J. Statist. Comput. Simul., 1997, Vol. 59, pp. 101–121
Reprints available directly from the publisher
Photocopying permitted by license only

DATA DRIVEN SMOOTH TESTS FOR COMPOSITE HYPOTHESES: COMPARISON OF POWERS

WILBERT C. M. KALLENBERG* and TERESA LEDWINA

University of Twente, Technical University of Wrocław

(Received 29 November 1995; In final form 12 May 1997)

Rao's score statistic is a standard tool for constructing statistical tests. If departures from the null model are described by some k-dimensional exponential family the resulting score test is called also smooth test or Neyman's smooth test with k components. An important practical question in applying a smooth test in the goodness-of-fit problem is how large k should be taken. Since a wrong choice may give a considerable loss of power, it is important to make a careful selection. Renewed research in this area shows that the simple question has no simple *deterministic* answer. Therefore, Ledwina introduced, for testing a simple goodness-of-fit hypothesis, a data driven version of Neyman's smooth test. First, Schwarz's rule is applied to find a suitable dimension and then the smooth test statistic in the "right" dimension finishes the job. Simulation results and some theoretical considerations show that this data driven version of smooth tests performs well for a wide range of alternatives, and is competitive with other recently introduced (data driven) procedures. This data-dependent choice of the number of components is extended in this paper to testing the goodness-of-fit problem with composite null hypothesis, being of more practical interest.

A k-dimensional exponential family for modelling departures from the null hypothesis is given and the related Rao's score test is described. A suitable version of Schwarz's rule is proposed and some simplifications of it are discussed.

To check validity of the proposed construction, the method is applied to testing exponentiality and normality. In the extensive simulation study, presented in this paper, it turns out that the data driven version of smooth tests compares well for a wide range of alternatives with other, more specialized, commonly used tests.

Keywords: Goodness-of-fit; smooth test; Neyman's test; Rao's score test; Schwarz's rule; data driven procedure; Monte Carlo study

*Corresponding author: Faculty of Applied Mathematics, University of Twente, P. O. Box 217, 7500 AE Enschede, The Netherlands.

By taking orthonormal versions $\phi_0, \phi_1, \ldots, \phi_k$ of $\phi_0, \tilde{\phi}_1, \ldots, \tilde{\phi}_k$ with

$$\tilde{\phi}_j(x) = \{1_{(d_{j-1},d_j]}(x) - (d_j - d_{j-1})\}(d_j - d_{j-1})^{-1/2}$$

we get Pearson's chi-square test with partition $0 = d_0 < d_1 < \cdots < d_k = 1$. (By $1_A(x)$ the indicator function of the set A is denoted.)

However, changing k in the last example leads to another orthonormal system and therefore a slightly more complicated structure than considered in the rest of the paper.

Testing H_0 reduces in the family (1.1) to testing $H: \theta = 0$. When β is known, the so-called smooth test statistics are given by

$$T_k(\beta) = \sum_{j=1}^{k} \left\{ n^{-(1/2)} \sum_{i=1}^{n} \phi_j[F(X_i; \beta)] \right\}^2, \quad k = 1, 2, \ldots \tag{1.2}$$

For the terminology (in particular the term "smooth"), a motivated introduction and for some properties of $T_k(\beta)$ we refer to Rayner and Best (1989). Here we mention only that $T_k(\beta)$ can be seen as Rao's score statistic for the model (1.1), cf. Theorem 4.2.1 in Rayner and Best (1989). It can also be interpreted as the density-based test statistic with k being the smoothing parameter, cf. Eubank and LaRiccia (1992). Although Neyman's (1937) pioneering paper is considered to be ingenious (cf. Le Cam and Lehmann (1974) no.3 p. ix), smooth tests arosed little attention for several years. For an overview of the renewed interest in smooth tests we refer to Rayner and Best (1990), who conclude in reviewing several tests of fit: "don't use those other methods—use a smooth test!", a conclusion also derived in Milbrodt and Strasser (1990) p. 14. The reasons that Neyman's smooth tests based on $T_k(\beta)$ have been somewhat overlooked might be their lack of consistency for large sets of alternatives and the lack of rule for selecting k.

