Pré-Publicações do Departamento de Matemática
Universidade de Coimbra
Preprint Number 18–53

ON THE AUTOMATIC SELECTION OF THE TUNING PARAMETER APPEARING IN CERTAIN FAMILIES OF GOODNESS-OF-FIT TESTS

CARLOS TENREIRO

ABSTRACT: The situation, common in the current literature, is that of a whole family of location-scale/scale invariant test statistics, indexed by a parameter $\lambda \in \Lambda$, is available to test the goodness of fit of F, the underlying distribution function of a set of independent real-valued random variables, to a location-scale/scale family of distribution functions. The power properties of the tests associated with the different statistics usually depend on the parameter λ, called the "tuning parameter", which is the reason that its choice is crucial to obtain a performing test procedure. In this paper, we address the automatic selection of the tuning parameter when Λ is finite, as well as the calibration of the associated goodness-of-fit test procedure. Examples of existing and new tuning parameter selectors are discussed, and the methodology presented of combining different test statistics into a single test procedure is applied to well known families of test statistics for normality and exponentiality. A simulation study is carried out to access the power of the different tests under consideration, and to compare them with the fixed tuning parameter procedure, usually recommended in the literature.

KEYWORDS: Goodness-of-fit tests; data-based tuning parameter selection; calibration; normality tests; empirical characteristic function; exponentiality tests; empirical Laplace transform.
AMS SUBJECT CLASSIFICATION (2010): 62G10; 62G20.

1. Introduction

Given a sample $X_1, \ldots, X_n$ of independent and identically distributed real-valued random variables from a distribution function F, assume that $T_{n,\lambda} = T_{n,\lambda}(X_1, \ldots, X_n)$, for $\lambda \in \Lambda$, is a finite family of statistics for testing the hypothesis

$$H_0: \ F \in \mathcal{F}, \tag{1}$$

against a general alternative hypothesis, where $\mathcal{F}$ is either a location-scale family,

$$\mathcal{F} = \left\{ F_0((\cdot - b)/a) : a > 0, b \in \mathbb{R} \right\} \tag{2}$$

Received December 7, 2018.

or a scale family of distribution functions,

$$\mathcal{F} = \big\{F_0(\cdot/a) : a > 0\big\}, \tag{3}$$

with F_0 a known distribution function on $\mathbb{R}$. If the test statistics $T_{n,\lambda}$ are location-scale invariant in case (2), that is, $T_{n,\lambda}(\nu X_1 + \mu, \ldots, \nu X_n + \mu) = T_{n,\lambda}(X_1, \ldots, X_n)$, for each $\nu > 0$ and $\mu \in \mathbb{R}$, or scale invariant in case (3), that is, $T_{n,\lambda}(\nu X_1, \ldots, \nu X_n) = T_{n,\lambda}(X_1, \ldots, X_n)$, for each $\nu > 0$, the distribution of $T_{n,\lambda}$ under H_0 does not depend on F. Therefore, if large values of $T_{n,\lambda}$ are significant, each one of the tests with critical regions $\{T_{n,\lambda}(X_1, \ldots, X_n) > c_{n,\lambda}(\alpha)\}$, has a level of significance at most equal to α, that is, $P_F\big(T_{n,\lambda}(X_1, \ldots, X_n) > c_{n,\lambda}(\alpha)\big) \leq \alpha$, for all $F \in \mathcal{F}$, where $\alpha \in]0, 1[$, and $c_{n,\lambda}(\alpha) = F^{-1}_{T_{n,\lambda}}(1-\alpha)$ denotes the quantile of order $1-\alpha$ of $T_{n,\lambda}$ under H_0. Of course, if the distribution functions of all $T_{n,\lambda}$ under H_0, are continuous on $\mathbb{R}$, the test procedures associated with the previous critical regions have a level of significance exactly equal to α. The power properties of the previous test procedures usually depend on the parameter λ which is the reason that its choice is crucial to obtain a performing test procedure.

The previous situation, where a finite family of test statistics is available for testing the hypothesis H_0, is now common in the current literature as evidenced by the works of Epps and Pulley (1983), Baringhaus and Henze (1991), Henze (1993), Güntler and Henze (2000), Klar (2001), Henze and Meintanis (2002), Meintanis (2004, 2004a), and Meintanis et al. (2014), where goodness-of-fit tests for the normal, exponential, Cauchy, Laplace, or logistic distributions, based on the empirical characteristic function, the probability weighted characteristic function, the integrated empirical distribution function or the Laplace transform, are proposed (for related work, see also Henze and Zirkler, 1990, Fan, 1998, Tenreiro, 2005, 2007). In all these situations the test statistics $T_{n,\lambda}$, $\lambda \in \Lambda$, are either location-scale invariant in case (2) or scale invariant in case (3). More precisely, they can be written in the form $T_{n,\lambda}(X_1, \ldots, X_n) = \bar{T}_{n,\lambda}(Y_1, \ldots, Y_n)$ with

$$(Y_1, \ldots, Y_n) = g(X_1, \ldots, X_n), \tag{4}$$

where g is a known function given by $g(X_1, \ldots, X_n) = \big(\frac{X_1 - \hat{b}_n}{\hat{a}_n}, \ldots, \frac{X_n - \hat{b}_n}{\hat{a}_n}\big)$, in case (2), and by $g(X_1, \ldots, X_n) = \big(\frac{X_1}{\hat{a}_n}, \ldots, \frac{X_n}{\hat{a}_n}\big)$, in case (3), where $\hat{a}_n = \hat{a}_n(X_1, \ldots, X_n)$ is a location invariant and scale equivariant estimator of the scale parameter a, i.e., $\hat{a}_n(\nu X_1 + \mu, \ldots, \nu X_n + \mu) = \nu\, \hat{a}_n(X_1, \ldots, X_n)$, for each $\nu > 0$ and $\mu \in \mathbb{R}$, and $\hat{b}_n = \hat{b}_n(X_1, \ldots, X_n)$ is a location-scale equivariant estimator of location parameter b, i.e., $\hat{b}_n(\nu X_1 + \mu, \ldots, \nu X_n + \mu) = \nu\, \hat{b}_n(X_1, \ldots, X_n) + \mu$, for each $\nu > 0$ and $\mu \in \mathbb{R}$.

Focusing our attention on the normality test introduced by Epps and Pulley (1983), the considered test statistic is defined as a weighted L_2-distance between the empirical characteristic function of the scaled residual $Y_j = (X_j - \bar{X}_n)/S_n$, $j = 1, \ldots, n$, given by $\varphi_n(t) = \frac{1}{n}\sum_{j=1}^n \exp\big(itY_j\big)$, $t \in \mathbb{R}$, and the characteristic function $\varphi(t) = \exp(-t^2/2)$ of the standard Gaussian density $\phi(x) = (2\pi)^{-1/2}\exp(-x^2/2)$, $x \in \mathbb{R}$, where $\bar{X}_n = n^{-1}\sum_{j=1}^n X_j$ and $S_n^2 = n^{-1}\sum_{j=1}^n (X_j - \bar{X}_n)^2$, are the sample mean and the sample variance, respectively. The weight function is given by $t \mapsto \exp(-\lambda^2 t^2)$, where λ a strictly positive real number that needs to be chosen by the user. Therefore the Epps-Pulley test statistic is given by

$$T_{n,\lambda} = n\int_{\mathbb{R}} |\varphi_n(t) - \varphi(t)|^2 \exp(-\lambda^2 t^2)dt = 2\pi\frac{1}{n}\sum_{i,j=1}^n Q(Y_i, Y_j; \lambda), \tag{5}$$

with $Q(u, v; \lambda) = \phi_{(2\lambda^2)^{1/2}}(u-v) - \phi_{(1+2\lambda^2)^{1/2}}(u) - \phi_{(1+2\lambda^2)^{1/2}}(v) + \phi_{(2+2\lambda^2)^{1/2}}(0)$, for $u, v \in \mathbb{R}$, $\lambda \in]0, +\infty[$, and $\phi_h(\cdot) = \phi(\cdot/h)/h$, $h > 0$. The simplicity of the previous expression shows the attractive feature of the considered weight function (see Henze and Zirkler, 1990, and Fan, 1998, for the relation between the Epps-Pulley test statistic and the Bickel-Rosenblatt test statistic). From a practical point of view, it is well known that the finite sample performance of the Epps-Pulley test is very sensitive to the choice of λ. Choosing a small value of λ, which means letting the weight function decay slowly, will produce a test sensitive to short tailed or high moment alternatives, whereas large values of λ, which means putting most of the mass of the weight function near zero, are adequate for detecting alternative distributions with long tails, symmetric or asymmetric (cf. Tenreiro, 2009). Note that for large values of λ the considered weight function puts most of its mass near the origin, and then, the previous behaviour can be seen as a consequence of the fact that the tail behaviour of a probability distribution is reflected by the behaviour of its characteristic function at the origin (cf. Kawata, 1972, pp. 419–420).

