Available online at www.sciencedirect.com
SCIENCE DIRECT®
ELSEVIER
Linear Algebra and its Applications 417 (2006) 124–133
LINEAR ALGEBRA AND ITS APPLICATIONS
www.elsevier.com/locate/laa

Towards data driven selection of a penalty function for data driven Neyman tests

Tadeusz Inglot, Teresa Ledwina *

Institute of Mathematics, Polish Academy of Sciences, ul. Kopernika 18, 51-617 Wrocław, Poland

Received 20 December 2004; accepted 24 October 2005
Available online 15 December 2005
Submitted by A. Markiewicz

Abstract

The data driven Neyman statistic consists of two elements: a score statistic in a finite dimensional submodel and a selection rule to determine the best fitted submodel. For instance, Schwarz BIC and Akaike AIC rules are often applied in such constructions. For moderate sample sizes AIC is sensitive in detecting complex models, while BIC works well for relatively simple structures. When the sample size is moderate, the choice of selection rule for determining a best fitted model from a number of models has a substantial influence on the power of the related data driven Neyman test. This paper proposes a new solution, in which the type of penalty (AIC or BIC) is chosen on the basis of the data. The resulting refined data driven test combines the advantages of these two selection rules.

Keywords: Akaike criterion; Asymptotic optimality; Data driven test; Goodness of fit; Neyman test; Power comparison; Schwarz selection rule

1. Introduction

Data driven Neyman tests are based on two elements: Neyman's smooth statistic (or equivalently, score statistic) in a finite dimensional submodel and a selection rule to choose the appropriate submodel. Ledwina [20] introduced such a construction for the case of testing uniformity, proposing to use Schwarz [24] BIC criterion as the selection rule. There was clear motivation for such a choice. The smooth statistic can be naturally related to some testing problem in a

* Corresponding author. Tel.: +48 71 372 8895; fax: +48 71 348 1098.
E-mail address: ledwina@impan.pan.wroc.pl (T. Ledwina).

0024-3795/$ - see front matter © 2005 Elsevier Inc. All rights reserved.
doi:10.1016/j.laa.2005.10.023

finite dimensional exponential family, while the Schwarz rule was simply designed to select the dimension of the exponential model. Obviously, some extensions of the original Schwarz rule were available at that time, allowing for a large range of penalties (cf. [9, Remark 1.2]). However, an appealing feature of Schwarz's solution was the large penalty, which results in the selection rule behaving nicely for "small" models, using the handy formulation of Shen and Ye [25]. This guarantees that the construction of the related test is consistent with the following, often very natural, assumption. A researcher has some knowledge about the phenomenon under investigation and this is reflected by the form of the null model. Consequently, small, rather than large departures, from the null model are expected. This postulate is silently assumed in many simulation studies. For illustration, almost every alternative to normality listed in [22] can be considered as a small departure (as explained in Sections 2 and 5) from the null model. In contrast, large departures from normality could be defined by multimodal mixtures or heavy tailed distributions.

In recent years, many other penalized statistics with smaller penalties, such as Akaike [2] or modified forms of BIC, have been proposed. In particular, such an approach appeared in goodness-of-fit tests for some regression problems. See [1,7,8] for some examples and further references. Such a choice is motivated by the greater complexity of the underlying models and imprecise knowledge regarding the null and alternative distributions. It seems that the terminology *lack of fit test*, often used in the context of regression, instead of *goodness of fit test*, typically associated with fitting distributions, also reflects the greater uncertainty related to some regression problems. Since small penalties are associated with detecting "large" models, again using the terminology of Shen and Ye [25], the effect is that the related tests are less powerful for small departures (from the null hypothesis) than those related to BIC.

This paper aims to propose a solution that has the advantages of both types of penalty. For conciseness we restrict attention to testing uniformity and two (simplified) selection rules: BIC and AIC. Roughly speaking, the basic idea is, given the data, to decide which penalty should be used to select the number of terms in the score statistic. The new solution, together with some optimality properties, is described in Section 2. Section 3 presents a simulation study showing the advantages of the refined version of the data driven test. Section 4 briefly discusses the asymptotic behaviour of new test statistic under the null hypothesis. In Section 5 some possible extensions are indicated.

2. The new test

We start with some notation and discussion. For brevity we only consider Neyman's original test related to the Legendre system. Let $b_1, b_2, \ldots$ be orthonormal Legendre polynomials with respect to Lebesgue measure defined on $[0, 1]$. Let $X_1, \ldots, X_n$ be i.i.d. each distributed according to a continuous distribution with density $p(x)$ with respect to Lebesgue measure on $[0, 1]$. The null hypothesis H_0 asserts that $p = p_0$, where $p_0(x) \equiv 1$ for $x \in [0, 1]$. Denote the distribution corresponding to p_0 by P_0.

