PROBABILITY
AND
MATHEMATICAL STATISTICS
Vol. 35, Fasc. 1 (2015), pp. 91–106

REFINED DATA DRIVEN TESTS FOR UNIVARIATE SYMMETRY

BY

TADEUSZ INGLOT (WROCŁAW) AND DAWID KUJAWA (WROCŁAW)

Abstract. We propose a modification of the data driven score rank tests studied recently in Inglot et al. (2012) by an appropriate choice of the orthonormal system. The simulation study confirms much better performance of the new tests for alternatives with dominating asymmetry in the tails and comparable sensitivity for other types of alternatives. In effect we obtain omnibus tests for symmetry which are equal to the best existing procedures for typical alternatives and overtake them significantly for atypical ones.

2000 AMS Mathematics Subject Classification: Primary: 62G10; Secondary: 62G30, 65C05, 65C60.

Key words and phrases: Testing symmetry, non-smooth functions, data driven score test, effective score, Schwarz-type rule, model selection, rank test, Monte Carlo study.

1. INTRODUCTION

Let $X_1, \ldots, X_n$ be i.i.d. real random variables with a continuous distribution function $F(x)$ and with the median 0 (or with a known median by which X_i's have already been centered). We are going to test

$$H_0 : F(x) = 1 - F(-x), \quad x \in \mathbb{R},$$

i.e. to test the symmetry of F about a known center 0. So, we consider a little different problem than testing symmetry about 0, namely we restrict our attention to the class of distributions with the median 0. Therefore, our solution is not expected to be powerful for alternatives with nonzero median.

Nowadays many tests of symmetry about 0 or about a known median are available and the problem takes a constant interest of many statisticians. For some overview of the literature we refer to [2]. Among a wide variety of constructions the Modarres and Gastwirth test (see [4]) has proved to be particularly powerful. It is a test of symmetry about 0 and can detect a nonzero median and asymmetry in the tails and does not have an omnibus character. The data driven score rank tests proposed in [2] are able to detect any type of asymmetry. However, they have a lower sensitivity for detecting asymmetry in the tails.

The aim of the present note is to refine the data driven tests proposed in [2]. We do it by an appropriate choice of an orthonormal system on the unit interval. Some attempt in this direction was made by Józefczyk [3]. We propose another orthonormal system than that used by Józefczyk, which seems to be better fitted to detecting different types of asymmetry. The paper is strongly related to [2] and applies some results of that paper. So, when possible, we avoid repeating similar considerations but simultaneously keep the paper self-contained.

In Sections 2 and 3 we construct test statistics and establish their asymptotic distribution. The main results are given in Section 4 where we present empirical performance of the new tests. Proofs are provided in Section 6.

2. TEST STATISTICS

Denote by $F_s(x) = \frac{1}{2}\big(F(x) + 1 - F(-x)\big)$ the distribution function of the symmetric part of F and put $F_a = F - F_s$. Transform the data into the unit interval using F_s to obtain $U_1, \ldots, U_n$ with $U_i = F_s(X_i),\ i = 1, \ldots, n$. Since F is absolutely continuous with respect to F_s, the transformed data U_i have an absolutely continuous distribution function $F \circ F_s^{-1}(t) = t + A(t)$, $t \in [0,1]$, and a density of the form

$$p(t) = 1 + a(t), \qquad t \in [0,1], \tag{2.1}$$

where $a(t)$ is an antisymmetric – with respect to $t = 1/2$ – derivative of $A(t)$. So, testing H_0 is equivalent to testing that $a = 0$. Observe that $|a(t)| \leqslant 1$ a.s. and contains all information about an asymmetry of F.

Let $d(n) \geqslant 1$ be a (possibly unbounded) nondecreasing sequence of natural numbers. For every $n \geqslant 1$ consider a triangular array

$$g^k = (g_{k1}, g_{k2}, \ldots, g_{kk}) = (g_{k1}^{(n)}, g_{k2}^{(n)}, \ldots, g_{kk}^{(n)}), \qquad k = 1, 2, \ldots, d(n), \tag{2.2}$$

of bounded rowwise orthonormal functions in $L_2[0,1]$, antisymmetric with respect to $1/2$ such that for each g_{kj} there exists a finite partition of the unit interval into l_{kj} intervals on which g_{kj} is absolutely continuous. In [2], g^k consisted of the first k odd Legendre polynominals while in [3] systems of indicator functions were taken into account. Our setting includes them as special cases and allows for more flexible solutions. For example, one can select various subsets of $g^{d(n)}$ to form consecutive rows of a triangular array or replace some functions from $g^{d(n)}$ by other ones when forming successive rows (cf. [3]).

For $1 \leqslant k \leqslant d(n)$ consider the sequence of exponential families of densities on the interval $[0,1]$,

$$c_k(\vartheta) \exp\Big\{ \sum_{j=1}^{k} \vartheta_j g_{kj}(t) \Big\}, \qquad k = 1, 2, \ldots, d(n), \tag{2.3}$$

where $\vartheta = (\vartheta_1, \ldots, \vartheta_k)^T \in \mathbb{R}^k$, v^T stands for the transposition of a vector v and $c_k(\vartheta)$ is the normalizing constant.

Suppose that $p(t) = 1 + a(t)$ can be treated approximately as a member of the family (2.3). Then H_0 reduces to $H_0' : \vartheta = 0$. By the orthonormality of the system g^k, the score statistic for such a parametric problem takes the form

$$\sum_{j=1}^{k} \left\{ \frac{1}{\sqrt{n}} \sum_{i=1}^{n} g_{kj}\bigl(F_s(X_i)\bigr) \right\}^2. \tag{2.4}$$

Let $F_{ns}(x)$ be the empirical distribution function of the pooled sample $Z = (X_1, \ldots, X_n, -X_1, \ldots, -X_n)$. Then $F_{ns}(x) = R_i/(2n)$, where R_i is the rank of X_i in Z. Estimating an unknown distribution function F_s by F_{ns} and taking into account the usual continuity correction, we obtain the statistic (2.4) in the form

$$T_k = \sum_{j=1}^{k} \widehat{g}_{kj}^2,$$

where

$$\widehat{g}_{kj} = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} g_{kj} \left(F_{ns}(X_i) - \frac{1}{4n} \right) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} g_{kj} \left(\frac{2R_i - 1}{4n} \right) \tag{2.5}$$

are linear rank statistics, thus invariant in the class of symmetric distributions, which implies that T_k is a distribution-free statistic for testing symmetry.

