n \xrightarrow{w} ||B||_\infty$ and this, combined with (2.1), gives a proof of the first statement in Theorem 2.2. The same method works for D_n^+.

The use of the Skorohod space is not the only means of circumventing the difficulty posed by the nonmeasurability of the empirical process. A different approach to the problem could be based on the following scheme. If we could define, on a rich enough probability space, a sequence of i.i.d. random variables uniformly distributed on $(0,1)$ with an associated empirical process $\alpha_n^*(t)$ and a Brownian bridge $B(t)$ such that

$$\sup_{0\leq t\leq 1} |\alpha_n^*(t) - B(t)| \xrightarrow{P} 0, \tag{2.2}$$

then we would easily obtain that, for any functional H defined on $\mathcal{D}[0,1]$ and continuous on $\mathcal{C}[0,1]$, $H(\alpha_n^*) \xrightarrow{P} H(B)$, obtaining a new proof of Theorem 2.2. The study of results of type (2.2), generically known as strong approximations, began with the Skorohod embedding, consisting of imitating the partial sum process by using a Brownian motion evaluated at random times (see Breiman 1968). Successive refinements of this idea became one of the most important methodologies in the research related to empirical processes.

Returning to the applications of Theorem 2.3 in the asymptotic theory of tests of fit, we should note that the functional $x \mapsto \int_0^1 x(t)^2 dt$ is also continuous for the Skorohod topology outside a set of B-measure zero. We can use this fact to obtain the asymptotic distribution of the Cramér-von Mises statistic. Namely,

$$W_n^2 \xrightarrow{w} \int_0^1 B(t)^2 dt.$$

Then, a Karhunen-Loève expansion of $B(t)$ allows us to easily compute the characteristic function of $\int_0^1 B(t)^2 dt$ and the inversion of this characteristic function allows us to tabulate the asymptotic distribution of W_n^2 (see, e.g., Shorack and Wellner 1986, pp. 215 for details). This methodology makes the arguments used by Smirnov to derive the asymptotic distribution of W_n^2 unnecessary. A recent full account of all the presently available information concerning the exact and limiting distributions of W_n^2, as linked by an asymptotic expansion, is given by Csörgő and Faraway (1996), with a comparable theory for Watson's circularly invariant version referred to at Theorem 3.4 below, where many errors that have accumulated on this topic are also corrected.

A little extra effort allows us to extend this method for deriving the asymptotic distribution of other statistics of Cramér-von Mises type. As a consequence of the Law of the Iterated Logarithm for the Brownian motion, Anderson and Darling (1952) showed that, provided

$$\int_0^\delta \Psi(t)t \log\log \frac{1}{t} dt \qquad \text{and} \qquad \int_\delta^1 \Psi(t)(1-t)\log\log\frac{1}{1-t}dt$$

are finite for some $\delta \in (0,1)$, the functional $x \mapsto \int_0^1 \Psi(t)x(t)^2 dt$ is continuous with respect to the Skorohod distance, outside a set of B-measure zero and, consequently, $W_n^2(\Psi) \xrightarrow{w} \int_0^1 \Psi(t)B(t)^2 dt$. This result covers the Anderson-Darling statistic A_n^2.

Although all the limit theorems for goodness-of-fit tests that we have described so far are based on the weak convergence of the empirical process considered as a random element taking values in the space of càdlàg functions, with the Skorohod topology plus the continuity of a suitable functional, there is a more natural way to study the asymptotic properties of statistics of Cramér-von Mises type and, more generally, of integral functionals of the empirical process.

The uniform empirical process can be viewed as a random element taking values in the separable Hilbert space $L_2((0,1),\Psi)$ of all real, Borel measurable functions f on $(0,1)$, such that $\int_0^1 \Psi(t)f(t)^2dt$ is finite, where we consider the norm given by

$$||f||_{2,\Psi}^2 = \int_0^1 \Psi(t)f(t)^2 dt.$$

In this setup $W_n^2(\Psi) = ||\alpha_n||_{2,\Psi}^2$. The theory of probability in Banach spaces, developed in the 60's and 70's, turned the problem of studying the asymptotic distribution of $W_n^2(\Psi)$ into an easier task, because the C.L.T. for random elements taking values in $L_2((0,1),\Psi)$ (see, e.g., Araujo and Giné 1980, pp. 205, ex. 14) asserts that a sequence $\{Y_n(t)\}_n$ of i.i.d. $L_2(0,1)$-valued random elements satisfies

$$\frac{1}{\sqrt{n}}\sum_{i=1}^n Y_i(t) \xrightarrow{w} Y(t),$$

if and only if $\int_0^1 E(Y_1(t))^2\Psi(t)dt < \infty$ and, in that case, Y is a Gaussian random element with the same covariance function as Y_1.

Therefore, if we set $Y_i(t) = I_{\{U_i \leq t\}} - t,\ i = 1,\ldots,n$, then $\alpha_n(t) = n^{-1/2}\sum_{i=1}^n Y_i(t)$ and $Y_1(t)$ has the same covariance function as the Brownian bridge $B(t)$. Hence, $\alpha_n \xrightarrow{w} B$ in $L_2((0,1),\Psi)$ if and only if $\int_0^1 t(1-t)\Psi(t)dt < \infty$.

A further application of Hoffmann-Jorgensen's inequality (see del Barrio 2000) allows us to conclude that $||\alpha_n||^2_{2,\Psi}$ has a limiting distribution if and only if $\int_0^1 t(1-t)\Psi(t)dt < \infty$, proving the following result.

Theorem 2.4. (Asymptotic distribution of statistics of the Cramér-von Mises type). *$W_n^2(\Psi)$ has a limiting distribution if and only if $\int_0^1 t(1-t)\Psi(t)dt < \infty$. In that case*

$$W_n^2(\Psi) \xrightarrow{w} \int_0^1 \Psi(t)B(t)^2 dt.$$

While the development of probability in Banach spaces provides this final result for statistics of the Cramér-von Mises type, the use of strong approximations produces a similar result for supremum norm statistics. Chibisov (1964) and O'Reilly (1974) used the Skorohod embedding and a special representation of the uniform empirical process in terms of a Poisson process (see, e.g. Shorack and Wellner 1986, pp. 339) to obtain necessary and sufficient conditions for the weak convergence of the empirical process to the Brownian bridge in weighted uniform metrics. If Ψ is a positive function on $(0,1)$ nondecreasing in a neighborhood of 0 and nonincreasing in a neighborhood of 1 and we consider the norm given by $||x||_\Psi = \sup_{0<t<1}(|x(t)|/\Psi(t))$ on $\mathcal{D}[0,1]$, then $\alpha_n \xrightarrow{w} B$ in $||\cdot||_\Psi$ norm (with the necessary modifications in the definition of weak convergence to avoid measurability problems), if and only if

$$\int_0^1 \frac{1}{t(1-t)} \exp\left(-\epsilon \frac{\Psi(t)^2}{t(1-t)}\right) dt < \infty, \tag{2.3}$$

for every $\epsilon > 0$. An immediate corollary of the Chibisov-O'Reilly theorem is that (2.3) is a sufficient condition for ensuring the convergence

$$K_n(\Psi) \xrightarrow{w} \sup_{0<t<1} \frac{|B(t)|}{\Psi(t)}.$$

A modification of the so-called Hungarian construction due to Komlós, Major and Tusnády (1975, 1976) and to Csörgő and Révész (1978) was used in Csörgő et al. (1986) to give the following final result for statistics of the Kolmogorov-Smirnov type.

Theorem 2.5. (Asymptotic distribution of statistics of the Kolmogorov-Smirnov type). *If Ψ is a positive function on $(0,1)$, nondecreasing in a* *neighborhood of* 0 *and nonincreasing in a neighborhood of* 1*, then* $K_n(\Psi)$ *converges in distribution to a nondegenerate limit law if and only if*

$$\int_0^1 \frac{1}{t(1-t)} \exp\left(-\epsilon \frac{\Psi(t)^2}{t(1-t)}\right) dt < \infty,$$

for some $\epsilon > 0$*. In that case,*

$$K_n(\Psi) \xrightarrow{w} \sup_{0<t<1} \frac{|B(t)|}{\Psi(t)}.$$

3 Testing fit to a family of distributions

We consider in this section the problem of testing whether the underlying distribution function of the sample, F, belongs to a given family of distribution functions, $\mathcal{F}$. We will assume $\mathcal{F}$ is a parametric family, i.e.,

$$\mathcal{F} = \{F(\cdot, \theta) : \theta \in \Theta\},$$

where Θ is some open set in $\mathbb{R}^d$. $F^{-1}(\cdot, \theta)$ is the quantile function associated with $F(\cdot, \theta)$.

Perhaps the most interesting case occurs when $\mathcal{F}$ is the Gaussian family. It seems that the first statistics for detecting possible departures from normality were introduced in Fisher (1930), Pearson (1930) and Williams (1935), and were based on the study of the standardized third and fourth moments, usually denoted by $\sqrt{b_1}$ and b_2, respectively.

To strengthen these procedures, some composite tests, intended to take into account both features simultaneously, were proposed. For instance, in Pearson, D'Agostino and Bowman (1977) the K^2 and the R tests, consisting of handling two suitable functions of the $\sqrt{b_1}$ and b_2 statistics, namely, $K^2 = K(\sqrt{b_1}, b_2)$ and $R = R(\sqrt{b_1}, b_2)$, were introduced. In that paper a Monte Carlo study comparing those tests to the most popular normality tests was accomplished. The authors select many alternative distributions and the power of both tests seems to be similar to that of the competing ones.

However, tests based on kurtosis and skewness are not too reliable because they are based on properties which do not characterize Gaussian distributions. For instance, Ali (1974) exhibits a sequence of distributions

$\{P_k\}$ which converges to the standard Gaussian distribution while the kurtosis of P_k goes to infinity. Thus, if we consider a random sample obtained from P_k, the bigger the index k, the greater chance to reject the normality of the sample. On the other hand, some examples of symmetric distributions, with shapes very far from normality (some of them even multimodal), and $\beta_2 = 3$ are known (see, for instance, Balanda and McGillivray 1988; Kale and Sebastian 1996). As a consequence, none of the $\sqrt{b}_1$, b_2, K^2 or R tests detects the non-normality of the parent distribution in all cases.

Other tests of normality are the u-test (see David, Hartley and Pearson 1954), based on the ratio between the range and the standard deviation in the sample, and the a-test (see Geary 1947), which studies the ratio of the sample mean to the standard deviation. These tests are broadly considered as not being too powerful against a wide range of alternatives (although it is known that the u-test has good power against alternatives with light tails; see Shapiro, Wilk and Chen 1968; in fact, according to Uthoff 1970, 1973, the u-test is the most powerful against the uniform distribution while the a-test is the most powerful against the double exponential distribution).

For these reasons, other tests, focusing on features that characterize completely (or, at least, more completely) the family under consideration, have been proposed. These tests can be divided, broadly speaking, into three categories. The first, more general category consists of tests that adapt other tests devised in the fixed-distribution setup. When we specialize on location scale families, new types of tests that try to take advantage of the particular structure of $\mathcal{F}$, can be employed. Tests based on the analysis of probability plots, usually referred to as correlation and regression tests, lie in this class. A third category, whose representatives combine some of the most interesting features exhibited by goodness-of-fit tests lying in the first two categories, is composed of tests based on a suitable L_2-distance between the empirical quantile function and the quantile functions of the distributions in $\mathcal{F}$, the so-called *Wasserstein distance.*

Tests based on Wasserstein distance are related to tests in the first category in the sense that all of them depend on functional distances. On the other hand, it happens that the study of Wasserstein-tests gives some hints about several properties of the probability plot-tests. These two facts have led us to present them separately. Our approach will try to show that tests based on Wasserstein distance provide the right setup to apply the empirical and quantile process theory to study probability plot-based tests.

3.1 Adaptation of tests coming from the fixed-distribution setup

All the procedures considered in Section 2 were based on measuring the distance between a distribution obtained from the sample and a fixed distribution. One way to adapt this idea for the new setup consists of choosing some adequate estimator $\hat{\theta}$ of θ (assuming the null hypothesis is true) and replacing the fixed distribution by $F(\cdot, \hat{\theta})$. This simple idea was suggested by Pearson for his χ^2-test. That is, Pearson suggested using the statistic

$$\hat{\chi}^2 = \sum_{j=1}^{k} \frac{(O_j - np_j(\hat{\theta}))^2}{np_j(\hat{\theta})},$$

where $p_j(\theta)$ denotes the probability, under $F(\cdot, \theta)$, that X_1 falls into cell j.

Pearson, however, did not realize the change in the asymptotic distribution of $\hat{\chi}^2$ due to the estimation of parameters. It was Fisher, in the 20's, who pointed out that the limiting distribution of $\hat{\chi}^2$ depends on the method of estimation and showed that, under regularity conditions, if $\hat{\theta}$ is the maximum likelihood estimator of θ from the grouped data $(O_1, \ldots, O_k)$, then $\hat{\chi}^2$ has asymptotic χ^2_{k-d-1} distribution (see, e.g., Cochran 1952, for a detailed review of Pearson's and Fisher's contributions).

Fisher also observed that estimating θ from the grouped data instead of using the complete sample (e.g., by estimating θ from the complete likelihood) could produce a loss of information resulting in a lack of power. Further, estimating θ from the original data is often computationally simpler. Fisher studied the asymptotic distribution of $\hat{\chi}^2$ when θ is unidimensional and $\hat{\theta}$ is its maximum likelihood estimator from the ungrouped data. His result was extended by Chernoff and Lehmann (1954) for a general d-dimensional parameter showing that, under regularity conditions (essentially conditions to ensure the consistency and asymptotic normality of the maximum likelihood estimator),

$$\hat{\chi}^2 \xrightarrow{w} \sum_{j=1}^{k-d-1} Z_j^2 + \sum_{j=k-d}^{k-1} \lambda_j Z_j^2, \tag{3.1}$$

where Z_j are i.i.d. standard normal random variables and $\lambda_j \in [0,1]$ may depend on the parameter θ. This dependence shows a serious drawback to the use of $\hat{\chi}^2$ for testing fit to some families of distributions, the normal family being one of them (see Chernoff and Lehmann 1954).

The practical use of $\hat{\chi}^2$ for testing fit presented another difficulty: the choice of cells. The asymptotic χ^2_{k-1} distribution of Pearson's statistic was a consequence of the asymptotic normality of the cell frequencies. A cell with a very low expected frequency would cause a very slow convergence to normality and this could result in a poor approximation of the distribution of χ^2. This (somewhat oversimplifying) observation led to the diffusion of rules of thumb such as "use cells with at least 10 observations". Hence, combining neighboring cells with few observations became common practice (see, e.g., Cochran 1952).

From a more theoretical point of view, in the setup of testing fit to a fixed distribution, Mann and Wald (1942) and Gumbel (1943) suggested using equally likely intervals under the null hypothesis as a reasonable way to reduce the arbitrariness in the choice of cells (this choice offers some good properties; for instance, it makes the χ^2 test unbiased, see, e.g., Cohen and Sackrowitz 1975). Trying to adapt this idea to the case of testing fit to parametric families poses the problem that different distributions in the null hypothesis lead to different partitions into equiprobable cells. A natural solution involves choosing, for cells, equally likely intervals under $F(\cdot, \hat{\theta})$, where $\hat{\theta}$ is some suitable estimator of θ. A consequence of this procedure is that, again, the cells are chosen at random.

Allowing the cells to be chosen at random introduces a deep modification to the statistical structure of χ^2 because the distribution of the random vector $(O_1, \ldots, O_k)$ is no longer multinomial; remarkably, however, it can, in some important cases, eliminate the dependence on the parameter θ of the asymptotic distribution in (3.1). Watson (1957, 1958) noted that if $\hat{\theta}$ is the maximum likelihood estimator of θ (from the ungrouped data) and cell j has boundaries $F^{-1}((j-1)/k, \hat{\theta})$ and $F^{-1}(j/k, \hat{\theta})$, then (3.1) remains true. Further, if $\mathcal{F}$ is a location scale family, then the λ_j's do not depend on θ, but only on the family $\mathcal{F}$. As a consequence, an improved χ^2 method could be used for testing normality or exponentiality.

The development of the theory of weak convergence in metric spaces provided valuable tools for further insights in χ^2-testing. Using the weak convergence of the empirical process in $\mathcal{D}[0,1]$, Moore (1971) obtained a short rigorous proof of Watson's result which was also valid for multivariate observations and random rectangular cells. Later, Pollard (1979), using a general C.L.T. for empirical measures due to Dudley (1978), extended the result to very general random cells under the mild assumption that these random cells were chosen from a Donsker class.

Despite the fact that all these theoretical contributions have widely spread the applicability and reliability of χ^2-tests, the limitations of this procedure, noted when testing fit to a fixed distribution, carry over to the case of testing fit to a family (see, e.g., Stephens 1974, or 1986a).

The use of supremum or quadratic statistics based on the empirical distribution function with parameters estimated from the data could provide more powerful tests, just as in the fixed distribution setup. The adaptation of W_n^2 or K_n to this situation can be easily carried out. Let $\hat{\theta}_n$ be some estimator of θ. We can define the statistics

$$\hat{W}_n^2(\Psi) = n \int_{-\infty}^{\infty} \Psi(F(x,\hat{\theta}_n))(F_n(x) - F(x,\hat{\theta}_n))^2\, dF(x,\hat{\theta}_n),$$

and

$$\hat{K}_n(\Psi) = \sqrt{n} \sup_{-\infty<x<\infty} \frac{|F_n(x) - F(x,\hat{\theta}_n)|}{\Psi(F(x,\hat{\theta}_n))},$$

and use them as statistical tests, rejecting the null hypothesis when large values of $\hat{W}_n^2(\Psi)$ or $\hat{K}_n(\Psi)$ are observed. However, it took a long time for these statistics to be considered as serious competitors to the χ^2-test; little was known about these versions of Cramér-von Mises or Kolmogorov-Smirnov tests until the 50's (see, e.g., Cochran 1952).

The property exhibited by W_n^2 and K_n of being distribution-free does not carry over to $\hat{W}_n^2(\Psi)$ or $\hat{K}_n(\Psi)$. If we set $Z_i = F(X_i,\hat{\theta}_n)$ and let $\hat{G}_n(t)$ denote the empirical distribution function associated with $Z_1,\ldots,Z_n$ then, obviously,

$$\hat{W}_n^2(\Psi) = n \int_0^1 \Psi(t)(\hat{G}_n(t) - t)^2 dt, \tag{3.2}$$

$$\hat{K}_n(\Psi) = \sqrt{n} \sup_{0<t<1} \frac{|\hat{G}_n(t) - t|}{\Psi(t)}, \tag{3.3}$$

but, unlike in the fixed distribution case, $Z_1,\ldots,Z_n$ are not i.i.d. uniform random variables.

However, in some important cases the distribution of $Z_1,\ldots,Z_n$ does not depend on θ, but only on $\mathcal{F}$. In these cases, the distribution of $\hat{W}_n^2(\Psi)$ or $\hat{K}_n(\Psi)$ is parameter-free. This happens if $\mathcal{F}$ is a location scale family and

$\hat{\theta}_n$ is an equivariant estimator, a fact noted by David and Johnson (1948). Therefore $\hat{W}_n^2(\Psi)$ or $\hat{K}_n(\Psi)$ can be used in a straightforward manner as test statistics in this situation. Lilliefors (1967) took advantage of this property and, from a simulation study, constructed his popular table for using the Kolmogorov-Smirnov statistic when testing normality.

The first attempt to derive the asymptotic distribution of any statistic of $\hat{W}_n^2(\Psi)$ or $\hat{K}_n(\Psi)$ type was due to Darling (1955). His study concerned the Cramér-von Mises statistic

$$\hat{W}_n^2 = n\int_{-\infty}^{\infty}(F_n(x)-F(x,\hat{\theta}_n))^2\,dF(x,\hat{\theta}_n) = n\int_0^1(\hat{G}_n(t)-t)^2dt,$$

assuming that θ was one-dimensional. Let us define

$$\begin{aligned} H_n &:= n\int_{-\infty}^{\infty}\left(F_n(x)-F(x,\theta)-(\hat{\theta}_n-\theta)\frac{\partial}{\partial\theta}F(x,\theta)\right)^2 dF(x,\theta) \\ &= \int_0^1\left(\sqrt{n}(G_n(t)-t)-T_ng(t)\right)^2 dt, \end{aligned}$$

where $T_n = \sqrt{n}(\hat{\theta}_n-\theta)$, and

$$g(t) = g(t,\theta) = \left.\frac{\partial}{\partial\theta}F(x,\theta)\right|_{x=F^{-1}(t,\theta)}. \tag{3.4}$$

Darling's approach was based on showing that, when the underlying distribution of the sample is $F(\cdot,\theta)$ and $\mathcal{F}$ and $\hat{\theta}$ satisfy some adequate regularity conditions, then

$$\hat{W}_n^2 - H_n = o_P(1). \tag{3.5}$$

Thus, the asymptotic distribution of $\hat{W}_n^2$ can be studied through that of H_n. Darling showed that the finite dimensional distributions of $\sqrt{n}(G_n(t)-t)-T_ng(t)$ converge weakly to those of a Gaussian process $Y(t)$ with covariance function $K(s,t) = s\wedge t - st - \psi(s)\psi(t)$, where $\psi(t) = \sigma g(t)$ and σ^2 is the asymptotic variance of T_n. He showed, further, that under some additional assumptions on $\hat{\theta}_n$, Donsker's invariance principle could be applied to conclude that

$$\hat{W}_n^2 \xrightarrow{w} \int_0^1 (Y(t))^2dt,$$

and, as in the fixed distribution case, a Karhunen-Loève expansion for $\int_0^1 (Y(t))^2 dt$ can provide a good way to tabulate the limiting distribution of $\hat{W}_n^2$. Sukhatme (1972) extended Darling's result to multidimensional parameters and gave very valuable information for the Karhunen-Loève expansion of the limiting Gaussian process.

Instead of considering the process $\{\sqrt{n}(G_n(t) - t) - T_n g(t)\}_t$, a direct study of the estimated empirical process, $\{\sqrt{n}(\hat{G}_n(t) - t)\}_t$, could yield the asymptotic distribution of general $\hat{W}_n^2(\Psi)$ and $\hat{K}_n(\Psi)$ statistics (recall (3.2) and (3.3)) without having to rely on a different asymptotic equivalence as in (3.5) for every different statistic. Kac, Kiefer and Wolfowitz (1955) were the first to study this estimated empirical process in a particular case: if we are testing fit to the family of normal distributions $N(\mu, \sigma^2)$ and we estimate $\theta = (\mu, \sigma^2)$ by $\hat{\theta}_n = (\bar{X}_n, S_n^2)$, then the finite dimensional distributions of $\{\sqrt{n}(\hat{G}_n(t) - t)\}_t$ converge weakly to those of a centered Gaussian process $Z(t)$ with covariance function

$$\begin{aligned} K(s,t) &= s \wedge t - st - \phi(\Phi^{-1}(s))\phi(\Phi^{-1}(t)) \\ & -\frac{1}{2}\Phi^{-1}(s)\phi(\Phi^{-1}(s))\Phi^{-1}(t)\phi(\Phi^{-1}(t)), \end{aligned} \tag{3.6}$$

where Φ^{-1} is the quantile inverse of Φ (note that the difference between Darling's result and (3.6) is the introduction of an extra term corresponding to the second parameter to be estimated). Although they did not prove weak convergence of the estimated empirical process itself, they used this result (combined with a particular invariance result due to Kac) to conclude that $\hat{W}_n^2 \xrightarrow{w} \int_0^1 (Z(t))^2 dt$, providing, therefore, the asymptotic distribution of the Cramér-von Mises test of normality.

A general study of the weak convergence of the estimated empirical process was carried out by Durbin (1973) using the theory of weak convergence in $\mathcal{D}[0,1]$. Durbin's result can be essentially summarized as follows. Assume $\hat{\theta}_n$ satisfies

$$\sqrt{n}(\hat{\theta}_n - \theta) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} l(X_i, \theta) + \epsilon_n,$$

where $\epsilon_n \xrightarrow{P} 0$ and $l(X_1, \theta)$ is centered and has covariance matrix $L = L(\theta)$. Assume further that $F(x, \theta)$ is continuous in x for all θ. Set $h(t, \theta) = \int_{-\infty}^{F^{-1}(t,\theta)} l(x, \theta) dF(x, \theta)$ and assume that the vector (with the same dimension as θ) $g(t, \theta)$, as defined in (3.4), is continuous in (t, θ). Then we have

Theorem 3.1. *Under the null hypothesis and provided the above assumptions hold, the estimated empirical process, $\sqrt{n}(\hat{G}_n(t)-t)$, converges weakly in $\mathcal{D}[0,1]$ to a centered Gaussian process $Z(t)$ with covariance function*

$$K(s,t) = s\wedge t - st - h(s)'g(t) - h(t)'g(s) + g(s)'Lg(t). \qquad (3.7)$$

When $\mathcal{F}$ satisfies some regularity conditions and $\hat{\theta}_n$ is an efficient estimator (in the sense given in Durbin 1973), then L in (3.7) is the inverse of the information matrix, $I = I(\theta)$, and $h(t) = I^{-1}g(t)$. In this case (3.7) simplifies to

$$K(s,t) = s\wedge t - st - g(s)'I^{-1}g(t).$$

Note that this covariance function can be expressed as $s\wedge t - st - \sum_{j=1}^d \phi_j(s)\phi_j(t)$ for some real functions ϕ_j. A very complete study of the Karhunen-Loève expansion of Gaussian processes with this type of covariance function was carried out in Sukhatme (1972). Note also that a variant of Durbin's theorem in the form of weak approximation, given by Theorem 3.1(a) in Burke et al. (1979), proved useful later from the technical point of view.