Consider the following probabilistic model for departures from p_0:

$$p_k(x;\theta) = p_0(x)\left[c_k(\theta)\exp\left\{\sum_{j=1}^{k}\theta_j b_j(x)\right\}\right], \tag{1}$$

where $\theta = (\theta_1, \ldots, \theta_k)$ and $c_k(\theta)$ is the normalizing factor. Obviously, $E_{P_0} b_j(X_1) = 0$ and $E_{P_0} b_i(X_1) b_j(X_1) = \delta_{ij}$, $i, j = 1, 2, \ldots$, where δ_{ij} is the Kronecker delta.

The score statistic for testing $\theta_1 = \cdots = \theta_k = 0$ in (1) is of the form

$$N_k = \sum_{j=1}^{k} \{\sqrt{n}\hat{b}_j\}^2, \quad \text{where } \hat{b}_j = \frac{1}{n}\sum_{i=1}^{n} b_j(X_i).$$

Obviously, the same score test would be obtained when considering any other type of departure with approximate structure $1 + \sum_{i=1}^{k} \theta_j b_j(x)$ for $\theta_1, \ldots, \theta_k$ close to 0.

The simplified BIC and AIC are defined by

$$S1 = \min\left\{1 \leqslant k \leqslant d(n) : N_k - k\log n \geqslant N_j - j\log n,\ j = 1, \ldots, d(n)\right\}$$

and

$$A1 = \min\left\{1 \leqslant k \leqslant d(n) : N_k - 2k \geqslant N_j - 2j,\ j = 1, \ldots, d(n)\right\},$$

where $d(n)$ is the number of models in the list. As seen, the simplification relies on replacing the maximized loglikelihood for (1) by $(1/2)N_k$, which is the standard approximation resulting from the delta method. The original versions of AIC and BIC could be considered as well, with the outcome expected to be similar, but we prefer to consider the simplest set-up. $A1$ and $S1$ are examples of so called score-based selection rules, while N_{A1} and N_{S1} are data driven Neyman's tests.

The asymptotic properties of N_{S1} were studied by Inglot [10] and Inglot and Ledwina [12]. In these papers the notation $S2$ was used for the simplified Schwarz rule. Kallenberg [15] considered a large class of data driven tests corresponding to a variety of penalties, including $A1$ and $S1$ as special cases. In particular, Kallenberg [15] proved that tests based on N_{A1} and N_{S1} are locally optimal in the sense of vanishing shortcoming (cf. his Theorem 4.7 and the comment following it). However, local asymptotic optimality does not exclude the situation that for moderate sample sizes, fixed alternatives and significance levels, the powers of N_{A1} and N_{S1} may be substantially different. Some evidence is presented in Tables 2–4 of Section 3. An explanation is given below. Note that, as typical for penalized criteria, in many cases the argument is understood in relation to the actual sample size.

Akaike's small penalty results in the inconsistency of the criterion (cf. [27]) and, as a consequence, in large critical values (see Table 1 for illustration). Hence, for "small" models, such as those defined mainly by three or four not very large, low order Fourier coefficients, N_{A1} is much weaker than N_{S1}, as the critical value is overestimated in the context of real data (remember that N_k is the score statistic in (1)). On the other hand, Schwarz's large penalty causes oversmoothing for moderate n. Hence, the power of N_{S1} for models with relatively large higher order Fourier coefficients is much smaller than that of N_{A1}, as large components are not included in the test statistic for relatively small n's. However, a large penalty is profitable in the sense that the critical value is small in comparison to the critical value obtained using AIC. This enables attaining high power for models with relatively large low order Fourier coefficients. For illustration, see the cases $j = 1$ and $j = 8$ in Table 2.

We propose to balance these two extreme tendencies as follows: use $A1$ only when an alternative is very distant from the null distribution, otherwise use $S1$. Using such an approach, the only missing element is an indication of how to decide whether we are close to the null model or not. To address this question, we propose to use the simple threshold rule described below.