The important practical question of how large the number k of components should be in $T_k(\beta)$ can be tackled by three different approaches. The first one takes k (and the orthonormal system) in such a way that, among the broad range of alternatives, some alternatives of particular interest are represented fairly well using only the first k components of the orthonormal system for testing H_0 with k as small as possible. It forces the user to think on what type of alternatives are of particular interest, which may be seen more as an

advantage than a disadvantage of the method. A criterion of a simple structure to implement the idea is presented and extensively motivated in Inglot *et al.* (1994a).

The second approach is a more data-analytic approach. If the null hypothesis is rejected, the components are used informally to suggest the nature of the departure from the null hypothesis. For more details we refer to Rayner and Best (1989).

Returning to formal testing theory, the third approach concerns an automatical choice of k, based on the data. Numerical results show that a considerable loss of power may occur, when a wrong choice of k is made (see e.g., Inglot *et al.* (1994a), Kallenberg and Ledwina (1995a,1997)). Therefore a good procedure for choosing k based on the data is very welcome. Schwarz's (1978) selection rule provides such a choice. The finishing touch comes from the smooth test statistic in the "right" dimension. This data driven approach has also the extra advantage lying behind the idea of the second approach, since, when rejecting the null hypothesis, automatically a well defined alternative model is provided for the data at hand. There is now a lot of interest in this kind of procedures as is seen in the papers of Bickel and Ritov (1992), Eubank and Hart (1992), Eubank and LaRiccia (1992), Eubank *et al.* (1993), Bowman and Foster (1993), Ledwina (1994), Inglot *et al.* (1994b), Kallenberg and Ledwina (1995a,b), Inglot and Ledwina (1996), Bogdan (1995) and Fan (1996).

When β is known, or, equivalently, for testing uniformity, the simulation results and the theoretical support for the data driven version of smooth goodness-of-fit tests show that the test has a stable and relatively high power for a broad range of alternatives (cf. Ledwina (1994), Kallenberg and Ledwina (1995a,b), Inglot and Ledwina (1996)). Moreover, in Kallenberg and Ledwina (1995a) consistency of data driven smooth tests is proved under essentially all alternatives.

In applications testing a composite null hypothesis is far more important. Smooth tests with fixed k when nuisance parameters are present are extensively discussed in Rayner and Best (1989).

In view of the good performance of the data driven smooth test in the simple hypothesis case, it is natural to define and investigate a data driven version of the smooth test for composite hypotheses. This is the topic of the present paper.

First of all, Schwarz's rule is extended to the situation where β is unknown. Also some modifications of Schwarz's rule are presented.

Secondly, the score statistic for testing the (composite) hypothesis $H: \theta = 0$ against $A: \theta \neq 0$ in the model (1.1) is applied, using the dimension given by Schwarz's rule. Simulation results for testing normality and for testing exponentiality show that the test works well for a wide range of alternatives and is competitive to well-known tests as Shapiro-Wilk's test in case of normality and Gini's test for exponentiality.

The paper is organized as follows. In Section 2 the test statistics are formally defined. Section 3 provides the simulated critical values and powers of the data driven smooth tests.

2. TEST STATISTICS

To define the test statistics consider an orthonormal system

$$\phi_0, \phi_1, \phi_2, \ldots$$

in $L_2([0,1])$ with bounded functions $\phi_1, \phi_2, \ldots$ and $\phi_0(x) \equiv 1$. The functions $\phi_1, \phi_2, \ldots$ are not necessarily uniformly bounded. Let $\{f(x;\beta): \beta \in B\}$ be a given set of densities with corresponding distribution functions $\{F(x;\beta): \beta \in B\}$, where $B \subset \mathbb{R}^q$. For $k = 1, 2, \ldots$ define exponential families by their density

$$g(x;\theta,\beta) = \exp\{\theta \cdot \phi[F(x;\beta)] - \psi_k(\theta)\} f(x;\beta), \tag{2.1}$$