Theorem 3.1 provides, as an easy corollary, the asymptotic distribution of a variety of $\hat{W}_n^2(\Psi)$ and $\hat{K}_n(\Psi)$ statistics under the null hypothesis. In fact, Durbin's results also give a valuable tool for studying its asymptotic power because they include too the asymptotic distribution of the estimated empirical process under contiguous alternatives. A survey of results connected to Theorem 3.1 as well as a simple derivation of it based on Skorohod embedding can be found in Shorack and Wellner (1986). Among the statistics whose asymptotic distribution can be derived from Theorem 3.1, three representatives have deserved special attention in the literature: the Cramer-von Mises statistic, and

$$\hat{K}_n = \sqrt{n} \sup_{-\infty<x<\infty} |F_n(x) - F(x,\hat{\theta}_n)|,$$

and

$$\hat{A}_n^2 = n\int_{-\infty}^{\infty} \frac{(F_n(x)-F(x,\hat{\theta}_n))^2}{F(x,\hat{\theta}_n)(1-F(x,\hat{\theta}_n))}\, dF(x,\hat{\theta}_n),$$

which are known, as in the fixed distribution setup, as Kolmogorov-Smirnov and Anderson-Darling statistics respectively. Also, as in the fixed distribution case, quadratic statistics offer in general better power properties than

$\hat{K}_n$, with $\hat{A}_n^2$ outperforming $\hat{W}_n^2$. Any of these statistics offers considerable gain in power with respect to the χ^2 test (see, e.g., Stephens 1974 or 1986a).

Let us conclude this subsection by commenting, briefly, that the achievements of subsequent advances in the theory of empirical processes have allowed the development of other goodness-of-fit procedures.

For instance, in Feuerverger and Mureika (1977) the asymptotic distribution of the empirical characteristic function is obtained; see also Csörgő (1981a). Analogous versions of Durbin's theorem for empirical characteristic and quantile functions were developed by Csörgő (1981b) and LaRiccia and Mason (1986). This was applied in Murota and Takeuchi (1981), Hall and Welsh (1983), Epps and Pulley (1983) and Csörgő (1986a, 1989) to propose new normality tests. Simulations in Hall and Welsh (1983) suggest that these tests have good behaviour against symmetric alternatives. Related ideas for testing for the broader model of all stable distributions are in Csörgő (1986b) and references therein, and these tests were recently simulated by Koutrouvelis and Meintanis (1999).

A different way to adapt the fixed-distribution tests is the *minimum distance method.* Assume that $\delta(F,G)$ is a distance between distribution functions. Set $\Delta(F_n,\mathcal{F}) := \inf_\theta \delta(F_n, F(\cdot,\theta))$. $\Delta(F_n,\mathcal{F})$ is a reasonable measure of the discrepancy between the sample distribution and the family $\mathcal{F}$ that can also be used for testing fit to $\mathcal{F}$. Dudley's theory of weak convergence of empirical processes can be used for deriving the limiting distribution of $\Delta(F_n,\mathcal{F})$ when $\delta(F,G) = ||F-G||$ with $||\cdot||$ being some norm on $\mathcal{D}[0,1]$ or $\mathcal{D}[-\infty,\infty]$ (see, e.g., Pollard 1980). An alternative derivation can be based on Skorohod embedding (see Shorack and Wellner 1986, pp. 254-257).

3.2 Correlation and regression tests

In this and in the next subsection we will assume that $\mathcal{F}$ is a location scale family, i.e., given a distribution function H_0, we will assume that $\mathcal{F}$ is the family of distribution functions obtained from H_0 by location or scale changes. We will assume H_0 to be standardized.

Goodness-of-fit tests in this subsection focus on the analysis of the popular probability plot. Some reviews on this subject have appeared recently (see, for instance, Lockhart and Stephens 1998, or Stephens 1986b). The idea behind the probability plot is as follows.

Let $X_1, \ldots, X_n$ be a random sample whose common distribution function belongs to $\mathcal{F}$ and has mean μ and variance σ^2. Let $X_0 = (X_{(1)}, \ldots, X_{(n)})$ be the corresponding ordered statistic. Let $Z_0 = (Z_{(1)}, \ldots, Z_{(n)})$ be an ordered sample with underlying distribution function H_0 and let $m' = (m_1, \ldots, m_n)$ and $V = (v_{ij})$ be, respectively, the mean vector and the covariance matrix of Z_0, that is, $m_i = EZ_{(i)}$ and $v_{ij} = E(Z_{(i)} - m_i)(Z_{(j)} - m_j)$. Then,

$$X_{(i)} = \mu + \sigma Z_{(i)}, \text{ in distribution, } i = 1, \ldots, n. \tag{3.8}$$

Thus, the plot of the ordered values $X_{(1)}, \ldots, X_{(n)}$ against the points $m_1, \ldots, m_n$ should be approximately linear; lack of linearity in this plot suggests that the distribution function of X_1 does not belong to $\mathcal{F}$. Checking this linearity is often done "by eye", but, some analytical procedures have been devised. They were proposed according to two different criteria, which essentially lead to equivalent tests, the main difference being the point of view employed by the proposer to justify his/her proposal.

The first criterium relies on the idea of selecting an estimator $\hat{\sigma}^2$ of σ^2, assuming the linear model (3.8) is correct, and comparing it with S_n^2 which, in any case, is a consistent estimator of σ^2. Under the null hypothesis $\hat{\sigma}^2/S_n^2$ should take values close to 1. Hence, values of $\hat{\sigma}^2/S_n^2$ far from 1 would lead to rejection of the null hypothesis. These procedures are called *regression tests*.

A second class consists of tests assessing the linearity in (3.8) through the correlation coefficient between vectors X_0 and m, $\rho(m, X_0)$ (notice that here we have no real correlation coefficient because m is not random). When model (3.8) is true, we expect $\rho^2(m, X_0)$ to take values close to 1 and, consequently, small values of $\rho^2(m, X_0)$ would indicate that the null hypothesis is not true. Tests of this kind are called *correlation tests*. The vector m can be replaced by other vectors $\beta = (\beta_1, \ldots, \beta_n)$ satisfying, under the null hypothesis, some approximate linear relation with X_0. Coordinates of the vector β are usually known as plotting positions.

The first example of these tests was the Shapiro-Wilk W-test of normality, proposed in Shapiro and Wilk (1965). There, the authors state that they are trying to provide an analytical procedure "to summarize formally indications of probability plots" (pp. 591). The best linear unbiased estimators, BLUE, of μ and σ in model (3.8) are

$$\hat{\mu} = \bar{X}_n \text{ and } \hat{\sigma} = \frac{m' V^{-1} X_0}{m' V^{-1} m}$$

(this holds for any symmetric H_0). Hence, under the null hypothesis, $\hat{\sigma}^2 / S_n^2$ should take values close to 1. The Shapiro-Wilk statistic, W, is a normalized version of $\hat{\sigma}^2 / S_n^2$, namely,

$$W = \frac{\left(m' V^{-1} X_0 \right)^2}{m' V^{-1} V^{-1} m \sum_i (X_i - \bar{X})^2}. \tag{3.9}$$

The normalization ensures that W always takes values between 0 and 1 (since W equals $\rho^2(V^{-1}m, X_0)$). Small values of W would lead to rejection of the null hypothesis. This is a regression test, since it is based on the comparison of $\hat{\sigma}$ and S_n^2, but, obviously, it can also be seen as a correlation test with plotting positions $V^{-1}m$. According to simulations (provided, for instance, in Shapiro, Wilk and Chen 1968) it seems that the W-test is one of the most powerful normality tests against a wide range of alternatives. This fact has made the test very popular, and it can be considered the gold standard for comparisons. However, employing W for testing normality presents several difficulties of different kinds.

One problem concerns computational aspects. Computation of W requires previous computation of m and V^{-1}. This task is difficult and, in fact, when W was introduced, elements in V were tabulated only for $n \leq 20$. For this reason some numerical approximations that allowed the computation of W quite accurately for sample sizes up to 50 were proposed in Shapiro and Wilk (1965).

An equally important concern regarding W was the tabulation of its null distribution. Except in case $n = 3$, when the W-test is equivalent to the u-test (see Shapiro and Wilk 1965) the exact distribution of W is unknown. Percentiles of W were computed by simulation in Shapiro and Wilk (1965) for sample sizes up to 50. However, the asymptotic distribution of W remained unknown for a long time. In fact, it was not obtained until 20 years later, in Leslie, Stephens and Fotopoulos (1986) who showed the asymptotic equivalence, under normality, of W and another correlation test whose distribution was already known at this time (see the considerations concerning the de Wet-Venter test below).

Some transformations of W that made its distribution approximately Gaussian were proposed (see Shapiro and Wilk 1968, or Royston 1982). It is curious to notice that, in Shapiro and Wilk (1968), the authors employ normal probability plots, whose fit is addressed "by eye", to analyze the goodness of the proposed approximation. However, these results must be used with some caution because, as shown in Leslie (1984), they rely on approximations which do not hold with the necessary accuracy.

An additional weakness of the Shapiro-Wilk test is that the procedure may be not consistent for testing fit to non-normal families of distributions. For instance, if $\mathcal{F}$ is the exponential location scale family then the Shapiro-Wilk statistic becomes

$$W_E = \frac{(\bar{X}_n - X_{(1)})^2}{(n-1)S_n^2},$$

which is a function of the coefficient of variation. There are some families of distributions with the same coefficient of variation as the exponential family (see Sarkadi 1975; Spinelli and Stephens 1987). Thus, the W_E-test is not consistent when testing for exponentiality. In particular, simulations in Spinelli and Stephens (1987) suggest that the power of the W_E-test against the beta $(1/4, 5/12)$ distribution decreases with the sampling size.

The limitations of the Shapiro-Wilk test led to the introduction of modifications of W, which aimed to ease them. The first examples were the D'Agostino test (see D'Agostino 1971) and the Shapiro-Francia test (see Shapiro and Francia 1972). They were intended to replace the W-test for sample sizes greater than 50. Both tests are easier to compute than the W-test. The D'Agostino test employs an estimator of σ proposed in Downton (1966) to get the statistic

$$D = \frac{\sum_i (i - (n+1)2^{-1}) X_{(i)}}{n^2 S_n}.$$

The Shapiro-Francia test is based on an idea suggested (without proof) in Gupta (1952) (see also Stephens 1975) according to which the matrix V^{-1} in (3.9) can be replaced with the identity I, obtaining the statistic

$$W' = \frac{\left(m' X_0\right)^2}{m'm \sum (X_i - \bar{X})^2}.$$

Both tests are correlation tests. The plotting positions are $(1, 2, \ldots, n)$ for the D-test and m for the W'-test. Simulation studies in D'Agostino (1971) and Shapiro and Francia (1972), respectively, suggest that the proposed tests are approximately equivalent to the W-test. The D-test has the advantage of being asymptotically normal and its distribution can be approximated by a Cornish-Fisher expansion for moderate sample sizes.

Apart from the ease of computation, an interesting feature of the W'-Shapiro-Francia test is its consistency for testing fit to any location scale family with finite second order moment, a fact shown in Sarkadi (1975). However, it is curious to notice that this consistency disappears if one you employ the asymptotic distribution. This happens, as shown in sub-section 3.3.3, because if the family under testing has tails a bit heavier than those of the Gaussian distribution (this includes, for instance, the exponential family), then the asymptotic distribution only depends on the tails of the distribution. Therefore, if we have a distribution in the alternative with the same tails than a distribution in the family, then the asymptotic distributions of W'-Shapiro-Francia test under the null hypothesis and under the distribution in the alternative coincide (see also the comments about the power of the Shapiro-Francia test below).

A further simplification of the W'-test was proposed in Weisberg and Bingham (1975) by replacing m by the vector $\tilde{m} = (\tilde{m}_1, \ldots, \tilde{m}_n)$, where

$$\tilde{m}_i = \Phi^{-1}\left(\frac{i - 3/8}{n + 1/4}\right), \qquad i = 1, \ldots, n.$$

This statistic is easier to compute than W', and a Monte Carlo study in Weisberg and Bingham (1975) suggests that the two tests are equivalent.

Another modification of W was proposed by de Wet and Venter (1972). It seems that the concept of the correlation test was introduced for the first time in that paper. The de Wet and Venter test is the correlation test with plotting positions

$$\beta = \left(\Phi^{-1}\left[\frac{1}{n+1}\right], \ldots, \Phi^{-1}\left[\frac{n}{n+1}\right]\right),$$

or, equivalently, the test which rejects normality when large values of

$$W^* = \sum_i \left(\frac{X_{(i)} - \bar{X}_n}{S_n} - \Phi^{-1}[i/(n+1)]\right)^2$$

are observed.

Subsequent tests continued this approach. For instance, Filliben (1975) proposed a correlation test with the medians of the ordered statistic Z_0 as plotting positions. Some simulations comparing this and the W and W' tests were given. The distribution of this statistic was also computed via the Monte Carlo method.

An interesting feature of the W^*-test is that it was the first correlation normality test with known asymptotic distribution. To be precise, it was shown in de Wet and Venter (1972) that, if $\{Z_i\}$ is a sequence of independent standard Gaussian random variables, then

$$W^* - a_n \xrightarrow{w} \sum_{i=3}^{\infty} \frac{Z_i^2 - 1}{i}$$

for a certain sequence of constants $\{a_n\}$. The key to the proof relied on showing, through rather involved calculations, the asymptotic equivalence, under normality, of W^* and a certain quadratic form and using the asymptotic theory for quadratic forms given in de Wet and Venter (1973).

Since the publication of de Wet and Venter (1972), the possibility of obtaining the asymptotic distribution of other correlation tests of normality by showing their asymptotic equivalence with the W^*-test has been considered. An important paper in this program was that of Verrill and Johnson (1987), where the asymptotic equivalence of correlation tests under some general conditions (satisfied by most of the correlation tests in the literature) is shown. In particular, it is shown that the Shapiro-Francia, the Weisberg-Bingham and the Filliben tests are asymptotically equivalent to the de Wet-Venter test, having consequently the same asymptotic distribution.

The asymptotic distribution of the Shapiro-Wilk test could, then, be obtained using its asymptotic equivalence with the Shapiro-Francia, shown in Leslie, Stephens and Fotopoulos (1986). This solved an important problem that had existed for around twenty years. It would be unfair not to mention Leslie (1984), which proved the validity of the key step in previous heuristic reasonings based on assuming that the vector m is an "asymptotic eigenvector" of V^{-1}. More precisely, the main result of that paper is that there exists a constant C which does not depend on n, such that

$$\|V^{-1}m - 2m\| \leq C(\log n)^{-1/2},$$

where, given the matrix $B = (b_{ij})$, then $\|B\|^2 = \sum_{ij} b_{ij}^2$.

The possibility of extending the use of correlation tests to cover goodness-of-fit to other families of distributions has also been explored, for instance, in Smith and Bain (1976), for the exponential distribution, or in Gerlach (1979) for the extreme value distributions. In this setup, correlation tests do not present the same nice properties exhibited when testing normality. In Lockhart (1985) the asymptotic normality of the Shapiro-Francia test when applied to the exponential family is obtained. The rate of convergence is extremely slow: $(\log n)^{1/2}$. This result was generalized in McLaren and Lockhart (1987) to cover extreme-value and logistic distributions with the same rate and the same asymptotic distribution as in the exponential case. However, the asymptotic efficiency of the Shapiro-Francia test in these situations was found to be 0 when compared with tests based on the empirical distribution function, since it was possible to find a sequence of contiguous alternatives such that the asymptotic power coincides with the nominal level of significance of the test (on this question, see also Lockhart 1991).

3.3 Tests based on Wasserstein distance

A different approach to correlation tests was suggested in del Barrio, Cuesta-Albertos, Matrán and Rodríguez-Rodríguez (1999) and will be widely developed in the remainder of this work. The methodology consists of analyzing the L_2-Wasserstein distance between a fixed distribution and a location scale family of probability distributions in $\mathbb{R}$. Our study will cover different kinds of distribution tails, including as key examples the uniform, normal, exponential and a more heavily tailed law.

Let $\mathcal{P}_2(\mathbb{R})$ be the set of probabilities on $\mathbb{R}$ with a finite second moment. For probabilities P_1 and P_2 in $\mathcal{P}_2(\mathbb{R})$ the L_2-Wasserstein distance between P_1 and P_2 is defined as

$$\mathcal{W}(P_1, P_2) := \inf\left\{ \left[E\,(X_1 - X_2)^2\right]^{1/2} : \mathcal{L}(X_1) = P_1, \mathcal{L}(X_2) = P_2 \right\}.$$

For simplification of notation we will identify probability laws with their distribution functions. In particular, if F_i, $i = 1, 2$, are the distribution functions associated with the probability laws $P_i \in \mathcal{P}_2(\mathbb{R})$, we will say that $F_i \in \mathcal{P}_2(\mathbb{R})$, $i = 1, 2$ and write $\mathcal{W}(F_1, F_2)$ instead of $\mathcal{W}(P_1, P_2)$.

An important fact, which makes $\mathcal{W}$ useful for univariate statistics (the multivariate setting is very different), is that it can be explicitly obtained in terms of quantile functions. If $F_i \in \mathcal{P}_2(\mathbb{R})$, $i = 1, 2$, then (see, e.g., Vallender 1973; Bickel and Freedman 1981)

$$\mathcal{W}(F_1, F_2) = \left[\int_0^1 \left(F_1^{-1}(t) - F_2^{-1}(t) \right)^2 dt \right]^{1/2}. \tag{3.10}$$

Some relevant well-known properties of the Wasserstein distance are included in the following proposition for future reference. The reader interested in properties and uses of the general L_p-Wasserstein distance can refer to Bickel and Freedman (1981), Cuesta-Albertos, Matrán, Rachev and Rüschendorf (1996) or Rachev and Rüschendorf (1998).

Proposition 3.1.

(a) Let $F_i \in \mathcal{P}_2(\mathbb{R})$, $i = 1, 2$. Call m_i the mean value of F_i and F_i^* the distribution function defined by $F_i^*(x) = F_i(x - m_i)$. Then

$$\mathcal{W}^2(F_1, F_2) = \mathcal{W}^2(F_1^*, F_2^*) + (m_1 - m_2)^2.$$

(b) Let $\{F_n\}_n$ be a sequence in $\mathcal{P}_2(\mathbb{R})$. The following statements are equivalent:

i. $F_n \to F \in \mathcal{P}_2(\mathbb{R})$ in $\mathcal{W}$-distance (i.e. $\mathcal{W}(F_n, F) \to 0$).

ii. $F_n \overset{w}{\to} F$ and $\int |t|^2 dF_n \to \int |t|^2 dF < \infty$.

iii. $F_n^{-1} \to F^{-1}$ a.s. and in $L^2(0, 1)$.

As in Subsection 3.2 we assume $\mathcal{F}$ to be a location scale family of distribution functions, that is, $\mathcal{F} = \{H : H(x) = H_0((x - \mu)/\sigma), \mu \in \mathbb{R}, \sigma > 0\}$ for some $H_0 \in \mathcal{P}_2(\mathbb{R})$ which we choose, for simplicity, with zero mean and unit variance (thus, given $H(x) = H_0((x - \mu)/\sigma)$ in $\mathcal{F}$, μ and σ are its mean and its standard deviation, respectively).

Note that the quantile function associated with $H(x) = H_0((x - \mu)/\sigma)$ satisfies $H^{-1}(t) = \mu + \sigma H_0^{-1}(t)$. Therefore, if F is a distribution function in $\mathcal{P}_2(\mathbb{R})$ with mean μ_0 and standard deviation σ_0, (3.10) and Proposition 3.1 (a) imply that

$$\begin{aligned} \mathcal{W}^2(F,\mathcal{F}) &= \inf\{\mathcal{W}^2(F,H),\ H\in\mathcal{F}\} \\ &= \inf_{\sigma>0}\left\{\int_0^1 \left(F^{-1}(t)-\mu_0-\sigma H_0^{-1}(t)\right)^2 dt\right\} \\ &= \inf_{\sigma>0}\left\{\sigma_0^2+\sigma^2-2\sigma\int_0^1 \left(F^{-1}(t)-\mu_0\right)H_0^{-1}(t)dt\right\} \\ &= \sigma_0^2-\left(\int_0^1 \left(F^{-1}(t)-\mu_0\right)H_0^{-1}(t)dt\right)^2 \\ &= \sigma_0^2-\left(\int_0^1 F^{-1}(t)H_0^{-1}(t)dt\right)^2. \qquad (3.11)\end{aligned}$$

Thus, the law in $\mathcal{F}$ closest to F is given by $\mu=\mu_0$ and $\sigma=\int_0^1 F^{-1}(t)$ $H_0^{-1}(t)dt$, which is the covariance between F^{-1} and H_0^{-1} when seen as random variables defined on $(0,1)$. The ratio $\mathcal{W}^2(F,\mathcal{F})/\sigma_0^2$ is not affected by location or scale changes of F. Hence, it can be considered as a measure of dissimilarity between F and $\mathcal{F}$. For example, the best $\mathcal{W}$-approximation to F in the set $\mathcal{F}_\mathcal{N}$ of normal laws will be the normal law with mean μ_0 and standard deviation $\int_0^1 F^{-1}(t)\Phi^{-1}(t)\,dt$, and the ratio

$$\frac{\mathcal{W}^2(F,\mathcal{F}_\mathcal{N})}{\sigma_0^2}=1-\frac{\left(\int_0^1 F^{-1}(t)\Phi^{-1}(t)dt\right)^2}{\sigma_0^2}$$

measures the non-normality of F.

The invariance of $\mathcal{W}^2(F,\mathcal{F})/\sigma_0^2$ against location or scale changes of F suggests that it is convenient to use a sample version of it for testing fit to the location scale family $\mathcal{F}$. More precisely, if $X_1,X_2,\ldots,X_n$ is a random sample with underlying distribution function F,

$$\mathcal{R}_n := \frac{\mathcal{W}^2(F_n,\mathcal{F})}{S_n^2}=1-\frac{\hat{\sigma}_n^2}{S_n^2},$$

where $\hat{\sigma}_n=\int_0^1 F_n^{-1}(t)H_0^{-1}(t)dt$, can be used as a test statistic for the null hypothesis $F\in\mathcal{F}$. Large values of $\mathcal{R}_n$ would lead to the rejection of the null hypothesis.

This testing procedure belongs to the class of minimum distance tests described in Subsection 3.1. A nice feature of Wasserstein tests is that we have an explicit expression for the minimum distance estimators and, consequently, for the minimum distance statistic, unlike what happens for other metrics (e.g., those leading to Kolmogorov-Smirnov or Cramér-von Mises statistics).

The connection between $\mathcal{R}_n$ and correlation and regression tests can be clearly seen by noting that large values of $\mathcal{R}_n$ correspond to small values of $\hat{\sigma}_n^2/S_n^2$, which, with the notation employed in Subsection 3.2, can be expressed as

$$\rho^2(\nu, X_0) = \frac{\left(\sum_{i=1}^n \nu_i X_{(i)}\right)^2}{n\nu'\nu S_n^2},$$

where $\nu = (\nu_1, \ldots, \nu_n)'$ and $\nu_i = \int_{(i-1)/n}^{i/n} H_0^{-1}(t)\, dt$, $i = 1, \ldots, n$ (observe that ν is centered and $\lim_n \nu'\nu = 1$ since H_0 is assumed to be standardized). Hence, the $\mathcal{R}_n$-test is equivalent to a correlation test with plotting positions $\nu = (\nu_1, \ldots, \nu_n)'$. In fact, the plotting positions in the Shapiro-Wilk, Shapiro-Francia or de Wet-Venter tests are approximations to the Wasserstein plotting positions. This was noticed, in the context of normal probability plots, by Brown and Hettmansperger in (1996), which considered the problem of finding the optimum plotting positions. That paper presented a heuristic explanation, based on an orthogonal expansion of $\mathcal{R}_n$, of the power properties of the $\mathcal{R}_n$ normality test against general alternatives, observed by Stephens (1975). Our results (Theorems 3.5 and 3.6) will justify those heuristic considerations.

We now consider the problem of obtaining the asymptotic distribution of $\mathcal{R}_n$ under the null hypothesis. The invariance of $\mathcal{R}_n$ with respect to location or scale changes of F allows us to assume that $F = H_0$. By the convergence of $S_n^2 \to \sigma^2(H_0) = 1$ a.s., we will be able to study the asymptotic behaviour of $\mathcal{R}_n$ through that of $S_n^2\mathcal{R}_n$ which in turn (recalling that H_0 is standardized and $F = H_0$) permits the following decomposition

$$\begin{aligned} 0 \le \mathcal{R}_n^* &:= S_n^2\mathcal{R}_n = S_n^2 - \left(\int_0^1 F_n^{-1}(t)H_0^{-1}(t)dt\right)^2 \\ &= \int_0^1 (F_n^{-1}(t))^2 dt - (\overline{X}_n)^2 - \left(\int_0^1 F_n^{-1}(t)F^{-1}(t)dt\right)^2 \\ &= \int_0^1 (F_n^{-1}(t) - F^{-1}(t))^2 dt - \left(\int_0^1 (F_n^{-1}(t) - F^{-1}(t))dt\right)^2 \end{aligned}$$

$$
\begin{aligned}
&- \left(\int_0^1 (F_n^{-1}(t) - F^{-1}(t)) F^{-1}(t) dt \right)^2 \\
=: \quad & \mathcal{R}_n^{(1)} - \mathcal{R}_n^{(2)} - \mathcal{R}_n^{(3)}. \qquad (3.12)
\end{aligned}
$$

Let us remark that $n\mathcal{R}_n^{(2)} = (n^{1/2}\bar{X}_n)^2$ which, since $F \in \mathcal{P}_2(\mathbb{R})$, has a χ_1^2 asymptotic law. On the other hand,

$$
n\mathcal{R}_n^{(3)} = \left(n^{1/2} \left(\int_0^1 F_n^{-1}(t) F^{-1}(t) dt - 1 \right) \right)^2 = (n^{1/2}(\hat{\sigma}_n - 1))^2,
$$

which, under not-too-restrictive conditions, has a scaled χ_1^2 asymptotic law (but see (3.26)). Finally note that, in the normal case, $n\mathcal{R}_n^{(1)}$ is similar to the statistic L_n^0 of de Wet and Venter. However, the derivation of the asymptotic distribution of $\mathcal{R}_n^*$ requires a joint treatment of $(\mathcal{R}_n^{(1)}, \mathcal{R}_n^{(2)}, \mathcal{R}_n^{(3)})$.