Up to now, the dimension k was arbitrarily chosen, but we want to fit it to the data at hand. To this end we apply a Schwarz-type selection rule (cf. [2]) defined by the formula

$$S = \min\{k : \ 1 \leqslant k \leqslant d(n),\ T_k - k \log n = \max_{1 \leqslant j \leqslant d(n)} (T_j - j \log n)\}$$

and denote the corresponding data driven statistic by T_S.

An alternative to S is a less conservative selection rule L, we take here, which was introduced in Inglot and Janic [1] and applied to testing symmetry in [2]. Let $1 \leqslant D_n < d(n)$ be a natural number and $\delta_n > 0$ a small number. Define thresholds $c_{jn},\ j = 1, \ldots, D_n$, by the equations

$$\bigl(2 - 2\Phi(c_{jn})\bigr)^j = \delta_n D_n^{-1} \binom{d(n)}{j}^{-1},$$

where Φ denotes the standard normal distribution function. Next, order $\widehat{g}_{d(n)1}^2, \ldots, \widehat{g}_{d(n)d(n)}^2$ from the smallest to the largest, obtaining $\mathcal{G}_{(1)}^2, \ldots, \mathcal{G}_{(d(n))}^2$, and consider the event $E_n = \{\mathcal{G}_{(d(n))}^2 \geqslant c_{1n}^2\} \cup \ldots \cup \{\mathcal{G}_{(d(n)-D_n+1)}^2 \geqslant c_{D_n n}^2\}$. Then define the data dependent penalty $\pi(j, n) = j \log n \cdot \mathbf{1}_{E_n^c} + 2j \cdot \mathbf{1}_{E_n}$, where

$\mathbf{1}_E$ denotes the indicator of a set E and E^c denotes the complement of E, the corresponding selection rule L,

$$L = \min\big\{1 \leqslant k \leqslant d(n) : T_k - \pi(k,n) = \max_{1\leqslant j\leqslant d(n)} \big(T_j - \pi(j,n)\big)\big\},$$

and the data driven statistic $T_L = T_L(D_n, \delta_n)$. By the above considerations, T_S and T_L can be applied as test statistics of upper-tailed distribution-free data driven tests for testing H_0.

Now, we propose for future use a particular triangular array g^k, $1 \leqslant k \leqslant d(n)$. For $0 < \Delta < 1/4$ transform the odd Legendre polynomials on $[0,1]$ linearly onto the set $I_0 = \big[\Delta, \frac{1}{2} - \Delta\big] \cup \big[\frac{1}{2} + \Delta, 1 - \Delta\big]$, put value 0 outside this set, normalize and denote the resulting functions by $b_1, b_3, \ldots$ Next, define the function $h_c(t) = \operatorname{sign}(2t-1)(2\Delta)^{-1/2}\mathbf{1}_{[0,2\Delta]}(|2t-1|)$. For an interval $I = [u,v] \subset [1/2, 1]$ define the antisymmetric trapezoid function

$$h_I(t) = C[|4t-2| + 2 + v - 5u]\mathbf{1}_{[2u,2v]}(1 + |2t-1|), \quad t \in [0,1], \tag{2.6}$$

with $C = \operatorname{sign}(2t-1)\sqrt{3/\big(62(v-u)^3\big)}$.

For $n = 100$ we take $d(n) = 6$ and $\Delta = 1/16$. Putting $I_1 = [1-\Delta, 1]$ and $h_{I_1} = h_1$, consider the orthonormal system $b_3, h_1, b_1, h_c, b_5, b_7$. Now, we take g^k, $k = 1, \ldots, 6$, as the first k functions of this system and use such a triangular array and the corresponding data driven tests in our simulation study.

To give the reader some hints how to modify the above triangular array for other sample sizes we propose $d(n)$ to be slowly increasing with n and $\Delta = \Delta_{d(n)}$ be depending only on $d(n)$ and equal approximately to $3/\big(8d(n)\big)$. We propose to take b_3, h_1, b_1, h_c for $d(n) = 4$ while b_3, h_1, b_1, h_c, b_5 for $d(n) = 5$. For $d(n) > 6$ set $d_1(n)$ equal approximately to $\frac{2}{3}d(n)$ and put $d_2(n) = d(n) - d_1(n)$. Divide the interval $[1 - \Delta_{d(n)}, 1]$ onto $d_2(n) - 1$ subintervals $I_1, \ldots, I_{d_2(n)-1}$ of equal length and consider the corresponding functions $h_{I_j} = h_{d_2(n)-j}$, $j = 1, \ldots, d_2(n) - 1$. Then we propose to take

$$g^{d(n)} = (b_3, h_1, b_1, h_2, b_5, h_3, b_7, \ldots, b_{2d_1(n)-1}), \tag{2.7}$$

where the functions b_j and h_j are taken alternately until all the functions h_j have been exhausted. Additionally, we place the function h_c (based on the actual $\Delta_{d(n)}$) approximately on the position $d_1(n)$. Obviously, we obtain the orthonormal system. Having defined $g^{d(n)}$ we take g^k as the first k functions of this system. Note that the functions h_j are designed to detect asymmetry on the tails while b_j measure asymmetry in the middle part of a distribution. The function h_c measures asymmetry in the very center and is useful especially for distributions which are bimodal or have a density close to 0 in the center. We propose the above triangular array after some trials which convinced us that we obtain a good testing procedure for different sample sizes. However, we have no justification that such choice is optimal and one may seek for some further improvements. The same concerns an ordering suggested in (2.7). When $d_2(n) > 2$ the functions $h_j,\ 1 < j < d_2(n)$, may be replaced by the indicator functions of the corresponding intervals equipped with an appropriate sign.

3. ASYMPTOTIC BEHAVIOUR OF THE TEST STATISTICS

In this section we present asymptotic results for the test statistics T_S and T_L constructed in Section 2.

First, we make assumptions on a triangular array $g^k, k = 1, \ldots, d(n)$, introduced in (2.2). Assume that there exist constants $\eta \geqslant 0$, $\zeta \geqslant 0$ and $\kappa > 0$ such that for some positive constant c the following conditions hold true:

$$\max_{1\leqslant k\leqslant d(n)} \max_{1\leqslant j\leqslant k} \sup_{t\in[0,1]} |g_{kj}(t)| \leqslant c[d(n)]^{\eta}, \tag{3.1}$$

$$\max_{1\leqslant k\leqslant d(n)} \max_{1\leqslant j\leqslant k} l_{kj}(t) \leqslant c[d(n)]^{\zeta}, \tag{3.2}$$

$$\max_{1\leqslant k\leqslant d(n)} \sum_{j=1}^{k} \Big(\int_0^1 |g'_{kj}(t)|dt\Big)^2 \leqslant c[d(n)]^{\kappa}. \tag{3.3}$$

To obtain asymptotic results for our test statistics we adopt the idea of [3] and approximate each g_{kj} by an absolutely continuous function on $[0,1]$, normalize it, and use the results from the Appendix in [2]. Details are given in Section 6.