where

$$\theta = (\theta_1, \ldots, \theta_k),\ \phi = (\phi_1, \ldots, \phi_k),\ \psi_k(\theta) = \log \int_0^1 \exp\{\theta \cdot \phi(y)\} dy$$

and $\cdot$ stands for the inner product in $\mathbb{R}^k$. When there is no confusion, the dimension k is sometimes suppressed in the notation. Let $'$ denote the transpose of a matrix or vector. Writing

$$Y_n(\beta) = (\bar{\phi}_1(\beta), \ldots, \bar{\phi}_j(\beta))' = n^{-1} \sum_{i=1}^{n} (\phi_1[F(X_i;\beta)], \ldots, \phi_j[F(X_i;\beta)])'$$

with j depending on the context, the likelihood of the independent r.v.'s $X_1, \ldots, X_n$, each having density (2.1) is given by

$$\exp\{n[Y_n(\beta) \cdot \theta - \psi_k(\theta)]\} \prod_{i=1}^{n} f(X_i; \beta).$$

For each $\beta \in B$ Schwarz's (1978) rule for choosing submodels corresponding to successive dimensions yields

$$S(\beta) = \min\{k : 1 \leq k \leq d(n), L_k(\beta) \geq L_j(\beta),\ j = 1, \ldots, d(n)\}, \quad (2.2)$$

where

$$L_k(\beta) = n \sup\{\theta \cdot Y_n(\beta) - \psi_k(\theta) : \theta \in \mathbb{R}^k\} - \frac{1}{2} k \log n.$$

Although it is not mentioned in the notation, $S(\beta)$ depends of course on the upper bound $d(n)$ of the exponential families under consideration.

Let $\hat{\beta}$ be the maximum likelihood estimator of β under H_0. Define

$$S = S(\hat{\beta}). \quad (2.3)$$

Properties of this extension of Schwarz's rule to the situation where nuisance parameters are estimated, are given in Inglot *et al.* (1994b).

When β is known (and k is fixed), the test statistic (1.2) is the score statistic for the model (1.1), cf. Theorem 4.2.1 in Rayner and Best (1989). Therefore, when β is unknown (and k is fixed) we also use the score statistic for testing the (now composite) hypothesis $H: \theta = 0$ against $A: \theta \neq 0$ (cf. Javitz (1975), Kopecky and Pierce (1979), Thomas and Pierce (1979), Neyman (1980), Rayner and Best (1989)).

It is assumed in the rest of the paper that the set of densities

$$\{f(x; \beta) : \beta \in B\} \quad \text{is regular}$$

in the usual Cramér sense.

Denote by I the $k \times k$ identity matrix and by $E_{(0,\beta)}$ the expectation when X has density $f(x; \beta)$. Further define

$$I_\beta = \left\{ - E_{(0,\beta)} \frac{\partial}{\partial \beta_t} \phi_j[F(X; \beta)] \right\}_{t=1,\ldots,q,\ j=1,\ldots,k}$$

$$I_{\beta\beta} = \left\{ -E_{(0,\beta)} \frac{\partial^2}{\partial\beta_t \partial\beta_u} \log f(X;\beta) \right\}_{t=1,\ldots,q,\ u=1,\ldots,q}$$

$$R(\beta) = I'_\beta (I_{\beta\beta} - I_\beta I'_\beta)^{-1} I_\beta$$

$$W_k = n Y'_n(\hat{\beta})\{I + R(\hat{\beta})\} Y_n(\hat{\beta}). \tag{2.4}$$

The data driven smooth test statistic is now defined by

$$W_S = W_{S(\hat{\beta})} \tag{2.5}$$

with W_k given in (2.4), $S(\beta)$ given in (2.2) and $\hat{\beta}$ the maximum likelihood estimator of β under $H: \theta = 0$. The null hypothesis is rejected for large values of W_S.