A look at (3.12) suggests that this joint treatment can be based on the asymptotic theory of quantile processes. This will be our approach. If F has derivative f, the general quantile process, ρ_n, is defined by

$$
\rho_n(t) := n^{1/2} f(F^{-1}(t)) \left(F^{-1}(t) - F_n^{-1}(t) \right), \; 0 \leq t \leq 1.
$$

This general quantile process can be approximated, under certain regularity conditions, by Brownian bridges in a way that we will study in more detail below. This approximation can be used in the study of $\mathcal{R}_n^*$ since

$$
n\mathcal{R}_n^* = \int_0^1 \left(\frac{\rho_n(t)}{f(F^{-1}(t))} \right)^2 dt - \left(\int_0^1 \frac{\rho_n(t)}{f(F^{-1}(t))} dt \right)^2 - \left(\int_0^1 \frac{\rho_n(t) F^{-1}(t)}{f(F^{-1}(t))} dt \right)^2. \qquad (3.13)
$$

For the approximation of the general quantile process, we will assume the following regularity conditions on F.

Assumptions. *Let* $a = \sup\{x : F(x) = 0\}, b = \inf\{x : F(x) = 1\}, -\infty \leq a \leq b \leq \infty$. *We will assume that*

1. *F is twice differentiable on (a, b).*
2. *$F'(x) = f(x) > 0$, $x \in (a, b)$.*

3. For some $\gamma > 0$ we have

$$\sup_{0<t<1} t(1-t)|f'(F^{-1}(t))|/f^2(F^{-1}(t)) \leq \gamma.$$

The following strong approximation result for ρ_n (Theorem 6.2.1 in Csörgő and Horváth 1993) will enable us to use expression (3.13):

Theorem 3.2. *Under Assumptions 1, 2 and 3, we can define, on a rich enough probability space, a sequence of Brownian bridges $\{B_n(t), 0 \leq t \leq 1\}_n$ such that*

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

Under the additional assumption

$$\int_0^1 \frac{t(1-t)}{f(F^{-1}(t))^2} dt < \infty, \tag{3.14}$$

we can use Theorem 3.2 and the same techniques employed in the proof of Theorem 6.4.2 in Csörgő and Horváth (1993) to show that

$$\begin{aligned} &\int_{\frac{1}{(n+1)}}^{\frac{n}{n+1}} \left(\frac{\rho_n(t)}{f(F^{-1}(t))} \right)^2 dt - \left(\int_{\frac{1}{(n+1)}}^{\frac{n}{n+1}} \frac{\rho_n(t)}{f(F^{-1}(t))} dt \right)^2 - \left(\int_{\frac{1}{(n+1)}}^{\frac{n}{n+1}} \frac{\rho_n(t) F^{-1}(t)}{f(F^{-1}(t))} dt \right)^2 \\ &\xrightarrow{w} \int_0^1 \left(\frac{B(t)}{f(F^{-1}(t))} \right)^2 dt - \left(\int_0^1 \frac{B(t)}{f(F^{-1}(t))} dt \right)^2 - \left(\int_0^1 \frac{B(t) F^{-1}(t)}{f(F^{-1}(t))} dt \right)^2, \end{aligned} \tag{3.15}$$

where $B(t)$ is a Brownian bridge. Therefore, if the following conditions on the behaviour of the extremes hold

$$\begin{aligned} n \int_0^{\frac{1}{n}} (F_n^{-1}(t) - F^{-1}(t))^2 dt = n \int_0^{\frac{1}{n}} \left(X_{1n} - F^{-1}(t) \right)^2 dt &\xrightarrow{P} 0 \text{ and} \\ n \int_{1-\frac{1}{n}}^1 (F_n^{-1}(t) - F^{-1}(t))^2 dt = n \int_{1-\frac{1}{n}}^1 \left(X_{nn} - F^{-1}(t) \right)^2 dt &\xrightarrow{P} 0, \end{aligned} \tag{3.16}$$

taking into account that for every Borel set, $A \subset [0,1]$,

$$\begin{aligned}\int_A (F_n^{-1}(t) - F^{-1}(t))^2 dt &\geq \left(\int_A (F_n^{-1}(t) - F^{-1}(t))dt\right)^2 \text{ and}\\ \int_A (F_n^{-1}(t) - F^{-1}(t))^2 dt &\geq \left(\int_A (F_n^{-1}(t) - F^{-1}(t))F^{-1}(t)dt\right)^2 \quad (3.17)\end{aligned}$$

we can conclude from (3.13) that the asymptotic distribution of $n\mathcal{R}_n^*$ is given by the limiting expression in (3.15). The following theorem summarizes this fact.

Theorem 3.3. *Under Assumptions 1, 2 and 3, if $F \in \mathcal{F}$ and (3.14) and (3.16) hold, then*

$$n\mathcal{R}_n \xrightarrow{w} \int_0^1 \left(\frac{B(t)}{h(H_0^{-1}(t))}\right)^2 dt - \left(\int_0^1 \frac{B(t)}{h(H_0^{-1}(t))} dt\right)^2 - \left(\int_0^1 \frac{B(t)H_0^{-1}(t)}{h(H_0^{-1}(t))} dt\right)^2,$$

where h denotes the derivative of H_0.

We will now consider the application of Theorem 3.3 to several location scale families $\mathcal{F}$.

3.3.1 Uniform pattern

The hypotheses in Theorem 3.3 are easily checked for the uniform model. Here $H_0(t) = 12^{-1/2}(t + 3^{1/2})$ and $h(t) = 12^{-1/2}$ for $t \in (-3^{1/2}, 3^{1/2})$ and we trivially obtain the following result.

Theorem 3.4. (Uniform model). *If $\mathcal{F}$ is the family of uniform distributions on intervals, then*

$$n\mathcal{R}_n \xrightarrow{w} 12\left[\int_0^1 B^2(t)dt - \left(\int_0^1 B(t)dt\right)^2\right] - 144\left(\int_0^1 \left(t - \frac{1}{2}\right) B(t)dt\right)^2. \tag{3.18}$$

A principal components decomposition (see, e.g., Shorack and Wellner 1986) allows us to express this limiting distribution as a weighted sum of independent χ_1^2 random variables. The expression in square brackets has been studied in relation to the Watson statistic and admits an easy expansion (see, e.g., Shorack and Wellner 1986). On the other hand, Lockhart and Stephens (1998) have obtained in the expansion of (3.18) through the analysis of the covariance function of the Gaussian process

$$B(t) - \int_0^1 B(u)du - 12\left(t-\frac{1}{2}\right)\int_0^1\left(u-\frac{1}{2}\right)B(u)du,$$

resulting in the following expression for the limiting distribution in (3.18):

$$n\mathcal{R}_n \xrightarrow{w} 12\sum_{j=1}^{\infty}\lambda_j Y_j^2,$$

where Y_j are i.i.d. standard normal random variables and λ_j are the solutions of the following equation

$$1-\cos\left(\frac{1}{\sqrt{\lambda}}\right) = \frac{1}{2}\frac{1}{\sqrt{\lambda}}\sin\left(\frac{1}{\sqrt{\lambda}}\right). \tag{3.19}$$

We note that $\lambda = (\pi j)^{-2}$ with j a positive even integer is a solution of (3.19), but we do not have an explicit expression for all solutions of this equation.

3.3.2 Normal pattern

The normal model needs a more careful treatment. The main problem arises from the fact that the integral in (3.14) diverges. In fact we have (see Bickel and van Zwet 1978)

$$\int_{\frac{1}{n}}^{1-\frac{1}{n}} \frac{t(1-t)}{\phi(\Phi^{-1}(t))^2}dt = \log\log n + \log 2 + \gamma + o(1), \tag{3.20}$$

where $\gamma = \lim_{k\to\infty}\left(\sum_{j=1}^k j^{-1} - \log k\right)$ is Euler's constant. Since it is well known (see, e.g., Lemma 5.3.2 in Csörgő and Horváth 1993, or Corollary 2.2 in Csörgő, Horváth and Shao 1993) that

$$P\left[\int_0^1\left(\frac{B(t)}{f(F^{-1}(t))}\right)^2 dt < \infty\right] = \begin{cases} 1, & \text{if } \int_0^1 \frac{(t(1-t))}{f(F^{-1}(t))^2}dt < \infty \\ 0, & \text{if } \int_0^1 \frac{t(1-t)}{f(F^{-1}(t))^2}dt = \infty, \end{cases}$$

the limiting expression in Theorem 3.3 becomes $+\infty$ with probability 1 and a more precise argument is needed.

The asymptotic theory of extremes and the well known equivalence $\phi(x) \approx |x|\, \Phi(x)$ for $x \to -\infty$ enables us to prove (3.16) for the normal law. We state this result in the following proposition, which was proved in del Barrio, Cuesta-Albertos, Matrán and Rodríguez-Rodríguez (1999).

Proposition 3.2. *If $\{X_{in}, i = 1, \ldots, n\}$ is the ordered sample obtained from an (i.i.d.) random sample with standard normal law, then:*

$$n \int_0^{\frac{1}{n}} \left(X_{1n} - \Phi^{-1}(t)\right)^2 dt \xrightarrow{p} 0 \text{ and } n \int_{1-\frac{1}{n}}^{1} \left(X_{nn} - \Phi^{-1}(t)\right)^2 dt \xrightarrow{p} 0.$$

The next result reduces the problem of deriving the asymptotic distribution of $\mathcal{R}_n^*$ to the study of a certain functional of a Brownian bridge.

Proposition 3.3. *On an adequate probability space there exists a sequence $\{B_n(t)\}_n$ of Brownian bridges that fulfills*

$$n\mathcal{R}_n^* - \left[\int_{\frac{1}{(n+1)}}^{\frac{n}{n+1}} \left(\frac{B_n(t)}{\phi(\Phi^{-1}(t))} \right)^2 dt \;-\; \left(\int_{\frac{1}{(n+1)}}^{\frac{n}{n+1}} \frac{B_n(t)}{\phi(\Phi^{-1}(t))} dt \right)^2 \right.$$
$$\left. -\; \left(\int_{\frac{1}{(n+1)}}^{\frac{n}{n+1}} \frac{B_n(t)\Phi^{-1}(t)}{\phi(\Phi^{-1}(t))} dt \right)^2 \right] \xrightarrow{p} 0.$$

Proof. From Proposition 3.2 and (3.17) it follows that

$$n\mathcal{R}_n^* - \left[\int_{\frac{1}{(n+1)}}^{\frac{n}{n+1}} \left(\frac{\rho_n(t)}{\phi(\Phi^{-1}(t))} \right)^2 dt \;-\; \left(\int_{\frac{1}{(n+1)}}^{\frac{n}{n+1}} \frac{\rho_n(t)}{\phi(\Phi^{-1}(t))} dt \right)^2 \right.$$
$$\left. -\; \left(\int_{\frac{1}{(n+1)}}^{\frac{n}{n+1}} \frac{\rho_n(t)\Phi^{-1}(t)}{\phi(\Phi^{-1}(t))} dt \right)^2 \right] \xrightarrow{p} 0.$$

Therefore, the result follows from the fact that, on an adequate space, we can substitute the quantile process, ρ_n, with a sequence of Brownian bridges. This can be obtained by a careful use of Theorem 3.2 and the equivalence $\phi(x) \approx |x|\, \Phi(x)$ for $x \to -\infty$ (see details in del Barrio, Cuesta-Albertos, Matrán and Rodríguez-Rodríguez 1999). □

The convergence of $\mathcal{R}_n$ and the characterization of its limit law are easier problems. In the following theorem we establish the convergence in law of $\mathcal{R}_n$ through the analysis of an equivalent version based on the Brownian bridge. Note that the main difficulty is giving sense to

$$\int_0^1 \frac{B^2(t) - EB^2(t)}{\left(\phi(\Phi^{-1}(t))\right)^2} dt \tag{3.21}$$

because it follows from Lemma 2.2 in Csörgő, Horváth and Shao (1993) that the function defined by $t \mapsto (B^2(t) - EB^2(t))/(\phi(\Phi^{-1}(t)))^2$ is a.s. not integrable. Therefore, we cannot assume the a.s. existence of

$$\lim_n \int_{\frac{1}{n+1}}^{\frac{n}{n+1}} \frac{B^2(t) - EB^2(t)}{\left(\phi(\Phi^{-1}(t))\right)^2} dt.$$

However, this limit does exist in the L_2-sense, and we can define (3.21) as this L_2-limit.

Theorem 3.5. (Normal case) *Let $\{X_n\}_n$ be a sequence of i.i.d. normal random variables. Then*

$$n(\mathcal{R}_n - a_n) \xrightarrow{w} \int_0^1 \frac{B^2(t) - EB^2(t)}{\left(\phi(\Phi^{-1}(t))\right)^2} dt - \left(\int_0^1 \frac{B(t)}{\phi(\Phi^{-1}(t))} dt\right)^2 - \left(\int_0^1 \frac{B(t)\Phi^{-1}(t)}{\phi(\Phi^{-1}(t)} dt\right)^2,$$

where

$$a_n = \frac{1}{n} \int_{\frac{1}{n+1}}^{\frac{n}{n+1}} \frac{t(1-t)}{[\phi(\Phi^{-1}(t))]^2} dt.$$

Proof. As already observed, we can assume without loss of generality that the variables have the standard normal law, and that, by the asymptotic normality of the sample variance S_n^2 and (3.20), we have

$$\begin{aligned} n(\mathcal{R}_n - a_n) - n(\mathcal{R}_n^* - a_n) &= \frac{n}{S_n^2} \mathcal{R}_n^* (1 - S_n^2) \\ &= O_p(1)\sqrt{n}(\mathcal{R}_n^* - a_n + a_n) \xrightarrow{p} 0 \end{aligned}$$

provided $n(\mathcal{R}_n^* - a_n) = O_p(1)$. Hence, the result will be proved if we show that $n(\mathcal{R}_n^* - a_n)$ converges in distribution to the functional of the Brownian bridge involved in the statement of the theorem. By Proposition 3.3, it suffices to give a limit sense to

$$\int_0^1 \frac{B^2(t) - EB^2(t)}{(\phi(\Phi^{-1}(t)))^2} dt.$$

If we set

$$A_n = \int_{\frac{1}{n+1}}^{\frac{n}{n+1}} \frac{B^2(t) - EB^2(t)}{(\phi(\Phi^{-1}(t)))^2} dt,$$

then straightforward calculations show that

$$\begin{aligned} EA_n^2 &= \int_{\frac{1}{n+1}}^{\frac{n}{n+1}} \int_{\frac{1}{n+1}}^{\frac{n}{n+1}} \frac{2(s \wedge t - st)^2}{(\phi(\Phi^{-1}(s)))^2(\phi(\Phi^{-1}(t)))^2} dsdt \\ &\to \int_0^1 \int_0^1 \frac{2(s \wedge t - st)^2}{(\phi(\Phi^{-1}(s)))^2(\phi(\Phi^{-1}(t)))^2} dsdt < \infty. \qquad (3.22) \end{aligned}$$

This fact can be used to show that $E(A_n - A_m)^2 \to 0$ as $n, m \to \infty$, hence that A_n converges in L_2 to a random variable

$$A := \int_0^1 \frac{B^2(t) - EB^2(t)}{[\phi(\Phi^{-1}(t))]^2} dt.$$

□

The next theorem provides a series expansion of the limit law in Theorem 3.5. Note that, to some degree, the proof of Theorem 3.6 contains that of Theorem 3.5 because the key step in this theorem is statement (3.22) and the proof of Theorem 3.6 relies solely on a more careful analysis of the limit in (3.22).

Theorem 3.6. *Let $\{X_n\}_n$ be a sequence of i.i.d. normal random variables. Then*

$$n(\mathcal{R}_n - a_n) \xrightarrow{w} -\frac{3}{2} + \sum_{j=3}^{\infty} \frac{Z_j^2 - 1}{j},$$

where $\{Z_n\}_n$ is a sequence of independent $N(0,1)$ random variables and

$$a_n = \frac{1}{n} \int_{\frac{1}{n+1}}^{\frac{n}{n+1}} \frac{t(1-t)}{[\phi(\Phi^{-1}(t))]^2} dt.$$

Proof. It suffices to show that the functional of the Brownian bridge in Theorem 3.5 has the same distribution as

$$-\frac{3}{2} + \sum_{j=3}^{\infty} \frac{Z_j^2 - 1}{j}.$$

The operator $L : L_2(0,1) \to L_2(0,1)$ defined by

$$Lf(t) := \int_0^1 \frac{s \wedge t - st}{\phi(\Phi^{-1}(s))\phi(\Phi^{-1}(t))} f(s)ds$$

has eigenvalues $\lambda_j = 1/j$, $j = 1, 2, \ldots$, with associated eigenfunctions $H_j(\Phi^{-1}(t))$, H_j being the j-th Hermite polynomial. Since $\{H_j(\Phi^{-1}(t))\}_{j=1}^{\infty}$ is a complete orthonormal system in $L_2(0,1)$, we have that

$$\int_{\frac{1}{n+1}}^{\frac{n}{n+1}} \left(\frac{B(t)}{\phi(\Phi^{-1}(\Phi^{-1}(t)))} \right)^2 dt = \sum_{j=1}^{\infty} \left(\int_{\frac{1}{n+1}}^{\frac{n}{n+1}} \frac{B(t)}{\phi(\Phi^{-1}(t))} H_j(\Phi^{-1}(t))dt \right)^2 .$$

The first two Hermite polynomials are $H_1(x) = 1$ and $H_2(x) = x$. Hence,

$$\begin{aligned}
W_n \quad := & \quad \int_{\frac{1}{n+1}}^{\frac{n}{n+1}} \left(\frac{B(t)}{\phi(\Phi^{-1}(t))} \right)^2 dt - E\left(\int_{\frac{1}{n+1}}^{\frac{n}{n+1}} \left(\frac{B(t)}{\phi(\Phi^{-1}(t))} \right)^2 dt \right) \\
& - \left(\int_{\frac{1}{n+1}}^{\frac{n}{n+1}} \frac{B(t)}{\phi(\Phi^{-1}(t))} dt \right)^2 - \left(\int_{\frac{1}{n+1}}^{\frac{n}{n+1}} \frac{B(t)\Phi^{-1}(t)}{\phi(\Phi^{-1}(t))} dt \right)^2 \\
= & \quad \sum_{j=3}^{\infty} \left(\left(\int_{\frac{1}{n+1}}^{\frac{n}{n+1}} \frac{B(t)H_j(\Phi^{-1}(t))}{\phi(\Phi^{-1}(t))} dt \right)^2 - E\left(\int_{\frac{1}{n+1}}^{\frac{n}{n+1}} \frac{B(t)H_j(\Phi^{-1}(t))}{\phi(\Phi^{-1}(t))} dt) \right)^2 \right) \\
& -E\left(\int_{\frac{1}{n+1}}^{\frac{n}{n+1}} \frac{B(t)H_1(\Phi^{-1}(t))}{\phi(\Phi^{-1}(t))} dt \right)^2 - E\left(\int_{\frac{1}{n+1}}^{\frac{n}{n+1}} \frac{B(t)H_2(\Phi^{-1}(t))}{\phi(\Phi^{-1}(t))} dt \right)^2 \\
:= & \quad \sum_{j=3}^{\infty} (Z_j(n)^2 - EZ_j(n)^2) - EZ_1(n)^2 - EZ_2(n)^2,
\end{aligned}$$

where the random variables $\{Z_j(n)\}_{j=1}^{M}$ have, for every fixed M, a joint M-dimensional Gaussian law, and their variances, $\sigma_j^2(1/n)$, satisfy

$$\sigma_j^2(1/n) = \int_{\frac{1}{n+1}}^{\frac{n}{n+1}} \int_{\frac{1}{n+1}}^{\frac{n}{n+1}} \frac{(s \wedge t - st)}{\phi(\Phi^{-1}(s))\phi(\Phi^{-1}(t))} H_j(\Phi^{-1}(s))H_j(\Phi^{-1}(t))dsdt \to \lambda_j = \frac{1}{j}.$$

Moreover,

$$\mathrm{Cov}(Z_j(n), Z_k(n)) = \int_{\frac{1}{n+1}}^{\frac{n}{n+1}} \int_{\frac{1}{n+1}}^{\frac{n}{n+1}} \frac{(s \wedge t - st)}{\phi(\Phi^{-1}(s))\phi(\Phi^{-1}(t))} H_j(\Phi^{-1}(s))H_k(\Phi^{-1}(t))dsdt \to 0$$

as $n \to \infty$. Therefore for every fixed M:

$$\sum_{j=3}^{M} (Z_j(n)^2 - EZ_j(n)^2) - EZ_1(n)^2 - EZ_2(n)^2 \overset{w}{\to} -\frac{3}{2} + \sum_{j=3}^{M} \frac{Z_j^2 - 1}{j},$$

where $Z_1, Z_2, \ldots, Z_M$ are independent $N(0,1)$ random variables.

Let us observe that

$$\begin{aligned}
&\operatorname{Var}\left(\int_{\frac{1}{n+1}}^{\frac{n}{n+1}} \frac{B^2(t) - EB^2(t)}{(\phi(\Phi^{-1}(t)))^2} dt\right) \\
&\qquad = 2\int_{\frac{1}{n+1}}^{\frac{n}{n+1}} \int_{\frac{1}{n+1}}^{\frac{n}{n+1}} \left(\frac{s \wedge t - st}{\phi(\Phi^{-1}(s))\phi(\Phi^{-1}(t))}\right)^2 dsdt \\
&\qquad \to 2\int_0^1 \int_0^1 \left(\frac{s \wedge t - st}{\phi(\Phi^{-1}(s))\phi(\Phi^{-1}(t))}\right)^2 dsdt \\
&\qquad = 2\int_0^1 \sum_{j=1}^{\infty} \left(\int_0^1 \frac{s \wedge t - st}{\phi(\Phi^{-1}(s))\phi(\Phi^{-1}(t))} H_j(\Phi^{-1}(s))ds\right)^2 dt \\
&\qquad = 2\sum_{j=1}^{\infty} \int_0^1 \lambda_j^2 H_j^2(\Phi^{-1}(s))ds = 2\sum_{j=1}^{\infty} \frac{1}{j^2},
\end{aligned}$$

while

$$\operatorname{Var}\left(\sum_{j=1}^{M} (Z_j(n)^2 - EZ_j(n)^2)\right) \to \operatorname{Var}\left(\sum_{j=1}^{M} \frac{Z_j^2 - 1}{j}\right) = 2\sum_{j=1}^{M} \frac{1}{j^2}.$$

On the other hand, taking into account that, if X and Y are standardized random variables with a joint normal law and covariance ρ, then $\operatorname{Cov}(X^2, Y^2) = 2\rho^2$, we obtain that

$$\begin{aligned}
&\operatorname{Cov}\left(\int_{\frac{1}{n+1}}^{\frac{n}{n+1}} \frac{B^2(t) - EB^2(t)}{\phi^2(\Phi^{-1}(t))} dt, Z_k^2(n)\right) \\
= \quad & \sum_{j=1}^{\infty} \operatorname{Cov}\left(Z_j^2(n), Z_k^2(n)\right) = 2\sum_{j=1}^{\infty} \left[E(Z_j(n)Z_k(n))\right]^2 \\
= \quad & 2\sum_{j=1}^{\infty} \left(\int_{\frac{1}{n+1}}^{\frac{n}{n+1}} \int_{\frac{1}{n+1}}^{\frac{n}{n+1}} \frac{s \wedge t - st}{\phi(\Phi^{-1}(s))\phi(\Phi^{-1}(t))} H_j(\Phi^{-1}(s)) H_k(\Phi^{-1}(t)) dtds\right)^2 \\
= \quad & 2\int_{\frac{1}{n+1}}^{\frac{n}{n+1}} \left(\int_{\frac{1}{n+1}}^{\frac{n}{n+1}} \frac{s \wedge t - st}{\phi(\Phi^{-1}(s))\phi(\Phi^{-1}(t))} H_k(\Phi^{-1}(t))dt\right)^2 ds \\
\to \quad & 2\int_0^1 \left(\int_0^1 \frac{s \wedge t - st}{\phi(\Phi^{-1}(s))\phi(\Phi^{-1}(t))} H_k(\Phi^{-1}(t))dt\right)^2 ds \\
= \quad & 2\int_0^1 \lambda_k^2 H_k^2(\Phi^{-1}(t))dt = 2\frac{1}{k^2},
\end{aligned}$$

and, consequently,

$$\operatorname{Cov}\left(\int_{\frac{1}{n+1}}^{\frac{n}{n+1}} \frac{B^2(t)-EB^2(t)}{\phi^2(\Phi^{-1}(t))}dt, \sum_{j=1}^{M} Z_j^2(n)\right) \to 2\sum_{j=1}^{M}\frac{1}{j^2}.$$

It is obvious that

$$\operatorname{Var}\left(\sum_{j=M+1}^{\infty} (Z_j^2(n) - EZ_j^2(n))\right) \to 2\sum_{j=M+1}^{\infty}\frac{1}{j^2}$$

and, as a consequence,

$$W_n \overset{w}{\to} -\frac{3}{2} + \sum_{j=3}^{\infty}\frac{Z_j^2-1}{j}.$$

□

The asymptotic equivalence of $\mathcal{R}_n$ with the Shapiro-Wilk, Shapiro-Francia or de Wet-Venter statistics, which can be obtained from the results of Leslie, Stephens and Fotopoulos (1986), or of Verrill and Johnson (1987), can be combined with Theorem 3.6 to obtain a new derivation of the asymptotic distribution of these statistics.

Our approach can give some light on the role played by the estimation of parameters in a minimum Wasserstein distance test. Let us consider once more decomposition (3.12). We have seen that 1 and $\Phi^{-1}(t)$ are eigenfunctions of the integral operator L in the proof of Theorem 3.6. Hence, in the normal case, in the limit $n\mathcal{R}_n^{(2)}$ and $n\mathcal{R}_n^{(3)}$ simply cancel out the first two terms in the principal components expansion of the limit law of $n\mathcal{R}_n^{(1)}$ (hence, in a certain sense, we can say that the estimation of parameters μ and σ results in a loss of two degrees of freedom). The normal law is the only distribution for which this cancellation holds.