The exponentiality tests introduced in Henze and Meintanis (2002), is another example where a whole family of test statistics is available to the user. In this case the test statistics are based on a weighted L_2-distance between the empirical Laplace transform of the scaled data $Y_j = X_j/\bar{X}_n$, $j = 1, \ldots, n$, defined by $\psi_n(t) = \frac{1}{n}\sum_{j=1}^n \exp\big(-tY_j\big)$, $t \geq 0$, and the Laplace transform of the unit exponential distribution $\psi(t) = 1/(1+t)$, $t \geq 0$, with weight function $t \mapsto (1+t)^2\exp(-\lambda t)$, for $\lambda > 0$. Thus, the Henze-Meintanis test statistic is given by

$$\begin{aligned} T_{n,\lambda} &= n\int_0^\infty \big(\psi_n(t)-\psi(t)\big)^2(1+t)^2\exp(-\lambda t)dt \\ &= \frac{1}{n}\sum_{j,k=1}^n \frac{1+(Y_j+Y_k+\lambda+1)^2}{(Y_j+Y_k+\lambda)^3} - 2\sum_{j=1}^n \frac{1+Y_j+\lambda}{(Y_j+\lambda)^2} + \frac{n}{\lambda}, \end{aligned} \tag{6}$$

for $\lambda \in]0,+\infty[$. As the Epps-Pulley test for normality, the Henze-Meintanis test for exponentiality is very sensitive to the choice of λ. As remarked by Baringhaus and Henze (1991, p. 552) (see also Henze and Meintanis, 2002, p. 148), from Tauberian theorems on Laplace transform (cf. Feller, 1971, Chapter XIII.5), it is known that the tail behaviour of a probability distribution concentrated on $[0,\infty[$ is reflected by the behaviour ot its Laplace transform at zero and vice versa. Therefore, choosing a small value of λ, which means letting the weight function decay slowly, gives high power against alternative distributions having a point mass or infinite density at zero, and a large value of λ means putting most of the mass of the weight function near zero, which should give high power against alternatives that greatly differ in tail behaviour with respect to the exponential distribution.

As illustrated by the previous examples, the parameter λ acts as a tuning parameter, through which the user can increase the power of the test toward some particular direction along the alternative distribution set. However, as the formulation of a specific alternative hypothesis is, in general, impossible in a real situation, the usual practice is to evaluate the test power performance for λ varying in a finite set Λ, and then suggesting a selection of λ that produces a test with a reasonable power against a wide range of alternative distributions. However, this strategy of taking a fixed tuning parameter does not prevent the user from obtaining a test that achieves a very low power against some of the considered alternative distributions (cf. Henze and Meintanis, 2002).

Some efforts have been made in order to combine, based on the available data, test procedures associated to different values of the tuning parameter λ into a single test procedure that could show a good power performance against a wide range of alternative distributions. This was the case of the multiple test approach, considered in Klar (2001), Fromont and Laurent (2006) and Tenreiro (2011, 2017), which can be viewed as an improvement of the classical Bonferroni multiple test procedure. The proposed test leads to the rejection of the null hypothesis if one of the statistics $T_{n,\lambda}$, for $\lambda \in \Lambda$, is larger than its $(1-u)$ quantile under the null hypothesis, the level u being calibrated so that the resulting multiple test has a level of significance at most equal to α. Thus, the associated critical region is given by

$$\Big\{ \max_{\lambda\in\Lambda} \big(T_{n,\lambda} - c_{n,\lambda}(u)\big) > 0 \Big\}, \tag{7}$$

for some $u \in]0,1[$. This testing procedure is closely related to the single-step *minP* multiple testing procedure based on minima of unadjusted p-values (cf. Dudoit and van der Lann, 2008, pp. 117–121). Unlike classical Bonferroni multiple, that can be obtained by taking $u = \alpha/|\Lambda|$, where $|\Lambda|$ denotes the cardinality of Λ, the previous rejection region takes into account the dependence structure among the test statistics $T_{n,\lambda}$ for $\lambda \in \Lambda$. As the previous critical region can be written as $\big\{T_{n,\bar\lambda_u} > c_{n,\bar\lambda_u}(u)\big\}$, where

$$\bar\lambda_u = \bar\lambda_u(X_1,\ldots,X_n) = \arg\max_{\lambda\in\Lambda} \big(T_{n,\lambda}(X_1,\ldots,X_n) - c_{n,\lambda}(u)\big), \tag{8}$$

the previous multiple test procedure can be seen as a test based on a data-dependent procedure for selecting the tuning parameter λ: for a given sample of size n, one selects the value λ in Λ for which the test statistic $T_{n,\lambda}$ shows stronger evidence, at level u, against the null hypothesis.

More recently, Allison and Santana (2015) proposed an alternative data-dependent method, based on the bootstrap, for choosing the tuning parameter. They considered the test with critical region $\{T_{n,\lambda^*} > c_{n,\lambda^*}(\alpha)\}$, where $\lambda^* = \lambda^*(X_1,\ldots,X_n)$ is obtained by maximizing the bootstrap power, that is,

$$\lambda^* = \lambda^*(X_1,\ldots,X_n) = \arg\max_{\lambda\in\Lambda} \frac{1}{B}\sum_{k=1}^{B} I\big(\bar T_{n,\lambda}(Y^*_{k1},\ldots,Y^*_{kn}) > c_{n,\lambda}(\alpha)\big),$$

where $(Y^*_{k1},\ldots,Y^*_{kn})$ is a bootstrap random sample of size n drawn with replacement from the empirical distribution function of the transformed sample $(Y_1,\ldots,Y_n)$ defined by (4), $B \in \mathbb{N}$ is the considered number of bootstrap samples, and $I(A)$ denotes the indicator function of the set A. Unfortunately, the proposed method presents two important drawbacks. Firstly, the suggested bootstrap procedure, as based on a bootstrap random sample drawn from the empirical distribution function of the transformed sample (4), and not from the empirical distribution function of the original sample, does not always produce a good approximation for the power associated with the distribution of the observations. Secondly, by using the quantiles of order $1-\alpha$ of each one of the test statistics $T_{n,\lambda}$ to define the critical region, the proposed test is not correctly calibrated and may reach a level of significance much bigger than α (see Section 4, Figure 1). In order to overcome these problems, we consider in this paper the test with critical region $\{T_{n,\tilde{\lambda}_u} > c_{n,\tilde{\lambda}_u}(u)\}$, for $u \in]0,1[$, where the modified tuning parameter selector $\tilde{\lambda}_u$, is defined by

$$\tilde{\lambda}_u(X_1,\ldots,X_n) = \arg\max_{\lambda\in\Lambda} \frac{1}{B}\sum_{k=1}^{B} I\big(T_{n,\lambda}(X^*_{k1},\ldots,X^*_{kn}) > c_{n,\lambda}(u)\big), \tag{9}$$

with $X^*_{kj} = X_{U_{(k-1)n+j}}$, for $k=1,\ldots,B$ and $j=1,\ldots,n$, where U_l, for $l=1,\ldots,nB$, are independent copies of the discrete uniform distribution on $\{1,\ldots,n\}$, and u is calibrated so that the test has a level of significance at most equal to α. Although not assumed or discussed in this paper, if, for B and n large enough, the mean in (9) gives a good approximation for the probability $P_F(T_{n,\lambda}(X_1,\ldots,X_n) > c_{n,\lambda}(u))$, we might expect that $\tilde{\lambda}_u$ mimics the behaviour of the ideal tuning parameter

$$\lambda_u(F) = \arg\max_{\lambda\in\Lambda} P_F\big(T_{n,\lambda}(X_1,\ldots,X_n) > c_{n,\lambda}(u)\big), \tag{10}$$

this being the main motivation for the previous definition of $\tilde{\lambda}_u$. As $\tilde{\lambda}_u$ depends on $U=(U_l, l=1,\ldots,nB) \in \{1,\ldots,n\}^{nB}$, any statement on this tuning parameter selector should always be interpreted conditionally on U. However, by the law of large numbers, different choices of U, essentially lead to tuning parameter selectors with similar behaviours.

The paper is organised as follows. Sections 2 and 3 deal with the calibration and the consistency of the tests with critical region $\{T_{n,\hat{\lambda}_u} > c_{n,\hat{\lambda}_u}(u)\}$, where $\hat{\lambda}_u = \hat{\lambda}_u(X_1,\ldots,X_n)$ is a general family of measurable functions, indexed by $u \in]0,1[$, taking values in Λ, which are either location-scale invariant in case (2) or scale invariant in case (3). The cases of the tuning parameter selectors $\bar{\lambda}_u$ and $\tilde{\lambda}_u$, defined by (8) and (9), respectively, are analysed in detail. In Sections 4 and 5 we will restrict our attention to the cases where $T_{n,\lambda}$ is either the test statistic for normality of Epps and Pulley (1983), or the test statistic for exponentiality of Henze and Meintanis (2002). We conclude that the proposed calibration procedure is effective, and, as a result of a simulation study, we deduce that the tests based on $\bar{\lambda}_u$ and $\tilde{\lambda}_u$ are serious competitors for the tests based on a fixed tuning parameter usually recommended in the literature, and perhaps should be employed in practice, in the absence of any information about the type of deviation from the null model. All the proofs and some auxiliar results are deferred to Section 6. The simulations and plots in this article were carried out using the R software (R Core Team, 2014).