Schwarz's rule $S(\beta)$ as given in (2.2) compares (penalized) maximized likelihoods. It turns out (cf. Inglot *et al.* (1994b)) that the maximized likelihood (which is in fact the likelihood ratio statistic for testing $H : \theta = 0$ against $A : \theta \neq 0$ when β is known) is locally equivalent to $(1/2)n\| Y_n(\beta)\|^2$, where $\| \cdot \|$denotes the Euclidean norm.

Inserting the estimator $\hat{\beta}$, the extra term $R(\beta)$ appears in the inverse of the asymptotic covariance matrix. Taking this into account we get the modification

$$\begin{aligned} S1 = S1(\hat{\beta}) = &\min\{k : 1 \leq k \leq d(n), \\ &W_k - k \log n \geq W_j - j \log n,\ j = 1, \ldots, d(n)\}, \end{aligned} \tag{2.6}$$

which is easier to calculate. The corresponding test statistic is

$$W_{S1} = W_{S1(\hat{\beta})}. \tag{2.7}$$

An even more simple modification, which is still easier to calculate, is

$$\begin{aligned} S2 = S2(\hat{\beta}) = \min\{&k : 1 \leq k \leq d(n), n\| Y_n(\hat{\beta})\|^2_{(k)} \\ &- k \log n \geq n\| Y_n(\hat{\beta})\|^2_{(j)} - j \log n, \quad j = 1, \ldots, d(n)\}, \end{aligned} \tag{2.8}$$

where the index of the norm denotes the dimension. Here the maximized likelihood is replaced by its locally equivalent form without

an "adjustment" for inserting the estimator. Note that in $S(\hat{\beta})$ there is also no "adjustment" for inserting the estimator. In this sense $S(\hat{\beta})$ and $S2(\hat{\beta})$are similar.

The corresponding test statistic is

$$W_{S2} = W_{S2(\hat{\beta})}. \tag{2.9}$$

In case of a location-scale family $\{f(x;\beta) : \beta \in B\}$ we write $\beta = (\mu\ \sigma)$, $f(x;\beta) = \sigma^{-1} f((x-\mu)/\sigma)$ and $F(x;\beta) = F((x-\mu)/\sigma)$. Now $R(\beta)$ defined in (2.4) does not depend on β. The statistics $S(\hat{\beta}), S1(\hat{\beta}), S2(\hat{\beta})$, W_S, W_{S1} and W_{S2} all depend on $X_1, \ldots, X_n$ by means of

$$\frac{X_i - \hat{\mu}}{\hat{\sigma}}, \quad i = 1, \ldots, n,$$

where $(\hat{\mu}, \hat{\sigma}) = \hat{\beta}$. Since $(\hat{\mu}, \hat{\sigma})$ is location-scale equivariant, the distribution of

$$\left(\frac{X_1 - \hat{\mu}}{\hat{\sigma}}, \ldots, \frac{X_n - \hat{\mu}}{\hat{\sigma}}\right)$$

does not depend on the location-scale parameter if X_i comes from a location-scale family.

Therefore in case of a location-scale family $\{f(x;\beta) : \beta \in B\}$ the null distributions of $S(\hat{\beta}), S1(\hat{\beta}), S2(\hat{\beta})$, W_S, W_{S1} and W_{S2} do not depend on β. Moreover, if the alternative also belongs to a location-scale family, the distributions of $S(\hat{\beta}), S1(\hat{\beta}), S2(\hat{\beta})$, W_S, W_{S1} and W_{S2} do not depend on the location-scale parameter of that family.

The same remark applies to location families and to scale families. To facilitate the application of the test statistics we summarize how to calculate them.

1. Calculate the maximum likelihood estimator $\hat{\beta}$ of β under H_0.
2. Choose $d(n)$. (We recommend $d(n) = 6$, 7 and 10 for $n = 20$, 30 and 50, respectively).
3. Compute $Y_n(\hat{\beta})$ for dimension $j = 1, \ldots, d(n)$.
4. Compute W_j for $j = 1, \ldots, d(n)$ in case of W_{S1} and $L_j(\hat{\beta})$ for $j = 1, \ldots, d(n)$ in case of W_S (for computing the maximum likelihood estimator $\hat{\theta}$ see Ledwina (1994, Section 3.1)).
5. Calculate $S(\hat{\beta}), S1(\hat{\beta})$ and $S2(\hat{\beta})$ and compute W_S, W_{S1} and W_{S2}, respectively.