More precisely, let F be an arbitrary distribution function with finite variance and density function f. Under the hypothesis of the cancellation of the first term in the orthogonal series expansion of

$$\int_0^1 \frac{B^2(t)-EB^2(t)}{(f(F^{-1}(t)))^2}dt - \left(\int_0^1 \frac{B(t)}{f(F^{-1}(t))}dt\right)^2 - \left(\int_0^1 \frac{B(t)F^{-1}(t)}{f(F^{-1}(t))}dt\right)^2,$$

$h_0(t) \equiv 1$ should be an eigenfunction of the operator

$$Lh(t) := \int_0^1 \frac{s \wedge t - st}{f(F^{-1}(t))f(F^{-1}(s))} h(s)ds.$$

Let $h(t) = g(F^{-1}(t))$ be an eigenfunction. By differentiating the equation above twice we obtain that g must satisfy

$$g''(x) + l(x)g'(x) + l'(x)g(x) = -\frac{1}{\lambda}g(x), \tag{3.23}$$

where $l(x) = \frac{d}{dx}(\log f(x))$. If $g(x) = 1$ is a solution of (3.23), then $l'(x) = -\lambda^{-1}$ and $l(x) = -\lambda^{-1}x + b$, from which $\log f(x) = -\lambda^{-1}x^2 + bx + c$, and necessarily (under the additional hypothesis of standardization) $f(x) = (2\pi)^{-1/2}e^{-x^2/2}$.

3.3.3 Heavy-tailed patterns

For distributions whose tails are heavier than the Gaussian the situation is more complex. An interesting fact already observed by several authors (Csörgő 1983; Stephens 1986b; McLaren and Lockhart 1987; Lockhart 1991) is the bad behaviour of correlation tests for heavy-tailed distributions. In fact, let us assume that a normalizing sequence, $b_n = o(n)$, is necessary to achieve a nondegerate limit law for $\mathcal{R}_n^*$; i.e., let us assume that $b_n\mathcal{R}_n^* - c_n \xrightarrow{w} V$, for some $c_n \in \mathbb{R}$. Then, by Theorem 3.2, for every fixed $\delta > 0$ we have the approximation

$$\left| \int_\delta^{1-\delta} \left(\frac{\rho_n(t)}{f(F^{-1}(t))} \right)^2 dt - \int_\delta^{1-\delta} \left(\frac{B_n(t)}{f(F^{-1}(t))} \right)^2 dt \right| \to 0$$

and thus

$$\frac{b_n}{n} \int_\delta^{1-\delta} \left(\frac{\rho_n(t)}{f(F^{-1}(t))} \right)^2 dt \to 0.$$

Therefore (recall inequalities (3.17) and also that $n\mathcal{R}_n(2) \xrightarrow{w} \chi_1^2$), the statistic $b_n\mathcal{R}_n^* - c_n$ has the same asymptotic behaviour as

$$\frac{b_n}{n}\left(\int_{[\delta,1-\delta]^c} \left(F_n^{-1}(t) - F^{-1}(t)\right)^2 dt - \left(\int_{[\delta,1-\delta]^c} \left(F_n^{-1}(t) - F^{-1}(t)\right) F^{-1}(t)dt \right)^2 \right) - c_n.$$

Hence, the asymptotic distribution of $\mathcal{R}_n^*$ depends only on the tails of the distribution, so that a sample with underlying distribution function different from F but with the same tails would be indistinguishable through this statistic. This simple observation offers a useful hint for the explanation of the poor performance of correlation tests when testing fit of heavy-tailed families of distributions. Correlation tests might still be of some use for these heavy-tailed families if one is interested in assessing departures from the null hypothesis concerning the shape of the tails.

The asymptotic distribution in the heavy-tailed case has been considered in several papers. For instance, Lockhart (1985) and McLaren and Lockhart (1987) have obtained the asymptotic normality of correlation tests for testing fit to the exponential, extreme value and logistic distributions at rate $\sqrt{\log n}$. We note that the exponential case could be almost trivially handled in our setup, using Theorem 5.4.3 ii) in Csörgő and Horváth (1993).

To conclude, we will provide an example showing that for heavy-tailed distributions we can obtain non-normal limit laws for $\mathcal{R}_n^*$. This fact, as far as we know, was previously unknown.

Example 3.1. Let

$$Q(x) = \begin{cases} (\sqrt{x}\log(x))^{-1}, & 0 < x < e^{-3}, \\ -(\sqrt{1-x}\log(1-x))^{-1}, & 1 - e^{-3} < x < 1. \end{cases}$$

We can assume that Q is also defined in $[e^{-3}, 1-e^{-3}]$, in such a way that it is a nondecreasing function of $\mathcal{C}^2$ in $(0,1)$ and satisfying $Q(1-x) = -Q(x)$ and $Q'(x) > 0$ for every $x \in (0,1)$ and $\int_0^1 Q^2(t)dt = 1$.

If we define $F = Q^{-1}$, then F is a distribution function (and Q its quantile function) with variance 1, which (as it can be easily checked) satisfies our Assumptions 1, 2 and 3. We will denote its density function by f.

We will analyze the behaviour of $\mathcal{R}_n^*$ for this example in the following propositions.

Proposition 3.4. *Let $\{X_n\}_n$ be a sequence of i.i.d. random variables with distribution function F, defined above. If $\{\psi_n^1\}_n$ and $\{\psi_n^2\}_n$ are two independent sequences of i.i.d. random variables with exponential distribution* *with* $E\psi_j^i = 1$, *and* $S_i(x) = \sum_{1\le j<x+1} \psi_j^i$, $i = 1, 2$, $x \ge 1$ *then*

$$\begin{aligned}(\log n)^2 &\left(\int_0^1 \left(F_n^{-1}(t) - F^{-1}(t)\right)^2 dt - \frac{2}{\log(n+1)} \right) \\ \xrightarrow{w} \Gamma := \quad &\frac{1}{\psi_1^1} - \frac{4}{\sqrt{\psi_1^1}} + \int_1^\infty \frac{1}{u}\left(\left(\frac{S_1(u)}{u}\right)^{-1/2} - 1 \right)^2 du \\ &+ \frac{1}{\psi_1^2} - \frac{4}{\sqrt{\psi_1^2}} + \int_1^\infty \frac{1}{u}\left(\left(\frac{S_2(u)}{u}\right)^{-1/2} - 1 \right)^2 du. \quad (3.24)\end{aligned}$$

Proof. Let $\{B_n(t)\}_n$ be the sequence of Brownian bridges of Theorem 3.2. Then

$$\left| \left(\frac{(\log n)^2}{n} \int_{\frac{\log n}{n}}^{1-\frac{\log n}{n}} \frac{\rho_n^2(t)}{(f(Q(t))^2} dt \right)^{1/2} - \left(\frac{(\log n)^2}{n} \int_{\frac{\log n}{n}}^{1-\frac{\log n}{n}} \frac{B_n^2(t)}{(f(Q(t))^2} dt \right)^{1/2} \right|$$

$$\le \left(\sup_{\frac{1}{n+1}\le t \le \frac{n}{n+1}} \frac{|\rho_n(t) - B_n(t)|}{(t(1-t))^{1/2}} \right) \left(\frac{(\log n)^2}{n} \int_{\frac{\log n}{n}}^{1-\frac{\log n}{n}} \frac{t(1-t)}{(f(Q(t)))^2} dt \right)^{1/2} \xrightarrow{p} 0.$$

This convergence follows from the fact that the first term in the bounding expression is $O_p(1)$, while

$$\frac{(\log n)^2}{n} \int_{\frac{\log n}{n}}^{\delta} \frac{t(1-t)}{4t^3(\log t)^2} dt \le \frac{(\log n)^2}{n} \int_{\frac{\log n}{n}}^{\delta} \frac{1}{4t^2(\log t)^2} dt \to 0,$$

which, in turn, implies that

$$E\left(\frac{(\log n)^2}{n} \int_{\frac{\log n}{n}}^{1-\frac{\log n}{n}} \frac{B_n^2(t)}{(f(Q(t))^2} dt \right) = \frac{(\log n)^2}{n} \int_{\frac{\log n}{n}}^{1-\frac{\log n}{n}} \frac{t(1-t)}{(f(Q(t)))^2} dt \to 0,$$

and, consequently, that

$$(\log n)^2 \int_{\frac{\log n}{n}}^{1-\frac{\log n}{n}} \left(F_n^{-1}(t) - F^{-1}(t)\right)^2 dt \xrightarrow{p} 0.$$

Using symmetry, in the remainder of the proof we will consider only the left tail integral, which we have split into two pieces.

Since Q varies regularly at 0 with exponent $-1/2$, the asymptotic theory of extremes (see, e.g., Galambos 1987, pp. 56) asserts that

$$\frac{\log n}{\sqrt{n}} X_{1n} \overset{w}{\rightarrow} L_{1,2} \tag{3.25}$$

(here, $L_{1,2}$ is the extreme value distribution defined, with the same notation, in Galambos 1987).

Using L'Hôpital's rule and (3.25) we can easily obtain

$$\begin{aligned} &(\log n)^2 \left(\int_0^{\frac{1}{n+1}} \left(F_n^{-1}(t) - F^{-1}(t)\right)^2 dt - \frac{1}{\log(n+1)} \right) \\ = \quad &(\log n)^2 \left(\frac{X_{in}^2}{n+1} + \int_0^{\frac{1}{n+1}} \frac{1}{t(\log t)^2} dt - 2X_{1n} \int_0^{\frac{1}{n+1}} \frac{1}{\sqrt{t}\log t} dt - \frac{1}{\log(n+1)} \right) \\ = \quad &(\log n)^2 \left(\frac{X_{in}^2}{n+1} - 2X_{1n} \int_0^{\frac{1}{n+1}} \frac{1}{\sqrt{t}\log t} dt \right) \overset{w}{\rightarrow} L_{1,2}^2 + 4L_{1,2}. \end{aligned}$$

On the other hand, the following equality

$$\lim_{t\to\infty} \frac{|t|f(t)}{F(t)} = \lim_{t\to 0} \frac{|Q(t)|f(Q(t))}{t} = \lim_{t\to 0} \frac{|Q(t)|}{tQ'(t)} = 2,$$

allows us to apply Theorem 6.4.5 ii) in Csörgő and Horváth (1993) (take $p = \gamma = 2, \tau = 0, L \equiv 1$) to obtain

$$(\log n)^2 \int_{\frac{1}{n+1}}^{\frac{\log n}{n}} \left(F_n^{-1}(t) - F^{-1}(t)\right)^2 dt \overset{w}{\rightarrow} \int_0^\infty \frac{1}{u} \left(\left(\frac{\tilde{S}(u)}{u} \right)^{-\frac{1}{2}} - 1 \right)^2 du,$$

where $\tilde{S}(u) := \sum_{1\leq j<u} \psi_j,\ u \geq 1$, for a sequence $\{\psi_j\}_j$, of i.i.d. exponentially distributed random variables with $E\psi_j = 1$.

Finally, taking into account the simultaneous character of the approximations used to prove the convergences above (based on Lemma 3.0.1 in Csörgő and Horváth 1993), with standard arguments about the asymptotic independence of functions of order statistics like Rossberg's lemma (see e.g. Lemma 5.1.4 in Csörgő and Horváth 1993), and some elementary calculus on distributions, we obtain (3.24).

As already observed, $n\mathcal{R}_n(2)\xrightarrow{w}\chi_1^2$, so that $(\log n)^2\mathcal{R}_n(2) \xrightarrow{P} 0$. On the other hand, the computations in the last proof, Schwarz's inequality and inequalities (3.17) easily show that

$$\begin{aligned}(\log n)^2\mathcal{R}_n(3) &= (\log n)^2 \left(\int_0^1 \left(F_n^{-1}(t) - F^{-1}(t)\right) F^{-1}(t)dt \right)^2 \qquad (3.26)\\ &\approx (\log n)^2 \left[\int_0^{\frac{1}{n}} \left(F_n^{-1}(t) - F^{-1}(t)\right) F^{-1}(t)dt \right.\\ &\quad \left. + \int_{1-\frac{1}{n}}^1 \left(F_n^{-1}(t) - F^{-1}(t)\right) F^{-1}(t)dt \right]^2 \to 4.\end{aligned}$$

This completes the proof of the following consequence of Proposition 3.4.

Proposition 3.5. *With the notation and hypotheses of Proposition 3.4 we have*

$$(\log n)^2 \left(S_n^2\mathcal{R}_n - \frac{2}{\log(n+1)} \right) \xrightarrow{w}\Gamma - 4. \qquad (3.27)$$

Obtaining the asymptotic behaviour of $\mathcal{R}_n$ from $S_n^2\mathcal{R}_n$ is not as easy now as in cases considered previously. What is obvious from (3.27) is that

$$(\log n)^2 \left(\mathcal{R}_n - \frac{2}{S_n^2 \log(n+1)} \right) \xrightarrow{w}\Gamma - 4, \qquad (3.28)$$

but the analysis of the asymptotic behavior of $S_n^2 \log(n+1)$ is not completely trivial. The conclusion, given in the following proposition, is amusing: the inclusion of S_n^2 contributes to the asymptotic law of $\mathcal{R}_n$ just canceling the -4 summand contributed by $\mathcal{R}_n^{(3)}$, and retrieving the original asymptotic law of $\mathcal{R}_n^{(1)}$.

Proposition 3.6. *With the notation and hypotheses of Proposition 3.4 we have*

$$(\log n)^2 \left(\mathcal{R}_n - \frac{2}{\log(n+1)} \right) \xrightarrow{w}\Gamma \qquad (3.29)$$

Proof. First note that, for t small enough, the quantile function, Q_{X^2}, associated with X_i^2 satisfies

$$\begin{aligned} Q_{X^2}(1-t) &= \inf\{x : t \geq P(X_i^2 > x)\} = \inf\left\{x : F(\sqrt{x}) \geq 1 - \frac{t}{2}\right\} \\ &= \left(Q(1-\frac{t}{2})\right)^2 = \frac{2}{t(\log t - \log 2)^2}. \end{aligned}$$

Therefore, $Q_{X^2}(1-t)$ is regularly varying at 0 with exponent -1 and the Central Limit Theory allows us to claim that

$$\frac{(\log n)^2}{n}\sum_{i=1}^n X_i^2 - b_n \xrightarrow{w} \gamma$$

for some distribution γ, where $b_n = (\log n)^2 EX_i^2 I_{\{X_1^2 \leq n/(\log n)^2\}}$. From this it is obvious that

$$\log n(S_n^2 - 1) \approx \frac{1}{\log n}\left(\frac{(\log n)^2}{n}\sum_{i=1}^n X_i^2 - b_n\right) + \frac{b_n}{\log n} - \log n \approx_p \frac{b_n}{\log n} - \log n.$$

Observe now that

$$\frac{b_n}{\log n} - \log n = -\log n\left(EX_i^2 I_{\{X_1^2 > \frac{n}{\log n)^2}\}}\right) = -(\log n)2\int_0^{\frac{1}{n}} \frac{1}{x(\log x)^2}dx \to -2.$$

Hence, $\log n(S_n^2 - 1) \xrightarrow{p} -2$ and

$$(\log n)^2\left(\frac{2}{S_n^2\log(n+1)} - \frac{2}{\log(n+1)}\right) = \frac{2(\log n)^2}{S_n^2\log(n+1)}(1 - S_n^2) \xrightarrow{p} 4,$$

which, combined with (3.28), shows (3.29).

□

Acknowledgements

Authors want to thank Professor Sándor Csörgő and Professor Richard Lockhart for pointing us some mistakes and omissions in the text.

References

Ali, M.M. (1974). Stochastic ordering and kurtosis measure. *Journal of the American Statistical Association*, **69**, 543-545.

Anderson, T.W. and D.A. Darling (1952). Asymptotic theory of certain "goodness of fit" criteria based on stochastic processes. *Annals of Mathematical Statistics*, **23**, 193-212.

Araujo, A. and E. Giné (1980). *The Central Limit Theorem for Real and Banach Valued Random Variables.* Wiley, New York.

Balanda, K.P. and H.L. McGillivray (1988). Kurtosis: A critical review. *American Statistician*, **42**, 111-119.

Bickel, P. and D. Freedman (1981). Some asymptotic theory for the bootstrap. *Annals of Statistics*, **9** 1196-1217.

Bickel, P. and W.R. van Zwet (1978). Asymptotic expansions for the power of distribution free tests in the two-sample problem. *Annals of Statistics*, **6**, 937-1004.

Billingsley, P. (1968). *Convergence of Probability Measures.* Wiley, New York.

Breiman, L. (1968). *Probability.* Addison-Wesley, Reading.

Brown, B. and T. Hettmansperger (1996). Normal scores, normal plots, and test for normality. *Journal of the American Statistical Society*, **91**, 1668-1675.

Burke, M.D., M. Csörgő, S. Csörgő and P. Révész (1979). Approximations of the empirical process when parameters are estimated. *Annals of Probability*, **7**, 790-810.

Chernoff, H. and E.L. Lehmann (1954). The use of maximum likelihood estimates in χ^2 tests of goodness of fit. *Annals of Mathematical Statistics*, **25**, 579-586.

Chibisov, D.M. (1964). Some theorems on the limiting behavior of empirical distribution functions. *Selected Translations in Mathematical Statistics and Probability*, American Mathematical Society, Providence, **6**, 147-156.

Cochran, W.G. (1952). The χ^2 test of goodness of fit. *Annals of Mathematical Statistics*, **23**, 315-345.

Cohen, A. and H.B. Sackrowitz (1975). Unbiasedness of the chi-square, likelihood ratio and other goodness of fit tests for the equal cell case. *Annals of Statistics*, **3**, 959-964.

Cramér, H. (1928). On the composition of elementary errors. Second paper: Statistical applications. *Skand. Aktuartidskr.*, **11**, 141-180.

Csörgő, S. (1981a). Limit behaviour of the empirical characteristic function.

Annals of Probability, **9**, 130-144.

Csörgő, S. (1981b). The empirical characteristic process when parameters are estimated. In: *Contributions to Probability* (Eugene Lukacs Festschrift; J. Gani and V.K. Rohatgi, eds.), pp. 215-230. Academic Press, New York.

Csörgő, M. (1983). *Quantile Processes with Statistical Applications.* SIAM.

Csörgő, S. (1986a). Testing for normality in arbitrary dimension. *Annals of Statistics*, **14**, 708-723.

Csörgő, S. (1986b). Testing for stability. In: *Colloquia Mathematica Societatis János Bolyai*, **45**. *Goodness of Fit* (P. Révész, K. Sarkadi and P.K. Sen, eds.), pp. 101-132. North-Holland, Amsterdam.

Csörgő, S. (1989). Consistency of some tests for multivariate normality. *Metrika*, **36**, 107-116.

Csörgő, M., S. Csörgő, L. Horváth and D.M. Mason (1986). Weighted empirical and quantile process. *Annals of Probability*, **14**, 31-85.

Csörgő, S. and J.J. Faraway (1996). The exact and asymptotic distributions of Cramér−von Mises statistics. *Journal of the Royal Statistical Society, B*, **58**, 221-234.

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

Csörgő, M., L. Horváth and Q.-M. Shao (1993). Convergence of integrals of uniform empirical and quantile processes. *Stochastic Processes and their Applications*, **45**, 283-294.

Csörgő, M. and P. Révész (1978). Strong approximations of the quantile process. *Annals of Statistics*, **6**, 882-894.

Cuesta-Albertos, J.A., C. Matrán, S.T. Rachev and L. Rüschendorf (1996). Mass transportation problems in Probability Theory. *The Mathematical Scientist*, **21**, 34-72.

D'Agostino, R.B. (1971). An omnibus test of normality for moderate and large sample sizes. *Biometrika*, **58**, 341-348.

Darling, D.A. (1955). The Cramér-Smirnov test in the parametric case. *Annals of Mathematical Statistics*, **26**, 1-20.

David, F.N. and N.L. Johnson (1948). The probability integral transformation when parameters are estimated from the sample. *Biometrika*, **35**, 182-190.

David, H.A., H.O. Hartley and E.S. Pearson (1954). The distribution of the ratio, in a single normal sample, of range to standard deviation. *Biometrika*, **41**, 482-493.

del Barrio, E. (2000). Asymptotic distribution of statistics of Cramér-von Mises type. *Preprint.*

del Barrio, E., J.A. Cuesta-Albertos, C. Matrán and J. Rodríguez-Rodríguez (1999). Tests of goodness of fit based on the L2-Wasserstein distance. *Annals of Statistics*, **27**, 1230-1239.

de Wet, T. and J. Venter (1972). Asymptotic distributions of certain test criteria of normality. *South African Statistics Journal*, **6**, 135-149.

de Wet, T. and J. Venter (1973). Asymptotic distributions for quadratic forms with applications to test of fit. *Annals of Statistics*, **2**, 380-387.

Donsker, M.D. (1951). An invariance principle for certain probability limit theorems. *Memoirs of the American Mathematical Society*, **6**.

Donsker, M.D. (1952). Justification and extension of Doob's heuristic approach to the Kolmogorov-Smirnov theorems. *Annals of Mathematical Statistics*, **23**, 277-281.

Doob, J.L. (1949). Heuristic approach to the Kolmogorov-Smirnov theorems. *Annals of Mathematical Statistics*, **20**, 393-403.

Downton, F. (1966). Linear estimates with polynomial coefficients. *Biometrika*, **53**, 129-141.

Dudley, R.M. (1978). Central limit theorems for empirical measures. *Annals of Probability*, **6**, 899-929.

Durbin, J. (1973). Weak convergence of the sample distribution function when parameters are estimated. *Annals of Statistics*, **1**, 279-290.

Epps, T.W. and L.B. Pulley (1983). A test for normality based on the empirical characteristic function. *Biometrika*, **70**, 723-726.

Feller, W. (1948). On the Kolmogorov-Smirnov limit theorems for empirical distributions. *Annals of Mathematical Statistics*, **19**, 177-189.

Feuerverger, A. and R.A. Mureika (1977). The empirical characteristic function and its applications. *Annals of Statistics*, **5**, 88-97.

Filliben, J.J. (1975). The probability plot correlation coefficient test for normality. *Technometrics*, **17**, 111-117.

Fisher, R.A. (1930). The moments of the distribution for normal samples of measures of departure from normality. *Proceedings of the Royal Society, A*, **130**, 16.

Galambos, J. (1987). *The Asymptotic Theory of Extreme Order Statistics*, second edition. Krieger, Melbourne, Florida.

Geary, R.C. (1947). Testing for normality. *Biometrika*, **34**, 209-242.

Gerlach, B. (1979). A consistent correlation-type goodness-of-fit test; with application to the two parameter Weibull distribution. *Mathematische Operationsforschung und Statistik Series Statistics*, **10**, 427-452.

Gumbel, E.J. (1943). On the reliability of the classical chi-square test. *Annals of Mathematical Statistics*, **14**, 253-263.

Gupta, A.K. (1952). Estimation of the mean and the standard deviation of a normal population from a censored sample. *Biometrika*, **39**, 260-273.

Hall, P. and A.H. Welsh (1983). A test for normality based on the empirical characteristic function. *Biometrika*, **70**, 485-489.

Kac, M., J. Kiefer, and J. Wolfowitz (1955). On tests of normality and other tests of goodness of fit based on distance methods. *Annals of Mathematical Statistics*, **26**, 189-211.

Kale, B.K. and G. Sebastian (1996). On a class of symmetric nonnormal distributions with kurtosis of three. In *Statistical Theory and Applications: Papers in Honor of Herbert A. David.* (H.H. Nagaraja, P.K. Sen and D. Morrison, eds.) Springer Verlag, New York.

Kolmogorov, A.N. (1933). Sulla determinazione empirica di una legge di distribuzione. *Giornale del Istituto Italiano degli Attuari*, **4**, 83-91.

Kolmogorov, A.N. and Y.V. Prohorov (1949). On sums of a random number of random terms. *Uspekhi Matematicheskikh Nauk*, Rossijskaya Akademiya Nauk, Moscow (in Russian), 4, 168-172.

Komlós, J., P. Major and G. Tusnády (1975). An approximation of partial sums of independent RV's and the sample DF. Part I. *Zeitschrift fWahrscheinlichkeitstheorie und Verwandte Gebiete*, **32**, 111-131.

Komlós, J., P. Major and G. Tusnády (1976). An approximation of partial sums of independent RV's and the sample DF. Part II. *Zeitschrift fWahrscheinlichkeitstheorie und Verwandte Gebiete*, **34**, 33-58.

Koutrouvelis, I.A. and S.G. Meintanis (1999). Testing for stability based on the empirical characteristic function with applications to financial data. *Journal of Statistical Computation and Simulation*, **64**, 275-300.

LaRiccia, V. and D.M. Mason (1986). Cramér – von Mises statistics based on the sample quantile function and estimated parameters. *Journal of Multivariate Analysis*, **18**, 93-106.

Leslie, J.R. (1984). Asymptotic properties and new approximations for both the covariance matrix of normal order statistics and its inverse. *Colloquia Mathematica Societatis János Bolyai*, **45**, 317-354. (P. Révész, K. Sarkadi and P.K. Sen. eds.) Elsevier, Amsterdam.

Leslie, J.R., M.A. Stephens and S. Fotopoulos (1986). Asymptotic distribution of the Shapiro-Wilk W for testing for normality. *Annals of Statistics*, **14**, 1497-1506.

Lilliefors, H.W. (1967). On the Kolmogorof-Smirnov test for normality with mean and variance unknown. *Journal of the American Statistical Association*, **62**, 399-402.

Lockhart, R.A. (1985). The asymptotic distribution of the correlation coefficient in testing fit to the exponential distribution. *Canadian Journal of Statistics*, **13**, 253-256.

Lockhart, R.A. (1991). Overweight tails are inefficient. *Annals of Statistics*, **19**, 2254-2258.

Lockhart, R.A. and M.A. Stephens (1998). The probability plot: Test of fit based on the correlation coefficient. In *Order statistics: applications*. Handbook of Statistics, **17**, 453–473. North-Holland, Amsterdam.

Mann, H.B. and A. Wald (1942). On the choice of the number of class intervals in the application of the chi square test. *Annals of Mathematical Statistics*, **13**, 306-317.

McLaren, C.G. and R.A. Lockhart (1987). On the asymptotic efficiency of certain correlation tests of fit. *Canadian Journal of Statistics*, **15**, 159-167.

Moore, D.S. (1971). A chi-square statistic with random cell boundaries. *Annals of Mathematical Statistics*, **42**, 147-156.

Moore, D.S. (1986). Tests of chi-squared type. In *Goodness-of-Fit Techniques* (R.B. d'Agostino and M.A. Stephens, eds.) North-Holland, Amsterdam. 63-96.

Murota, K. and K. Takeuchi (1981). The studentized empirical characteristic function and its application to test for the shape of distribution. *Biometrika*, **68**, 55-65.