2. The calibration procedure

In this section we will denote by $T_{n,\lambda}$, for $\lambda \in \Lambda$, a finite family of test statistics for testing the hypothesis (1), whose large values are considered significant. We will also assume that such test statistics are either location-scale invariant in case (2) or scale invariant in case (3). This assumption enables us to consider that the quantiles of order $1-u$ of $T_{n,\lambda}$ under H_0, denoted, as before, by $c_{n,\lambda}(u) = F^{-1}_{T_{n,\lambda}}(1-u)$, are known quantities as they can be approximated by performing Monte Carlo experiments under the null hypothesis.

Given a general invariant tuning parameter selector $\hat{\lambda}_u = \hat{\lambda}_u(X_1, \ldots, X_n)$, that is, a family of measurable functions indexed by $u \in]0,1[$, taking values in Λ, which are either location-scale invariant in case (2) or scale invariant in case (3), we have

$$\big\{T_{n,\hat{\lambda}_u} > c_{n,\hat{\lambda}_u}(u)\big\} \subset \Big\{\max_{\lambda\in\Lambda}\big(T_{n,\lambda} - c_{n,\lambda}(u)\big) > 0\Big\} = \bigcup_{\lambda\in\Lambda}\big\{T_{n,\lambda} > c_{n,\lambda}(u)\big\}, \quad (11)$$

from which we conclude that it is always possible to choose $u \in]0,1[$ such that the test with critical region $\{T_{n,\hat{\lambda}_u} > c_{n,\hat{\lambda}_u}(u)\}$, has a level of significance at most equal to the nominal level α:

Theorem 1. *Given an invariant tuning parameter selector $\hat{\lambda}_u$, and $\alpha \in]0,1[$, then for all $0 < u \leq \alpha/|\Lambda|$, we have*

$$P_F\big(T_{n,\hat{\lambda}_u} > c_{n,\hat{\lambda}_u}(u)\big) \leq \alpha, \ \textit{for all } F \in \mathcal{F},$$

where the probability $P_F\big(T_{n,\hat{\lambda}_u} > c_{n,\hat{\lambda}_u}(u)\big)$, we denote by $\psi_{\hat{\lambda}}(u)$, is independent of F, for $F \in \mathcal{F}$.

Note that the previous assumptions are fulfilled by the tuning parameter selector $\bar{\lambda}_u$ given by (8), and, conditionally on $U = (U_l, l = 1, \ldots, nB)$, by $\tilde{\lambda}_u$ defined by (9).

2.1. A first calibration stage. Although important, as it assures that the Type I error of the test with critical region $\{T_{n,\hat\lambda_u} > c_{n,\hat\lambda_u}(u)\}$, may be put under a preassigned level of significance α through an appropriate choice of u, Theorem 1 does not provide a criterium for such a choice. Taking into account that the test should have a level of significance not only less than or equal to α but also as close to α as possible, the practical selection of u will be performed by considering a regular grid $G_p = \{u_k, k \in I_p\}$ on the interval $]0,1[$, where $u_1 = p$ and $u_{k+1} = u_k + p$, for some $0 < p \le \alpha/|\Lambda|$, and taking $u = u^{\hat\lambda}_{n,\alpha,p}$, where $u^{\hat\lambda}_{n,\alpha,p}$ is the largest value of G_p satisfying $\psi_{\hat\lambda}(u) := P_{F_0}\big(T_{n,\hat\lambda_u} > c_{n,\hat\lambda_u}(u)\big) \le \alpha$, that is,

$$u^{\hat\lambda}_{n,\alpha,p} = \max\{u \in G_p : \psi_{\hat\lambda}(u) \le \alpha\}. \tag{12}$$

We present in the next theorem a set of sufficient conditions on $\psi_{\hat\lambda}$ assuring that the test with critical region $\{T_{n,\hat\lambda_u} > c_{n,\hat\lambda_u}(u)\}$, for $u = u^{\hat\lambda}_{n,\alpha,p}$, has a level of significance as close to α as possible, when p tends to zero.

Theorem 2. *Given an invariant tuning parameter selector $\hat\lambda_u$, let $\alpha \in]0,1[$, and $u^{\hat\lambda}_{n,\alpha,p}$ given by (12). If $\psi_{\hat\lambda}$ is increasing on $]0,1[$, with $\lim_{u\uparrow 1}\psi_{\hat\lambda}(u) = 1$, we have $\lim_{p\downarrow 0} u^{\hat\lambda}_{n,\alpha,p} = u^{\hat\lambda}_{n,\alpha}$, where*

$$u^{\hat\lambda}_{n,\alpha} = \sup\{u \in]0,1[\, : \psi_{\hat\lambda}(u) \le \alpha\}, \tag{13}$$

is such that $\alpha/|\Lambda| \le u^{\hat\lambda}_{n,\alpha} < 1$. Moreover, if $\psi_{\hat\lambda}$ is continuous on $]0,1[$, we have

$$\lim_{p\downarrow 0}\sup_{F\in\mathcal{F}} P_F\big(T_{n,\hat\lambda_{u_p}} > c_{n,\hat\lambda_{u_p}}(u_p)\big) = \sup_{F\in\mathcal{F}} P_F\big(T_{n,\hat\lambda_u} > c_{n,\hat\lambda_u}(u)\big) = \alpha,$$

where u_p and u stand for $u^{\hat\lambda}_{n,\alpha,p}$ and $u^{\hat\lambda}_{n,\alpha}$, respectively.

The previous general result allows us to present a set of sufficient conditions on the null distribution of the statistics $T_{n,\lambda}$, weaker that those considered in Tenreiro (2011, Theorem 1), assuring that the test associated with the critical region $\{T_{n,\bar\lambda_u} > c_{n,\bar\lambda_u}(u)\}$, with $\bar\lambda_u$ given by (8) and $u = u^{\bar\lambda}_{n,\alpha,p}$ given by (12), has a level of significance not only inferior, but also as close to α as possible, when p tends to zero.

Theorem 3. *If the distribution function of $T_{n,\lambda}$ under H_0 is continuous, for all $\lambda \in \Lambda$, then, for $\alpha \in]0,1[$ we have*

$$\lim_{p \downarrow 0} \sup_{F \in \mathcal{F}} P_F\big(T_{n,\bar{\lambda}_{u_p}} > c_{n,\bar{\lambda}_{u_p}}(u_p)\big) = \sup_{F \in \mathcal{F}} P_F\big(T_{n,\bar{\lambda}_u} > c_{n,\bar{\lambda}_u}(u)\big) = \alpha,$$

where u_p and u stand for $u^{\bar{\lambda}}_{n,\alpha,p}$ and $u^{\bar{\lambda}}_{n,\alpha}$ given by (12) and (13), respectively, with $\alpha/|\Lambda| \le u^{\bar{\lambda}}_{n,\alpha} \le \alpha$.

2.2. A second calibration stage. Under the assumptions of Theorem 3 on the null distribution of the statistics $T_{n,\lambda}$, it can be proved that $\lim_{u \uparrow 1} \psi_{\hat{\lambda}}(u) = 1$, for any invariant tuning parameter selector $\hat{\lambda}_u$ (see Section 6, Proposition 1). However, the same set of assumptions does not necessarily assure that $\psi_{\hat{\lambda}}$ is increasing and continuous on $]0,1[$. Therefore, under the assumptions of Theorem 3 the test with critical region $\{T_{n,\hat{\lambda}_u} > c_{n,\hat{\lambda}_u}(u)\}$, with $u = u^{\hat{\lambda}}_{n,\alpha,p}$ given by (12), has a level of significance no bigger than α, but we cannot affirm that its level of significance becomes as close to α as possible, when p tends to zero. Next, we will see that this goal can be achieved if a second calibration stage is performed.

For a fixed $u \in]0,1[$, consider the family of critical regions $\big\{T_{n,\hat{\lambda}_u} > c_{n,\hat{\lambda}_u}(v)\big\}$, indexed by $v \in]0,1[$. Although it is natural to take for u the value $u^{\hat{\lambda}}_{n,\alpha,p}$ obtained in the first calibration stage, this is not assumed in what follows. Define

$$v^{\hat{\lambda}_u}_{n,\alpha,q} = \max\{v \in G_q : \psi_{\hat{\lambda}_u}(v) \le \alpha\}, \tag{14}$$

where $\psi_{\hat{\lambda}_u}$ is given by $\psi_{\hat{\lambda}_u}(v) := P_{F_0}\big(T_{n,\hat{\lambda}_u} > c_{n,\hat{\lambda}_u}(v)\big)$, and $G_q = \{v_k, k \in I_q\}$, is a regular grid on the interval $]0,1[$, with $v_1 = q$ and $v_{k+1} = v_k + q$, for some $0 < q < u$.