Under some regularity conditions all three test statistics have a chi-square distribution with 1 degree of freedom as asymptotic null distribution. Accurate critical values or p-values can either be obtained by simulation (cf. also Section 3) or by the second order null approximation given in Kallenberg and Ledwina (1997).

Finally we mention that the tests are consistent against essentially all alternatives. For details we refer to Inglot *et al.* (1994b).

3. SIMULATED CRITICAL VALUES AND SIMULATED POWERS

All programs used in this paper have been written by Krzysztof Bogdan under the MEN Grant 341 046 and KBN Grant 665/2/91.

The simulations are performed in a simular way as described in Ledwina (1994). Attention here is focussed on testing exponentiality and normality.

We start this section with presenting simulated critical values of W_S, W_{S1} and W_{S2} for testing exponentiality with significance level $\alpha = 0.05$ and $\alpha = 0.10$. In the simulation study as orthonormal system we take the orthonormal Legendre polynomials on [0, 1]

It is seen from Table I that the critical values do not vary much for different values of $d(n)$ in the range 2(3) to 12. Although the introduction of W_S, $W_{S1}$and W_{S2} suggests much similarity, the differences between the corresponding critical values are not quite neglectable for sample sizes $n \leq 50$. On the other hand the critical values of W_S and W_{S2} come closer to each other when n becomes larger. The selection rules concentrate on dimension 1 under H_0, cf. Inglot *et al.* (1994b) for an extensive discussion. Therefore one might expect that the critical values were close to the chi-square-one α-points, being 3.841 for $\alpha = 0.05$ and 2.706 for $\alpha = 0.10$. However, the simulated critical values are substantially different. The same phenomenon occurs in the simple hypothesis case. An accurate approximation when testing a simple hypothesis is given in Kallenberg and Ledwina (1995b). A similar approach for the composite null hypothesis is given in Kallenberg and Ledwina (1997).

We proceed with presenting simulated critical values of W_S, W_{S1} and W_{S2} for testing normality with significance level $\alpha = 0.05$ and $\alpha = 0.10$ in Table II.

TABLE I 5% and 10% critical values of W_S, W_{S1} and W_{S2} for testing exponentiality; each case is based on 10000 samples

n	α		$d(n)$											
			1	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*	*11*	*12*
20	0.05	W_S	3.626	4.903	6.029	6.697	7.117	7.326	7.491	–	–	–	–	–
		W_{S1}	3.626	5.137	6.029	6.508	6.732	6.835	6.863	6.896	6.912	6.984	7.009	7.010
		W_{S2}	3.626	4.919	5.925	6.381	6.633	6.766	6.807	6.839	6.874	6.886	6.896	6.914
	0.10	W_S	2.587	3.338	3.924	4.238	4.445	4.589	4.663	–	–	–	–	–
		W_{S1}	2.587	3.798	4.109	4.247	4.305	4.329	4.346	4.351	4.358	4.372	4.381	4.385
		W_{S2}	2.587	3.273	3.751	3.944	4.006	4.038	4.066	4.075	4.088	4.094	4.104	4.104
30	0.05	W_S	3.646	4.850	5.570	5.986	6.128	6.239	6.289	6.322	–	–	–	–
		W_{S1}	3.646	5.224	5.772	6.137	6.201	6.245	6.248	6.248	6.248	6.289	6.289	6.289
		W_{S2}	3.646	4.850	5.376	5.593	5.628	5.690	5.690	5.690	5.690	5.696	5.696	5.696
	0.10	W_S	2.637	3.146	3.451	3.605	3.664	3.722	3.733	3.747	–	–	–	–
		W_{S1}	2.637	3.719	3.933	4.008	4.052	4.067	4.067	4.067	4.067	4.075	4.075	4.075
		W_{S2}	2.637	3.099	3.316	3.423	3.435	3.443	3.443	3.433	3.443	3.445	3.445	3.445
50	0.05	W_S	3.592	4.445	4.790	4.960	5.020	5.063	5.086	5.086	5.086	5.086	5.086	–
		W_{S1}	3.592	5.003	5.438	5.604	5.646	5.678	5.683	5.684	5.684	5.687	5.687	5.691
		W_{S2}	3.592	4.492	4.800	4.891	4.935	4.947	4.947	4.947	4.947	4.947	4.947	4.947
	0.10	W_S	2.562	2.783	2.886	2.937	2.955	2.959	2.967	2.967	2.967	2.967	2.967	–
		W_{S1}	2.562	3.335	3.549	3.609	3.619	3.640	3.642	3.651	3.651	3.652	3.652	3.652
		W_{S2}	2.562	2.781	2.864	2.905	2.921	2.937	2.937	2.937	2.937	2.937	2.937	2.937