O'Reilly, N.E. (1974). On the weak convergence of empirical processes in sup-norm metrics. *Annals of Probability*, **2**, 642-651.

Pearson, E.S. (1930). A further development of test for normality. *Biometrika*, **22**, 239-249.

Pearson, E.S., R.B. D'Agostino and K.O. Bowman (1977). Tests for departure from normality: Comparison of powers. *Biometrika*, **64**, 231-46.

Pollard, D. (1979). General chi-square goodness-of-fit tests with data-dependent cells. *Zeitschrift fWahrscheinlichkeitstheorie und Verwandte Gebiete*, **50**, 317-331.

Pollard, D. (1980). The mimnimum distance method of testing. *Metrika*, **27**, 43-70.

Prohorov, Y.V. (1953). Probability distributions in functional spaces. *Uspekhi Matematicheskikh Nauk*, Rossijskaya Akademiya Nauk, Moscow (in Russian), **8**, 165-167.

Prohorov, Y.V. (1956). The convergence of random processes and limit theorems in probability. *Theory of Probability and its Applications*, **1**, 157-214.

Rachev, S.T. and L. Rüschendorf (1998). *Mass transportation problems*, 2 Vols. Springer, New York.

Royston, J.P. (1982). An extension of Shapiro and Wilk's W test for normality to large samples. *Applied Statistics*, **31**, 115-124.

Sarkadi, K. (1975). The consistency of the Shapiro-Francia test. *Biometrika*, **62**, 445-450.

Shapiro, S.S. and R.S. Francia (1972). An approximate analysis of variance test of normality. *Journal of the American Statistical Association*, **67**, 215-216.

Shapiro, S.S. and M.B. Wilk (1965). An analysis of variance test for normality (complete samples). *Biometrika*, **52**, 591-611.

Shapiro, S.S. and M.B. Wilk (1968). Approximations for the null distribution of the W statistic. *Technometrics*, **10**, 861-866.

Shapiro, S.S., M.B. Wilk and H.J. Chen (1968). A comparative study of various tests for normality. *Journal of the American Statistical Association*, **63**, 1343-1372.

Shorack, G.R. and J.A. Wellner (1986). *Empirical Processes With Applications to Statistics.* Wiley, New York.

Skorohod, A.V. (1956). Limit theorems for stochastic processes. *Theory of Probability and its Applications*, **1**, 261-290.

Smirnov, N.V. (1936). Sur la distribution de ω^2 (Critérium de M.R. von Mises). *Comptes Rendus de l'Académie des Sciences Paris*, **202**, 449-452.

Smirnov, N.V. (1937). Sur la distribution de ω^2 (Critérium de M.R. von Mises). *Matematicheskij Sbornik*, Rossijskaya Akademiya Nauk, Moscow (in Russian with French summary), **2**, 973-993.

Smirnov, N.V. (1939). Sur les écarts de la courbe de distribution empirique. *Matematicheskij Sbornik*, Rossijskaya Akademiya Nauk, Moscow (in Russian with French summary), **6**, 3-26.

Smirnov, N.V. (1941). Approximate laws of distribution of random variables from empirical data. *Uspekhi Matematicheskikh Nauk*, Rossijskaya Akademiya Nauk, Moscow (in Russian), **10**, 179-206.

Smith, R.M. and L.J. Bain (1976). Correlation-type goodness-of-fit statistics with censored sampling. *Communications in Statistics-Theory and Methods*, **5**, 119-132.

Spinelli, J.J and M.A. Stephens (1987). Tests for exponentiality when origin and scale parameters are unknown. *Technometrics*, **29**, 471-476.

Stephens, M.A. (1974). EDF statistics for goodness of fit and some comparisons. *Journal of the American Statistical Association*, **69**, 730-737.

Stephens, M.A. (1975). Asymptotic properties for covariance matrices of order statistics. *Biometrika*, **62**, 23-28.

Stephens, M.A. (1986a). Tests based on EDF statistics. In *Goodness-of-Fit Techniques* (R.B. D'Agostino and M.A. Stephens, eds.) Marcel Dekker, New York. 97-193.

Stephens, M.A. (1986b). Tests based on regression and correlation. In *Goodness-of-Fit Techniques* (R.B. D'Agostino and M.A. Stephens, eds.) Marcel Dekker, New York. 195-233.

Sukhatme, S. (1972). Fredholm determinant of a positive definite kernel of a special type and its application. *Annals of Mathematical Statistics*, **43**, 1914-1926.

Uthoff, V.A. (1970). An optimum test property of two well known statistics. *Journal of the American Statistical Association*, **65**, 1597.

Uthoff, V.A. (1973). The most powerful scale and location invariant test of the normal versus the double exponential. *Annals of Statistics*, **1**, 170-174.

Vallender, S. (1973). Calculation of the Wasserstein distance between probability distributions on the line. *Theory of Probability Applications*, **18**, 785-786.

Verrill, S. and R. Johnson (1987). The asymptotic equivalence of some modified Shapiro-Wilk statistics - Complete and censored sample cases. *Annals of Statistics*, **15**, 413-419.

von Mises, R. (1931). *Wahrscheinlichkeitsrechnung.* Wein, Leipzig.

Watson, G.S. (1957). The χ^2 goodness-of-fit test for normal distributions. *Biometrika*, **44**, 336-348.

Watson, G.S. (1958). On chi-square goodness-of-fit tests for continuous distributions. *Journal of the Royal Statistical Society, B*, **20**, 44-61.

Weisberg, S. and C. Bingham (1975). An approximate analysis of variance test for non-normality suitable for machine calculation. *Technometrics*, **17**, 133-134.

Williams, P. (1935). Note on the sampling distribution of $\sqrt{\beta_1}$, where the population is normal. *Biometrika*, **27**, 269-271.

DISCUSSION

Sándor Csörgő
University of Szeged, Hungary

Comment: Testing for Weibull scale families as a test case for Wasserstein correlation tests

It is a real pleasure to have the privilege to congratulate the authors in the first round on a brilliant and stimulating contribution. The contribution is to two realms. One of these is the repertoire of basic statistical ideas, to which an appealing procedure is added for the time-honored problem of testing goodness of fit to a location-scale family of distributions, which is new in its generality and at the same time unifies and explains scattered results for testing normality, and which is based on the sample version of a normed minimal L_2-Wasserstein distance that is scale and shift invariant. The other is the probabilistic methodology of that considerable part of large-sample statistical theory which is based on the asymptotic behavior of empirical and quantile processes. Here, to determine the asymptotic distribution of their basic Wasserstein correlation test statistic, the authors' fine technique enables the method of weighted approximations to do most of the work; the powerful method is due to M. Csörgő, S. Csörgő, Horváth and Mason (1986), here the version for general quantile processes by M. Csörgő and Horváth (1993) is used.

Most fittingly, the twofold advance in the paper is nicely embedded into an historical sketch of the development of empirical and quantile process theory and that of the asymptotic theory of tests of fit, with particular emphasis on the interaction between the two, for which the present paper is an outstanding example indeed. History itself is well told in the first two sections of the paper and in Section 3.1. I only have one historical correction, noticed in the last minute: treating it as part of the physical sciences, Hilbert actually did refer to probability –through an article on insurance mathematics–, calling for an axiomatic foundation of "the theory of probabilities" as part of his Problem 6; the English translation of Hilbert's Paris lecture in 1900 is Hilbert (1902) as an additional reference below. The rest of the historical review in the paper is scholarly and insightful.

In the present note, praise is intended by emulation in the last four sections, following some lighter remarks in the first.

1 Testing for normality and uniformity

The masterly review in Section 3.2 of the traditional correlation and regression tests, in particular for normality, is a good introduction to the Wasserstein-distance procedures. On the other hand, given its asymptotic equivalence to the celebrated Shapiro-Wilk, Shapiro-Francia and de Wet-Venter tests, the resulting new test for normality described by del Barrio, Cuesta-Albertos, Matrán and Rodríguez-Rodríguez (1999) and in Section 3.3.2 here, will likely not only share the well-known "unreasonable effectiveness" of the former tests, but in fact the source it comes from, i.e. the underlying Wasserstein motivation also explains that the powerful nature of the earlier tests may very well be reasonable. Hence the stereotyped call for a simulation study (by the authors' eager students) in such discussions, both for the precision of the asymptotic distribution under normality and for the power under alternatives, is probably in order here. It would indeed be of interest to compare the test not only with those already beaten by its successful early versions of the Shapiro-Wilk type, but with later consistent tests such as those of Epps and Pulley (1983) and Csörgő (1986a) for example, and the recent adaptive versions of Neyman's smooth tests, as applied to normality, by Kallenberg and Ledwina (1997) and their references. The proof of Theorem 3.6 is a piece of beauty along with the added discussion concerning the structure and relationship between the two representations of the distributional limit, shedding extra new light on the probably high power for normality and the possible lack of it against alternatives for other location-scale families. To see how "robust" the power of the test for testing normality, it would also be of interest to extend simulations to the tests for Weibull scale families, entertained in Section 4 below, for shape parameters α not too far from 2.

One should also emphasize the attractiveness of the test for uniformity in the present paper *on some interval*, neither the center nor the length of which is fixed in advance. It is interesting that the structure of the limiting random variable nicely represents the "evolution" of the problem: the term $12(\int_0^1 B(t)dt)^2$ subtracted in the limiting random variable $\mathcal{R}$, say, in (3.18) is Watson's 'price' to make the original Cramér-von Mises statistic circularly invariant, i.e. for estimating the location of the interval of uniformity with a fixed length, while the term $144(\int_0^1\{t-2^{-1}\}B(t)dt)^2$ subtracted is the present 'price' paid for estimating also the length. An important problem arises here, under the condition of Theorem 3.4: For some coefficient functions $\psi_1(\cdot), \psi_2(\cdot), \ldots$, is there a complete asymptotic expansion of the form $P\{n\mathcal{R}_n \le x\} = P\{\mathcal{R} \le x\} + \sum_{j=1}^{r} \psi_{2j}(x) n^{-j} + O(n^{-r-1})$, $x \in \mathbb{R}$, for any fixed $r \in \mathbb{N}$? While it appears a safe bet that Götze's (1979) Hilbert-space techniques with later improvements will produce an asymptotic expansion of the usual form $P\{n\mathcal{R}_n \le x\} = P\{\mathcal{R} \le x\} + \sum_{j=1}^{r} \psi_j(x) n^{-j/2} + O(n^{-(r+1)/2})$, $x \in \mathbb{R}$, the real question is whether $\psi_1(\cdot) = \psi_3(\cdot) = \cdots = 0$, as in the asymptotic expansions both for the original Cramér-von Mises statistic and for Watson's modification, recently discussed by Csörgő and Faraway (1996). In particular: Is the rate of convergence $O(1/n)$ rather than the more customary $O(1/\sqrt{n})$? This problem is not for an easy rejoinder.

2 Three questions

Clearly, in relation with del Barrio, Cuesta-Albertos, Matrán and Rodríguez-Rodríguez (1999), the main aim of the paper under discussion, besides the historical overview, is to understand the behavior of their normality test in a broader picture and thus to see how far their correlation tests for testing for a location-scale family, suggested by minimized L_2-Wasserstein distances, or simply the *Wasserstein correlation tests*, may go beyond testing for normality. It is made plain by the paper that statistically reasonable versions of the resulting test procedures are rather demanding on the tails. Indeed, underlying distributions with slightly more than finite second moment must be termed in Section 3.3.3 "heavy tailed" from the point of view of the asymptotic distribution of the main test statistic $\mathcal{R}_n$. Theorem 3.3, the umbrella result for the statistically reasonable best versions where no centering sequence is needed for $n\mathcal{R}_n$, is of course readily applicable to testing for uniformity, where the support is finite and hence the tails are the lightest possible, but already the normal tail requires some adjustment in the form of a centering sequence which goes to infinity (though at the very slow rate of $\log\log n$). Hence the first question: is there a statistically meaningful domain between the uniform and the normal distributions to which Theorem 3.3 (or its variant, Theorem 1 below for pure scale families) still applies directly? Next, the normal distribution appears on the boundary of reasonable asymptotic behavior in terms of that of $n\mathcal{R}_n$. Is the normal tail the only one for which the type of behavior in Theorem 3.5 obtains, or is there a whole range of tail orders resulting in such a behavior? Finally, the jump from good behavior under normality to the truly amazing asymptotic extreme-value theoretic limiting distributions, with a very slow stochastic order of convergence, under "heavy tails" appears too harsh. Is there something in between, possibly another range of tail grades, connecting the two?

We figured that the best way to appreciate this very fine paper and to help disseminate its beautiful ideas is to try and contribute within its framework by typifying answers to these three questions. This is done in Section 4 below on a single class of Weibull scale families, indexed by a shape parameter, which class, at least from the point of view of reasonable asymptotic behavior under the null hypothesis, is entertained here as a test case or testing ground for the Wasserstein correlation tests themselves. First we must reformulate Theorem 3.3 in order to adjust the general framework to scale families.

3 Goodness of fit to a scale family

In a distinctive class of fitting problems location as nuisance is not significant: it is not that we necessarily know the mean, but that we know one endpoint of the support. This is the case with life distributions where the beginning of time is either known or directly set by the experimenter, which is then usually convenient to regard as time zero, as for the Weibull families in the next section. Scale families are in fact simpler than location-scale families treated in the paper, so the reader may find their present discussion instructive in understanding the main Wasserstein ideas of the authors.

Let $G(x)$, $x \in \mathbb{R}$, be any fixed distribution function with a finite second moment $\mu_2(G) = \int_{\mathbb{R}} x^2 dG(x)$, and consider the scale family $\mathcal{G} = \{G_s(x) = G(x/s), x \in \mathbb{R} : s > 0\}$ generated by G, which is supposed to have a density function $g = G'$ on $\mathbb{R}$. Assuming throughout that $\mu_2(G) > 0$ and considering any distribution function F for which $0 < \mu_2(F) < \infty$ is also satisfied, since $G_s^{-1}(\cdot) = sG^{-1}(\cdot)$ the argument in (3.11) reduces to

$$\begin{aligned} d^2(F,\mathcal{G}) &:= \frac{\mathcal{W}^2(F,\mathcal{G})}{\mu_2(F)} = \frac{\inf_{s>0} \mathcal{W}^2(F,G_s)}{\mu_2(F)} \\ &= \frac{1}{\mu_2(F)} \inf_{s>0} \int_0^1 \left[F^{-1}(t) - sG^{-1}(t)\right]^2 dt \end{aligned}$$

$$= \frac{1}{\mu_2(F)} \inf_{s>0} \left\{ \mu_2(F) + s^2\mu_2(G) - 2s \int_0^1 F^{-1}(t)G^{-1}(t)dt \right\}$$

$$= 1 - \frac{\left[\int_0^1 F^{-1}(t)G^{-1}(t)dt\right]^2}{\mu_2(F)\mu_2(G)},$$

so that $d^2(F_c, \mathcal{G}) = d^2(F, \mathcal{G})$ for any $c > 0$ whether F is a deterministic or random distribution function, in particular also when F is replaced by the sample distribution function $F_n(x) = \sum_{k=1}^n I\{X_k \leq x\}/n$, $x \in \mathbb{R}$, pertaining to a sample $X_1, \ldots, X_n$ from F. Thus the distribution of the natural test statistic

$$T_n := d^2(F_n, \mathcal{G}) = 1 - \frac{\left[\int_0^1 F_n^{-1}(t)G^{-1}(t)dt\right]^2}{\mu_2(F_n)\mu_2(G)} = 1 - \frac{\left[\int_0^1 F_n^{-1}(t)G^{-1}(t)dt\right]^2}{\mu_2(G)\,\frac{1}{n}\sum_{k=1}^n X_k^2}$$

under the null-hypothesis $F \in \mathcal{G}$ is the same whichever way this hypothesis is satisfied, and hence one may assume that $F = G$ when deriving this distribution. Since, in this case, when $f = F' = G' = g$, we have $\mu_2(F_n)\mu_2(G) \to \mu_2^2(G)$ almost surely, all asymptotic relations being understood as $n \to \infty$ unless otherwise specified, one now begins with

$$\begin{aligned} T_n^* &:= \mu_2(F_n)\mu_2(G)\,T_n \\ &= \left[\int_0^1 \{F_n^{-1}(t)\}^2 dt\right]\left[\int_0^1 \{F^{-1}(t)\}^2 dt\right] - \left[\int_0^1 F_n^{-1}(t)F^{-1}(t)dt\right]^2 \\ &= \mu_2(G)\int_0^1 \{F^{-1}(t) - F_n^{-1}(t)\}^2 dt - \left[\int_0^1 \{F^{-1}(t) - F_n^{-1}(t)\}F^{-1}(t)dt\right]^2, \end{aligned}$$

obtained as an analogue of (3.12), and hence (3.13) presently reduces to

$$nT_n^* = \int_0^1 \frac{\rho_n^2(t)}{g^2(G^{-1}(t))}\,dt - \left[\int_0^1 \rho_n(t)\,\frac{G^{-1}(t)}{g(G^{-1}(t))}\,dt\right]^2$$

with the general quantile process $\rho_n(\cdot) = \sqrt{n}\,f(F^{-1}(\cdot))\{F^{-1}(\cdot) - F_n^{-1}(\cdot)\}$ on $(0,1)$. Hence the variant of Theorem 3.3 for a scale family is the following

Theorem 1. *Under Assumptions 1, 2, 3, (3.14) and (3.16), if $F \in \mathcal{G}$, then*

$$nT_n \xrightarrow{w} \frac{1}{\mu_2^2(G)}\left\{ \int_0^1 \left[\frac{B(t)}{g(G^{-1}(t))}\right]^2 dt - \left[\int_0^1 B(t)\,\frac{G^{-1}(t)}{g(G^{-1}(t))}\,dt\right]^2 \right\}.$$

As is clear from the paper, the direct applicability of this result is just as limited as that of Theorem 3.3: condition (3.14) is often violated, or, equivalently, the first term of the limiting distribution here blows up almost surely; even underlying normal tails are not light enough to suffice! The following section is thought to exhibit a kind of a gentle, "smooth" transition first between Sections 3.3.1 (uniformity) and 3.3.2 (normality) and then from Section 3.3.2 towards Section 3.3.3 (heavy tails) in the paper, demonstrated on a single class of examples that is of some traditional importance in modelling life distributions.

4 Weibull scale families

For every $\alpha > 0$, let the scale family $\mathcal{G}_\alpha = \{G_{\alpha,s}(x) = G_\alpha(x/s), x \in \mathbb{R} : s > 0\}$ be generated by $G_\alpha(x) = 1 - e^{-x^\alpha}$, $x \geq 0$, the distribution function of the power of order $1/\alpha$ of an exponentially distributed random variable with mean one, with density function $g_\alpha(x) = \alpha x^{\alpha-1} e^{-x^\alpha}$, $x > 0$. With the usual gamma function $\Gamma(u) = \int_0^\infty x^{u-1} e^{-x} dx$, $u > 0$, the moment of order $\beta > 0$ of $G_{\alpha,s} \in \mathcal{G}_\alpha$ is

$$\mu_\beta(G_{\alpha,s}) = \int_{-\infty}^{\infty} x^\beta dG_{\alpha,s}(x) = s^\beta \, \Gamma\Big(1 + \frac{\beta}{\alpha}\Big),$$

while

$$G_{\alpha,s}^{-1}(t) = s \log^{\frac{1}{\alpha}} \frac{1}{1-t}, \quad 0 \leq t < 1,$$

for its quantile function, and the reciprocal quantile-density functions are given by

$$\begin{aligned} g_\alpha\big(G_\alpha^{-1}(t)\big) &= g_{\alpha,1}\big(G_{\alpha,1}^{-1}(t)\big) \\ &= s g_{\alpha,s}\big(G_{\alpha,s}^{-1}(t)\big) = \alpha\,(1-t) \log^{1-\frac{1}{\alpha}} \frac{1}{1-t}, \qquad 0 \leq t < 1, \end{aligned}$$

for all $s > 0$, where $\log^u x = (\log x)^u$, $x > 0$, for any $u \in \mathbb{R}$; analogous notation will be used for powers of other functions.

Note first that Assumptions 1 and 2 are satisfied on the common support

$(0,\infty)$ and

$$\begin{aligned}
\sup_{0<t<1} \frac{t(1-t)\big|g'_\alpha\big(G_\alpha^{-1}(t)\big)\big|}{g_\alpha^2\big(G_\alpha^{-1}(t)\big)} &= \sup_{0<t<1} \frac{t(1-t)\big|g'_{\alpha,s}\big(G_{\alpha,s}^{-1}(t)\big)\big|}{g_{\alpha,s}^2\big(G_{\alpha,s}^{-1}(t)\big)} \\
&= \sup_{0<t<1} \left|\frac{\alpha-1}{\alpha}\frac{t}{\log\frac{1}{1-t}} - t\right| < \infty,
\end{aligned}$$

so that Assumption 3 is also satisfied for any $\alpha, s > 0$. Thus the weighted approximation method based on Theorem 3.2 is applicable in principle for all $\alpha > 0$. Next, considering the order statistics $X_{1n} < \cdots < X_{nn}$ of a sample $X_1, \ldots, X_n$ from $G_\alpha(\cdot)$, we have

$$\begin{aligned}
L_n^{(\alpha)} &:= \int_0^{\frac{1}{n+1}} \left[X_{1n} - \log^{\frac{1}{\alpha}}\frac{1}{1-t}\right]^2 dt \\
&= \frac{X_{1n}^2}{n+1} - 2X_{1n}\int_1^{1+\frac{1}{n}} \frac{\log^{\frac{1}{\alpha}} x}{x^2}\,dx + \int_1^{1+\frac{1}{n}} \frac{\log^{\frac{2}{\alpha}} x}{x^2}\,dx,
\end{aligned}$$

indicating the way we like to work, so, since $P\{n^{1/\alpha}X_{1n} \le x\} = G_\alpha(x)$ for all $x \in \mathbb{R}$,

$$nL_n^{(\alpha)} = O_P\!\left(\frac{1}{n^{2/\alpha}}\right) \quad \text{for all} \quad \alpha > 0. \tag{1}$$

Hence, as expected in view of the finite starting point 0 of the support, no problem arises for any $\alpha > 0$ for that half of condition (3.16) which concerns the left tail.

However, it is well known and an amusing exercise to derive directly that

$$P\Big\{\alpha(\log n)^{1-\frac{1}{\alpha}}X_{nn} - \alpha\log n \le x\Big\} \to e^{-e^{-x}} \quad \text{for all} \quad x \in \mathbb{R},$$

and hence we have $|X_{nn} - \log^{\frac{1}{\alpha}} n| = O_P(1/\log^{1-\frac{1}{\alpha}} n)$. This is for the term

$$\begin{aligned}
R_n^{(\alpha)} &:= \int_{1-\frac{1}{n+1}}^{1} \left[X_{nn} - \log^{\frac{1}{\alpha}}\frac{1}{1-t}\right]^2 dt \\
&\le \frac{\big[X_{nn} - \log^{\frac{1}{\alpha}} n\big]^2}{n} - 2I_{1,n}(\alpha)\big[X_{nn} - \log^{\frac{1}{\alpha}} n\big] + I_{2,n}(\alpha)
\end{aligned}$$

responsible for the right tail, where $I_{\ell,n}(\alpha) = \int_n^\infty [\log^{\frac{1}{\alpha}} x - \log^{\frac{1}{\alpha}} n]^\ell x^{-2}\, dx$, $\ell = 1, 2$. Integrating by parts and using Karamata's theorem to obtain the asymptotic equality,

$$I_{1,n}(\alpha) = \frac{1}{\alpha} \int_n^\infty \frac{\log^{\frac{1}{\alpha}-1} x}{x^2}\, dx \sim \frac{1}{\alpha} \frac{\log^{\frac{1}{\alpha}-1} n}{n},$$

and if we now integrate by parts four times and then use Karamata's theorem, we obtain

$$\begin{aligned} I_{2,n}(\alpha) &= \frac{2}{\alpha}\left[\frac{2}{\alpha} - 1\right] \int_n^\infty \frac{\log^{\frac{2}{\alpha}-2} x}{x^2}\, dx \\ &- \frac{2}{\alpha}\left[\frac{1}{\alpha} - 1\right] (\log n)^{\frac{1}{\alpha}} \int_n^\infty \frac{\log^{\frac{1}{\alpha}-2} x}{x^2}\, dx \sim \frac{2}{\alpha^2} \frac{\log^{\frac{2}{\alpha}-2} n}{n}. \end{aligned}$$

Therefore,

$$nR_n^{(\alpha)} = O_P\left(\frac{1}{\log^\gamma n}\right) \qquad \text{for all } \alpha > 0, \tag{2}$$

where

$$\gamma = \gamma(\alpha) := 2 - \frac{2}{\alpha} \begin{cases} > 0, & \text{if } \alpha > 1, \\ = 0, & \text{if } \alpha = 1, \\ < 0, & \text{if } \alpha < 1. \end{cases}$$

Since the only inequality in the derivation is $\int_{1-\frac{1}{n+1}}^1 \leq \int_{1-\frac{1}{n}}^1$ for the original integral, this stochastic order is precise, and hence we see in particular that the other half of condition (3.16), regarding the right tail, is satisfied if and only if $\alpha > 1$. Furthermore,

$$\begin{aligned} \int_0^1 \frac{t(1-t)}{g_\alpha^2(G_\alpha(t))}\, dt &= \frac{1}{\alpha^2} \int_0^1 \frac{t(1-t)}{(1-t)^2 \log^{2-\frac{2}{\alpha}} \frac{1}{1-t}}\, dt \\ &= \frac{1}{\alpha^2} \int_1^\infty \frac{x-1}{x} \frac{1}{x \log^{2-\frac{2}{\alpha}} x}\, dx, \end{aligned}$$

which is finite, and hence condition (3.14) is satisfied, if and only if $\alpha > 2$. This is the case, therefore, to which Theorem 1 is applicable, implying part (i) of Theorem 2 below.