Theorem 4. *Given an invariant tuning parameter selector $\hat{\lambda}_u$, where u is assumed to be fixed in $]0,1[$, if the distribution function of $T_{n,\lambda}$ under H_0 is strictly increasing (on the set $\{t \in \mathbb{R} : 0 < F_{T_{n,\lambda}}(t) < 1\}$) and continuous, for all $\lambda \in \Lambda$, then, for $\alpha \in]0,1[$, we have $\lim_{q \downarrow 0} v^{\hat{\lambda}_u}_{n,\alpha,q} = v^{\hat{\lambda}_u}_{n,\alpha}$, where*

$$v^{\hat{\lambda}_u}_{n,\alpha} = \sup\{v \in]0,1[: \psi_{\hat{\lambda}_u}(v) \le \alpha\}, \tag{15}$$

is such that $\max(u, \alpha/|\Lambda|) \leq v_{n,\alpha}^{\hat{\lambda}_u} < 1$, *whenever* $\psi_{\hat{\lambda}}(u) \leq \alpha$, *and* $\alpha/|\Lambda| \leq v_{n,\alpha}^{\hat{\lambda}_u} \leq u$, *whenever* $\psi_{\hat{\lambda}}(u) > \alpha$. *Moreover, we have*

$$\lim_{q \downarrow 0} \sup_{F \in \mathcal{F}} P_F\big(T_{n,\hat{\lambda}_u} > c_{n,\hat{\lambda}_u}(v_{u,q})\big) = \sup_{F \in \mathcal{F}} P_F\big(T_{n,\hat{\lambda}_u} > c_{n,\hat{\lambda}_u}(v_u)\big) = \alpha,$$

where $v_{u,q}$ *and* v_u *stand for* $v_{n,\alpha,q}^{\hat{\lambda}_u}$ *and* $v_{n,\alpha}^{\hat{\lambda}_u}$, *respectively.*

This general result gives us a set of sufficient conditions on the null distribution of the statistics $T_{n,\lambda}$, assuring that, conditionally on $U = (U_l, l = 1, \ldots, nB)$, the test with critical region $\{T_{n,\tilde{\lambda}_u} > c_{n,\tilde{\lambda}_u}(v_u)\}$, with $u = u_{n,\alpha,p}^{\tilde{\lambda}}$ and $v_u = v_{n,\alpha,q}^{\tilde{\lambda}_u}$ given by (12) and (14), respectively, has a level of significance not only less than or equal but also as close to α as possible, when q tends to zero.

3. Consistency against fixed alternatives

Under some general assumptions, the test procedures considered in the previous section detect an alternative $F \notin \mathcal{F}$, if such an alternative is detected by all the test statistics $T_{n,\lambda}$, for $\lambda \in \Lambda$. Next we will restrict our attention to the tests that use a single calibration stage. However, similar results can be derived for the test procedures considered in subsection 2.2, where a second calibration stage is used.

Theorem 5. *Under the conditions of Theorem 2, let* $F \notin \mathcal{F}$, *and assume that* $T_{n,\lambda} \xrightarrow{p} +\infty$, *under* F, *for all* $\lambda \in \Lambda$. *If* $T_{n,\lambda} \xrightarrow{d} T_{\infty,\lambda}$, *under* H_0, *with* $F_{T_{\infty,\lambda}}$ *strictly increasing (on the set* $\{t \in \mathbb{R} : 0 < F_{T_{\infty,\lambda}}(t) < 1\}$*), for all* $\lambda \in \Lambda$, *then*

$$P_F\big(T_{n,\hat{\lambda}_u} > c_{n,\hat{\lambda}_u}(u)\big) \longrightarrow 1, \text{ as } n \to \infty,$$

where u *stands for* $u_{n,\alpha}^{\hat{\lambda}}$ *given by (13).*

In the particular case of the tuning parameter selector $\bar{\lambda}_u$ given by (8), it is interesting to note that the previous consistency result may be obtained under weaker assumptions. In fact, the test with critical region $\{T_{n,\bar{\lambda}_u} > c_{n,\bar{\lambda}_u}(u)\}$, with $u = u_{n,\alpha}^{\bar{\lambda}}$, detects any alternative $F \notin \mathcal{F}$ that is detected by at least one of the tests based on $T_{n,\lambda}$, for $\lambda \in \Lambda$. This attractive property of the tuning parameter selector $\bar{\lambda}_u$ is stated in the following result.

Theorem 6. *Under the conditions of Theorem 3, let* $F \notin \mathcal{F}$, *and assume there exists* $\lambda \in \Lambda$ *such that* $T_{n,\lambda} \xrightarrow{p} +\infty$, *under* F. *If* $T_{n,\lambda} \xrightarrow{d} T_{\infty,\lambda}$, *under*

H_0, where $F_{T_{\infty,\lambda}}$ is strictly increasing (on the set $\{t \in \mathbb{R} : 0 < F_{T_{\infty,\lambda}}(t) < 1\}$), then

$$P_F\big(T_{n,\bar{\lambda}_u} > c_{n,\bar{\lambda}_u}(u)\big) \longrightarrow 1, \text{ as } n \to \infty,$$

where u stands for $u^{\bar{\lambda}}_{n,\alpha}$ given by (13).

4. Combining the EP and HM test statistics

In this and the following section we will restrict our attention to the families of test statistics of Epps and Pulley (1983) and of Henze and Meintanis (2002), given by (5) and (6), respectively (henceforth denoted by EP and HM). In the former case, the parametric family $\mathcal{F}$ is given by (2), with F_0 the distribution function of the standard Gaussian distribution, whereas in the latter it is given by (3), with F_0 the distribution function of the unit exponential distribution. We start by showing that both test statistic families satisfy the assumptions of Theorems 3 and 4 stated in the previous section.

Theorem 7. *The null distribution functions of the EP test statistics (for $n \geq 3$), and of the HM test statistics (for $n \geq 2$), are strictly increasing (on the set $\{t \in \mathbb{R} : 0 < F_{T_{n,\lambda}}(t) < 1\}$) and continuous, for all $\lambda \in]0, +\infty[$.*

Taking into account Theorems 3 and 4, the previous result enables us to conclude that, either for the EP test statistic family ($n \geq 3$), or for the HM test statistic family ($n \geq 2$), the test with critical region $\big\{T_{n,\bar{\lambda}_u} > c_{n,\bar{\lambda}_u}(u)\big\}$, with $\bar{\lambda}_u$ given by (8) and $u = u^{\bar{\lambda}}_{n,\alpha}$ given by (13), has an exact α level of significance. The same is true for the test with critical region $\big\{T_{n,\tilde{\lambda}_{u_p}} > c_{n,\tilde{\lambda}_{u_p}}(v_{u_p})\big\}$, with $\tilde{\lambda}_u$, $u_p = u^{\tilde{\lambda}}_{n,\alpha,p}$, and $v_{u_p} = v^{\tilde{\lambda},u_p}_{n,\alpha}$, given by (9), (12) and (14), respectively. Moreover, from the results presented in Section 3, and those of Baringhaus and Henze (1988, Theorems 3 and 4), and Henze and Meintanis (2002, Theorems 2.3 and 2.7), we deduce that the previous tests are consistent against each nondegenerated non-normal distribution with finite variance, when $T_{n,\lambda}, \lambda \in \Lambda$, is the EP test statistic family, and they are consistent against each nonnegative non-exponential distribution not degenerated at zero, when $T_{n,\lambda}, \lambda \in \Lambda$, is the HM test statistic family.

In order to implement the previous test procedures in practice, where the values $u^{\bar{\lambda}}_{n,\alpha}$ and $v^{\tilde{\lambda}_{u_p}}_{n,\alpha}$, are replaced by the approximations $u^{\bar{\lambda}}_{n,\alpha,p}$ and $v^{\tilde{\lambda}_{u_p}}_{n,\alpha,q}$, respectively, the values $\psi_{\bar{\lambda}}(u)$, $\psi_{\tilde{\lambda}}(u)$ and $\psi_{\tilde{\lambda}_{u_p}}(v)$, with $u_p = u^{\tilde{\lambda}}_{n,\alpha,p}$, need to be approximated by Monte Carlo experiments under the null hypothesis, for

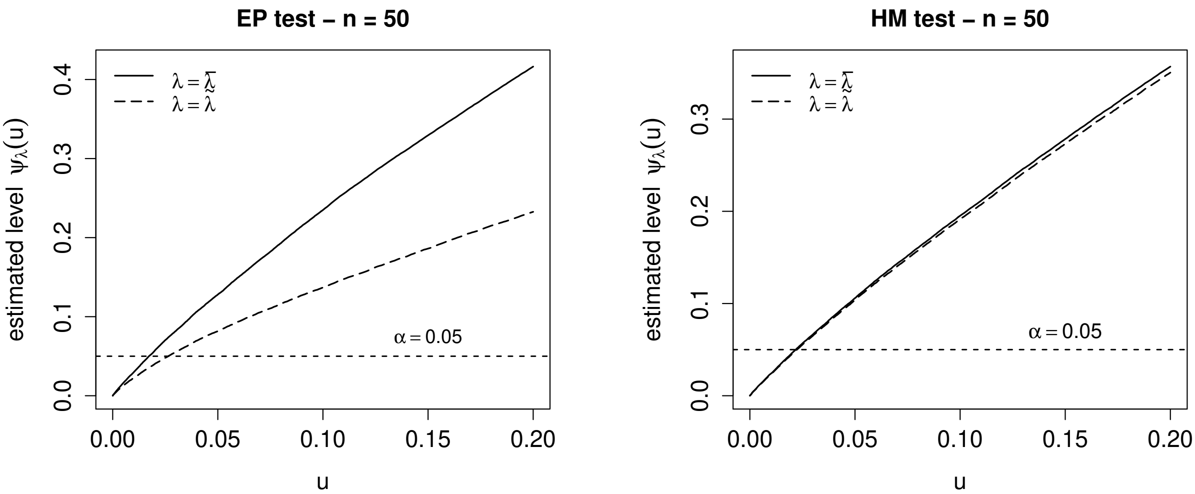

FIGURE 1. *Estimates of* $\psi_{\bar\lambda}(u)$ *and* $\psi_{\tilde\lambda}(u)$, *for* $u \in]0, 0.2[$ *and* $n = 50$, *for the EP and HM test statistic families with* $\Lambda = \{0.1, 0.25, 0.5, 0.75, 1, 1.5, 2, 2.5, 3.5, 5\}$. *These plots describe the level of significance of the test procedures based on the tuning parameter selectors* $\bar\lambda_u$ *and* $\tilde\lambda_u$, *as a function of* u.