Set

$$\rho = \max(\kappa, 2\eta + 2\zeta + 1). \tag{3.4}$$

Note that the system defined by (2.7) satisfies (3.1)–(3.3) with $\eta = 1/2$, $l_{kj} \leqslant 5,\ \zeta = 0$ and any $\kappa > 3/2$. Hence for this system we have $\rho = 2$.

The following theorem, proved in Section 6, establishes an asymptotic behaviour of T_S and T_L under the null hypothesis.

THEOREM 3.1. *Suppose that H_0 is true and* (3.1)–(3.3) *are satisfied and $d(n) = O(n^{\tau})$ for some $\tau < 1/(1+2\rho)$ with ρ given in* (3.4).

(1) *Then $S \xrightarrow{P} 1$ and $T_S \xrightarrow{D} \chi_1^2$ as $n \to \infty$, where χ_k^2 denotes a random variable with the central chi-square distribution with k degrees of freedom.*

(2) *If $1 \leqslant D_n \leqslant D < d(n)$, where D is a fixed number and $\delta_n > 0$ is such that* $\log(1/\delta_n) = o(n^{1/(1+2\rho)})$ *and* $\log(1/\delta_n)/d(n) \to \infty$, *then*

$$P(L = S) \to 1 \quad \textit{and} \quad T_L \xrightarrow{D} \chi_1^2 \quad \textit{as } n \to \infty.$$

The second theorem concerns the asymptotic behaviour of T_S and T_L under alternatives.

THEOREM 3.2. *Suppose* (3.1)–(3.3) *are satisfied*, $d(n) = O(n^\tau)$ *for some* $\tau < 1/(1+2\rho)$ *with* ρ *given by* (3.4) *and* F *is a fixed asymmetric distribution function such that*

$$\omega_n^4 = \frac{n}{[d(n)]^\rho \log^2 n} \Big| \int_0^1 g^{d(n)}(t)\, a(t) dt \Big|_{d(n)}^2 \to \infty \quad \text{as } n \to \infty, \tag{3.5}$$

where a *is defined in* (2.1) *and* $|v|_k = (v_1^2 + v_2^2 + \ldots + v_k^2)^{1/2}$ *denotes the Euclidean norm of a vector* $v = (v_1, \ldots, v_k)^T$. *Then* $T_S \to \infty$ *and* $T_L \to \infty$ *in probability. Consequently, for* D_n *and* δ_n *as in Theorem* 3.1 (2) *the tests based on* T_S *and* T_L *are consistent in the family of alternatives satisfying* (3.5).

Observe that (3.5) is a weak condition. For example, if $d(n) \to \infty$ and rows g^k consist of k first functions of a complete orthonormal system, then, by Parseval's inequality, (3.5) holds trivially for any $a \neq 0$. Since the set $I_0 = I_{0n}$ increases to $[0, 1]$ when $d(n) \to \infty$, the system defined in (2.7) also satisfies (3.5) for any $a \neq 0$ provided $d(n) \to \infty$.

4. SIMULATION STUDY

In this section we present results of an extensive simulation study in which we compare performance of our tests based on statistics T_S and T_L with some tests which proved to be powerful for various asymmetric distributions. For notational convenience we shall denote here the new tests by TS and TL. We restrict our attention only to the case $n = 100$ and the typical significance level 0.05. As was said at the end of Section 2 we take $d(n) = 6$ and $g^6 = (b_3, h_1, b_1, h_c, b_5, b_7)$ as the orthonormal system. For the selection rule L we take $D_n = 3,\ \delta_n = 0.05$. All computations were performed by using R. Every Monte Carlo experiment was repeated at least 10,000 times.

Critical values. Due to slow convergence of the test statistics T_S and T_L to their asymptotic distribution we use simulated critical values (see, e.g., [2] for more explanations). In Table 1 we provide empirical critical values for some choices of n. For $n = 50$ we took $d(50) = 5, D_n = 2$, while for $n = 400$ we took $d(400) = 8, D_n = 3$ and $g^8 = (b_3, h_1, b_1, h_2, b_5, h_c, b_7, b_9)$.

TABLE 1. Simulated critical values of TS and TL. $\alpha = 0.05$, $n = 50, 100, 400$, $d(50) = 5$, $d(100) = 6$, $d(400) = 8$; 30,000 MC runs

Test	$n = 50$	$n = 100$	$n = 400$
TS	5.590	5.290	4.524
TL	6.177	6.362	5.986

In power simulations we used critical values from Table 1.

Tests for comparisons. As competitors of TS and TL we consider here the tests which showed the best performance in simulations presented in [2]. They are as follows:

• The Modarres and Gastwirth hybrid test, denoted by MG, with $p = 0.8$ and $\alpha_1 = 0.01,\ \alpha_2 = 0.0404$ as was suggested by the authors. For detailed description see [4].

• The test based on the function $h_1(t)$ (which, in our case, is $h_{[15/16,1]}(t)$, cf. (2.6)) denoted here by H. Because the test statistic $n\widehat{h}_1^2$ tends fast to the asymptotic chi-square distribution with one degree of freedom, we use the asymptotic critical value 3.841.

• The Inglot et al. [2] data driven tests NS and $NL3$ based on the system of the odd Legendre polynominals, denoted here by NS and NL.

Alternatives. We have considered a broad spectrum of alternatives including the popular Tukey family (denoted by Tuk(λ_3, λ_4)) and the generalized lambda family (denoted by Lamb(λ_3, λ_4)). Most of them have been described in [2] or [3], but some are new. For the reader's convenience we provide a full list of alternatives divided into three groups according to a structure of their asymmetry. We took 16 alternatives from the first group, 8 from the second and 4 from the third.