For $\alpha \le 2$ one cannot avoid going through first the whole procedure of the weighted Gaussian approximation of $\rho_n(\cdot)$ by $B_n(\cdot)$ from Theorem 3.2 in the middle term

$$nM_n^{(\alpha)} := \frac{1}{\alpha^2}\left\{\int_{\frac{1}{n+1}}^{\frac{n}{n+1}} \frac{\rho_n^2(t)}{(1-t)^2 \log^{2-\frac{2}{\alpha}} \frac{1}{1-t}}\, dt - \left[\int_{\frac{1}{n+1}}^{\frac{n}{n+1}} \frac{\rho_n(t)}{(1-t)\log^{1-\frac{2}{\alpha}} \frac{1}{1-t}}\, dt\right]^2\right\}$$

of the present Weibull version of $nT_n^* = nT_n^*(\alpha) = nL_n^{(\alpha)} + nM_n^{(\alpha)} + nR_n^{(\alpha)}$, to see whether this can be replaced by

$$Y_n^{(\alpha)} := \frac{1}{\alpha^2}\left\{\int_{\frac{1}{n+1}}^{\frac{n}{n+1}} \frac{B_n^2(t)}{(1-t)^2 \log^{2-\frac{2}{\alpha}} \frac{1}{1-t}}\, dt - \left[\int_{\frac{1}{n+1}}^{\frac{n}{n+1}} \frac{B_n(t)}{(1-t)\log^{1-\frac{2}{\alpha}} \frac{1}{1-t}}\, dt\right]^2\right\}$$

for the determination of the asymptotic distribution of the basic Weibull test statistic

$$T_n = T_n(\alpha) = 1 - \frac{\left[\int_0^1 F_n^{-1}(t) \log^{\frac{1}{\alpha}} \frac{1}{1-t}\, dt\right]^2}{\Gamma(1+\frac{2}{\alpha})\ \frac{1}{n}\sum_{k=1}^n X_k^2} = 1 - \frac{\left[\sum_{k=1}^n X_{kn} \int_{\frac{n}{n-k+1}}^{\frac{n}{n-k}} \frac{\log^{1/\alpha} x}{x^2}\, dx\right]^2}{\Gamma(1+\frac{2}{\alpha})\ \frac{1}{n}\sum_{k=1}^n X_k^2}.$$

Adjusting to our Weibull situation the whole proof in del Barrio, Cuesta-Albertos, Matrán and Rodríguez-Rodríguez (1999) step by step, on the probability space of Theorem 3.2 we obtain that

$$\left|nM_n^{(\alpha)} - Y_n^{(\alpha)}\right| = O_P\!\left(\frac{1}{\log^\gamma n}\right) \qquad \text{for all } \alpha \in (0,2], \tag{3}$$

with

$$\gamma = 2 - \frac{2}{\alpha}\begin{cases} > 0, & \text{if } \alpha > 1, \\ = 0, & \text{if } \alpha = 1, \\ < 0, & \text{if } \alpha < 1, \end{cases}$$

remarkably the same as (2). In the derivation we separate the three cases $\alpha \in (1,2]$, $\alpha = 1$ and $\alpha \in (0,1)$, and use details similar to those for (2), such as the Karamata theorem. (The winning rate is that of the term corresponding to $L_n^{(1)}$ in that proof, being the common order of our versions of $A_n^{(1)}$ and $A_n^{(2)}$. For the term corresponding to $L_n^{(3)}$ there the proof must be modified, avoiding the analogue of (2.9) there, and it turns out that this term is $O(n^{\nu-\frac{1}{2}})$ for any $\nu \in (0,1/2)$, and hence dominated by the other term.)

Recalling $\gamma = \gamma(\alpha) = 2 - 2/\alpha$, everything depends now on the finiteness of the integral

$$I(\alpha) := \int_0^1 \int_0^1 \frac{2[\min(u,v) - uv]^2}{g_\alpha^2(G_\alpha^{-1}(u))\, g_\alpha^2(G_\alpha^{-1}(v))}\, du\, dv$$

$$= \frac{2}{\alpha^4} \int_0^1 \left[\frac{1}{\log^\gamma \frac{1}{1-u}} \int_0^u \frac{v^2}{(1-v)^2 \log^\gamma \frac{1}{1-v}}\, dv + \frac{u^2}{(1-u)^2 \log^\gamma \frac{1}{1-u}} \int_u^1 \frac{1}{\log^\gamma \frac{1}{1-v}}\, dv \right] du$$

$$= \frac{2}{\alpha^4} \int_1^\infty \left[\frac{1}{x^2 \log^\gamma x} \int_1^x \left\{ \frac{y-1}{y} \right\}^2 \frac{1}{\log^\gamma y}\, dy + \left\{ \frac{x-1}{x} \right\}^2 \frac{1}{\log^\gamma x} \int_x^\infty \frac{1}{y^2 \log^\gamma y}\, dy \right] dx$$

because, in the present situation, this is what ensures that the centering idea of Theorem 3.5 works. Since it is trivial that $I(1) = \infty$ for the exponential distribution when $\gamma = 0$ (in which case $nR_n^{(1)} = O_P(1)$ and so we already know that, with the magnifying factor n in front of $T_n^*(1)$, the right tail cannot disappear anyway), we must restrict α to the interval $(1, 2]$, so that $0 < \gamma \leq 1$. Some detail is necessary here since the resulting turning point in Theorem 2 below is not only unexpected, but in fact is hardly believable at first sight. This is particularly so in view of the fact that the bounds in (1)–(3) all go to zero for any $\gamma \in (0, 1]$ and hence, were $I(\alpha)$ finite for all of this range, would allow every $\alpha \in (1, 2]$ in part (ii) of the theorem. Cutting both of the outer integrals into two pieces at $x = e$, and then also the inner integral at $y = e$ in one of the integrals obtained from the first, it is easy to see that $I(\alpha) < \infty$ if and only if both

$$J_1(\alpha) := \int_e^\infty \left[\frac{1}{x^2 \log^\gamma x} \int_e^x \frac{1}{\log^\gamma y}\, dy \right] dx$$

and

$$J_2(\alpha) := \int_e^\infty \left[\frac{1}{\log^\gamma x} \int_x^\infty \frac{1}{y^2 \log^\gamma y}\, dy \right] dx$$

are finite. Clearly, $J_1(\alpha) = \infty$ if $2\gamma \leq 1$. On the other hand, if $2\gamma > 1$, it is equally obvious that $J_2(\alpha) < \infty$ and a somewhat more involved analysis shows that $J_1(\alpha) < \infty$ as well. Thus $I(\alpha) < \infty$ if and only if $1/2 < \gamma \leq 1$, which happens if and only if $4/3 < \alpha \leq 2$. In this case the ingredients (1)–(3) above and the main line of the proof of Theorem 3.5 (Theorem 2 in del Barrio, Cuesta-Albertos, Matrán and Rodríguez-Rodríguez 1999) give part (ii) of Theorem 2 below, once we ascertain that

$c_n(\alpha)|\mu_2(F_n) - \mu_2(G_\alpha)| \xrightarrow{P} 0$, which is required to transfer the limit theorem for $T_n^*(\alpha)$ to that for $T_n(\alpha)$. Since $|\mu_2(F_n) - \mu_2(G_\alpha)| = O_P(1/\sqrt{n})$ by the central limit theorem, this follows in view of the asymptotic order of the centering sequence $c_n(\alpha)$. The exact asymptotic behavior of $c_n(\alpha)$ comes by elementary analysis, particularly enjoyable for $\alpha = 2$. With part (ii) we see that it is not just the case $\alpha = 2$, giving a right tail closest to a normal tail (already slightly longer than that), which exhibits the behavior in Theorem 3.5, but a whole weird range of shape parameters.

So, the first random integral in the limit of part (ii) makes sense now in $L_2(\Omega, \mathcal{A}, P)$, if $(\Omega, \mathcal{A}, P)$ denotes the underlying probability space for $B(\cdot)$, while, since

$$\int_0^1 \frac{\sqrt{t(1-t)}}{1-t} \log^{\frac{2}{\alpha}-1} \frac{1}{1-t}\, dt = \int_1^\infty \sqrt{\frac{x-1}{x}}\, \frac{\log^{\frac{2}{\alpha}-1} x}{x^{\frac{3}{2}}}\, dx < \infty \quad \text{for all } \alpha > 0, \tag{4}$$

the subtracted squared integral still makes sense almost surely by an application of Lemma 5.3.2 in Csörgő and Horváth (1993).

The first integral in (ii) blowing up for $\alpha = 4/3$ even in $L_2(\Omega, \mathcal{A}, P)$, adjustment by mere centering is exhausted. Thus the next idea, for $\alpha \leq 4/3$, is to consider an extra normalization as well, that is, to look at, on the probability space of Theorem 3.2,

$$\frac{nT_n^*(\alpha) - m_n(\alpha)}{d_n(\alpha)} = \frac{Z_n(\alpha) - m_n(\alpha)}{d_n(\alpha)} + O_P\left(\frac{1}{d_n(\alpha)\log^{\gamma(\alpha)} n} + \frac{1}{d_n(\alpha)}\right) \tag{5}$$

for $m_n(\alpha) := E(Z_n(\alpha))$ and $d_n^2(\alpha) := \mathrm{Var}(Z_n(\alpha))$, where $\gamma = \gamma(\alpha) = 2 - 2/\alpha$ as before and

$$\begin{aligned} Z_n(\alpha) &:= \int_0^{\frac{n}{n+1}} \frac{B_n^2(t)}{(1-t)^2 \log^{2-\frac{2}{\alpha}} \frac{1}{1-t}}\, dt \\ &\overset{\mathcal{D}}{=} \int_0^{\frac{n}{n+1}} \frac{W^2\left(\frac{t}{1-t}\right)}{\log^\gamma \frac{1}{1-t}}\, dt = \int_0^n \frac{W^2(u)}{(1+u)^2 \log^\gamma (1+u)}\, du. \end{aligned}$$

Here we used all of the findings in (1)–(4), and an extra little consideration when changing the original lower limit $1/(n+1)$ of the integral to 0 in the expression for $Z_n(\alpha)$, while the distributional equality is by Doob's transformation, which states that for a standard Wiener process $W(u)$, $u \geq 0$, the process $(1-t)W(t/(1-t))$, $0 \leq t < 1$, is a Brownian bridge. Of course, $m_n(\alpha) = \alpha^2\Gamma^2(1+2/\alpha)c_n(\alpha) + o(1)$, and the asymptotic expression for $c_n(\alpha)$ in case (ii) remains true for all α in the whole interval $(0, 4/3]$.

First we need to have a close look at $d_n(\alpha)$. Straightforward calculation yields

$$\begin{aligned} d_n^2(\alpha) &= 2\int_1^{n+1}\left[\frac{1}{x^2\log^\gamma x}\int_1^x\left\{\frac{y-1}{y}\right\}^2\frac{1}{\log^\gamma y}\,dy\right. \\ &\qquad \left. +\left\{\frac{x-1}{x}\right\}^2\frac{1}{\log^\gamma x}\int_x^{n+1}\frac{1}{y^2\log^\gamma y}\,dy\right]dx, \end{aligned}$$

an approximate form of the integral $I(\alpha)$ above. Hence $d_n^2(\alpha) \to \infty$ for all $\alpha \in (0, 4/3]$, which at once implies that the error term in (5) goes to zero for $\alpha \in [1, 4/3]$, when $\gamma(\alpha) \geq 0$. But to see whether or not $d_n(\alpha)\log^{\gamma(\alpha)} n \to \infty$ for $\alpha < 1$, when $\gamma = \gamma(\alpha) = 2 - 2/\alpha < 0$, we need to know the speed at which $d_n(\alpha) \to \infty$. This is what decides whether the Brownian-bridge approximation in (5) fully determines the asymptotic behavior for all $\alpha \in (0,1)$, or the extremes assessed in (2) will also start to contribute below some other critical value of α, which at this point might even be 1. It takes time but is routine to show that $d_n^2(\alpha) \sim 2J_{1,n}^{(\delta)}(\alpha) + 2J_{2,n}^{(\delta)}(\alpha)$, where

$$J_{1,n}^{(\delta)}(\alpha) = \int_{e^\delta}^{n+1}\left[\frac{1}{x^2\log^\gamma x}\int_{e^\delta}^x\frac{1}{\log^\gamma y}\,dy\right]dx$$

and

$$J_{2,n}^{(\delta)}(\alpha) = \int_{e^\delta}^{n+1}\left[\frac{1}{\log^\gamma x}\int_x^{n+1}\frac{1}{y^2\log^\gamma y}\,dy\right]dx,$$

provided any of the two integrals goes to ∞ at the same rate for each fixed $\delta > 1$. Of course the case $\gamma = 0$, obtained for $\alpha = 1$, can be calculated directly. The trick in general is to see via integrating by parts that $J_{2,n}^{(\delta)}(\alpha) = J_{1,n}^{(\delta)}(\alpha) + C_\delta(\alpha) + o(1)$ for some constant $C_\delta(\alpha) \in \mathbb{R}$. Bounding $J_{1,n}^{(\delta)}(\alpha)$ from above and $J_{2,n}^{(\delta)}(\alpha)$ from below the natural way when $\gamma < 0$, and in the opposite directions for $\gamma > 0$, we see that $J_{1,n}^{(\delta)}(4/3) \sim \log\log n \sim J_{2,n}^{(\delta)}(4/3)$, and hence $d_n^2(4/3) \sim 4\log\log n$, for $\gamma = 1/2$, while $J_{1,n}^{(\delta)}(\alpha) \sim \log^{1-2\gamma} n/(1-2\gamma) \sim J_{2,n}^{(\delta)}(\alpha)$, and hence $d_n^2(\alpha) \sim 4\log^{1-2\gamma} n/(1-2\gamma)$, for all $\gamma < 1/2$, i.e. for all $\alpha \in (0, 4/3)$. All these asymptotic equalities hold invariably for every fixed $\delta > 1$. In particular, the first error term in (5) is $O_P(1/\sqrt{\log n})$

for all $\alpha \in (0, 4/3)$, forcing the whole error term to go to zero in probability for all $\alpha \in (0, 4/3]$. Thus the extremes are still negligible in the whole range of part (iii) of Theorem 2 below.

To switch back to $T_n(\alpha)$, using simple algebra and again that $|\mu_2(F_n) - \Gamma(1+2/\alpha)| = O_P(1/\sqrt{n})$ we see that the incoming error term is of a smaller order than the one in (5), and hence, starting out now from any probability space for the observations, we find that

$$\Gamma^2\left(1+2/\alpha\right)\frac{nT_n(\alpha)-a_n(\alpha)}{d_n(\alpha)} \overset{\mathcal{D}}{=} \frac{1}{d_n(\alpha)}\int_0^n \frac{W^2(t)-t}{(1+t)^2\log^\gamma(1+t)}\,dt + \varepsilon_n(\alpha),$$

where $a_n(\alpha) = m_n(\alpha)/\Gamma^2(1+2/\alpha) = \alpha^2 c_n(\alpha) + o(1)$ and

$$\varepsilon_n(\alpha) = \begin{cases} O_P\left(1/\sqrt{\log\log n}\,\right), & \text{if } \alpha = 4/3, \\ O_P\left(1/\log^{\frac{2}{\alpha}-\frac{3}{2}} n\right), & \text{if } 1 \le \alpha < 4/3, \\ O_P\left(1/\sqrt{\log n}\,\right), & \text{if } 0 < \alpha \le 1, \end{cases}$$

giving the precise order of the error term in the main statement of case (iii) in the following

Theorem 2. *Suppose that $F \in \mathcal{G}_\alpha$ for some $\alpha > 0$.*

(i) If $\alpha > 2$, then

$$nT_n(\alpha) \xrightarrow{w} \frac{1}{\alpha^2\,\Gamma^2\left(1+\frac{2}{\alpha}\right)}\left\{\int_0^1 \frac{B^2(t)}{(1-t)^2\log^{2-\frac{2}{\alpha}}\frac{1}{1-t}}\,dt - \left[\int_0^1 \frac{B(t)}{(1-t)\log^{1-\frac{2}{\alpha}}\frac{1}{1-t}}\,dt\right]^2\right\}.$$

(ii) If $4/3 < \alpha \le 2$, then

$$nT_n(\alpha) - c_n(\alpha) \xrightarrow{w} \frac{1}{\alpha^2\,\Gamma^2\left(1+\frac{2}{\alpha}\right)}\left\{\int_0^1 \frac{B^2(t)-E\left(B^2(t)\right)}{(1-t)^2\log^{2-\frac{2}{\alpha}}\frac{1}{1-t}}\,dt - \left[\int_0^1 \frac{B(t)\log^{\frac{2}{\alpha}-1}\frac{1}{1-t}}{1-t}\,dt\right]^2\right\},$$

where

$$c_n(\alpha) = \frac{\int_{\frac{1}{n+1}}^{\frac{n}{n+1}} \frac{E\big(B^2(t)\big)}{(1-t)^2 \log^{2-\frac{2}{\alpha}} \frac{1}{1-t}}\, dt}{\alpha^2\,\Gamma^2\big(1+\frac{2}{\alpha}\big)} = \frac{\int_{1+\frac{1}{n}}^{n+1} \frac{x-1}{x} \frac{1}{x \log^{2-\frac{2}{\alpha}} x}\, dx}{\alpha^2\,\Gamma^2\big(1+\frac{2}{\alpha}\big)}$$

$$= \begin{cases} \frac{\log\log n}{4} + \frac{\int_0^\infty e^{-x} \log x\, dx}{4} + o(1), & \text{if } \alpha = 2, \\ \frac{\log^{\frac{2}{\alpha}-1} n}{\alpha(2-\alpha)\Gamma^2\left(1+\frac{2}{\alpha}\right)} - \frac{\int_1^\infty x^{-2} \log^{\frac{2}{\alpha}-1} x\, dx}{\alpha(2-\alpha)\,\Gamma^2\left(1+\frac{2}{\alpha}\right)} + o(1), & \text{if } \alpha < 2. \end{cases}$$

(iii) If $0 < \alpha \le 4/3$, then

$$\Gamma^2\left(1+2/\alpha\right) \frac{nT_n(\alpha) - a_n(\alpha)}{d_n(\alpha)} \overset{\mathcal{D}}{=} \frac{1}{d_n(\alpha)} \int_0^n \frac{W^2(t) - t}{(1+t)^2 \log^{2-\frac{2}{\alpha}}(1+t)}\, dt + o_P(1),$$

where $\{W(t) : t \ge 0\}$ is a standard Wiener process,

$$a_n(\alpha) = \frac{\int_0^n \frac{t}{(1+t)^2 \log^{2-\frac{2}{\alpha}}(1+t)}\, dt}{\Gamma^2\left(1+\frac{2}{\alpha}\right)}$$

$$= \frac{\alpha\, \log^{\frac{2}{\alpha}-1} n}{(2-\alpha)\Gamma^2\left(1+\frac{2}{\alpha}\right)} - \frac{\alpha \int_1^\infty x^{-2} \log^{\frac{2}{\alpha}-1} x\, dx}{(2-\alpha)\Gamma^2\left(1+\frac{2}{\alpha}\right)} + o(1)$$

and

$$d_n(\alpha) = \sqrt{\mathrm{Var}\left(\int_0^n \frac{W^2(t)}{(1+t)^2 \log^{2-\frac{2}{\alpha}}(1+t)} dt\right)} \sim \begin{cases} 2\sqrt{\log\log n} & \text{if } \alpha = \frac{4}{3} \\ \frac{2\sqrt{\alpha}}{\sqrt{4-3\alpha}} \log^{\frac{2}{\alpha}-\frac{3}{2}} n & \text{if } \alpha < \frac{4}{3}. \end{cases}$$

In particular, for $\alpha = 1$,

$$\frac{4n\,T_n(1) - \log n}{2\sqrt{\log n}} \xrightarrow{\mathrm{w}} \mathcal{N}(0,1), \qquad (6)$$

where $\mathcal{N}(0,1)$ is a standard normal random variable.

For $\alpha \in (1, 4/3]$ and $\alpha \in (0,1)$, the problem of the asymptotic distribution of the standardized random variable $\int_0^n [W^2(t) - t][(1+t)^2 \log^{2-\frac{2}{\alpha}}(1+t)]^{-1} dt / d_n(\alpha)$ in part (iii) is left open here. While it is in the “bad” statistical domain, it is still well motivated, and in itself it seems challenging as a pure probability problem for Brownian motion.

For $\alpha = 1$, when testing for an exponential distribution is at hand, the remark in Section 3.3.3 of the paper appears somewhat hasty: even the closer reference to Lemma 5.3.4, rather than to Theorem 5.4.3 (ii) in Csörgő and Horváth (1993), would call for a little extra work. (Incidentally, we must point out that the factor 1/2 in (5.3.17) in Csörgő and Horváth (1993) should be 2, but they probably compute with the correct value since the statement of their Lemma 5.3.4 is luckily correct.) Given the main statement in case (iii) of Theorem 2 for the special case $\alpha = 1$, a more direct reference is to Lemma 5.3.3 in Csörgő and Horváth (1993). As translated from an Ornstein-Uhlenbeck process to a Wiener process, for $p = 2$ that says that

$$\frac{1}{2\sqrt{\log n}}\left[\int_1^n \frac{W^2(t)}{t^2}\,dt - \log n\right] \xrightarrow{\mathrm{w}} \mathcal{N}(0,1).$$

This implies (6) since, by the Bunyakovski-Schwarz inequality

$$\frac{1}{\sqrt{\log n}}\,E\!\left(\left|\int_1^{n+1} \frac{W^2(t) - W^2(t-1)}{t^2}\,dt\right|\right) \leq \frac{\sqrt{3}}{\sqrt{\log n}} \int_1^\infty \frac{1}{t^{3/2}}\,dt\,,$$

and so the integral under the expectation divided by $\sqrt{\log n}$ goes to zero in probability.

The statement in (6) for the exponential case is probably equivalent to Lockhart's (1985) theorem; having tripled the originally allotted space and writing three days past the deadline for the submission of this discussion, I did not check this. It is also likely that the result of McLaren and Lockhart (1987) for $\alpha = 1$, concerning zero asymptotic relative efficiency, generalizes for $nT_n(\alpha)$ for all $\alpha \in (0, 4/3]$, regardless of the nature of the missing limiting distributions, already following from the present statement of Theorem 2 (iii).

Even though the exponent of the logarithm in $d_n(\alpha)$ may be arbitrary large if the shape exponent $\alpha > 0$ is sufficiently small, case (iii) here is not yet the whole way down to the heavy tails in Section 3.3.3 of the paper. The scale family generated by the Pareto distribution function $G_\alpha(x) = 1 - x^{-\alpha}$, $x \geq 1$, for some tail parameter $\alpha > 2$, replacing the Weibull above, will very likely connect the two. Assumptions 1, 2 and 3 still hold, so Brownian-bridge approximation is still possible, but the respective terms $nR_n^{(\alpha)}$ and $Y_n^{(\alpha)}$ turn out to be of the same stochastic order. So, while the Brownian-bridge approximation will likely be helpful still to delineate what parts of the integrals in the corresponding $nM_n^{(\alpha)}$ matter, we conjecture that in this Pareto case $[nT_n(\alpha) - a_n^*(\alpha)]/n^{2/\alpha}$, for some centering sequence $a_n^*(\alpha)$, has a non-degenerate limiting distribution related to extremes as in Proposition 3.6. It would be interesting to see the details of such a result. If it is true, the size of the norming constants $n^{2/\alpha}$ would go up almost to that of $n/\log^2 n$ in Proposition 3.6, thereby, modulo logarithmic factors, practically completing the picture.

5 Conclusion and further questions

Taking the expected kind of performance for testing normality as a yardstick of good behavior, we expect similar good behavior of the Wasserstein correlation test statistic nT_n for testing goodness of fit to a scale or location-scale family generated by a distribution function whose tails are dominated by $e^{-|x|^\alpha}$, for $|x|$ large enough, as long as $\alpha > 4/3$. We expect poor performance otherwise, when domination by such a function may be achieved only for some $\alpha \leq 4/3$, or not at all.

For testing for the Weibull family $\mathcal{G}_\alpha$ for a given $\alpha > 4/3$, it would be of definite interest to determine the limiting distributions in cases (i) and (ii) of Theorem 2 in the respective forms $\sum_{j=1}^{\infty} \lambda_j Y_j^2$ and $\sum_{j=1}^{\infty} \mu_j (Y_j^2 - 1)$ of Sections 3.3.1 and 3.3.2, for some eigenvalues λ_j and μ_j, $j \in \mathbb{N}$. Even if one doubts the practical import of the testing problem $F \in \mathcal{G}_\alpha$ for a given $\alpha > 4/3$, the probabilistic sports value of seeing whether the elegant Hilbert-space methods in these sections work for these cases is not negligible.

The Wasserstein correlation test procedures appear to be tailored for location-scale and pure scale families. While on the Weibull scale families, it is inevitable for the question to pop up: how about not specifying the shape parameter α? It did not escape our attention that minimizing $\mathcal{W}^2(F_n, \mathcal{G}_\alpha)$ may yield estimators for $\alpha > 0$ with attractive properties. But can the procedure be somehow modified to produce reasonable statistics for testing the composite hypothesis that $F \in \cup_{\{\alpha > 4/3\}} \mathcal{G}_\alpha$? Or at least that $F \in \cup_{\{\alpha > 2\}} \mathcal{G}_\alpha$?

Acknowledgements

Research supported in part by the Hungarian National Foundation for Scientific Research, Grant T 032025.

References

Götze, F. (1979). Asymptotic expansions for bivariate von Mises functionals. *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, **50**, 333-355.

Hilbert, D. (1902). Mathematical problems. *Bulletin of the American Mathematical Society, Ser. 2*, **8**, 437-479.

Kallenberg, W.C.M. and T. Ledwina (1997). Data-driven smooth tests when the hypothesis is composite. *Journal of the American Statistical Association*, **92**, 1094-1104.

Carles M. Cuadras
Universitat de Barcelona, Spain

I really appreciate the opportunity to discuss this paper and congratulate the authors for their very interesting contribution.

The goodness of fit assessment is a major subject in statistics as it focuses on the central issue of model choice. On the other hand, the study of goodness of fit tests is a stimulating opportunity to use a wide range of statistical and mathematical tools, including empirical processes, multivariate analysis, special functions, asymptotic expansions and distance analysis.

As this paper is well written and presented, I wish to comment and relate this contribution to other areas.