Set

$$I_n(c) = \mathbf{1}\left(\max_{1\leqslant j\leqslant d(n)} |\sqrt{n}\hat{b}_j| \leqslant \sqrt{c\log n}\right), \tag{2}$$

where $\mathbf{1}(\bullet)$ is the indicator of the set $\bullet$. Observe that under the null hypothesis $\sqrt{n}\hat{b}_1, \ldots, \sqrt{n}\hat{b}_n$ are uncorrelated and approximately $N(0,1)$. Moreover, asymptotically they have i.i.d. $N(0,1)$ distributions. The threshold $\sqrt{c\log n}$, $c \geqslant 2$, is an adaptation of the standard solution for a white noise sequence of independent and identically distributed $N(0,1)$ variables $Z_1, \ldots, Z_n$ for which

$$Pr\left(\max_{1\leqslant j\leqslant n} |Z_i| > \sqrt{2\log n}\right) \to 0, \quad \text{as } n \to \infty$$

(cf. [5, p. 445], for such a choice of threshold). Therefore, an intuitive conclusion is that $I_n(c) = 1$ when $c \approx 2$, $c > 2$, indicates that the distribution of observations is close to the null model. Some simple, more formal, argument is provided in Appendix A. On the other hand, consider an alternative distribution P, absolutely continuous with respect to Lebesgue measure, as assumed before. Then, there exists $K = K(P)$, such that $E_P b_1(X_1) = \cdots = E_P b_{K-1}(X_1) = 0$ and $E_P b_K(X_1) \neq 0$, cf. [23, pp. 178, 191–192 with $\theta(x) = p(x) - 1$]. Simultaneously, $\left\{\max_{1\leqslant j\leqslant d(n)} |\sqrt{n}\hat{b}_j| \leqslant \sqrt{c\log n}\right\} \supset \left\{|\sqrt{n}\hat{b}_K| \leqslant \sqrt{c\log n}\right\}$, provided that $d(n) \geqslant K$. Therefore, by WLLN, for any $\{d(n)\}$, such that $d(n) \to \infty$ as $n \to \infty$ and any positive c, under any arbitrary alternative P, it holds that

$$\lim_{n\to\infty} P(I_n(c) = 0) = 1. \tag{3}$$

In our simulation study we took $c = c_0 = 2.4$. For some explanation of this choice see Section 3.

Now define the new penalty by

$$\pi(j,n) = \{j\log n\}\{I_n(c_0)\} + \{2j\}\{1 - I_n(c_0)\}, \tag{4}$$

and the new selection rule by

$$T1 = \min\left\{1 \leqslant k \leqslant d(n) : N_k - \pi(k,n) \geqslant N_j - \pi(j,n),\ j = 1, \ldots, d(n)\right\}.$$

The new data driven statistic N_{T1} is defined analogously to N_{S1} and N_{A1}.

Since $S1 \leqslant T1 \leqslant A1$ holds for $n \geqslant 8$, one gets $N_{S1} \leqslant N_{T1} \leqslant N_{A1}$. Therefore, local optimality in the sense of vanishing shortcoming for N_{T1} can be established by exploiting some known auxiliary results for N_{A1} and N_{S1} (see [15]). We shall not derive and formulate a precise result. Instead, in the next section we present a fragment of a simulation study, we performed to investigate the power behaviour of N_{T1} for finite samples and to compare it with a recently introduced test based on the likelihood ratio (see [28]).

3. The performance of $T1$ and N_{T1} in the simulation study

Let us start with some comments on the choice of c in the definition of $I_n(c)$ (cf. (2)). Obviously, the choice of c is critical to the performance of $T1$ and N_{T1}, both under the null and alternative hypotheses, when the sample size is moderate. Roughly speaking, under the alternative, the Schwarz rule is almost always selected for large c, while for small c Akaike penalty is frequently preferred. There is also empirical evidence that $T1$ and N_{T1} change smoothly as c increases. For illustration see Figs. 1 and 2. In Fig. 2, $c = 0$ corresponds to the application of AIC, while $c = \infty$ means BIC was used.

Our choice of the regularization is subjective. We focused on a value of c that allows AIC to act only if really large departures from the null model are present. Our option $c = c_0 = 2.4$