u and v varying on the regular grid $G_p = \{w_k, k \in I_p\}$, on the interval $]0,1[$, where $w_1 = p$ and $w_{k+1} = w_k + p$, for some $0 < p \le \alpha/|\Lambda|$. For that, we use 100,000 simulations under the null hypothesis of the involved test statistics $T_{n,\lambda}$, $\lambda \in \Lambda$, and the R function quantile(·,type=7) for estimating the $1-u$ quantiles $c_{n,\lambda}(u)$, for u varying on G_p with $p = 0.0001$. Further 100,000 simulations are used for estimating the probabilities $\psi_{\bar\lambda}(u)$, $\psi_{\tilde\lambda}(u)$ and $\psi_{\tilde\lambda_{u_p}}(v)$, for u and v varying on G_p. In the evaluation of $\tilde\lambda_u$, $B = 100$ bootstrap samples are used.

We always take $\Lambda = \{0.1, 0.25, 0.5, 0.75, 1, 1.5, 2, 2.5, 3.5, 5\}$, a set of tuning parameters that includes the range of values for λ considered by Epps and Pulley (1983) and Henze and Meintanis (2002). Although the choice of the set Λ, of relevant values for the tuning parameter λ, may be based on some preliminar information, the previous set Λ is meant for the most common situation in practice where no relevant information about the alternative hypothesis is available. For $n = 50$ we show in Figure 1 the graphics of the functions $\psi_{\bar\lambda}(u)$ and $\psi_{\tilde\lambda}(u)$, for $u \in]0, 0.2[$, that describe the estimated levels of significance of the test procedures based on the tuning parameter

	$\bar{\lambda}_u$		$\tilde{\lambda}_u$		$\bar{\lambda}_u$		$\tilde{\lambda}_u$	
n	$u^{\bar{\lambda}}_{n,\alpha,p}$	EL	$u^{\tilde{\lambda}}_{n,\alpha,p}$	EL	$u^{\bar{\lambda}}_{n,\alpha,p}$	EL	$u^{\tilde{\lambda}}_{n,\alpha,p}$	EL
	Normality test				**Exponentiality test**			
	$\alpha = 0.01$							
20	0.0032	0.0092	0.0049	0.0096	0.0040	0.0101	0.0042	0.0101
50	0.0029	0.0098	0.0037	0.0095	0.0037	0.0099	0.0038	0.0099
100	0.0029	0.0099	0.0039	0.0102	0.0038	0.0096	0.0039	0.0095
	$\alpha = 0.05$							
20	0.0183	0.0487	0.0325	0.0484	0.0220	0.0487	0.0225	0.0501
50	0.0172	0.0497	0.0264	0.0494	0.0217	0.0497	0.0223	0.0499
100	0.0173	0.0502	0.0277	0.0509	0.0218	0.0499	0.0224	0.0499

TABLE 1. *Estimates of* $u^{\bar{\lambda}}_{n,\alpha,p}$ *and* $u^{\tilde{\lambda}}_{n,\alpha,p}$*, for a preassigned level* α*, based on regular grids of size* 0.0001 *on the interval* $]0,1[$*, and estimates of the nominal levels of significance (EL) for the tests based on the EP and HM families of test statistics. For the estimation of the nominal levels, the number of replications for each case is* 100,000.

selectors $\bar{\lambda}_u$ and $\tilde{\lambda}_u$, respectively, as a function of u. As observed in Section 1, from these graphics we clearly see that choosing $u = \alpha$ for $\tilde{\lambda}_u$, as suggested by Allison and Santana (2015), leads to a badly calibrated test procedure with a level of significance bigger than α. Similar graphics have been observed for other sample sizes. The suggested continuity of $\psi_{\tilde{\lambda}}(u)$ explains the similar results observed in practice for the test with critical region $\{T_{n,\tilde{\lambda}_{u_p}} > c_{n,\tilde{\lambda}_{u_p}}(v_{u_p})\}$, that includes two calibration stages, and the test with critical region $\{T_{n,\tilde{\lambda}_{u_p}} > c_{n,\tilde{\lambda}_{u_p}}(u_p)\}$, that includes a single calibration stage. For this reason, and because it is less time-consuming than the test with two calibration stages, only the test with a single calibration stage is henceforth considered.

For $\alpha = 0.01, 0.05$, and sample sizes $n = 20, 50, 100$, we present in Table 1 estimates of the levels $u^{\bar{\lambda}}_{n,\alpha,p}$ and $u^{\tilde{\lambda}}_{n,\alpha,p}$, for the preassigned level of significance α, based on regular grids of size 0.0001 on the interval $]0,1[$, and estimates of the nominal levels of significance for the tests based on the EP and HM families of test statistics. The estimation of the nominal levels was based on 100,000 simulations under the null hypotheses. With some few exceptions, the preassigned level α is inside its approximate 95% confidence interval, revealing the effectiveness of the calibration procedures.

5. Finite sample power analysis

In order to study the performance of the tests based on the data-based tuning parameter selectors $\bar{\lambda}_u$ and $\tilde{\lambda}_u$, a simulation study is conducted for each one of the Epps-Pulley and Henze-Meintanis families of test statistics. Other than to assess their empirical power, the simulation study is also meant to compare the previous tests with the fixed tuning parameter procedures, usually recommended in the literature. For the Epps-Pulley test of normality, we take $\lambda = \lambda_{\mathrm{EP}} := 1/\sqrt{2}$, which is one of the two values for λ recommended in the pioneering work of Epps and Pulley (1983), and also considered in other studies like those of Baringhaus et al. (1989) and Arcones and Wang (2006) (see also Tenreiro, 2009). For the Henze-Meintanis test of exponentiality, we take $\lambda = \lambda_{\mathrm{HM}} := 1$, which is one of the two values for λ recommended in Henze and Meintanis (2002). As before the nominal levels $\alpha = 0.01, 0.05$ and the sample sizes $n = 20, 50, 100$ are considered. All the power estimates are based on $10,000$ samples from the considered alternatives.

Although the following conclusions are based on a simulation study carried out for large sets of alternative distributions usually considered in power studies for testing normality (see Epps and Pulley, 1983, Romão et al., 2010, Yap and Sim, 2011), and exponentiality (see Henze, 1993, Henze and Meintanis, 2002, 2005), we limit ourselves to present in Tables 2 and 3 the empirical power results for some of the considered alternatives. In these tables: LN(θ) denotes the lognormal distribution with density $(\theta x)^{-1}(2\pi)^{-1/2}\exp(-(\log x)^2/(2\theta^2))I(x \geq 0)$; W($\theta$) denotes the Weibull distribution with density $\theta x^{\theta-1}\exp(-x^\theta)I(x \geq 0)$; LF($\theta$) denotes the linear increasing failure rate distribution with density $(1+\theta x)\exp(-x-\theta x^2/2)I(x \geq 0)$; and PW($\theta$) denotes the power distribution with density $\theta^{-1}x^{1/\theta-1}I(0 \leq x \leq 1)$, where $x \in \mathbb{R}$.