1 Orthogonal expansions and goodness-of-fit tests

The expansions of the limit distribution of Anderson-Darling A^2 and Cramér-von Mises W^2 statistics, have some points in common with the orthogonal expansion of a random variable X with cdf F

$$\begin{aligned} X &= x_0 + \textstyle\sum_{j=1}^{\infty} h_j(b)(X_j - h_j(x_0)), \\ X &= x_0 + \textstyle\sum_{j=1}^{\infty} (X_j{}^2 - h_j(x_0)h_j(b)), \end{aligned}$$

where the convergence is in the mean-square sense, $h_j\left(x\right) = \int_a^x \psi_j\left(s\right) ds$, $X_j = h_j(X)$ and (λ_j, ψ_j) is the countable orthonormal set of eigenvalues and eigenfunctions of the integral operator $\mathcal{K}$ with kernel $K(s,t) = F(s) \wedge F(t) - F(s)F(t)$

$$(\mathcal{K}\varphi)(t) = \int_a^b K\,(s,t)\,\varphi(s)ds,$$

i.e., (λ_j, ψ_j) satisfies $\mathcal{K}\psi_j = \lambda_j \psi_j$. Then each $X_j = h_j(X)$ is a principal dimension for X, which can be obtained via the Karhunen-Loève expansion or continuous scaling. See Cuadras and Fortiana (1995, 2000), Cuadras and Lahlou (2000). Let me give two examples.

1) If X is $[0,1]$ uniform, setting $x_0 = 0$, X can be expanded as $X = \sum_{j=1}^{\infty} \lambda_j U_j^2$, where $U_j = \sqrt{2}(1 - \cos j\pi X)$, $j \geq 1$, is a countable set of uncorrelated equally distributed random variables and $\lambda_j = (j\pi)^{-2}$. Thus the analogy with $W^2 = \sum_{j=1}^{\infty} \lambda_j Y_j^2$ is clear, but also there is some analogy with the limit distribution of $n\mathcal{R}_n$ under the uniform model.

2) If X is standard logistic, $F(x) = 1/(1+\exp(-x))$, then the expansion of X can be expressed in terms of $P_j(2F(x) - 1)$, with P_j being the Legendre polynomials of degree j, and $\lambda_j = (j(j+1))^{-1}$, also obtaining a formal analogy with the expansion of A^2. See Shorack and Wellner (1985, pp. 225).

Such an analogy may be due to the function Ψ in defining the statistics

$$W_n^2(\Psi) = n \int_{-\infty}^{\infty} \Psi(F_0(x))(F_n(x) - F_0(x))^2 dF_0(x),$$

which is $\Psi^{-1} = 1$, the uniform density, for the Cramér-von Mises statistic, and $\Psi^{-1} = t(1-t)$, giving the logistic density $f = F(1 - F)$, for the Anderson-Darling statistic. This suggests that by setting $\Psi^{-1} =$ a probability density function, we obtain a general form for $W_n^2(\Psi)$.

The eigenvalues λ_j satisfy (Cuadras and Fortiana 1995)

$$\frac{1}{2}E[\mid X - X' \mid] = \mathrm{tr}(\mathcal{K}) = \sum_{j=1}^{\infty} \lambda_j,$$

where X, X' are iid and $\lambda_j =$Var(X_j). Thus, each eigenvalue accounts for the so-called geometric variability $V = E[\mid X - X' \mid]/2$ of X with respect to distance $\delta(x, x') = (\mid x - x' \mid)^{1/2}$.

2 Bounds for the Wasserstein distance

The authors propose and study a test based on the Wasserstein distance between two distributions F_1, F_2. If the means μ_1, μ_2 and variances σ_1^2, σ_2^2 are finite, this distance can be expressed as

$$\mathcal{W}^2(F_1, F_2) = \sigma_1^2 + \sigma_2^2 - 2\sigma_1\sigma_2\rho^+(F_1, F_2) + (\mu_1 - \mu_2)^2,$$

where $\rho^+(F_1, F_2)$ is the maximum Hoeffding correlation between F_1, F_2, i.e., by considering the Fréchet upper bound $F_1 \wedge F_2$. Let me present some results derived from this distance and test.

A. Suppose F_1, F_2 are absolutely continuous with densities f_1, f_2. Assuming $\sigma_1 \leq \sigma_2$, from $0 < \rho^+(F_1, F_2) \leq 1$ and the following inequality (Cuadras 1996)

$$\rho_0 = \inf_x\{f_1(x)/f_2(x)\} \leq \frac{\sigma_1}{\sigma_2}\rho^+(F_1, F_2), \tag{1}$$

we can easily prove that

$$\sigma_1^2 + \sigma_2^2 - 2\sigma_1\sigma_2 + (\mu_1 - \mu_2)^2 \leq \mathcal{W}^2(F_1, F_2) \leq \sigma_1^2 + \sigma_2^2(1 - 2\rho_0) + (\mu_1 - \mu_2)^2.$$

Note that under normal distribution $\rho_0 = \sigma_1/\sigma_2$ and the equality holds (see also Cuesta-Albertos et al. 1996).

B. From now on we suppose F_2 uniform on $[0, 1]$. Writing F_0 for F_2, it can be proved that

$$\rho^+(F_1, F_0) = \frac{\sqrt{3}}{\sigma_1}V, \tag{2}$$

where V is the geometric variability of $X \sim F_1$ defined above. Thus we obtain the identity

$$V + \mathcal{W}^2(F_1, F_0) = \sigma_1^2 + \frac{1}{12} + (\mu_1 - 1/2)^2.$$

C. As $f_2 = 1$, from (1) and (2), we have the inequality

$$\frac{1}{6}\inf_x\{f_1(x)\} \leq V \leq \sigma_1/\sqrt{3}.$$

As a consequence of this we obtain the following bounds

$$\alpha - \sigma_1/\sqrt{3} \leq \mathcal{W}^2(F_1, F_0) \leq \alpha - \frac{1}{6}\inf_x\{f_1(x)\},$$

where $\alpha = \sigma_1^2 + (1/12) + (\mu_1 - 1/2)^2$. There is equality iff F_1 is also uniform on $[0, 1]$.

3 Tests based on the maximum correlation

One of the tests proposed by the authors is based on

$$\mathcal{R}_n = \mathcal{W}^2(F_n, \mathcal{F})/S_n^2 = 1 - \frac{\widehat{\sigma}_n^2}{S_n^2} = 1 - \rho^+(F_n, H_0)^2, \quad F \in \mathcal{F},$$

where H_0 is assumed to be standardized. This test is therefore dependent on $\rho^+(F_n, F)$, the maximum Hoeffding correlation between the sample and the theoretical distribution F. The use of $\rho^+(F_n, F)$ was suggested by Cuadras and Fortiana (1993, 1994), who emphasised the need to explore the data from a multivariate analysis point of view, rather than investigating the sampling distribution of this correlation. This is in fact undertaken by Fortiana and Grané (2000), who propose suitable modifications of $\rho^+(F_n, F)$, e.g., the statistics $\sqrt{12}S_n\rho^+(F_n, F_0)$ and $S_n\rho^+(F_n, F)/\overline{X}_n$, for the uniform and exponential model, respectively, where $\overline{X}_n$ is the sample mean. The authors made a similar modification when they studied the asymptotic distribution of $S_n^2\mathcal{R}_n = S_n^2 - S_n^2\rho^+(F_n, H_0)^2$.

Returning to the orthogonal expansion of a random variable X, if $\rho_1, \rho_2, \ldots$ and $r_1, r_2, \ldots$ are the theoretical and sampling correlations between X and the principal dimensions, and between the sample $\mathcal{X}_n$ and the principal dimensions, the following expansion can be proved

$$\rho^+(F_n, F) = \sum_{j=1}^{\infty} \rho_j r_j,$$

where F is any distribution function. This suggests that the representation and comparison of the principal dimensions $h_j(X), h_j(\mathcal{X}_n)$ may be a graphical test for indicating how well the sample $\mathcal{X}_n$ fits to F. This approach is useful for distinguishing similar distributions, such as logistic and normal (Cuadras and Lahlou 2000).

Finally, the eigenvalues λ_j of $\mathcal{K}$ are of interest in studying the asymptotic distribution of some statistics related to Rao's quadratic entropy (Liu 1991, Rao 1982), which is also a weighted sum of independent chi-square random variables. As the geometric variability V is a particular case of quadratic entropy, and V is related to $\mathcal{R}_n$, we gain an additional insight into the distribution of $\mathcal{R}_n$.

References

Cuadras, C.M. (1996). A distribution with given marginals and given regression curve. In *Distributions with Fixed Marginals and Related Topics.* IMS Lecture Notes-Monograph Series, **28**, pp. 76-83 (L. Rüschendorf, B. Schweizer and M.D. Taylor, eds.) Hayward, California.

Cuadras, C.M. and J. Fortiana (1993). Continuous metric scaling and prediction. In *Multivariate Analysis, Future Directions 2*, pp. 47-66. (C.M. Cuadras and C.R. Rao eds.) Elsevier Science Publishers B.V., North–Holland, Amsterdam.

Cuadras, C.M. and J. Fortiana (1994). Ascertaining the underlying distribution of a data set. In *Selected Topics on Stochastic Modelling*, pp. 223-230 (R. Gutierrez and M.J. Valderrama eds.) World-Scientific, Singapore.

Cuadras, C.M. and J. Fortiana (1995). A continuous metric scaling solution for a random variable. *Journal of Multivariate Analysis*, **52**, 1-14.

Cuadras, C.M. and J. Fortiana (2000). The importance of geometry in multivariate analysis and some applications. In *Statistics for the 21st Century*, pp. 93-108 (C.R. Rao and G. Szekely, eds.) Marcel Dekker, New York.

Cuadras, C.M. and Y. Lahlou (2000). Some orthogonal expansions for the logistic distribution. *Communications in Statistics-Theory and Methods*, to appear.

Cuesta-Albertos, J.A., C. Matrán-Bea and A. Tuero-Diaz (1996). On lower bounds for the L^2–Wasserstein metric in a Hilbert space. *Journal of Theoretical Probability*, **9**, 263-283.

Fortiana, J. and A. Grané (2000). A scale-free goodness-of-fit statistic for the exponential distribution based on maximum correlations. Submitted.

Liu, Z. (1991). Bootstrapping one way analysis of Rao's quadratic entropy. *Communications in Statistics-Theory and Methods*, **20**, 1683-1703.

Rao, C.R. (1982). Diversity: its measurement, decomposition, apportionment and analysis. *Shankhya, A*, **44**, 1-21.

Shorack, G.R. and J.A. Wellner (1986). *Empirical Processes with Applications to Statistics.* Wiley, New York.

Tertius de Wet

University of Stellenbosch, South Africa

It was a pleasure to read this very timely paper and I would like to thank the Editor for the opportunity to comment on it. It is in the nature of an expository paper that it concentrates on the authors' particular preferences. This gives commentators the opportunity to add their own preferences, not covered by the authors. The fields of quantile and empirical processes and goodness-of-fit are so rich that any expository paper leaves ample room for additions. I will discuss a number of additions and extensions.

1. The authors discuss fairly extensively the new approach to constructing goodness-of-fit tests, based on the Wasserstein distance measure. This is a very natural measure to use and it produces asymptotic results similar to earlier quadratic type statistics. The authors find that it seems to work out "correctly" in testing for normality, in the sense of "loosing degrees of freedom". Using the test for other distributions, does not have this "nice" property (see the discussion at the end of paragraph 3.3.2). However, one can have this property, at least for a single parameter (location or scale), by considering a weighted Wasserstein distance. Taking the weight function identically equal to one, turns out to be the "right" choice in the Gaussian case. In addition, the Gaussian case leads to the same weight function for location and scale parameters.

   Let us consider a scale parameter family of distributions. We want to test the null hypothesis:

   $$H : F_\theta(\cdot) = F(\cdot/\theta),$$

   with F specified and θ an unknown scale parameter. Due to scale invariance in what follows, we take the true $\theta = 1$ without loss of generality.

   For W a weight function on $(0,1)$, define the weighted Wasserstein distance as:

   $$\omega^2(\theta) = \int_0^1 (F_n^{-1}(t) - \theta F^{-1}(t))^2 W(t)dt.$$

   A minimum distance estimator for θ is given by

   $$\begin{aligned}\hat{\theta} &= \arg\min \omega^2(\theta) \\ &= \int_0^1 (F_n^{-1}(t)F^{-1}(t))W(t)dt \Big/ \int_0^1 F^{-1}(t)^2 W(t)dt.\end{aligned}$$

Let $H = F^{-1}$. Following Chernoff, Gastwirth and Johns (1967) and de Wet and Venter (1973), define

$$W(t) = J(t)/H(t),$$

where, with $f = F'$,

$$\begin{aligned} J(t) &= I^{-1}L'(H(t)) \\ L(y) &= -1 - yf(y)/f'(y) \\ I &= \int_{-\infty}^{\infty} L^2(y)f(y)dy = \int_0^1 L'(H(u))H(u)du, \end{aligned}$$

where we make the necessary assumptions in order for the derivatives and integrals to exist (see Chernoff, Gastwirth and Johns 1967 in this regard).

Note that

$$\int_0^1 J(u)H(u)du = 1$$

and thus

$$\hat{\theta} = \int_0^1 F_n^{-1}(t)J(t)dt.$$

As test statistic for H, we use

$$\begin{aligned} \omega^2(\hat{\theta}) &= \int_0^1 (F_n^{-1}(t) - \hat{\theta}H(t))^2(J(t)/H(t))dt \\ &= \int_0^1 (F_n^{-1}(t) - H(t))^2(J(t)/H(t))dt - (\hat{\theta} - 1)^2. \end{aligned}$$

This is similar to the authors' (3.12) (corresponding to the first and third terms there). Furthermore, proceeding as in Theorems 3.3 and 3.5 (we potentially have the heavy tailed case, as in Theorem 3.5), we obtain

$$\begin{aligned} n(\omega^2(\hat{\theta}) - a_n) &\xrightarrow{w} \int_0^1 (B(t)^2 - EB(t)^2)H'(t)^2(J(t)/H(t))dt \\ &\quad - \left(\int_0^1 B(t)H'(t)J(t)dt\right)^2, \end{aligned}$$

for appropriate constants $\{a_n\}$ and $B(\cdot)$ the Brownian Bridge process. Let $K(s,t)$ be the covariance of the limiting stochastic process for the fixed distribution case (i.e. θ known). This process is clearly

$$B(t)H'(t)(J(t)/H(t))^{\frac{1}{2}},$$

and has covariance function

$$K(s,t) = (s \wedge t - st)H'(s)H'(t)[J(s)J(t)/H(s)H(t)]^{\frac{1}{2}}.$$

Let $\{h_j\}$ be a complete orthonormal system of eigenfunctions for K. Then, as in Theorem 3.6, we have, with $\alpha_n = (n+1)^{-1}$,

$$\begin{aligned}&\int_{\alpha_n}^{1-\alpha_n} \left[B(t)H'(t)(J(t)/H(t))^{\frac{1}{2}}\right]^2 dt \\ &= \sum_{j=1}^{\infty} \left(\int_{\alpha_n}^{1-\alpha_n} B(t)H'(t)(J(t)/H(t))^{\frac{1}{2}} h_j(t) dt\right)^2 .\end{aligned}$$

The second term in the limiting value of $\omega^2(\hat{\theta})$, is

$$\begin{aligned}&\left(\int_0^1 B(t)H'(t)J(t)dt\right)^2 \\ &= \left(\int_0^1 B(t)H'(t)(J(t)/H(t))^{\frac{1}{2}}(J(t)H(t))^{\frac{1}{2}} dt\right)^2 .\end{aligned}$$

It can be shown (see de Wet 1999 for this, as well as details of the above) that $(J(t)H(t))^{1/2}$ is an eigenfunction of K, with corresponding eigenvalue I^{-1}. From this it follows that in the above sum, we loose one term (due to estimation of θ), leading to a "loss of one degree of freedom".

Remark 1.

(a) Choosing the "correct" weight function we have shown in previous work leads to optimality in terms of approximate Bahadur slopes, at least in a limited number of cases (Gaussian, exponential). See e.g. de Wet (1980) for this. We conjecture that under certain contiguous alternatives, the above weight function will have optimal approximate Bahadur slope.

(b) The above argument also goes through in the case of a location parameter, but with a different choice of the weight function (see, for details, de Wet 1999).

2. Some work has been done in recent years on applying Cramér-von Mises type statistics in time series situations, and finding the asymptotic distribution theory. (See, e.g., Anderson and Stephens 1993). This is an area in which much more work needs to be done. The Wasserstein distance seems an ideal candidate for application in this setting.

3. What can be said of goodness-of-fit tests in a multivariate setting, and in particular testing for multivariate normality? What results do we get from the interaction with developments in empirical/quantile processes in a multivariate setting? One such proposal was made and studied in de Wet, Venter and van Wyk (1979). In this use was made of the fact that $\boldsymbol{X}$ has a $\mathcal{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ distribution if and only if $\boldsymbol{a}'\boldsymbol{X}$ is $\mathcal{N}_1(\boldsymbol{a}'\boldsymbol{\mu}, \boldsymbol{a}'\boldsymbol{\Sigma}\boldsymbol{a})$ for all non-zero p-dimensional real vectors $\boldsymbol{a}$. If $T(V_1, \ldots, V_n)$ is a correlation type test statistic for the one dimensional case with observations $V_1, \ldots, V_n$, then the proposed test statistic is

$$T_n = \sup_{\boldsymbol{a} \neq \boldsymbol{0}} T(\boldsymbol{a}'\boldsymbol{X}_1, \ldots, \boldsymbol{a}'\boldsymbol{X}_n).$$

Its limiting distribution was obtained, but was (at that time) unfortunately not amenable to computation. However, with the growth in results on multivariate empirical and quantile processes, as well as computing power during the last few years, it may be worth revisiting this problem.

I would like to thank the authors for reviewing two fields which have interactively had major growth during the last few decades and in which many exciting problems still remain.

References

Anderson, T.W. and M.A. Stephens (1993). The modified Cramér-von Mises goodness-of-fit criterion for time series. *Sankhya, A*, **55**, 357-369.

Chernoff, H., J.L. Gastwirth and M.V. Johns (1967). Asymptotic distribution of linear combinations of functions of order statistics with application to estimation. *Annals of Mathematical Statistics*, **38** , 52-53.

de Wet, T. and J.H. Venter (1973). A goodness-of-fit test for a scale parameter family of distributions. *South African Statistical Journal*, **7**, 35-46.

de Wet, T., J.H. Venter and J.W.J. van Wyk (1979). The null distributions of some test criteria of multivariate normality. *South African Statistical Journal*, **13**, 153-176.

de Wet, T. (1980). Cramér-von Mises tests for independence. *Journal of Multivariate Analysis*, **10**, 38-50.

de Wet, T. (1999). Goodness-of-fit tests based on a weighted Wasserstein distance. I) Scale parameter family. II) Location parameter family. Unpublished Manuscript.

Evarist Giné

University of Connecticut, USA

This nice survey paper consists basically of three parts. The first part is historical and covers the asymptotic theory of the classical Kolmogorov-Smirnov and Cramér-von Mises tests, regular as well as weighted, and with or without estimated parameters. The asymptotics for these statistics motivated several important developments in Probability Theory, including invariance principles, probability in Banach spaces, the modern theory of empirical processes and the theory of strong approximations. It is not surprising that, but it is important to point out that, the asymptotic theory of the Cramér-von Mises type tests (as defined in the above article) follows from the central limit theorem in Hilbert space. On the other hand, limit theorems such as Kolmogorov-Smirnov or Chibisov-O'Reilly, involving sup norms, require weak convergence results in $C[0,1]$ or $D[0,1]$, or, more generally and perhaps more simply, the modern theory of empirical processes indexed by general classes of functions. In this connection, I would like to mention that the Chibisov-O'Reilly theorem is a very simple consequence of the bracketing central limit theorem in Andersen, Giné, Ossiander and Zinn (1988, example 4.9, pp. 296–297). This proof does not use almost sure representations or embeddings. Strong approximations constitute an excellent tool for proving a.s. convergence result, but in general they should not be needed for proving weak convergence theorems, which are usually easier. However, I must admit that they are very useful and, sometimes, their use seems unavoidable: see, e.g., del Barrio, Giné and Matrán (1999), where the central limit theorem for the L_1-Wasserstein distance between the empirical and the true distributions in the case of infinite variance random variables in the domain of attraction of a normal law seemed to require a useful result on weighted approximation of the empirical process by Brownian bridges due to Mason (1991) and to Csörgő and Horváth (1986), an easier version of Theorem 3.2 above. The weighted approximation by Brownian bridges was used in order to infer weak convergence of a function of the empirical discrepancy from weak convergence of the corresponding statistic for the Brownian bridge or, equivalently, the Orstein-Uhlenbeck process. (It was not needed, however, to prove uniform tightness).

The second part covers the more recent theory of the Shapiro-Wilk and other correlation tests, and constitutes a very useful guide to the literature on the subject.

The third part develops the approach to correlation tests from del Barrio, Cuesta-Albertos, Matrán and Rodríguez-Rodríguez (1999), based on the L_2-Wasserstein distance between fixed distributions and location-scale families. This is a very interesting article, particularly relevant because it provides a structured proof of the Shapiro-Wilk test. Weighted approximation by Brownian bridges is used in this proof in a way similar to the above mentioned article of del Barrio, Giné and Matrán. The present paper ends with a very nice example showing that the L_2-Wasserstein test statistic may have non-normal limit laws for heavy-tailed distributions.

Since L_2 distances are easier to handle than L_p distances for $p \neq 2$, it is only natural to ask whether the asymptotics of Shapiro-Wilk type statistics can be handled by the more elementary central limit theorem for Hilbert space valued random variables, which goes back at least to Varadhan (1962) (as opposed to the more recent and much less elementary weighted approximations). Such a derivation has been announced at the BS-IMS Congress, May 2000, Guanajuato, by del Barrio and Matrán (Abstract No. 77).

Another abstract in the same Congress that is relevant to the present survey is that of A. Cabaña and E.M. Cabaña, "Consistent and focused goodness of fit for families of distribution functions", which, in particular, announces a test of normality that compares favorably to Shapiro-Wilk. The tests in this announcement are based on "transformed empirical processes" and their proofs use modern empirical process theory (see e.g., A. Cabaña and E. Cabaña 1997, and references therein for precursors of these tests in the case of simple hypotheses).

References

Andersen, L., E. Giné, M. Ossiander and J. Zinn (1988). The central limit theorem and the law of iterated logarithm under local conditions. *Probability Theory and Related Fields*, **77**, 271-301.

Cabaña, A. and E.M. Cabaña (1997). Transformed empirical processes and modified Kolmogorov-Smirnov tests for multivariate distributions. *Annals of Statistics*, **25**, 2388-2409.

Cabaña, A. and E.M. Cabaña (2000). Consistent and focused goodness of fit for families of distribution functions. *Program, Abstracts and Directory, BS/IMS Congress*, CIMAT, Guanajuato, Mexico. Abstract 44, pp. 56.

Csörgő, M. and L. Horváth (1986). Approximation of weighted empirical and quantile processes. *Statistics and Probability Letters*, **4**, 275-280.

del Barrio, E., E. Giné and C. Matrán (1999). Central limit theorems for the Wasserstein distance between the empirical and the true distributions. *Annals of Probability*, **27**, 1009-1071.

del Barrio, E. and C. Matrán (2000). An elementary derivation of the asymptotic distribution of Shapiro-Wilk type statistics. *Program, Abstracts and Directory, BS/IMS Congress*, CIMAT, Guanajuato, Mexico. Abstract 76, pp. 63.

Mason, D. (1991). A note on weighted approximations to the uniform empirical and quantile processes. In *Sums, Trimmed Sums and Extremes*, pp. 269-283. Birkhäuser.

Varadhan, S.R.S. (1962). Limit theorems for sums of independent random variables with values in a Hilbert space. *Sankhya*, **24**, 213-238.

Richard Lockhart

Simon Fraser University, Canada

Professors del Barrio, Cuesta-Albertos and Matrán are to be congratulated on a fine survey of empirical process theory applied to goodness-of-fit and particularly on their unifying discussion of tests based on the

Wasserstein metric. I want to raise some points about the importance of computable asymptotics and about power calculations under contiguous alternatives.

Statisticians often want to compute, exactly or approximately, the law $\mathcal{L}(T|F)$ of a statistic computed from data with distribution F. The standard asymptotic paradigm is to embed the problem in a sequence (T_n, F_n) indexed by some parameter n so that $\mathcal{L}(T|F)$ is $\mathcal{L}(T_{n_0}|F_{n_0})$. We then compute $\mathcal{L}_\infty = \lim_{n\to\infty} \mathcal{L}(T_n|F_n)$ and use $\mathcal{L}_\infty$ as an approximation to $\mathcal{L}(T|F)$. The approach will be useful for data analysis when $\mathcal{L}_\infty$ is computable.

When T is a functional of an empirical process α_n, say $T_n = g(\alpha_n)$, the process α_n converges weakly in some space to a process α_∞ and g is continuous (almost surely on the support of α) the limit law is $\mathcal{L}_\infty = \mathcal{L}(g(\alpha))$. For Gaussian processes α this limit will be computable for linear and quadratic functionals g and for some of the functionals of the weighted supremum type. Linear functionals g give normal limit laws which are certainly computable. Quadratic functionals give limit laws which are those of a linear combination of chi-squares; Stephens (1974, 1976) shows how to compute P-values effectively for many common covariance kernels. Statistics of the Kolmogorov-Smirnov kind are more problematic; the computation of the law of the weighted supremum of a Gaussian process is not easy.

Monte Carlo is a powerful competitor for asymptotic calculation of P-values. Consider, for instance, the situation in section 2 of the current paper. In this case the distribution of any statistic can be computed on the null hypothesis by repeated sampling from the null hypothesis, F_0. The calculation is simple and exact except for sampling error. Asymptotic calculations will be used if they are easier than Monte Carlo and achieve the same accuracy as Monte Carlo in roughly the same computing time. This will generally happen if the asymptotic calculations are particularly simple or the statistic is hard to evaluate or the sample size is very large.

Del Barrio et al note that the situation changes with composite hypotheses. If the distributional family to be tested is $\{F_\theta; \theta \in \Theta\}$ then the unknown parameter θ enters the problem in two ways. First, the process α_n has a definition depending on θ and must be replaced by an estimated empirical process $\hat{\alpha}_n$. Second, the law of $\hat{\alpha}_n$ generally depends on θ and n. The standard asymptotic paradigm is to compute $\mathcal{L}_\infty(\hat{\alpha}|F_\theta) = \lim_{n\to\infty} \mathcal{L}(\hat{\alpha}_n|F_\theta)$. At first glance the dependence of this limit on θ is a problem. In practice the limit law is estimated; we use $\mathcal{L}_\infty(\hat{\alpha}|F_{\hat{\theta}})$.