Let $\chi_k^2(x)$ denote the density of the chi-square distribution with k degrees of freedom, $\beta_{\xi,\eta}(x),\ \xi, \eta > 0$, the density of the beta distribution, $c(x) = 1/[\pi(1 + x^2)]$ the density of the Cauchy distribution, $\phi(x)$ the standard normal density function, U a random variable uniformly distributed on $[0, 1]$ and Z a random variable with the standard normal distribution. Additionally, define the density function en(x) by the formula

$$\mathrm{en}(x) = c[\phi(x+1)\mathbf{1}_{(-\infty,-1)}(x) + \phi(0)\mathbf{1}_{[-1,1]}(x) + \phi(x-1)\mathbf{1}_{(1,\infty)}(x)]$$

with $c = \big(1 + 2\phi(0)\big)^{-1}$. Each distribution, described below, is used as an alternative after centering by its median.

• **Alternatives with dominating asymmetry in the tails**:

Notation	**Description of a random variable or a density**
Tuk(λ_3,λ_4)	$X = (U^{\lambda_3} - 1)/\lambda_3 - \big((1-U)^{\lambda_4} - 1\big)/\lambda_4,\ \ \lambda_3, \lambda_4 > 0;$
Lamb(λ_3,λ_4)	$X = \mathrm{sgn}(\lambda_3)\big(U^{\lambda_3} - (1-U)^{\lambda_4}\big),\ \ \lambda_3 \cdot \lambda_4 > 0;$
IG(θ, λ)	$f(x) = \sqrt{\lambda/(2\pi x^3)}\exp\{-\lambda(x-\theta)^2/(2\theta^2 x)\},\ \ x > 0,$ $\theta, \lambda > 0;$
B(θ)	$\beta_{2,\theta}(x),\ \ \theta > 0;$
Chi(θ)	$\chi_\theta^2(x),\ \ \theta = 1, 2, \ldots;$
F(θ)	$f(x) = 0.5 + 2x\theta^{-2}(\theta - \lvert x\rvert)\mathbf{1}_{(-\theta,\theta)}(x),\ \ x \in [-1, 1],$ $\theta \in [0, 1];$
Lehm(θ)	$f(x) = \theta \cdot 0.5^\theta (x+1)^{\theta-1},\ \ x \in [-1, 1],\ \theta > 1;$
NFech(θ)	$f(x) = \phi\big(x/(1+\theta)\big)\mathbf{1}_{(-\infty,0]}(x) + \phi\big(x/(1-\theta)\big)\mathbf{1}_{(0,\infty)}(x),$ $x \in \mathbb{R},\ \theta \in (-1, 1);$
EV(θ)	$f(x) = \exp\{(x-\theta) - \exp(x-\theta)\},\ \ x \in \mathbb{R},\ \theta \in \mathbb{R};$
Ra(θ)	$f(x) = \theta^{-2}x\exp\{-x^2/(2\theta^2)\},\ \ x \geqslant 0,\ \theta > 0;$
ShAsh(∞, θ)	$X = 0.5\exp\{\mathrm{arcsinh}(Z)/\theta\} - 0.5,\ \ \theta > 0.$

- **Alternatives with asymmetry in the tails and in the center**:

Notation	**Density**
CFech(θ)	$f(x) = c\big(x/(1+\theta)\big)\mathbf{1}_{(-\infty,0]}(x) + c\big(x/(1-\theta)\big)\mathbf{1}_{(0,\infty)}(x),\ \ x \in \mathbb{R},$ $\theta \in (-1,1)$;
NB(θ)	$f(x) = 0.8\phi(x) + 0.2\beta_{3,3}(x+\theta),\ \ \theta \in \mathbb{R}$;
B2(θ)	$f(x) = 0.5\left(\beta_{2,\theta}(x+1) + \beta_{2,2}(x)\right),\ \ \theta > 0$;
Chi2(θ)	$f(x) = 0.5\left(\chi^2_\theta(-x) + \chi^2_6(x)\right),\ \ \theta = 1,2,\ldots$;
Sin(θ, j)	$f(x) = 0.5 + \theta\sin(\pi j x),\ \ x \in [-1,1],\ \theta \in [-0.5, 0.5],\ j \geqslant 1$;
B4(θ)	$f(x) = 0.2\beta_{3,3}(x) + 0.4\beta_{3,3}(x+1) + 0.1\beta_{2,5}(x+1) + 0.3\beta_{2,\theta}(x),$ $\theta \in \mathbb{R}$;
NC(θ)	$f(x) = 0.5\phi(x) + 0.5c(x-\theta),\ \ \theta \in \mathbb{R}$;
LC(θ)	$f(x) = 0.7\phi(x - \theta/0.7) + 0.3\phi(x + \theta/0.3),\ \ \theta \in \mathbb{R}$.

- **Alternatives with asymmetry only in the center**:

Notation	**Density**
B3(θ)	$f(x) = 0.1\beta_{1,2}(x+1) + 0.1\beta_{2,1}(x) + 0.8\beta_{1,\theta}(x+m),$ $\theta > 0,\ m-$ the median of $\beta_{1,\theta}(x)$;
ENB(θ)	$f(x) = 0.2\,\mathrm{en}(x) + 0.8m\beta_{\theta,2}(mx+m),\ \ \theta > 0,$ $m-$ the median of $\beta_{\theta,2}(x)$;
N2B2(θ)	$f(x) = 0.25\big(\phi(x-2) + \phi(x+2)\big) + \beta_{\theta,4}(4x+4) + 0.75\beta_{6,3}(3x),$ $\theta > 0$;
NC2(θ)	$f(x) = 0.3\phi(x) + 0.4c(x-\theta) + 0.3c(x+2\theta),\ \ \theta \in \mathbb{R}$.

Additionally, we wanted to verify how our tests perform for the problem of testing symmetry about 0, i.e. for alternatives with nonzero median. For this purpose we used alternatives Sin(θ, j), EV(θ), F(θ) and Lehm(θ) without centering by their medians. We denote them by the same symbol adding *. Moreover, we took the following two alternatives:

- **Alternatives with nonzero median**:

Notation	**Density**
Logis(θ)	$f(x) = \exp(x-\theta)/\big(1+\exp(x-\theta)\big)^2,\ \ x \in \mathbb{R},\ \theta \in \mathbb{R}$;
B3S(θ)	$f(x) = 0.3\big(\beta_{2,1}(x+1) + \beta_{1,2}(x)\big) + 0.4\beta_{1,\theta}(x+0.5),\ \ \theta > 0$.

Power comparisons. In Table 2 we present results for alternatives from the first group. As one could expect the directional tests MG and H attain the highest power. This is not surprising since TS and TL are omnibus tests. In spite of this, TS loses with respect to MG only ca. 3–4%. But NS and NL are distinctly weaker (ca. 11% on average with respect to TS and TL).