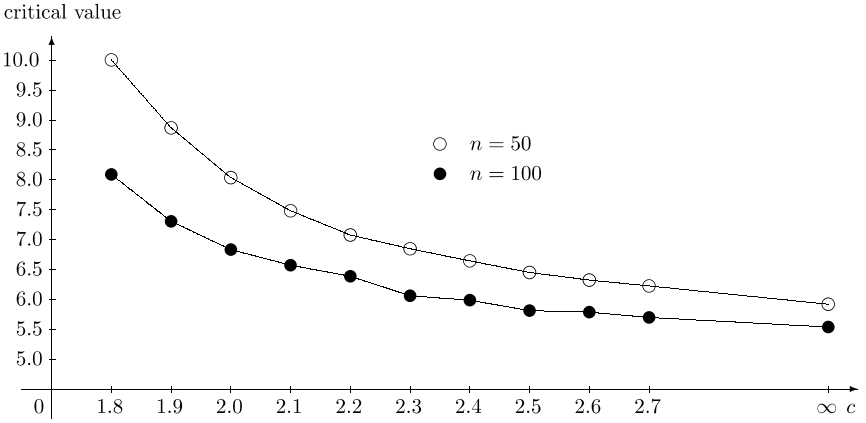

Fig. 1. The behaviour of simulated critical values of N_{T1} according to the switching constant c. $n = 100$, $\alpha = 0.05$, 10,000 MC.

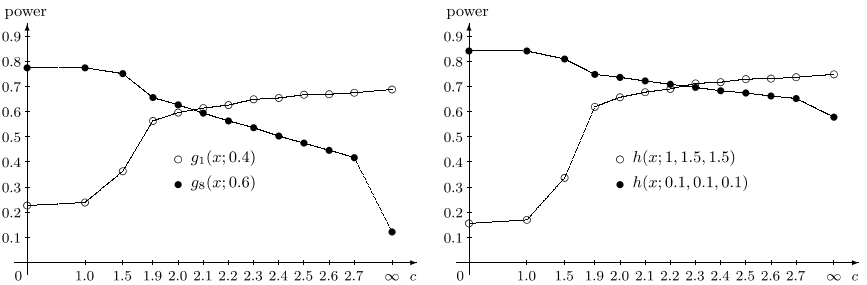

Fig. 2. The behaviour of simulated powers of N_{T1} according to the switching constant c. $n = 100$, $\alpha = 0.05$, 10,000 MC.

results in power comparable to that of N_{S1} when we have 'smooth' disturbances of the null model in the rough sense formulated by Neyman [21], i.e. when up to 3 or 4 of the initial components $\{\sqrt{n}\hat{b}_j\}^2$, $j = 1, 2, \ldots$ are significantly large. For illustration see Fig. 2 and Tables 2–4. Our choice also guarantees power comparable to that of N_{S1} if disturbances are small. On the other hand, for alternatives very far from the null model, such as, e.g., highly oscillating alternatives, the new solution essentially inherits the properties of N_{A1}. Obviously, as in the case of $A1$ and $S1$, when discussing $T1$ the complexity and size of the disturbances is understood relative to the sample size.

Throughout we consider $n = 100$, $d(n) = 12$ and significance level $\alpha = 0.05$. Each MC experiment was repeated 10,000 times.

Table 1 presents critical values of N_{A1}, N_{S1} and N_{T1} and empirical behaviour of the three selection rules when the null hypothesis H_0 is true. Note that in this experiment the number of cases in which $T1 = S1$ was 9858.

We also present some empirical powers under three types of departures: $g_j(x; \rho) = 1 + \rho\cos(\pi j x)$, $\rho \in (0, 1]$, $j = 1, 2, \ldots$, $p_k(x; \theta)$, $\theta \in R^k$, $k = 1, 2, \ldots$, given by (1) and $h(x; \epsilon, p, q) = 1 - \epsilon + \epsilon\beta_{p,q}(x)$, $\epsilon \in [0, 1]$, $p, q > 0$, where $\beta_{p,q}(x)$ stands for the beta density. To have some insight into the structure and magnitude of the alternatives, in each case we calculated (using an MC method) $d(n) = 12$ Fourier coefficients (in the Legendre basis) of the underlying

Table 1
Empirical behaviour of N_{A1}, N_{S1}, N_{T1} and selection rules under uniformity

Statistic	Critical value		Frequency in 10,000 simulations k											
			1	2	3	4	5	6	7	8	9	10	11	12
N_{A1}	15.684	$\{A1 = k\}$	7201	1086	549	353	211	163	131	93	86	77	50	0
N_{S1}	5.527	$\{S1 = k\}$	9613	323	47	14	1	1	0	0	1	0	0	0
N_{T1}	5.987	$\{T1 = k\}$	9536	295	57	25	13	17	13	13	11	12	8	0

$n = 100$, $d(n) = 12$, $\alpha = 0.05$, 10,000 MC runs.