TABLE II 5% and 10% critical values of W_S, W_{S1} and W_{S2} for testing normality; each case is based on 10000 samples

n	α		$d(n)$											
			1	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*	*11*	*12*
20	0.05	W_S	3.697	3.701	3.784	4.044	4.275	4.379	4.459	–	–	–	–	–
		W_{S1}	3.697	4.732	5.314	5.558	5.645	5.657	5.657	5.657	5.657	5.657	5.657	5.657
		W_{S2}	3.697	3.700	3.777	3.926	3.954	3.962	3.970	3.973	3.973	3.973	3.973	3.973
	0.10	W_S	2.669	2.669	2.679	2.785	2.827	2.889	2.904	–	–	–	–	–
		W_{S1}	2.669	3.609	3.794	3.864	3.877	3.892	3.892	3.892	3.892	3.892	3.892	3.892
		W_{S2}	2.669	2.669	2.676	2.721	2.741	2.745	2.746	2.750	2.750	2.750	2.750	2.750
50	0.05	W_S	3.817	3.820	3.825	3.847	3.854	3.854	3.854	3.855	3.855	3.855	3.855	–
		W_{S1}	3.817	5.034	5.471	5.649	5.703	5.742	5.752	5.752	5.752	5.752	5.752	5.752
		W_{S2}	3.817	3.820	3.825	3.836	3.840	3.840	3.840	3.840	3.840	3.840	3.840	3.840
	0.10	W_S	2.725	2.726	2.728	2.745	2.750	2.757	2.757	2.757	2.757	2.757	2.757	–
		W_{S1}	2.725	3.490	3.709	3.774	3.795	3.804	3.811	3.811	3.811	3.811	3.811	3.811
		W_{S2}	2.725	2.726	2.728	2.735	2.734	2.734	2.734	2.734	2.734	2.734	2.734	2.734

There is no (much) variation in the critical values of W_S and W_{S2}, but there is a difference between the critical values of W_{S1} and those of W_S, W_{S2}. Moreover, the critical values of W_S and W_{S2} are for $n=50$ much closer to the corresponding α-points of the chi-square-one distribution. For an explanation we refer to Kallenberg and Ledwina (1997). The critical values of W_{S1} do not change much in the range $d(n)=3$ to 12.

To see how well the tests, introduced in this paper, perform we present the results of an extensive Monte Carlo study of the power. The null hypothesis of exponentiality corresponds to

$$f(x;\beta)=\begin{cases}0, & x<0\\ \beta^{-1}\exp(-\beta^{-1}x), & x\geq 0\end{cases}$$

and $\hat{\beta}=n^{-1}\sum_{i=1}^{n}X_i$. For power comparison when testing exponentiality we consider the Gini statistic

$$G=\left[\sum_{i=1}^{n-1}\{i(n-i)(X_{(i+1)}-X_{(i)})\}\right]\Big/\left\{(n-1)\sum_{i=1}^{n}X_i\right\},$$

where $X_{(1)}\leq\cdots\leq X_{(n)}$ are the order statistics, cf. formula (2.1) on p. 351 of Gail and Gastwirth (1978). This test is called "powerful against a variety of alternatives" by Gail and Gastwirth (1978) and turned out to perform well in the study of Ascher (1990). It is also used for the sake of comparison by Rayner and Best (1989, p. 88) and LaRiccia (1991).