In Table 3 we show results for the second group. It is easily seen that now H becomes much weaker but the other five tests perform almost equally well. However, the tests NL and TL give a ca. 6% gain in average power to NS and TS, respectively, since the lighter penalty in L allows for better detection of asymmetry in the center.

TABLE 2. Empirical powers (in %) of MG, H, NS, NL, TS and TL. $\alpha = 0.05$, $n = 100$, $d(n) = 6$; 10,000 MC. Dominating asymmetry in the tails

Alternative	MG	H	NS	NL	TS	TL
Tuk$(0.1, 0.4)$	61	64	43	40	56	52
Tuk$(10, 0.9)$	68	77	49	47	73	68
Tuk$(7, 1.6)$	72	76	55	52	70	66
Tuk$(4, 6)$	67	70	53	50	62	59
Lamb$(0.025213, 0.094029)$	86	86	74	70	80	77
Lamb$(-0.0075, -0.03)$	96	96	89	87	93	91
Lamb$(-0.1, -0.18)$	49	49	39	35	41	37
IG$(0.05, 1)$	56	60	41	37	51	46
B(4)	78	83	60	57	77	73
Chi(9)	89	90	75	72	87	84
F(0.15)	57	64	38	37	53	50
Lehm(1.2)	66	75	47	46	67	64
NFech(0.4)	68	67	54	51	60	57
EV(0.367)	89	88	76	73	84	82
Ra(1)	74	79	55	52	72	68
ShAsh$(+\infty, 4.5)$	65	68	49	46	62	57
Average	71.3	74.5	56.1	53.4	68.0	64.4

TABLE 3. Empirical powers (in %) of MG, H, NS, NL, TS and TL. $\alpha = 0.05$, $n = 100$, $d(100) = 6$; 10,000 MC runs. Asymmetry in the tails and in the center

Alternative	MG	H	NS	NL	TS	TL
CFech(0.3)	49	36	56	52	48	46
NB(0.1)	56	40	69	66	56	59
B2(4)	47	32	55	57	55	70
Chi2(4)	50	38	49	49	44	54
Sin$(0.5, 8)$	44	50	44	98	56	70
B4(3)	46	38	38	36	36	41
NC(3.4)	57	34	83	82	66	69
LC(0.5)	67	59	61	57	58	56
Average	52.0	34.6	56.9	62.1	52.4	58.1

In Table 4, results for the third group are presented. In this case the tests MG and H perform poor. This could be expected since they are designed to detect asymmetry in the tails. All the data driven tests preserve good sensitivity. But the new tests TS and TL are slightly better than NS and NL and give a ca. 2–6% gain in average power.

In Table 5 we show empirical powers of the compared tests for alternatives with nonzero median. Although the assumptions of our model are not satisfied, the new tests perform comparably to MG. Here NS and NL are much better since nonzero median is well detected by the first Legendre polynomial.

Finally, in Table 6 we compare powers of the new tests and MG with the most powerful test (denoted by NP) for five selected alternatives. For each al-

TABLE 4. Empirical powers (in %) of MG, H, NS, NL, TS and TL. $\alpha = 0.05$, $n = 100$, $d(100) = 6$; 10,000 MC runs. Asymmetry only in the center

Alternative	MG	H	NS	NL	TS	TL
B3(2.5)	7	6	55	63	62	72
ENB(6)	7	5	25	30	40	41
N2B2(12)	11	7	67	72	46	69
NC2(1.4)	12	7	38	38	45	44
Average	9.3	6.3	46.3	50.8	48.3	56.5

TABLE 5. Empirical powers (in %) of MG, H, NS, NL, TS and TL. $\alpha = 0.05$, $n = 100$, $d(100) = 6$; 10,000 MC runs. Alternatives with nonzero median

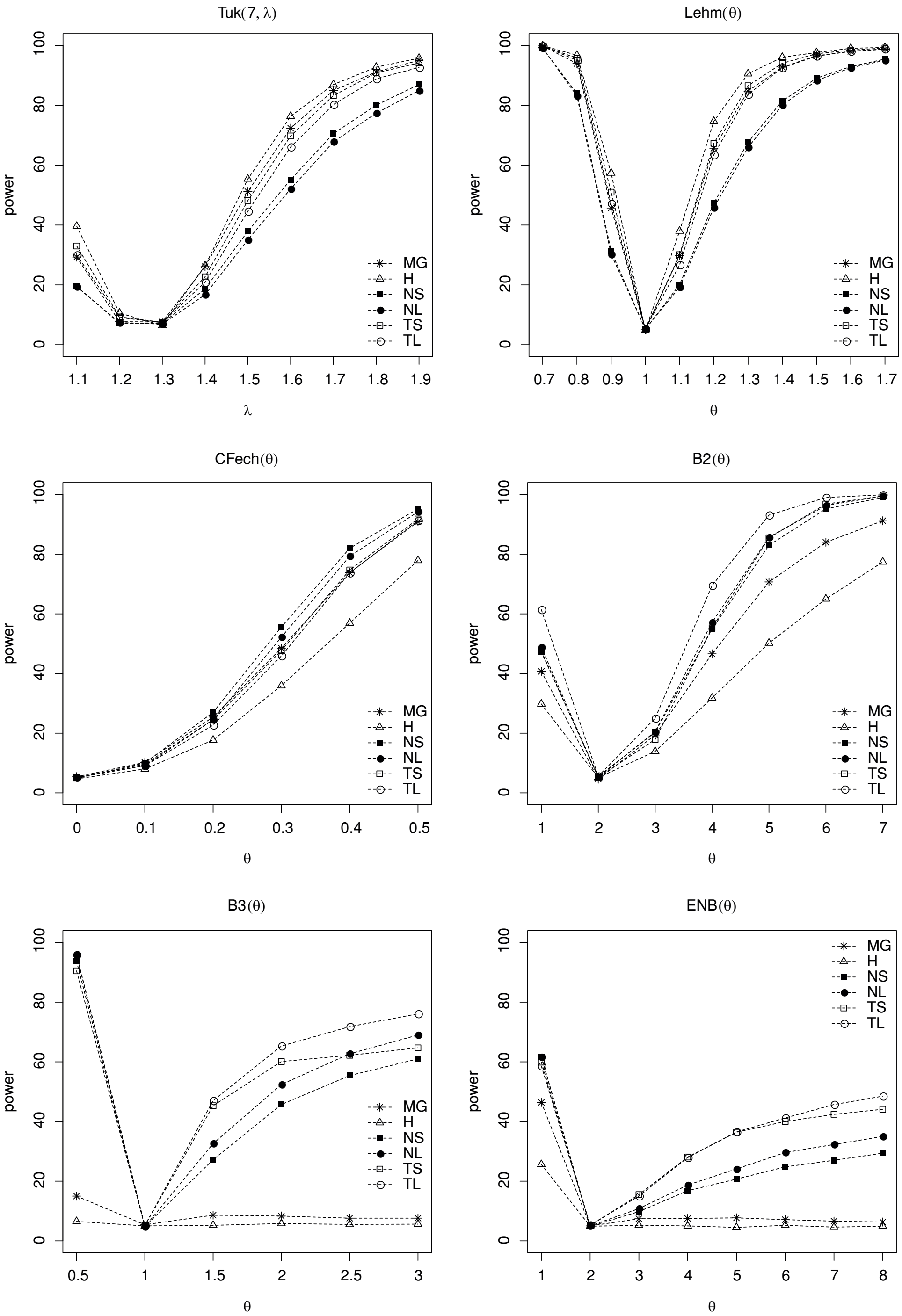
Alternative	MG	H	NS	NL	TS	TL
Sin*(0.3, 3.5)	42	33	54	72	47	64
EV*(0.6)	69	47	72	70	63	64
F*(0.2)	8	5	11	24	11	30
F*(0.4)	31	5	81	93	32	65
Lehm*(1.2)	21	22	27	23	21	20
Logis(0.4)	36	23	49	44	32	33
B3S(2)	40	5	56	55	40	48
Average	35.3	20.0	50.0	54.4	35.1	46.3