Table 2
Empirical powers of Zhang's test and tests based on N_{A1}, N_{S1} and N_{T1} under the alternative $g_j(x; \rho)$

Parameters j	ρ	The five largest (in absolute value) Fourier coefficients $\times 1000$					Empirical powers (as percentages)				Percentage of cases
							Z_A	N_{A1}	N_{S1}	N_{T1}	$T1 = S1$
1	0.45	[**1**]315	[**3**]39	[**8**]2	[**12**]1	[**2**]1	76	32	81	78	56
2	0.40	[**2**]273	[**4**]79	[**6**]6	[**1**]2	[**8**]2	18	34	70	68	71
3	0.50	[**3**]316	[**5**]149	[**1**]40	[**7**]23	[**9**]4	34	54	65	65	54
4	0.60	[**4**]335	[**6**]235	[**2**]102	[**8**]58	[**10**]8	17	86	64	71	37
5	0.70	[**7**]320	[**5**]317	[**3**]173	[**9**]108	[**1**]20	41	97	60	78	27
6	0.70	[**8**]347	[**6**]232	[**4**]209	[**10**]156	[**2**]53	14	98	46	77	27
7	0.75	[**9**]377	[**5**]238	[**11**]217	[**7**]147	[**3**]98	33	98	33	81	20
8	0.80	[**10**]385	[**12**]280	[**6**]245	[**4**]141	[**8**]50	13	99	34	90	11

$n = 100$, $d(n) = 12$, $\alpha = 0.05$, 10,000 MC runs.

distributions. The five largest (from these 12) Fourier coefficients are presented in Tables 2–4. We also display the percentage of cases in which $T1 = S1$.

The results are encouraging. The new solution outperforms the power behaviour of N_{A1} for smooth alternatives and is comparable to N_{S1} in these cases. For highly oscillating alternatives N_{T1} is much more powerful than N_{S1} and not much less powerful than N_{A1}. Also, N_{T1} outperforms the power behaviour of Z_A, recently introduced by Zhang [28], as an improved construction compared to traditional tests. For many cases considered in Tables 2–4, further comparisons with the powers of the solution proposed by Bickel and Ritov [3], the classical chi-square test, as well as many data driven versions of it, introduced and studied in [4], as well as in [11] are possible. Simulations for the test of Bickel and Ritov [3] can be found in [16]. Again, such inspection shows that the new test competes well with other procedures.

4. The limiting distribution of N_{T1} under uniformity and consistency under alternatives

From (A.1) of Appendix A, when $c \geqslant 2$ and $d(n) \log^3 n/n \to 0$ as $n \to \infty$, we have $P_0(T1 \neq S1) \to 0$ as $n \to \infty$. Since $P_0(S1 = 1) \to 1$ (cf. [12, (2.4)], e.g., as mentioned earlier $S1$ is $S2$ in that paper), we infer that the asymptotic null distribution of N_{T1} is chi-square with one degree of freedom. The convergence is rather slow. For example, in Table 1 the simulated critical value of N_{T1} is 5.987, while the limiting value is 3.841. A similar phenomenon was observed in the cases of the original and simplified Schwarz rules. Kallenberg and Ledwina [17] described a simple, nicely working approximation, which can serve to calculate p-values and which can be adopted to

Table 3
Empirical powers of Zhang's test and tests based on N_{A1}, N_{S1} and N_{T1} under the alternative $p_k(x;\theta)$

Parameters		The five largest (in absolute value) Fourier coefficients ×1000					Empirical powers (as percentages)				Percentage of cases
k	θ						Z_A	N_{A1}	N_{S1}	N_{T1}	$T1 = S1$
1	0.3	[**1**]295	[**2**]38	[**9**]2	[**12**]1	[**11**]1	70	28	74	71	64
2	(−0.2, −0.3)	[**2**]256	[**1**]151	[**3**]40	[**4**]31	[**5**]5	70	24	75	73	79
3	(0, 0, 0.4)	[**3**]395	[**6**]66	[**2**]47	[**4**]44	[**5**]14	53	69	87	87	26
4	(0, −0.3, 0, −0.2)	[**2**]223	[**4**]134	[**6**]42	[**8**]8	[**10**]2	66	14	52	48	89
4	(0.1, 0.15, −0.25, −0.35)	[**4**]335	[**3**]235	[**1**]150	[**2**]137	[**7**]68	47	86	85	86	39
5	(0, 0, 0, 0, 0.4)	[**5**]397	[**10**]67	[**2**]48	[**8**]41	[**4**]40	31	77	56	76	25
6	(0.1, 0, 0, 0.1, 0.2, 0.2)	[**5**]277	[**6**]272	[**4**]176	[**1**]165	[**7**]78	62	84	61	66	47
8	(0, 0, 0, 0, 0, 0, 0, −0.5)	[**8**]450	[**2**]55	[**4**]36	[**12**]26	[**6**]21	7	93	30	90	9

$n = 100$, $d(n) = 12$, $\alpha = 0.05$, 10,000 MC runs.