In the simulation study we consider the following broad class of alternatives (cf. Chambers *et al.* (1976), Pearson *et al.* (1977), Gail and Gastwirth (1978), Angus (1982), Baringhaus *et al.* (1989), Ascher (1990), Gan and Koehler (1990), Ebrahimi *et al.* (1992), Baringhaus and Henze (1992)). Here U denotes a $N(0;1)$ r.v., R denotes a uniform r.v., on (0,1) and Z a standard exponential distribution; R and Z are independent.

alternative	*density/definition*
Weibull $(b;k)$	$bk(bx)^{k-1}\exp\{-(bx)^k\}, \quad x>0$
Gamma $(p;q)$	$q^{-p}\{\Gamma(p)\}^{-1}x^{p-1}\exp(-x/q), \quad x>0$
$\chi^2_k = \text{Gamma}(\frac{1}{2}k;2)$	$\{2^{\frac{1}{2}k}\Gamma(k/2)\}^{-1}x^{\frac{1}{2}k-1}\exp(-\frac{1}{2}x), \quad x>0$
$LN(g;d)$	$d(x\sqrt{2\pi})^{-1}\exp\{-\frac{1}{2}(d\log x+g)^2\}, \quad x>0$
Beta $(p;q)$	$x^{p-1}(1-x)^{q-1}\{B(p,q)\}^{-1}, \quad 0\le x\le 1$
Uniform $(a;b)$	$(b-a)^{-1}, \quad a\le x\le b$
Shifted exp. $(l;b)$	$b\exp\{-(x-l)b\}, \quad x\ge l$
Pareto $(a;k)$	$ak^a x^{-a-1}, \quad x\ge k$
Shifted Pareto	$2(1+x)^{-3}, \quad x>0$
$SU(g;d)$	$U=g+d\sinh^{-1}(X), \quad -\infty<X<\infty$
$TU(l)$	$X=R^l-(1-R)^l, \quad -1\le X\le 1$
Logistic	$e^x(1+e^x)^{-2}, \quad -\infty<x<\infty$
$SC(p;d)$	$(2\pi)^{-\frac{1}{2}}[(p/d)\exp(-\frac{1}{2}x^2/d^2)+(1-p)\exp(-\frac{1}{2}x^2)], \quad -\infty<x<\infty$
$LC(p;m)$	$(2\pi)^{-\frac{1}{2}}[p\exp\{-\frac{1}{2}(x-m)^2\}+(1-p)\exp(-\frac{1}{2}x^2)], \quad -\infty<x<\infty$
$SB(g;d)$	$U=g+d\log\{X/(1-X)\}, \quad 0<X<1$
$S(a,b)$	$X=(X1/X2)(X3)$ with
	$X1=\sin\{a(R\pi-\frac{1}{2}\pi)+\frac{1}{2}\pi b(2-a)\}$
	$X2=\{\cos(R\pi-\frac{1}{2}\pi)\}^{\frac{1}{a}}$
	$X3=[Z^{-1}\cos\{(R\pi-\frac{1}{2}\pi)(1-a)-\frac{1}{2}\pi b(2-a)\}]^{\frac{1-a}{a}}$
	$a>1, 0\le b\le 1, -\infty<X<\infty$

Note that the Weibull alternative is a scale family w.r.t. b, the Gamma w.r.t. q, the lognormal LN w.r.t. $\exp(-g/d)$, the Uniform $(0,b)$ w.r.t. b, the Shifted exponential w.r.t. b^{-1}, the Pareto w.r.t. k.