TABLE 6. Empirical powers (in %) of NP and MG, TS and TL. $\alpha = 0.05$, $n = 100$, $d(100) = 6$; 10,000 MC runs

Alternative	NP	MG	TS	TL
Lehm*(1.2)	52	21	21	20
Sin*(0.3, 3.5)	100	42	47	64
LC(0.5)	86	67	58	56
B3(2.5)	100	7	62	72
ENB(6)	91	7	40	41
Average	85.8	28.8	45.6	50.6

ternative we take its symmetric part as the null distribution for constructing the most powerful test. One can see that MG performs unstably in opposite to the data driven tests (and particularly to TL) which, being stable, keep approximately a constant room in power to NP. On the average this room equals ca. 35% for TL. This observation reflects optimality properties of data driven tests discussed in Section 4.4.3 of [2].

For further illustration of the performance of the compared tests, in Figures 1–3 we present power curves for two alternatives selected from each group when changing the parameter θ (or λ in one case) of the underlying distribution. All figures confirm previous observations and a good performance of the new tests.



FIGURES 1–3. Empirical power curves (in %) of MG (-$*$-), H (-$\triangle$-), NS (-■-), NL (-•-), TS (-□-), TL (-○-) for alternatives from the first, second and third groups. $n = 100, \alpha = 0.05, d(100) = 6$; 10,000 MC runs

5. CONCLUSIONS

The presented simulation study shows that our main goal to refine the data driven tests NS and NL has been achieved. The newly introduced data driven tests TS and TL perform much better for alternatives with dominating asymmetry in the tails and slightly better for alternatives with asymmetry in the center. To assess the omnibus character of all compared tests we calculated average power over all 28 alternatives and obtained (in %): MG 56.9, H 53.4, NS 54.9, NL 55.4, TS 60.7, TL 61.5. If one expects (or wants to detect) asymmetry in the tails then the tests MG and H are the best ones but TS is only slightly weaker. When one expects also strong asymmetry in the center then we recommend TL as the best solution. Moreover, results presented in Table 5 show that TL preserves a good sensitivity for alternatives with a nonzero median. For larger samples the tests TS and TL perform better than for $n = 100$ in comparison with their competitors. For $n = 50$, taking $d(n) = 5, D_n = 2$, one gets practically the same picture as for $n = 100$. For smaller samples data driven score tests cannot give a profit from their construction. However, for $n = 25,\ d(25) = 4$ and $D_n = 2$ the tests TS and TL lose on average ca. 5% in power with respect to MG.

6. PROOFS

In all proofs c denotes some generic constant different in each case. To prove theorems stated in Section 3 we shall apply results from the Appendix in [2].

To this end, we modify each g_{kj} in both ends of every interval from a partition it determines and we obtain $\widetilde{\varphi}_{kj}^{(n)} = \widetilde{\varphi}_{kj}$ in accordance with the following principles. Let $[u, v] \in [0, 1]$ be one of the intervals from the partition determined by g_{kj}. If g_{kj} has no zeros on $\left[u, u + \frac{1}{2n}\right]$, then we take $\widetilde{\varphi}_{kj}(u) = 0$, $\widetilde{\varphi}_{kj}\left(u + \frac{1}{2n}\right) = g_{kj}\left(u + \frac{1}{2n}\right)$ and $\widetilde{\varphi}_{kj}$ linear on $\left(u, u + \frac{1}{2n}\right)$. Otherwise, we choose one of zeros in $\left[u, u + \frac{1}{2n}\right]$, say z, and put $\widetilde{\varphi}_{kj}(t) = 0$ on $[u, u + z]$ and $\widetilde{\varphi}_{kj}(t) = g_{kj}(t)$ on $\left[z, z + \frac{1}{2n}\right]$. The modification of g_{kj} on the interval $\left[v - \frac{1}{2n}, v\right]$ is carried over the same rule and $\widetilde{\varphi}_{kj} = g_{kj}$ on the rest of $[u, v]$. Additionally, the modification is made in such a way that $\widetilde{\varphi}_{kj}$ is antisymmetric with respect to $1/2$. Obviously, each $\widetilde{\varphi}_{kj}$ is absolutely continuous on $[0, 1]$. However, $\widetilde{\varphi}_{k1}, \ldots, \widetilde{\varphi}_{kk}$ may no longer be orthogonal and normalized.