It is, perhaps, not obvious that we are entitled to compute P-values as usual from this estimated limit; the problem often leads workers to focus, as del Barrio et al do here, on problems where T is distribution free — its law does not depend on θ. Two points arise. First, if T has a law free of θ then we can compute its law by Monte Carlo using $\mathcal{L}(T|\theta^*)$ for any θ^* we find convenient. In regression problems for instance we can simulate under the model with the slope set to 0. Again asymptotics will be useful only if calculations based on the asymptotic law are more convenient than Monte Carlo. The second point to make is that if $\mathcal{L}_\infty(\hat{\alpha}|\theta)$ depends continuously on θ then use of a consistent estimate $\hat{\theta}$ will give asymptotically valid P-values. In the empirical process setting moreover it is not really necessary to prove that $\mathcal{L}_\infty(\alpha|\theta)$ is continuous; instead if $T_n = g(\alpha_n)$ is the (presumably real valued) statistic to be used to test fit it suffices to prove continuity of $\mathcal{L}_\infty(g(\alpha)|\theta)$. Consider, for instance, the common case where the limit α is a mean 0 Gaussian process on the unit interval with a covariance function of the form $\rho_0(s,t) - g_\theta(s,t)$ with g_θ non-negative definite. In this case pointwise continuity of $\theta \mapsto g_\theta(s,t)$ implies continuity in θ of the law of $T = \int \alpha^2(t)dt$.

There are other uses to be made of asymptotic theory. An alternative to use of $\mathcal{L}_\infty$ is Monte Carlo calculation for finite n. When statistics are not distribution free this amounts to use of $\mathcal{L}(T_n|\hat{\theta}_n)$. Now continuity of the limit law is not enough to justify asymptotically the use of P-values obtained by this bootstrap method. Instead it is necessary to prove that the weak convergence results hold uniformly in θ. That is, you need to prove something like $\theta_n \to \theta$ implies $\mathcal{L}(\alpha_n|\theta_n) \to \mathcal{L}(\alpha|\theta)$. This can be significantly harder. Baringhaus and Henze (1992) for instance, have observed the need for this sort of result and given examples.

Asymptotic methods can also be used to make approximate power calculations and help choose good tests. Here I would like to strengthen the remarks made by del Barrio et al concerning correlation tests for “heavy tailed” distributions. (The exponential distribution does not have really “heavy tails” in the sense usually understood but they are heavy enough to cause problems for correlation tests of fit.) del Barrio et al prove that the Wasserstein test is asymptotically equivalent to a test based on a decreasing fraction of the tails of the sample. Lockhart (1991) actually proves that for *any* sequence (not just some sequence as suggested in the text) of alternative densities of the form $f(x)(1 + h_n(x)/n^{1/2})$ with h_n converging (in the appropriate L_2 sense) to some h the power of such a test must converge to the level of the test. Since most EDF tests will have limiting power strictly greater than the level for *every* such alternative sequence this must be regarded as a significant criticism of correlation tests for such distributions.

References

Baringhaus, L. and N. Henze (1992). A goodness of fit test for the Poisson distribution based on the empirical generating function. *Statistics and Probability Letters*, **13**, 269-274.

Stephens, M.A. (1974). EDF statistics for goodness of fit and some comparisons. *Journal of the American Statistical Association*, **69**, 730-737.

Stephens, M.A. (1976). Asymptotic results for goodness-of-fit statistics with unknown parameters. *Annals of Statistics*, **4**, 357-369.

Axel Munk

Ruhr-Universität Bochum and Universtität Siegen, Germany

First of all, I would like to congratulate the authors for this interesting paper which contributes to the theory of testing the goodness of fit of distributional assumptions based on quantile and empirical processes in two different ways. On the one hand in its first part this paper provides a very helpful survey on this area, particularly by relating many papers from the statistical literature to recent developments addressing probabilistic aspects of quantile and empirical processes. On the other hand the technique developed by the authors in their '99 paper is extended in the second part to heavy tailed distributions, exploring the limits of tests based on the Wasserstein distance.

I would like to focus in my discussion on some practical issues related to this paper – should one use any of these goodness of fit tests, and if so, which one? Before I address the first part of this question (which is the much more subtle problem), I would like to comment briefly on the second part.

There is certainly common agreement, at least among frequentists, that the proper choice of a goodness of fit test (besides of computational aspects, etc.) should be mainly driven by power considerations. Unfortunately, there is such a vast amount of literature providing theoretical power investigations (mainly computing various asymptotic efficiencies) as well as Monte Carlo studies, that it is difficult to get a homogeneous picture. However, one notable conclusion (cf. Neuhaus 1976, Milbrodt and Strasser 1990, Eubank and LaRiccia 1992, or Kallenberg and Ledwina 1997, among many others) is that essentially any g.o.f. test based on the empirical c.d.f and transforms of it (including the quantile process) are in a certain sense invalid to detect most alternatives for realistic sample sizes. For the particular case of quadratic statistics Q_n and testing the null $H_0 : F = \mathcal{N}(0,1)$ say, s.t. under H_0

$$Q_n \Longrightarrow \sum_{k=1}^{\infty} \lambda_k (U_k^2 - 1)$$

as $n \to \infty$ (here $U_1, U_2, \cdots$ denotes a sequence of *i.i.d.* normal random variables), Gregory (1980) provides explicit expressions of Pitman resp. Bahadur efficiency as the level of significance $\alpha \to 0$. This includes e.g. Cramer-von Mises or Shapiro-Wilks-type tests. One finds that the inferior power property of the Cramer von Mises test relies essentially on two facts. Firstly, the test performs only optimal for the L^2-direction $\cos \pi x$ (an alternative which occurs in practice rarely) and secondly the decay of eigenvalues λ_k is $O(k^{-2})$. This means that almost all of the subsequent directions can hardly be detected. In contrast, the Shapiro-Wilks test (and related statistics) has the property that the decay of eigenvalues is much slower, $O(k^{-1})$. Moreover, the first two principal components correspond to departures in location and scale, which are in many cases most important deviations to detect. The situation, however, may change drastically, when other distributions F_0 ore more complex models are to be tested. For $F_0 = (1 + \exp(-x))^{-1}$ logistic and for F_0 uniform on $[0, 1]$, the decay is $O(k^{-2})$, respectively. For the logistic case the two largest components come close to location and scale deviations, whereas for the uniform case $f_k(x) \sim \cos(\pi k x)$. This is highlighted by the fact that none of these tests is "adaptive" in that sense that the direction of alternative is chosen data driven, which leads in general to better omnibus properties (cf. Eubank and LaRiccia 1992).

Now I turn to the first part of the initial question, i.e. what can we draw from the result of *any* goodness of fit test designed for the hypothesis H_0 : "$F \in \mathcal{F}$" (say normal)? Besides of many fundamental arguments against the naive use of tests (for a discussion from a Bayesian perspective, see Berger 1985), in the case of testing the *goodness of fit* a specific problem occurs according to the paradigm: Choose the null such, that you consider rejection (albeit wrong) as the more serious error (type I). In many testing problems this yields a rather clear definition of the null (no treatment effect, no trend, ...). In goodness of fit problems, however, I claim that often the more serious error occurs if we do *not* reject the model although it is wrong because this will lead to a subsequent data analysis based on the wrong model assumption. However, when testing H_0, this is the type II error, which – at least to my experience – is never controlled when checking the goodness of fit in applications (e.g. by sample size adjustment or sequential procedures). Consequently, if testing H_0 no conclusion should be drawn in case of acceptance. Of course, practically nobody will act like this because hence one could never decide in favour of the model (recall the two options: rejecting $\mathcal{F}$ or making *no* decision). Therefore, in practice, large p-values associated with these tests are often taken as a measure for the "evidence" of the model. However, it is well know, that such an interpretation fails in general and has been criticised from various authors during the past, e.g. again from a Bayesian point of view (Berger and Selke 1987), albeit not directly in the context of goodness of fit testing. However, even without being a Bayesian there are simple reasons for methodological difficulties encountered with testing H_0: a large p-value could be caused by various other reasons besides of "the model $\mathcal{F}$ is (approximately) true". This might be: the number of observations is too small for rejection or the alternative lies in a different direction than those the test can effectively detect (recall the first paragraph). Particularly, for multivariate models this causes significant problems. Another problem in practice is, that if the sample size is large (in a clinical or econometrical study a few thousand observations quite often happens), *any* goodness of fit test for H_0 rejects due to the fact that arbitrary small deviations (although not relevant) from $\mathcal{F}$ will be detected. Therefore, in practice, often a significance level between 0.2 and 0.5 is chosen.

A radical proposal could be to reformulate the problem decision theoretically which leads, however, to serious mathematical and practical problems in infinite dimensional spaces (what is a proper loss function, how to choose the prior, etc.) Therefore, a compromise was suggested by Munk and Czado (1998), Dette and Munk (1998) and Munk (1999), which can be transferred to the present setting as follows. Choose a "suitable" metric

D and hypotheses $H_\pi : D(\mathcal{F}, F) > \pi$, where $D(\mathcal{F}, F)$ denotes the minimal distance between the model $\mathcal{F}$ and the "true" distribution F. Now, rejecting H_π and deciding in favour of a (neighborhood) of $\mathcal{F}$ can be performed at a controlled error rate. In fact, Munk and Czado (1998) suggested also a test based on the Wasserstein metric, sometimes denoted as Mallows distance. A variation of their result yields a central limit theorem if the true distribution F is not in $\mathcal{F}$ (Munk, Vogt and Freitag 2000), which is required for testing H_π. The additional difficulty – how to choose the distance π – can be circumvented by the simultaneous consideration of all p-values as a function of π, denoted as a p-value curve (Munk and Czado 1998, Czado and Munk 2000). The definition and interpretation of p-values becomes now simpler, because for smooth metrics (such as the Wasserstein metric) under H_π the asymptotic law is normal, instead of an infinite convolution of weighted and centered χ^2's. This leads to a simple graphical method which allows to visualize whether there is really evidence *for* the presence of $\mathcal{F}$, rather than simply the absence of evidence against $\mathcal{F}$. Currently, this method is investigated and extended to various other settings, including regression and survival analysis.

References

Berger, J.O. (1985). *Statistical Decision Theory and Bayesian Analysis*, 2nd. ed. Springer Verlag.

Berger, J.O. and T. Sellke (1987). Testing a point null hypothesis: the irreconcilability of p-values and evidence. *Journal of the American Statistical Association*, **82**, 112-122.

Czado, C. and A. Munk (2000). Noncanonical links in generalized linear models — when is the effort justified? *Journal of Statistical Planning and Inference*, to appear.

Dette, H. and A. Munk (1998). Validation of linear regression models. *Annals of Statistics*, **26**, 778-800.

Eubank, R.L. and V.N. LaRiccia (1992). Asymptotic comparison of Cramer-von Mises and nonparametric function estimation techniques for testing goodness-of-fit. *Annals of Statistics*, **20**, 2071-2086.

Gregory, G.G (1980). On efficiency and optimality of quadratic tests. *Annals of Statistics*, **8**, 116-131.

Kallenberg, W.C.M. and T. Ledwina (1997). Data-driven smooth tests when the hypothesis is composite. *Journal of the American Statistical Association*, **92**, 1094-1104.

Milbrodt, H. and H. Strasser (1990). On the asymptotic power of the two-sided Kolmogorov-Smirnov test. *Journal of Statistical Planning and Inference*, **26**, 1-23.

Munk, A. and C. Czado (1998). Nonparametric validation of similar distributions and the assessment of goodness of fit. *Journal of the Royal Statistical Society, B*, **60**, 223-241.

Munk, A. (1999). Testing the lack of fit in nonlinear regression models with random Toeplitz-forms. *Scandinavian Journal of Statistics*, under revision.

Munk, A., M. Vogt and G. Freitag (2000). Assessing structural relationships between distributions - a quantile process approach. In preparation.

Neuhaus, G. (1976). Asymptotic power properties of the Cramer-von Mises test under contiguous alternatives. *Journal of Multivariate Analysis*, **6**, 95-110.

Winfried Stute
University of Giessen, Germany

Frankly speaking, I very much enjoyed reading this paper. The authors did a nice job in reviewing some of the most important issues in goodness-of-fit testing. The material concentrates on i.i.d. real-valued data. In such a situation the empirical distribution function allows for a simple transformation to the uniform case and, as an alternative, also the quantile function may serve as a basic process.

Prior to Doob, Donsker, Kolmogorov and Smirnov, goodness-of-fit tests were mainly based on a comparison of frequencies and theoretical probabilities of finitely many cells. Mathematically, distributional approximations of finite-dimensional vectors were sufficient. With the 1940's, stochastic processes in the modern sense came into play. A priori any meaningful distance between the empirical distribution (or quantile) function and a hypothetical function may be considered. One possibility is to incorporate a weight function which, e.g., may serve to detect deviations in the tails. The quantile function is affine w.r.t. changes in location and scale. Therefore testing based on quantiles has always been popular for location-scale families. For composite models it is also worthwhile recalling that the need to estimate unknown parameters may have some serious impact on the distributional character of the underlying test process. Another important issue addressed in the paper is the possibility of the Karhunen-Loéve decomposition. This may be of some interest, when one wants to create directional and not just omnibus tests, by upweighting certain eigenfunctions and downweighting others.

A general strategy for deriving asymptotic properties of these tests is to study the underlying empirical or quantile process in an appropriate metric under the null and the alternative hypothesis, and then use the continuous mapping theorem. If, e.g., the weighting is overdone, some non-standard limit results come up, as was nicely pointed out in the last part of the paper.

Coming back to the underlying test (i.e., empirical or quantile) processes, the authors correctly point out the various approaches one can find in the literature:

2. The Vapnik-Chervonenkis Approach
3. The Hungarian Approach

Number 1 should be reserved for what could be called

1. The traditional Approach

(I prefer). This approach may be caracterized through issues like these: for computational feasibility, don't choose index families which are too complex, but restrict to intervals, rectangles of ellipsoids; don't go to asymptotics as soon as possible, but spend some time to check how good distributional approximations work for small to moderate sample size; sometimes it's useful not to prove everything in two steps (underlying process plus continuous mapping theorem), but to remember other techniques, like the elegant Hájek projection method, or to look for hidden martingale structures. Alternate techniques are becoming increasingly important if the data are no longer i.i.d. and 2. and 3. are unavailable.

I would also like to add a fourth approach, which is particularly powerful in goodness-of-fit testing:

4. The innovation process approach

This method is very successful when parameters need to be estimated. The idea is to transform the underlying process to the martingale part in its Doob-Meyer decomposition, which in the limit is a Brownian Motion in proper time. See Khmaladze (1981) for his key contribution to the subject.

It seems to be common use that discussants take the opportunity to also make some comments on their own contributions. Sections 1 and 2 and some parts of 3 are written in the spirit of Gaenssler and Stute (1979). In our monograph, Gaenssler and Stute (1987), we also discussed some goodness-of-fit problems in the multivariate setting. We also acknowledged the many contributions to a field which may be called

5. Combinatorial approach to goodness-of-fit testing

I would also like to add some comments on future directions in goodness-of-fit testing. The i.i.d. case for real-valued data is of course only the simplest case. In the multivariate setting the parametric bootstrap (see Stute et al. 1993) may at least be used to approximate the distributions, if the weight-functions are not too fancy. What is more interesting is goodness-of-fit testing in other situations:

a) When the data are incomplete (e.g., censored)

b) In time series, when it is required to fit the dynamics of the time series

c) In regression, when the target could be a parametric model for the regression function

d) The same as in b) and c), but the parametric model being replaced with a semiparametric model (like the Generalized Linear Model).

Some of my contributions to this area are reviewed in Stute (1997).

References

Gaenssler, P. and W. Stute (1979). Empirical processes: a survey of results for independent and identically distributed random variables. *Annals of Probability*, 7, 193-243.

Gaenssler, P. and W. Stute (1987). *Seminar on Empirical Processes.* Birkhaeuser, Basel.

Khmaladze, E. (1981). Martingale approach in the theory of goodness-of-fit tests. *Theory of Probability and its Applications*, **26**, 240-257.

Stute, W. (1997). Model checks in statistics: an innovation process approach. *IMS Lecture Notes*, **31**, 373-383. Hayward.

Stute, W., W. González Manteiga and M.P. Quindimil (1993). Bootstrap based goodness-of-fit tests. *Metrika*, **40**, 243-256.

Rejoinder by E. del Barrio, J.A. Cuesta and C. Matrán

When we began to write this paper, accepting the kind invitation of the editors of TEST, our goal was to contribute to this year of mathematical celebrations through a work relying on the mathematical evolution of empirical processes, one of the major developments in the recent theory of Statistics. Now, at the end, after the comments of the discussants our impression is that our goal has been clearly exceeded through this set of excellent comments.

Some of the discussants of the paper can be considered as major contributors to the theory of goodness-of-fit tests or to the theory of empirical processes. Therefore we would like to express our satisfaction for having the opportunity to share this work with such a distinguished group. In fact, in our opinion, the consequence of the discussants' comments is not only a wider scope than that of our initial paper. The new perspectives opened in the theory by this set of stimulating contributions are a major prize which was never expected by us.

Csörgő

Professor Csörgő poses several interesting questions concerning the limits of performance of Wasserstein correlation tests, the comparison of them to alternative procedures for testing fit to Gaussian or Weibull scale families or the rate of convergence to the asymptotic distribution in the uniform case. The different behavior of Wasserstein correlation tests depending on the tails of the family of distributions considered in the goodness-of-fit problem (powerful tests for families with "light tails" such as the uniform or the Gaussian; inefficient tests for families with "heavy tails" such as the exponential or the one considered in Subsubsection 3.3.3) motivates the exploration of how heavy can we allow tails to be if we want to maintain the good properties of the Gaussian case.

Professor Csörgő gives a brilliant answer to this question with his fine analysis of Wasserstein tests for the Weibull scale family. His Theorem 2 establishes clearly the border between tails for which Wasserstein tests have statistical interest and tails for which they do not (for those families considered in part (iii) of the theorem the weighted-Wasserstein-distance approach suggested by Professor de Wet could be a good alternative).

We would like to make some comments on the interesting problem posed by Professor Csörgő in part (iii) of Theorem 2, namely, finding the asymptotic distribution of

$$\int_0^n W^2(t)\big[(1+t)^2 \log^{2-\frac{2}{\alpha}}(1+t)\big]^{-1} dt, \tag{1}$$

for a Brownian motion $\{W(t)\}_t$. The equality in distribution between $\{W(t)/t^{1/2}\}_t$ and $\{V(2^{-1}\log(t(1-t)^{-1})\}_t$, where V denotes an Ornstein-Uhlenbeck process, allows us to reformulate the problem as the derivation of the asymptotic distribution of

$$\int_{-\infty}^{\frac{1}{2}\log n} V(y)^2 \left(\frac{e^{2y}}{1+e^{2y}}\right)^2 \frac{1}{\left(\log\left(\frac{1+e^{2y}}{e}{}^{2y}\right)+2y\right)^{2-\frac{2}{\alpha}}} dy,$$

which can be obtained through the study of

$$2^{\frac{2}{\alpha}-2} \int_0^{\frac{1}{2}\log n} V(y)^2 y^{\frac{2}{\alpha}-2} dy \tag{2}$$

(in fact the asymptotic behavior of both random integrals depends only on their right tails). An asymptotic result for integrals of the Ornstein-Uhlenbeck process with respect to measures other than Lebesgue (which corresponds to Mandl's result - Lemma 5.3.3 in Csörgő and Horváth 1993) is given by Theorem 6.8 in del Barrio, Giné and Matrán (1999). We quote here that result. Let Z denote a standard normal random variable. Then, if $\delta > -1$

$$\frac{1}{\sqrt{\frac{\delta \cdot 2^{-\delta}}{\delta+1} s^{(\delta+1)/2}}} \int_{-s/2}^{s/2} (|V(t)| - E|V(t)|)|t|^{\delta/2} dt \xrightarrow{w} \sqrt{1 + \frac{2\log 2}{\pi} - \frac{13}{3\pi}} Z,$$

while

$$\frac{1}{2(\log s)^{1/2}} \int_{-s/2}^{s/2} (|V(t)| - E|V(t)|)\, |t|^{-1/2} dt \xrightarrow{w} \sqrt{1 + \frac{2\log 2}{\pi} - \frac{13}{3\pi}} Z.$$

An adaptation of the proof of this result (based on expressing the integrals as sums of independent random variables from an infinitesimal array) might allow us to conclude that, for $\delta > -1$

$$\frac{1}{\sqrt{\frac{2^{-\delta}}{\delta+1} s^{(\delta+1)/2}}} \int_{0}^{s/2} (V(t)^2 - EV(t)^2) t^{\delta/2} dt \xrightarrow{w} Z$$

and

$$\frac{1}{2(\log s)^{1/2}} \int_{0}^{s/2} (V(t)^2 - EV(t)^2) t^{-1/2} dt \xrightarrow{w} Z.$$

This could give the asymptotic normality of the integral functional in (2) (hence of the one in (1)) as long as $2/\alpha - 2 \geq -1/2$, that is, as long as $\alpha \leq 4/3$, which would complete part (iii) of Theorem 2 in Professor Csörgő's comment. However, filling the gaps in the above mentioned adaptation does not seem to be straightforward.

Cuadras

Professor Cuadras discusses on some subjects related to the content of the paper. First, he introduces an orthogonal expansion of a random variable on principal directions which can be obtained via Karhunen-Loéve expansions. Then, he presents several bounds for the Wasserstein distance. Finally, he shows the relation between the statistic $\mathcal{R}_n$ and the so-called maximal Hoeffding correlation between the empirical and the theoretical distributions, which, in turn, is related to the previously introduced orthogonal expansion.

de Wet

It is a great pleasure for us to have Professor de Wet as a discussant since he was one of the authors who obtained the first asymptotic result in the field of regression tests. Professor de Wet points out three important possibilities to extend the Wasserstein tests of goodness of fit.

The first one consists of handling a weighted Wasserstein distance. Professor de Wet analyzes this possibility in the framework of a scale family and he nicely shows that if the weight is properly chosen (depending on the family under consideration), then some optimal cancellation properties are obtained. Moreover, this fact explains some particular properties obtained in the paper, because it turns out that the constant weight is the optimal one for the Gaussian family.

On the other hand, he suggests the possibility of using Wasserstein tests in the time series framework and, finally, he recalls a procedure which can be employed to construct a Wasserstein distance test to check fit to a multidimensional Gaussian distribution.

Giné

We agree with Professor Giné's point of view regarding the use of strong approximations or Banach spaces techniques for proving weak limit theorems. The latter are often the right choice for deriving an asymptotic distribution using more elementary methods but sometimes strong approximations seem to be the only way to handle the problem. In the particular case of the Chibisov-O'Reilly theorem we wonder if Empirical Processes Theory can also give Theorem 2.5, that is, necessary and sufficient conditions for the weak convergence of the supremum norm of the weighted empirical process (not for the weak convergence of the weighted empirical process itself), without using strong approximations. The answer to this question could give a new chance for assessing the strengths and weaknesses of both approaches to weak limit theorems.

Professor Giné corrects one important omission in our paper and we thank him for doing so. We are talking about A. Cabaña and E. Cabaña's approach to the goodness-of-fit problem based on the transformed empirical process. The main goal of this method is the derivation of tests of fit with maximum power for detecting a particular sequence of contiguous alternatives.

Lockhart

The first part of Professor Lockhart's discussion focuses on the interest, from the point of view of the practical applications, of asymptotic results. First, he fixes the framework by making some general considerations on the way in which statisticians use to work and the way in which asymptotic distributions apply in this field. After this, he gives some arguments supporting the Monte Carlo method as an alternative to the use of the asymptotic distributions in approximating the distribution of a given statistic. We must admit, without any doubt, that this is an important point we had missed in our work.

Professor Lockhart ends his discussion by specifying some details on his contributions on the subject of the asymptotic power of the correlation tests which were not clear enough in our paper.

Munk

Professor Munk's discussion focuses on two important aspects of goodness-of-fit tests. One of them is basic: What does a goodness-of- fit test really do? Professor Munk rightly exposes some doubts on the adequacy of choosing as null hypothesis the validity of the model under testing, based on the fact that, even if the data do not allow us to reject the hypothesized model, we have no guarantee that it is right. Then, he proposes a solution consisting of taking as null hypothesis that the model of interest does *not* hold approximately. Now, rejecting the null hypothesis means the data contain aspects supporting that the model is (approximately) correct.

The second aspect analyzed is which test should be chosen. Professor Munk assumes we are interested in checking fit to some fixed distribution and we have to choose between a Cramer-von Mises or a Shapiro-Wilk-type test. He makes a nice analysis of the circumstances under which each of them should be preferred. This analysis is based, mostly, on Karhunen-Loéve expansions of the limit distributions.

Stute

Professor Stute makes a very accurate comment on our work. He begins by listing the different techniques usually employed when handling empirical or quantile processes as well as pointing out the advantages of his (and ours) preferred approach: the traditional one.

On the other hand he also recalls three aspects missed in the work. The first one is methodological and consists of the possibility to employ the relatively new "innovation process approach". Then he suggests the possibility to employ the bootstrap to approximate the distributions of goodness-of-fit tests in the multivariate setting. Finally, he includes a short list of very important goodness-of-fit problems which are not mentioned in the paper.

We were not aware of the existence of the innovation process approach when writing the paper and, at this moment, we agree with Professor Stute on its usefulness to handle problems in which parameters have to be estimated.

Concerning other goodness-of-fit problems, we have fixed the scope of our work on the location-scale problem and the i.i.d. case because, otherwise, a work along the lines we have chosen had been, instead a paper, a quite bulky book. The same can be said about the bootstrap. However, we admit that, at least, we should have mentioned the existence of those possibilities. This fault has been corrected in his comment.

References

del Barrio, E., E. Giné and C. Matrán (1999). Central limit theorems for the Wasserstein distance between the empirical and the true distributions. *Annals of Probability*, **27**, 1009-1071.