For all the considered alternatives, the tests based on the data-dependent tuning parameter selectors performed similarly. Being clearly less time-consuming than the bootstrap-based method for choosing λ, our preference goes to $\bar{\lambda}_u$. For the generality of the considered alternatives, these data-dependent tuning parameter selectors compare well with the fixed tuning parameters λ_{EP} or λ_{HM}, none of them being the best over the considered set

	$\alpha = 0.01$			$\alpha = 0.05$		
	$n = 20$	$n = 50$	$n = 100$	$n = 20$	$n = 50$	$n = 100$
	# 1 Uniform distribution					
λ_{EP}	0.01	0.14	0.68	0.09	0.48	0.93
$\bar{\lambda}_u$	0.03	0.18	0.64	0.13	0.43	0.87
$\tilde{\lambda}_u$	0.04	0.15	0.55	0.14	0.39	0.80
	# 2 Logistic distribution					
λ_{EP}	0.05	0.07	0.12	0.13	0.19	0.27
$\bar{\lambda}_u$	0.04	0.07	0.12	0.11	0.18	0.26
$\tilde{\lambda}_u$	0.03	0.07	0.11	0.09	0.16	0.22
	# 3 LN(0.5) distribution					
λ_{EP}	0.30	0.76	0.98	0.50	0.90	1.00
$\bar{\lambda}_u$	0.27	0.74	0.98	0.46	0.89	1.00
$\tilde{\lambda}_u$	0.24	0.71	0.97	0.36	0.82	0.99
	# 4 Normal mixture $0.8N(0,1) + 0.2N(3,1)$					
λ_{EP}	0.11	0.45	0.86	0.28	0.69	0.95
$\bar{\lambda}_u$	0.08	0.34	0.78	0.23	0.59	0.92
$\tilde{\lambda}_u$	0.08	0.32	0.74	0.19	0.52	0.86
	# 5 Tukey(5) distribution					
λ_{EP}	0.02	0.03	0.05	0.08	0.10	0.16
$\bar{\lambda}_u$	0.04	0.12	0.34	0.12	0.27	0.56
$\tilde{\lambda}_u$	0.04	0.11	0.33	0.13	0.27	0.56

TABLE 2. *Empirical power results for the normality tests based on the Epps-Pulley test statistic family. The power estimates are based on* 10,000 *samples from the considered alternatives.*

of alternative distributions. This is illustrated by the alternative distributions # 1, # 2 and # 3 shown in Tables 2 and 3.

For the alternative distributions # 4 in both tables, the tests based on the fixed tuning parameters λ_{EP} or λ_{HM} are slightly more powerful than those based on the data-dependent tuning parameter selectors. For the normality test this is a consequence of small and large values of λ, as $\lambda = 0.1$ or $\lambda = 5$, included in the set Λ. In fact, for the considered normal mixture alternative the power of the Epps-Pulley tests seems to behave like an inverted U-shaped function of λ. In the case of the exponential test, the power of the Henze-Meintanis tests for the considered alternative, is very low for small values of λ. Therefore, the inclusion of the value $\lambda = 0.1$ in the set Λ, may explain the

	$\alpha = 0.01$			$\alpha = 0.05$		
	$n = 20$	$n = 50$	$n = 100$	$n = 20$	$n = 50$	$n = 100$
	# 1 Standard uniform distribution					
$\lambda_{\rm HM}$	0.22	0.73	0.99	0.53	0.92	1.00
$\bar\lambda_u$	0.13	0.76	1.00	0.46	0.96	1.00
$\tilde\lambda_u$	0.13	0.76	1.00	0.46	0.96	1.00
	#2 W(0.8) distribution					
$\lambda_{\rm HM}$	0.12	0.32	0.63	0.25	0.51	0.80
$\bar\lambda_u$	0.16	0.36	0.65	0.31	0.56	0.82
$\tilde\lambda_u$	0.16	0.35	0.64	0.31	0.55	0.82
	#3 LN(1.5) distribution					
$\lambda_{\rm HM}$	0.49	0.87	0.99	0.62	0.93	1.00
$\bar\lambda_u$	0.53	0.89	0.99	0.65	0.94	1.00
$\tilde\lambda_u$	0.52	0.89	0.99	0.65	0.94	1.00
	#4 LF(2) distribution					
$\lambda_{\rm HM}$	0.07	0.31	0.72	0.27	0.62	0.91
$\bar\lambda_u$	0.03	0.23	0.68	0.17	0.55	0.91
$\tilde\lambda_u$	0.03	0.23	0.68	0.17	0.55	0.91
	# 5 PW(2) distribution					
$\lambda_{\rm HM}$	0.07	0.15	0.30	0.17	0.30	0.50
$\bar\lambda_u$	0.25	0.54	0.84	0.44	0.73	0.94
$\tilde\lambda_u$	0.25	0.54	0.85	0.44	0.73	0.94

TABLE 3. *Empirical power results for the exponentiality tests based on the Henze-Mentainis test statistic family. The power estimates are based on* 10,000 *samples from the considered alternatives.*

inferior power attained for this alternative by the tests based on the data-dependent tuning parameter selectors. This is a price to pay for having tests with a reasonable power against a wide range of alternative distributions.

The power results reported for alternatives # 5 in Tables 2 and 3, illustrate the already mentioned weak point of the strategy of taking a fixed tuning parameter in the absence of any prior information on the underlying alternative distribution. The usually recommended tuning parameters $\lambda_{\rm EP}$ or $\lambda_{\rm HM}$ lead to tests that achieve a very low power against these alternatives, for which a smaller tuning parameter value would be a better choice. As values as small as $\lambda = 0.1$ or $\lambda = 0.25$, have been included in the set Λ, the tests based on the considered data-dependent tuning parameter selectors perform much better than the recommended fixed tuning parameters for these alternatives.

As any goodness-of-fit test, the tests considered in this paper have a preference for some finite-dimensional space of alternatives, and cannot pay equal attention to an infinite number of orthogonal alternatives (see Janssen, 2000). However, the considered tests based on the data-dependent tuning parameter selectors $\bar{\lambda}_u$ and $\tilde{\lambda}_u$ have shown to be serious competitors for the recommended tests based on a fixed tuning parameter, and perhaps should be used — especially $\bar{\lambda}_u$ because is less time-consuming than $\tilde{\lambda}_u$ —, in the absence of any information regarding the type of deviation from the null model.

6. Proofs

Proof of Theorem 1: The first part of Theorem 1 follows from (11) as it implies that

$$P_F\big(T_{n,\hat{\lambda}_u} > c_{n,\hat{\lambda}_u}(u)\big) \leq |\Lambda| u, \tag{16}$$

for all $u \in]0,1[$, where $|\Lambda|$ denotes the cardinality of Λ. In order to prove that the probability $P_F\big(T_{n,\hat{\lambda}_u} > c_{n,\hat{\lambda}_u}(u)\big)$, is independent of F, for $F \in \mathcal{F}$, it is enough to use the invariance properties of $T_{n,\lambda}$ and $\hat{\lambda}_u$. ■

Proof of Theorem 2: Let $I^{\hat{\lambda}}_{n,\alpha} = \{u \in]0,1[\colon \psi_{\hat{\lambda}}(u) \leq \alpha\}$. As we are assuming that $\psi_{\hat{\lambda}}$ is increasing, we have $I_{n,\alpha} =]0, u^{\hat{\lambda}}_{n,\alpha}[$ or $I^{\hat{\lambda}}_{n,\alpha} =]0, u^{\hat{\lambda}}_{n,\alpha}]$, where $u^{\hat{\lambda}}_{n,\alpha} = \sup I^{\hat{\lambda}}_{n,\alpha}$. Taking into account (16), we get $\psi_{\hat{\lambda}}(\alpha/|\Lambda|) \leq \alpha$, and therefore $\alpha/|\Lambda| \leq u^{\hat{\lambda}}_{n,\alpha}$. On the other hand, as $\lim_{u\uparrow 1} \psi_{\hat{\lambda}}(u) = 1$, we necessarily have $u^{\hat{\lambda}}_{n,\alpha} < 1$. Using this fact, take $p > 0$ small enough such that $G_p \cap]0, u^{\hat{\lambda}}_{n,\alpha}[\neq \emptyset$ and $G_p \cap]u^{\hat{\lambda}}_{n,\alpha}, 1[\neq \emptyset$, where G_q is the considered regular grid on the interval $]0,1[$. By the definition of $u^{\hat{\lambda}}_{n,\alpha,p}$, we have $\psi_{\hat{\lambda}}(u^{\hat{\lambda}}_{n,\alpha,p}) \leq \alpha$, that is, $u^{\hat{\lambda}}_{n,\alpha,p} \in I^{\hat{\lambda}}_{n,\alpha}$; thus $u^{\hat{\lambda}}_{n,\alpha,p} \leq u^{\hat{\lambda}}_{n,\alpha}$. On the other hand, as $\psi_{\hat{\lambda}}(u^{\hat{\lambda}}_{n,\alpha,p} + p) > \alpha$ we have $u^{\hat{\lambda}}_{n,\alpha,p} + p \notin I^{\hat{\lambda}}_{n,\alpha}$, and $u^{\hat{\lambda}}_{n,\alpha,p} + p \geq u^{\hat{\lambda}}_{n,\alpha}$. We have proved that $u^{\hat{\lambda}}_{n,\alpha,p} \leq u^{\hat{\lambda}}_{n,\alpha} < u^{\hat{\lambda}}_{n,\alpha,p} + p$, for $p > 0$ small enough, from which we deduce that $\lim_{p\downarrow 0} u^{\hat{\lambda}}_{n,\alpha,p} = u^{\hat{\lambda}}_{n,\alpha}$. Taking into account the continuity of $\psi_{\hat{\lambda}}$, we finish the proof by showing that $\psi_{\hat{\lambda}}(u^{\hat{\lambda}}_{n,\alpha}) = \alpha$. Let us consider $u_m \in I^{\hat{\lambda}}_{n,\alpha}$ such that $u_m \uparrow u^{\hat{\lambda}}_{n,\alpha}$. Therefore we get $\psi_{\hat{\lambda}}(u^{\hat{\lambda}}_{n,\alpha}) = \psi_{\hat{\lambda}}(\lim_m u_m) = \lim_m \psi_{\hat{\lambda}}(u_m) \leq \alpha$. Using the fact that $u^{\hat{\lambda}}_{n,\alpha} < 1$, let us consider a sequence $u_m \in]0,1[$ such that