In order to normalize $\widetilde{\varphi}_{kj}$ observe that $\widetilde{\varphi}_{kj} = g_{kj}$ outside the set of Lebesgue measure of at most l_{kj}/n and $\|\widetilde{\varphi}_{kj}\|_\infty \leqslant \|g_{kj}\|_\infty$, where for a bounded function v on $[0, 1]$ we put $\|v\|_\infty = \sup_{t \in [0,1]} |v(t)|$. From (3.1) and (3.2) and the construction of $\widetilde{\varphi}_{kj}$'s we also have

$$0 < 1 - \|\widetilde{\varphi}_{kj}\|^2 = \int_0^1 \left(g_{kj}^2(t) - \widetilde{\varphi}_{kj}^2(t)\right) dt \leqslant \frac{c[d(n)]^{2\eta+\zeta}}{n}, \tag{6.1}$$

where $\|v\|$ stands for the L_2-norm of a function v. Now, let us write

$$\varphi_{kj} = \frac{\widetilde{\varphi}_{kj}}{\|\widetilde{\varphi}_{kj}\|}, \quad j = 1, 2, \dots, k, \quad k = 1, 2, \dots, d(n),$$

and $\varphi^k = (\varphi_{k1}, \dots, \varphi_{kk})^T$.

For each $1 \leqslant k \leqslant d(n)$ let λ_k denote the largest eigenvalue of the covariance matrix $\Gamma_k = [\gamma_k(i,j)] = \int_0^1 \varphi^k(t)\big(\varphi^k(t)\big)^T dt$. Then we have the following lemma.

LEMMA 6.1. *If* (3.1) *and* (3.2) *are satisfied and* $d(n)^{2\eta+\zeta+1}/n \to 0$ *as* $n \to \infty$, *then for sufficiently large* n *we have*

$$\max_{1 \leqslant k \leqslant d(n)} \lambda_k \leqslant 1 + c\frac{[d(n)]^{2\eta+\zeta+1}}{n}.$$

P r o o f. To simplify the notation we shall write d instead of $d(n)$. Additionally, set $\xi_n = d^{2\eta+\zeta}$. Since φ_{kj} are normalized, we have $\gamma_k(i,i) = 1$. For $i \neq j$, using orthogonality of g_{kj}, (3.1), (3.2), (6.1) and the definition of $\widetilde{\varphi}_{kj}$ we get

$$|\gamma_k(i,j)| \leqslant \frac{1}{\|\widetilde{\varphi}_{ki}\|\,\|\widetilde{\varphi}_{kj}\|}\Big|\int_0^1 \widetilde{\varphi}_{ki}(t)\widetilde{\varphi}_{kj}(t)dt\Big| \leqslant c\frac{d^{2\eta+\zeta}}{n} = c\frac{\xi_n}{n}.$$

This enables us to write

$$\Gamma_k = I + \frac{\xi_n}{n}Q,$$

where elements q_{ij} of the matrix Q are uniformly bounded. Put $M = \max_{i,j}|q_{ij}|$. Then the elements of Q^2 are bounded by $\big|\sum_r q_{ir}q_{rj}\big| \leqslant M^2 k$, the elements of Q^3 by M^3k^2, and so on. Since

$$\Gamma_k^n = \left(I + \frac{\xi_n}{n}Q\right)^n = I + \xi_n Q + \left(\frac{\xi_n}{n}\right)^2\binom{n}{2}Q^2 + \ldots + \left(\frac{\xi_n}{n}\right)^n\binom{n}{n}Q^n,$$

the elements on the diagonal of Γ_k^n are bounded by

$$1 + \xi_n M + \ldots + \xi_n^n M^n k^{n-1}/n! < \exp\{Md\,\xi_n\}.$$

Hence, $\lambda_k^n \leqslant k\exp\{Md\,\xi_n\} \leqslant d\exp\{Md\,\xi_n\} \leqslant \exp\{(M+1)d\xi_n\}$. Using the assumption on d and the relation $\exp(u) \leqslant 1 + 2u$, being true for $u \in [0,1]$, we get for every k and sufficiently large n

$$\lambda_k \leqslant \exp\left\{(M+1)\frac{d\xi_n}{n}\right\} \leqslant 1 + c\frac{d\xi_n}{n},$$

which completes the proof of the lemma. ∎

In addition to $\widehat{g}_{kj}$ given by (2.5) define

$$\widehat{\varphi}_{kj} = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \varphi_{kj} \left(\frac{2R_i - 1}{4n} \right)$$

and $\widehat{\varphi}^k = (\widehat{\varphi}_{k1}, \ldots, \widehat{\varphi}_{kk})^T$. Additionally, put $\widehat{g}^k = (\widehat{g}_{k1}, \ldots, \widehat{g}_{kk})^T$. Then we have the following lemma.

LEMMA 6.2. *If the conditions* (3.1) *and* (3.2) *are satisfied*, *then*

$$\max_{1 \leqslant k \leqslant d(n)} |\widehat{g}^k - \widehat{\varphi}^k|_k^2 \leqslant c \frac{[d(n)]^{2\eta + 2\zeta + 1}}{n} \quad a.s. \tag{6.2}$$

By (6.1), the proof goes the same way as that of Lemma 3.1 in [3]. So, we omit it.

Before giving the proof of our theorems we state Theorems A.1 and A.2 from [2] in our present setting.

THEOREM A (Inglot et al. [2]). *Suppose H_0 is true.*

(1) *Then for each fixed* k, $1 \leqslant k \leqslant d(n)$,

$$|\widehat{\varphi}^k|_k^2 \xrightarrow{D} \chi_k^2 \quad as\ n \to \infty.$$

(2) *For any sequence $k(n)$ of natural numbers*, $1 \leqslant k(n) \leqslant d(n)$, *any* $\nu \in \left(0, \frac{1}{2}\right)$, *and every sequence x_n of positive numbers such that*

$$x_n \to 0,\ nx_n^2/\big(k(n)\lambda_{k(n)}\big) \to \infty,\ \textit{and}\ x_n^{2-4\nu}\psi^4\big(k(n)\big)/\lambda_{k(n)}^3 \to 0 \quad as\ n \to \infty$$

we have

$$P(|\widehat{\varphi}^{k(n)}|_{k(n)}^2 \geqslant nx_n^2) = \exp\left\{ -\frac{nx_n^2}{2\lambda_{k(n)}} + O\left(\frac{nx_n^{2+\nu}}{\lambda_{k(n)}} \right) + O\left(k(n) \log \frac{nx_n^2}{k(n)\lambda_{k(n)}} \right) \right\}, \tag{6.3}$$

where $\psi^2(k) = \sum_{j=1}^{k} \left(\int_0^1 |\varphi_{kj}'(t)| dt \right)^2$.