Table 4
Empirical powers of Zhang's test and tests based on N_{A1}, N_{S1} and N_{T1} under the alternative $h(x;\varepsilon,p,q)$

Parameters			The five largest (in absolute value) Fourier coefficients ×1000					Empirical powers (as percentages)				Percentage of cases
ε	p	q						Z_A	N_{A1}	N_{S1}	N_{T1}	$T1 = S1$
1	1.5	1.5	[**2**]280	[**4**]46	[**6**]18	[**8**]8	[**10**]5	75	17	75	72	73
0.25	2.0	10.0	[**1**]289	[**2**]129	[**4**]116	[**5**]95	[**6**]49	63	47	76	73	64
0.25	10.0	20.0	[**3**]225	[**2**]161	[**5**]161	[**1**]143	[**6**]97	38	57	59	58	79
0.50	0.8	1.5	[**1**]263	[**3**]62	[**2**]59	[**5**]36	[**4**]34	66	28	64	61	74
0.50	0.8	0.5	[**2**]217	[**1**]199	[**4**]158	[**3**]144	[**6**]130	79	71	71	71	62
0.60	0.5	0.5	[**2**]334	[**4**]254	[**6**]211	[**8**]184	[**10**]166	89	86	88	88	33
0.20	0.2	0.2	[**12**]300	[**10**]296	[**8**]287	[**6**]281	[**4**]270	98	93	80	85	25
0.10	0.1	0.1	[**12**]273	[**10**]260	[**8**]244	[**6**]227	[**4**]201	97	84	58	68	40

$n = 100$, $d(n) = 12$, $\alpha = 0.05$, 10,000 MC runs.

the present situation. The main idea of this approximation is also given in [18], where an extension to the case in which some nuisance parameters are present is also shown.

From (3), under any alternative P we have $P(T1 = A1) \to 1$ as $n \to \infty$. Therefore, the question of the consistency of N_{T1} is equivalent to establishing the consistency of N_{A1}. This problem was solved in [15], Section 3.

5. Discussion

In this paper we have proposed a method of extending the sensitivity of data driven Neyman tests defined using a (simplified) Schwarz selection rule to determine the number of components. Though, for brevity, we restricted attention to testing uniformity, it is clear that the same idea can be applied to other problems where similar constructions have been proposed or to other new applications. For illustration, let us mention three cases in which score statistics along with (simplified) Schwarz selection rules have been successfully applied. Testing goodness-of-fit when

some Euclidean nuisance parameters are present and a score based-selection rule applied has been treated, e.g., in [19, see statistic (2.7) therein]. A nonparametric two-sample problem was considered in [14, see (14) there]. Semiparametric linear regression was studied in [13, cf. Section 3.4 of that paper]. To make some more precise remarks, let us consider the first problem and testing normality as a particular application. Since nuisance parameters are present, the score statistic is of the following form:

$$W_k(\hat{\mathbf{v}}) = n\mathbf{V}_k(\hat{\mathbf{v}})\{\mathscr{M}_k(\hat{\mathbf{v}})\}\mathbf{V}_k^{\mathrm{T}}(\hat{\mathbf{v}}),$$

where $\mathbf{v} = (EX, \mathrm{Var}\, X)$ is the vector of nuisance parameters, $\hat{\mathbf{v}}$ is an appropriate estimator of $\mathbf{v}$, while $\mathbf{V}_k(\mathbf{v})$ is efficient score vector. $\mathscr{M}_k(\mathbf{v})$ is the inverse of the covariance matrix of $\mathbf{V}_k(\mathbf{v})$. T denotes transposition. For a location-scale family and appropriate estimators of $\mathbf{v}$ the matrix $\mathscr{M}_k(\mathbf{v})$ is independent of $\mathbf{v}$ and equals $\mathscr{M}_k$, say. In particular, this is the case for testing normality when MLEs are used. In such a case we have

$$W_k(\hat{\mathbf{v}}) = n\mathbf{V}_k(\hat{\mathbf{v}})\{\mathscr{M}_k\}\mathbf{V}_k^{\mathrm{T}}(\hat{\mathbf{v}}) = \sum_{j=1}^{k}\{\sqrt{n}U_j(\hat{\mathbf{v}})\}^2,$$