$u_m \downarrow u_{n,\alpha}^{\hat{\lambda}}$. As $u_{n,\alpha}^{\hat{\lambda}}$ is the supremum of $I_{n,\alpha}^{\hat{\lambda}}$ we have $\psi_{\hat{\lambda}}(u_m) > \alpha$ for all $m \in \mathbb{N}$, and therefore $\psi_{\hat{\lambda}}(u_{n,\alpha}^{\hat{\lambda}}) = \psi_{\hat{\lambda}}(\lim_m u_m) = \lim_m \psi_{\hat{\lambda}}(u_m) \geq \alpha$. ■

Proof of Theorem 3: For $u \in]0,1[$ we have

$$\big\{ \min_{\lambda\in\Lambda} p_{n,\lambda} < u \big\} \subset \big\{ T_{n,\bar{\lambda}_u} > c_{n,\bar{\lambda}_u}(u) \big\} \subset \big\{ \min_{\lambda\in\Lambda} p_{n,\lambda} \leq u \big\}, \tag{17}$$

where $p_{n,\lambda}$ denotes the p-value $p_{n,\lambda} = \inf\big\{\alpha \in]0,1[: T_{n,\lambda} > c_{n,\lambda}(\alpha)\big\}$, with $\inf\emptyset = 1$ (see Lehmann and Romano, 2005, Subsection 3.3). As $F_{T_{n,\lambda}}$ is continuous for all $\lambda \in \Lambda$, $p_{n,\lambda}$ is, under H_0, uniformly distributed on the interval $[0,1]$, and the distribution function of $\min_{\lambda\in\Lambda} p_{n,\lambda}$ under H_0, we denote by F_m, is continuous on $\mathbb{R}$. Finally from (17) we get $\psi_{\bar{\lambda}}(u) = F_m(u) \geq u$, for $u \in]0,1[$, and the stated result follows from Theorem 2, where $u_{n,\alpha}^{\bar{\lambda}} = \sup\{u \in]0,1[: F_m(u) \leq \alpha\} \leq \alpha$. ■

Before given de proof of Theorem 4, we state two auxiliary results on the function $\psi_{\hat{\lambda}}(u) := P_{F_0}\big(T_{n,\hat{\lambda}_u} > c_{n,\hat{\lambda}_u}(u)\big)$, defined for $u \in]0,1[$. Firstly, note that from inequality (16), we have $\lim_{u\downarrow 0} \psi_{\hat{\lambda}}(u) = 0$, for any invariant tuning parameter selector $\hat{\lambda}_u$. The limit of $\psi_{\hat{\lambda}}(u)$ when $u \uparrow 1$, is given in the following proposition, where $F_{T_{n,\lambda}}$ denotes the null distribution function of $T_{n,\lambda}$.

Proposition 1. *If $F_{T_{n,\lambda}}$ is continuous, for all $\lambda \in \Lambda$, then* $\lim_{u\uparrow 1} \psi_{\hat{\lambda}}(u) = 1$.

In the following proposition we analyse the case where $\hat{\lambda}_u = \dot{\lambda}$, for all $u \in]0,1[$, with $\dot{\lambda} = \dot{\lambda}(X_1,\ldots,X_n)$ a measurable function that takes values in Λ, which is either location-scale invariant in case (2) or scale invariant in case (3).

Proposition 2. *If $F_{T_{n,\lambda}}$ is strictly increasing (on the set $\{t \in \mathbb{R} : 0 < F_{T_{n,\lambda}}(t) < 1\}$) and continuous, for all $\lambda \in \Lambda$, and $\hat{\lambda}_u = \dot{\lambda}$, for all $u \in]0,1[$, then the function $\psi_{\hat{\lambda}}$ is increasing and continuous on* $]0,1[$.

Proof of Theorem 4: For u fixed in $]0,1[$ define the invariant tuning parameter selector $\hat{\lambda}_v$ by $\hat{\lambda}_v = \hat{\lambda}_u$, for all $v \in]0,1[$. Using Propositions 1 and 2, the stated result follow from Theorem 2, where $u_{n,\alpha,p}^{\hat{\lambda}} = v_{n,\alpha,q}^{\hat{\lambda}_u}$ and $u_{n,\alpha}^{\hat{\lambda}} = v_{n,\alpha}^{\hat{\lambda}_u}$. ■

Proof of Theorem 5: Let $F \notin \mathcal{F}$, such that $T_{n,\lambda} \xrightarrow{p} +\infty$ under F, for all $\lambda \in \Lambda$. As $\alpha/|\Lambda| \leq u_{n,\alpha}^{\hat{\lambda}} = u$, for each $\lambda \in \lambda$ we have $P_F\big(T_{n,\lambda} > c_{n,\lambda}(u)\big) \geq P_F\big(T_{n,\lambda} > c_{n,\lambda}(\alpha/|\Lambda|)\big)$, since $c_{n,\lambda}(u) \leq c_{n,\lambda}(\alpha/|\Lambda|)$. Moreover, from the continuity of $F^{-1}_{T_{\infty,\lambda}}$, and the convergence $F^{-1}_{T_{n,\lambda}}(t) \to F^{-1}_{T_{\infty,\lambda}}(t)$, for all $0 < t < 1$ (see Shorack and Wellner, 1986, pp. 3–10), we get $c_{n,\lambda}(\alpha/|\Lambda|) = F^{-1}_{T_{n,\lambda}}(1-\alpha/|\Lambda|) \to F^{-1}_{T_{\infty,\lambda}}(1-\alpha/|\Lambda|)$, which implies that $\sup_{n\in\mathbb{N}} c_{n,\lambda}(\alpha/|\Lambda|) < +\infty$. Therefore, we conclude that $P_F\big(T_{n,\lambda} > c_{n,\lambda}(u)\big) \to 1$, as $n \to \infty$, for all $\lambda \in \Lambda$. Finally, the stated result follows from the equality $P_F\big(T_{n,\hat{\lambda}_u} > c_{n,\hat{\lambda}_u}(u)\big) = \sum_{\lambda\in\Lambda} P_F\big(T_{n,\lambda} > c_{n,\lambda}(u), \hat{\lambda}_u = \lambda\big)$. ■

Proof of Theorem 6: Let $F \notin \mathcal{F}$, and take $\lambda \in \Lambda$ such that $T_{n,\lambda} \xrightarrow{p} +\infty$ under F. Proceeding as in the proof of Theorem 5, for $u = u^{\bar{\lambda}}_{n,\alpha}$ we conclude that $P_F\big(T_{n,\lambda} > c_{n,\lambda}(u)\big) \to 1$, as $n \to \infty$, and the stated result follows from the fact that $P_F\big(T_{n,\bar{\lambda}_u} > c_{n,\bar{\lambda}_u}(u)\big) \geq P_F\big(T_{n,\lambda} > c_{n,\lambda}(u)\big)$. ■

Proof of Theorem 7: The Epps and Pulley test statistics given by (5) are nonconstant whenever $n \geq 3$, and well defined on D_n, where $D_n^c = \big\{x \in \mathbb{R}^n : x_1 = \cdots = x_n\big\}$. Therefore, they are nonconstant and well defined with probability one, whenever $n \geq 3$, and F is such that $\mu^n_F(D_n) = 1$. This condition is fulfilled whenever F is absolutely continuous on $\mathbb{R}$, which occurs under the null hypothesis of normality. Regarding the Henze and Meintanis test statistics given by (6), they are nonconstant whenever $n \geq 2$, and well defined on the set $D_n = \{x \in \mathbb{R}^n : x_1, \ldots, x_n > 0\}$. Therefore, they are nonconstant and well defined with probability one, whenever $n \geq 2$, and F is such that $\mu^n_F(D_n) = 1$, where μ_F denotes the probability distribution of F and μ^n_F is the product measure. This condition is satisfied whenever F is such that $F(0) = 0$, which is true under the null hypothesis of exponentiality.

We start by proving that $F_{T_{n,\lambda}}$ is strictly increasing on the set $\{t \in \mathbb{R} : 0 < F_{T_{n,\lambda}}(t) < 1\}$. For that, let us take $s, t \in \mathbb{R}$ with $s < t$ and $0 < F_{T_{n,\lambda}}(s) \leq F_{T_{n,\lambda}}(t) < 1$. As $\mu^n_{F_0}\big(\{x \in D_n : T_{n,\lambda}(x) \leq s\}\big) > 0$ and $\mu^n_{F_0}\big(\{x \in D_n : T_{n,\lambda}(x) > t\}\big) > 0$, the sets $\{x \in D_n : T_{n,\lambda}(x) \leq s\}$ and $\{x \in D_n : T_{n,\lambda}(x) > t\}$ are nonempty, which implies, from the continuity of $T_{n,\lambda}$ and the fact that D_n is open and connected, that the set $\{x \in D_n : s < T_{n,\lambda}(x) < t\} = T^{-1}_{n,\lambda}(]s,t[)$ is a nonempty open subset of D_n. Taking into account that μ_{F_0} is either the standard Gaussian or the unit exponential probability distributions on $\mathbb{R}$, we conclude that $\mu^n_{F_0}\big(T^{-1}_{n,\lambda}(]s,t[)\big) > 0$, which finally implies that $F_{T_{n,\lambda}}(s) < F_{T_{n,\lambda}}(t)$.