The formula (6.3) has a slightly stronger form than (A.14) in [2]. However, its proof goes exactly the same way. The only difference is that we use a finer form of expansion of tails of multivariate Gaussian distributions. We need this stronger form to prove (6.4) below.

Proof of Theorem 3.1. Applying Theorem A (1) for $k = 1$ we obtain $\widehat{\varphi}_{11}^2 \xrightarrow{D} \chi_1^2$. By Lemma 6.2 and the assumption on $d(n)$ it immediately implies

$$T_1 = \widehat{g}_{11}^2 \xrightarrow{D} \chi_1^2.$$

Now, observe that by the construction of $\widetilde{\varphi}_{kj}$ we have for $k = 1, \ldots, d(n)$

$$\int_0^1 |\widetilde{\varphi}'_{kj}(t)|dt \leqslant \int_0^1 |g'_{kj}(t)|dt + \int_{\{\widetilde{\varphi}_{kj} \neq g_{kj}\}} |\widetilde{\varphi}'_{kj}(t)|dt \leqslant \int_0^1 |g'_{kj}(t)|dt + 2l_{kj}\|g_{kj}\|_\infty,$$

which, by (3.1)–(3.3) and (6.1), gives

$$\psi^2(k) = \sum_{j=1}^{k} \Big(\int_0^1 |\varphi'_{kj}(t)|dt\Big)^2 \leqslant c[d(n)]^\kappa + c[d(n)]^{2\eta+2\zeta+1} \leqslant c[d(n)]^\rho.$$

By the assumption of Theorem 3.1 and Lemma 6.1 we obtain $\lambda_k = 1 + o(1)$. Applying Theorem A (2) to $nx_n^2 = (k-1)\log n$ and some $\nu \in \big(0, \big(1 - \tau(1+2\rho)\big)/2\big)$ we see that the assumptions of this theorem are fulfilled. So, from (6.3) we get

(6.4)
$$P\big(|\widehat{\varphi}^k|_k^2 \geqslant (k-1)\log n\big) = \exp\Big\{ -\frac{k-1}{2}(\log n)\big(1+o(1)\big)\Big\} \leqslant n^{-\frac{1}{2}(1+o(1))}$$

as $n \to \infty$. Hence, by Lemma 6.2 and the assumption on $d(n)$ we have $P\big(|\widehat{g}^k|_k^2 \geqslant (k-1)\log n\big) \leqslant n^{-\frac{1}{2}(1+o(1))}$ as $n \to \infty$ and, consequently,

$$P(S \geqslant 2) \leqslant \sum_{k=2}^{d(n)} P\big(|\widehat{g}^k|_k^2 \geqslant (k-1)\log n\big) \leqslant d(n)n^{-\frac{1}{2}(1+o(1))},$$

which tends to zero again due to the assumption on $d(n)$, and therefore completes the proof of the first part of Theorem 3.1. The second part can be proved similarly to that of Theorem 3.1 in [3] after observing that Lemma 6.1 holds true for the covariance matrix of every part of the system φ^k. ■

P r o o f o f T h e o r e m 3.2. Let P denote the distribution of the sample $X_1, \ldots, X_n$ with a fixed asymmetric distribution function F such that the function a determined by F (cf. (2.1)) satisfies (3.5). Set

$$s_{na} = \int_0^1 \varphi^{d(n)}(t)a(t)dt.$$

Now, we shall use the results from the Appendix in [2]. Applying (A.21) with $nx_n^2 = \log^4 n$ and some $\sigma \in \big(\rho/(1+2\rho), 1/2\big)$, (A.22) with $nx_n^2 = \omega_n \log^2 n$, where ω_n is defined in (3.5), and (A.3), we get

$$\widehat{\varphi}^{d(n)} - \sqrt{n}s_{na} = O_P\big(\omega_n d(n)^{\rho/2}\log n\big).$$

By (3.1), (3.2), Lemma 6.2 and the construction of $\varphi^{d(n)}$ we obtain $\widehat{g}^{d(n)} - \widehat{\varphi}^{d(n)} = o_P(1)$ and $\sqrt{n}s_{na} - \sqrt{n}\int_0^1 g^{d(n)}(t)a(t)dt = o_P(1)$. This implies

$$\widehat{g}^{d(n)} - \sqrt{n}\int_0^1 g^{d(n)}(t)a(t)dt = \mathcal{R}_n = O_P\big(\omega_n d(n)^{\rho/2}\log n\big).$$

Hence, by the assumption (3.5),

$$P\big(|\widehat{g}^{d(n)}|^2_{d(n)} \geqslant 2d(n)\log n\big) \geqslant P\big(|\widehat{g}^{d(n)}|^2_{d(n)} \geqslant \omega_n^2 d(n)^\rho \log^2 n\big)$$
$$= P\Big(\Big|\mathcal{R}_n + \sqrt{n}\int_0^1 g^{d(n)}(t)a(t)dt\Big|^2_{d(n)} \geqslant \omega_n^2 d(n)^\rho \log^2 n\Big) \to 1$$

as $n \to \infty$. Since $T_L \geqslant T_S \geqslant |\widehat{g}^{d(n)}|^2_{d(n)} - d(n)\log n$ a.s. by the definition of S and L, the assertion of the theorem holds true. ■

REFERENCES

[1] T. Inglot and A. Janic, *How powerful are data driven score tests for uniformity*, Appl. Math. (Warsaw) 36 (2009), pp. 375–395.

[2] T. Inglot, A. Janic, and J. Józefczyk, *Data driven tests for univariate symmetry*, Probab. Math. Statist. 32 (2) (2012), pp. 323–358.

[3] J. Józefczyk, *Data driven score tests for univariate symmetry based on non-smooth functions*, Probab. Math. Statist. 32 (2) (2012), pp. 301–322.

[4] R. Modarres and J. L. Gastwirth, *Hybrid test for the hypothesis of symmetry*, J. Appl. Stat. 25 (1998), pp. 777–783.

Tadeusz Inglot
Department of Pure and Applied Mathematics
Faculty of Fundamental Problems of Technology
Wrocław University of Technology
Wybrzeże Wyspiańskiego 27
50-370 Wrocław, Poland
E-mail: Tadeusz.Inglot@pwr.edu.pl

Dawid Kujawa
Department of Pure and Applied Mathematics
Faculty of Fundamental Problems of Technology
Wrocław University of Technology
Wybrzeże Wyspiańskiego 27
50-370 Wrocław, Poland
E-mail: daw.kujawa@gmail.com

Received on 8.4.2014;
revised version on 3.6.2014