where $U_j(\hat{\mathbf{v}})$ is the jth component of $V_k(\hat{\mathbf{v}})\mathscr{M}_k^{1/2}$. Under the null model $\sqrt{n}U_1(\hat{\mathbf{v}}), \ldots, \sqrt{n}U_k(\hat{\mathbf{v}})$ are asymptotically i.i.d. $N(0, 1)$. The data driven score statistic is of the form $W_{S1(\hat{\mathbf{v}})}(\hat{\mathbf{v}})$, where $S1(\hat{\mathbf{v}})$ is defined as $S1$ in Section 2 with $W_k(\hat{\mathbf{v}})$ in place of N_k; cf. also (2.6) in [19]. Therefore, it is clear that $\sqrt{n}U_j(\hat{\mathbf{v}})$ can be used instead of $\sqrt{n}\hat{b}_j$ in (2), to obtain a refined version of $W_{S1(\hat{\mathbf{v}})}(\hat{\mathbf{v}})$.

To get some intuition regarding the behaviour of $U_j(\hat{\mathbf{v}})$, we inspected their (averaged over 10,000 MC runs) values in several cases taken from the extensive simulation study of Kallenberg and Ledwina [19]. This study covers, among others, alternatives from [22]. In all the cases we studied, except $TU(0.7)$, we observed that the averaged $U_j(\hat{\mathbf{v}})$'s follow the same pattern: the component $j = 2$ or $j = 3$ is dominant. For $TU(0.7)$ the dominant components are: $j = 6, 4, 8$ (in order of their magnitude). However, the empirical power when $n = 100$ and $\alpha = 0.05$ is 0.90 in this case. So, there is no room for much improvement for samples of that size or larger. $TU(0.7)$ is an example of a distribution with a heavier tail than the normal distribution. Higher order components are also dominant in cases of multimodal normal mixtures, for example. So, this class of departures leaves room for some improvement.

Anyway, it is clear that the construction applies to more complex situations as well, where some less regular disturbances may be expected. For example, Fan [6] constructed a two-sample test focused on detecting local characters such as sharp peaks. Guerre and Lavergne [7] proposed a test for regression, which is powerful in detecting highly oscillating alternatives. The refined method proposed in this paper allows us to construct more universal solutions in such cases. Such an application has been recently successfully introduced in [13]. For a practical comparison of Fan's test to related data driven tests with a score-based selection rule incorporated, see [14].

Acknowledgments

We are grateful to a reviewer for helpful comments. The programming work was done by A. Janic-Wróblewska. Her kind co-operation is gratefully acknowledged. The research was supported by Polish State Committee of Scientific Research Grant 5 P03A 03020.

Appendix A

Consider the event

$$A_n(c) = \left\{ \max_{1 \leqslant j \leqslant d(n)} |\sqrt{n}\hat{b}_j| > \sqrt{c \log n} \right\}$$

and assume throughout that $c \geqslant 2$ and $d(n) \log^3 n/n \to 0$ as $n \to \infty$. Since $|b_j(x)| \leqslant \sqrt{2j+1}$, $x \in [0,1]$, from Bernstein's inequality (cf. [26, p. 855]) applied with $\vartheta_j = 1$, $K = \sqrt{2j+1}$ and $\lambda = \sqrt{c \log n}$, we get

$$\begin{aligned} P_0(A_n(c)) &\leqslant \sum_{j=1}^{d(n)} P_0\big(|\sqrt{n}\hat{b}_j| \geqslant \sqrt{c \log n}\big) \\ &\leqslant 2 \sum_{j=1}^{d(n)} \exp\left[-\frac{1}{2} \cdot \frac{c \log n}{\{1 + (1/3)\sqrt{(2j+1)(c \log n)/n}\}} \right]. \end{aligned}$$

Hence, by the assumption $d(n) \log^3 n/n \to 0$, we obtain for large n and some positive constant C

$$P_0(A_n(c)) \leqslant \{C\}\{d(n)\}/\{n^{1+\epsilon}\}, \quad \epsilon = \left(\frac{c}{2} - 1\right). \tag{A.1}$$

(A.1) implies that $\lim_{n\to\infty} P_0(T1 = S1) = 1$.

Assuming additionally that $c > 2$, $d(n) = \mathrm{O}(n^\gamma)$, $\gamma = \gamma(c) < \epsilon$, we get $\sum_{n=1}^{\infty} P_0(A_n(c)) < \infty$. Therefore, in this case the Borel–Cantelli lemma yields

$$P_0\left(\max_{1 \leqslant j \leqslant d(n)} |\sqrt{n}\hat{b}_j| > \sqrt{c \log n} \text{ infinitely often} \right) = 0. \tag{A.2}$$

This shows that when $c > 2$, $I_n(c) = 0$ is highly unprobable under H_0 for large n. When $c \gg 2$, (A.2) implies a lot of room for some shift in the means of $\hat{b}_j$'s. Therefore, the result $I_n(c) = 0$ indicates that observations come from some distribution significantly different from P_0.