Next we will prove that $F_{T_{n,\lambda}}$ is continuous on $\mathbb{R}$. For that, we start by showing that is enough to prove that

$$\mu_{F_0}\big((T_{n,\lambda}^{x_2,\ldots,x_n})^{-1}(y)\big) = 0, \text{ for all } (x_2,\ldots,x_n) \in D_{n-1} \text{ and } y \in \mathbb{R}, \tag{18}$$

where $T_{n,\lambda}^{x_2,\ldots,x_n}$ is the section of $T_{n,\lambda}$ determined by $(x_2,\ldots,x_n)$, defined, for $z \in \mathbb{R}$, by $T_{n,\lambda}^{x_2,\ldots,x_n}(z) = T_{n,\lambda}(z, x_2, \ldots, x_n)$. In fact, from Fubini's theorem and the fact that $\mu_{F_0}^{n-1}(D_{n-1}^c) = 0$, we have

$$\begin{aligned}
\mu_{F_0}^n\big(T_{n,\lambda}^{-1}(y)\big) &= \mu_{F_0}^{n-1} \otimes \mu_{F_0}\big(T_{n,\lambda}^{-1}(y)\big) \\
&= \int \mu_{F_0}\Big(\big(T_{n,\lambda}^{-1}(y)\big)^{x_2,\ldots,x_n}\Big) d\mu_{F_0}^{n-1}(x_2,\ldots,x_n) \\
&= \int_{D_{n-1}} \mu_{F_0}\Big(\big(T_{n,\lambda}^{-1}(y)\big)^{x_2,\ldots,x_n}\Big) d\mu_{F_0}^{n-1}(x_2,\ldots,x_n) \\
&= \int_{D_{n-1}} \mu_{F_0}\big((T_{n,\lambda}^{x_2,\ldots,x_n})^{-1}(y)\big) d\mu_{F_0}^{n-1}(x_2,\ldots,x_n),
\end{aligned}$$

where $\big(T_{n,\lambda}^{-1}(y)\big)^{x_2,\ldots,x_n} = \{z \in \mathbb{R} : (z, x_2, \ldots, x_n) \in T_{n,\lambda}^{-1}(y)\} = (T_{n,\lambda}^{x_2,\ldots,x_n})^{-1}(y)$ is the section of $T_{n,\lambda}^{-1}(y)$ determined by $(x_2,\ldots,x_n)$. Therefore, under (18) we get $\mu_{F_0}^n\big(T_{n,\lambda}^{-1}(y)\big) = 0$, for all $y \in \mathbb{R}$, which enables us to conclude that $F_{T_{n,\lambda}}$ is continuous on $\mathbb{R}$.

Finally, in order to establish (18) it is enough to take into account that $T_{n,\lambda}^{x_2,\ldots,x_n}$ is a nonconstant analytic function on $\mathbb{R}$, for each $(x_2,\ldots,x_n) \in D_{n-1}$. In fact, for those functions the set $(T_{n,\lambda}^{x_2,\ldots,x_n})^{-1}(y)$, for $y \in \mathbb{R}$, is either empty or consists entirely of isolated points (cf. Carathéodory, 1983, p. 138), which implies that $\mu_{F_0}\big((T_{n,\lambda}^{x_2,\ldots,x_n})^{-1}(y)\big) = 0$, as μ_{F_0} is an absolutely continuous probability distribution on $\mathbb{R}$. ■

Acknowledgments

We thank our colleague Alexander Kovacec for drawing our attention to the result on nonconstant analytic functions used in the proof of Theorem 7.

Funding

Research partially supported by the Centre for Mathematics of the University of Coimbra – UID/MAT/00324/2013, funded by the Portuguese Government through FCT/MEC and co-funded by the European Regional Development Fund through the Partnership Agreement PT2020.

References

Allison, J., Santana, L. (2015). On a data-dependent choice of the tuning parameter appearing in certain goodness-of-fit tests. *J. Stat. Comput. Simul.* 85, 3276–3288.

Arcones, M.A., Wang, Y. (2006). Some new tests for normality based on U-processes. *Statis. Probab. Lett.* 76, 69–82.

Baringhaus, L., Danschke, R., Henze, N. (1989). Recent and classical tests for normality. A comparative study. *Comm. Statist. Simulation Comput.* 18, 363–379.

Baringhaus, L., Henze, N. (1988). A consistent test for multivariate normality based on the empirical characteristic function. *Metrika* 35, 339–348.

Baringhaus, L., Henze, N. (1991). A class of consistent tests for exponentiality based on the empirical Laplace transform. *Ann. Inst. Statist. Math.* 43, 197–192.

Carathéodory, C. (1983). *Theory of functions of a complex variable.* New York: Chelsea.

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

Dudoit, S., Lann, M.J. van der (2008). *Multiple testing procedures with applications to genomics.* Springer.

Feller, W. (1971). *An introduction to probability theory and its applications.* New York: John Wiley.

Fan, Y. (1998). Goodness-of-fit tests based on kernel density estimators with fixed smoothing parameters. *Econometric Theory* 14, 604–621.

Fromont, M., Laurent, B. (2006). Adaptive goodness-of-fit tests in a density model. *Ann. Statist.* 34, 680–720.

Güntler, N., Henze, N. (2000). Goodness-of-fit tests for the Cauchy distribution based on the empirical characteristic function. *Ann. Inst. Statist. Math.* 52, 267–286.

Henze, N. (1993). A new flexible class of omnibus tests for exponentiality. *Comm. Statist. Theory Methods* 22, 115–133.

Henze, N., Meintanis, S.G. (2002). Test of fit for exponentiality based on the empirical Laplace transform. *Statistics* 36, 147–161.

Henze, N., Meintanis, S.G. (2005). Recent and classical tests for exponentiality: a partial review with comparisons. *Metrika* 61, 29–45.

Henze, N., Zirkler, B. (1990). A class of invariant consistent tests for multivariate normality. *Comm. Statist. Theory Methods* 19, 3595–3617.

Janssen, A. (2000). Global power functions of goodness of fit tests. *Ann. Statist.* 28, 239–253

Kawata, T. (1972). *Fourier analysis in probability theory.* New York: Academic Press.

Klar, B. (2001). Goodness-of-fit tests for the exponential and the normal distribution based on the integrated distribution function. *Ann. Inst. Statist. Math.* 53, 338–353.

Lehmann, E.L., Romano, J.P. (2005). *Testing statistical hypotheses.* New York: Springer.

Meintanis, S.G. (2004). Goodness-of-fit tests for the logistic distribution based on empirical transforms. *Sankhya Ser. A* 66, 306–326.

Meintanis, S.G. (2004a). A class of omnibus tests for the Laplace distribution based on the empirical characteristic function. *Comm. Statist. Theory Methods* 33, 925–948.

Meintanis, S., Swanepoel, J., Allison, J. (2014). The probability weighted characteristic function and goodness-of-fit testing. *J. Statist. Plann. Inference* 146, 122–132.

R Core Team (2014). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/.

Romão, X., Delgado, R., Costa, A. (2010). An empirical power comparison of univariate goodness-of-fit tests for normality. *J. Stat. Comput. Simul.* 80, 545–591.

Shorack, G.R., Wellner, J.A. (1986). *Empirical processes with applications to statistics.* New York: John Wiley.

Tenreiro, C. (2005). On the role played by the fixed bandwidth in the Bickel-Rosenblatt goodness-of-fit test. *SORT* 29, 201–216.

Tenreiro, C. (2007). On the finite sample behaviour of fixed bandwidth Bickel-Rosenblatt test for univariate and multivariate uniformity. *Comm. Statist. Simulation Comput.* 36, 827–846.

Tenreiro, C. (2009). On the choice of the smoothing parameter for the BHEP goodness-of-fit test. *Comput. Statist. Data Anal.* 53, 1038–1053.

Tenreiro, C. (2011). An affine invariant multiple test procedure for assessing multivariate normality. *Comput. Statist. Data Anal.* 55, 1980–1992.

Tenreiro, C. (2017). A new test for multivariate normality by combining extreme and non-extreme BHEP test statistics *Comm. Statist. Simulation Comput.* 46, 1746–1759.

Yap, B.W., Sim, C.H. (2011). Comparisons of various types of normality tests. *J. Stat. Comput. Simul.* 81, 2141–2155.

Carlos Tenreiro
CMUC, Department of Mathematics, University of Coimbra, 3001–501 Coimbra, Portugal
E-mail address: tenreiro@mat.uc.pt