References

[1] M. Aerts, G. Claeskens, J.D. Hart, Testing lack of fit in multiple regression, Biomertika 87 (2000) 405–424.
[2] H. Akaike, A new look at statistical model identification, IEEE Trans. Automat. Control 19 (1974) 716–723.
[3] P.J. Bickel, Y. Ritov, Testing for goodness of fit: a new approach, in: A. Saleh (Ed.), Nonparametric Statistics and Related Topics, Amsterdam, North-Holland, 1992, pp. 51–57.
[4] M. Bogdan, Data driven versions of Pearson's chi-square test for uniformity, J. Statist. Comput. Simulation 52 (1995) 217–237.
[5] D.L. Donoho, I.M. Johnstone, Ideal spatial adaptation by wavelet shrinkage, Biometrika 81 (1994) 425–455.
[6] J. Fan, Test of significance based on wavelet thresholding and Neyman's truncation, J. Amer. Statist. Assoc. 91 (1996) 674–688.
[7] E. Guerre, P. Lavergne, Data-driven rate optimal specification testing in regression models, Ann. Statist. 33 (2005) 840–870.
[8] E.J. Hannan, B.G. Quinn, The determination of the order of autoregression, J. Roy. Statist. Soc., Ser. B 41 (1979) 190–195.
[9] D.M.A. Haughton, On the choice of a model to fit data from an exponential family, Ann. Statist. 16 (1988) 342–355.
[10] T. Inglot, Generalized intermediate efficiency of goodness-of-fit tests, Math. Methods Statist. 8 (1999) 487–509.
[11] T. Inglot, A. Janic-Wróblewska, Data driven chi-square test for uniformity with unequal cells, J. Statist. Comput. Simulation 73 (2003) 545–561.

[12] T. Inglot, T. Ledwina, Intermediate approach to comparison of some goodness-of-fit tests, Ann. Inst. Statist. Math. 53 (2001) 810–834.
[13] T. Inglot, T. Ledwina, Data driven score tests of fit for semiparametric homoscedstic linear regression model, submitted for publication.
[14] A. Janic-Wróblewska, T. Ledwina, Data driven-rank test for two-sample problem, Scand. J. Statist. 27 (2000) 281–297.
[15] W.C.M. Kallenberg, The penalty in data driven Neyman's tests, Math. Methods Statist. 11 (2002) 323–340.
[16] W.C.M. Kallenberg, T. Ledwina, Consistency and Monte Carlo simulation of a data driven version of smooth goodness-of-fit tests, Ann. Statist. 23 (1995) 1594–1608.
[17] W.C.M. Kallenberg, T. Ledwina, On data driven Neyman's tests, Probab. Math. Statist. 15 (1995) 409–426.
[18] W.C.M. Kallenberg, T. Ledwina, Data-driven smooth tests when the hypothesis is composite, J. Amer. Statist. Assoc. 92 (1997) 1094–1104.
[19] W.C.M. Kallenberg, T. Ledwina, Data driven smooth tests for composite hypotheses: Comparison of powers, J. Statist. Comput. Simulation 59 (1997) 101–121.
[20] T. Ledwina, Data-driven version of Neyman's smooth test of fit, J. Amer. Statist. Assoc. 89 (1994) 1000–1005.
[21] J. Neyman, 'Smooth test' for goodness of fit, Skand. Aktuarietidskr. 20 (1937) 159–199.
[22] E.S. Pearson, R.B. D'Agostino, K.O. Bowman, Tests for departures from normality: comparison of powers, Biometrika 64 (1977) 231–246.
[23] G. Sansone, Orthogonal Functions, Interscience, New York, 1959.
[24] G. Schwarz, Estimating the dimension of a model, Ann. Statist. 6 (1978) 461–464.
[25] X. Shen, J. Ye, Adaptive model selection, J. Amer. Statist. Assoc. 97 (2002) 210–221.
[26] G. Shorack, J.A. Wellner, Empirical Processes with Application to Statistics, Wiley, New York, 1986.
[27] M. Woodroofe, On model selection and the arc sin laws, Ann. Statist. 10 (1982) 1182–1194.
[28] J. Zhang, Powerful goodness-of-fit tests based on likelihood ratio, J. Roy. Statist. Soc. Ser. B 64 (2002) 281–294.