The following tables show the powers for testing exponentiality. It is indicated, where in literature the same alternatives occur, but note that several alternatives are used in more than one paper. In the referred papers one may find also simulated power for other tests for those alternatives. Many authors present simulation results for $n=20$ and $\alpha=0.05$. Although in our opinion this is an extreme situation when testing goodness-of-fit, for the sake of comparison Table III contains also $n=20$ and $\alpha=0.05$. Other authors take $n=20$ and $\alpha=0.10$. Table IV contains this case. The more realistic choice $n=50$ and $\alpha=0.05$ is presented also in Table III, while Table IV shows $n=50$ and $\alpha=0.10$. In this way the changes are seen, when n is growing.

The tests W_S and W_{S2} perform well and, although based on general ideas, they can compare even with 'special' tests for exponentiality,

TABLE III Estimated powers (in%) of G, W_S, W_{S1} and W_{S2} testing exponentiality; $\alpha = 0.05$, $n = 20$, 50; each case is based on 10000 samples; $d(20) = 6$ for S and 12 for $S1$, $S2$; $d(50) = 10$ for S and 12 for $S1$, $S2$

alt	G		W_S		W_{S1}		W_{S2}	
	$n=20$	$n=50$	$n=20$	$n=50$	$n=20$	$n=50$	$n=20$	$n=50$
Gail-Gastwirth (1978) and Agnus (1982)								
Weibull(1;0.8)	24	48	19	47	23	44	22	50
Weibull(1;1.5)	50	93	23	89	11	80	23	89
Uniform(0;2)	70	99	63	98	30	95	51	97
Pareto(2;0.5)	79	91	100	100	99	100	100	100
Shifted Pareto	47	81	38	77	42	77	39	77
Shifted exp.(0.2;1)	23	54	18	96	10	67	16	87
χ^2_4	48	90	22	91	11	79	23	91
Ascher(1990)								
Gamma(4;1)	99	100	90	100	78	100	92	100
Beta(2.1)	100	100	100	100	98	100	100	100
Gamma(0.7;1)	21	39	18	45	23	42	23	48
Beta(0.5;1)	6	5	41	74	42	84	47	76
Gan-Koehler (1990)								
Beta(0.5;0.5)	27	59	72	98	56	96	68	95
SB(0;0.5)	51	93	59	94	27	85	44	90
SB(0;0.707)	90	100	74	100	44	100	63	100
Beta(2;2)	99	100	93	100	82	100	91	100
Beta(3;2)	100	100	100	100	100	100	100	100
Weibull(1;4)	100	100	100	100	100	100	100	100
Weibull(1;3.6)	100	100	100	100	100	100	100	100
Weibull(1;2.2)	99	100	86	100	75	100	88	100
Weibull(1;2)	95	100	72	100	57	100	75	100
SB(1;2)	100	100	100	100	100	100	100	100
SB(0.5333;0.5)	4	6	18	32	13	44	16	30
SB(1;1)	76	100	40	100	21	99	39	100
Weibull(1;0.5)	91	100	90	100	92	100	93	100
χ^2_1	55	89	54	94	63	93	65	95
LN(0;1)	12	15	16	46	18	53	15	41
Ebrahim *et al.* (1992)								
Gamma(3;0.333)	89	100	63	100	45	100	67	100
LN(−0.3;0.775)	41	74	34	71	38	74	33	71
LN(−0.2;0.633)	73	98	63	97	66	96	64	97
Baringhaus-Henze (1992)								
Gamma(0.4;1)	76	99	79	100	85	100	86	100
Gamma(0.6;1)	35	68	33	75	41	72	41	78
Gamma(1.5;1)	18	44	8	44	4	29	7	44
Gamma(2.4;1)	69	99	38	99	22	96	40	99
Weibull(1;0.6)	74	98	69	99	73	98	74	99
Weibull(1;1.4)	35	79	15	73	6	58	15	72
Weibull(1;1.6)	63	98	31	96	17	92	32	96

TABLE IV Estimated powers (in%) of G, W_S, W_{S1} and W_{S2} testing exponentiality; $\alpha = 0.10$, $n = 20, 50$; each case is based on 10000 samples; $d(20) = 6$ for S and 12 for $S1$, $S2$; $d(50) = 10$ for S and 12 for $S1$, $S2$

alt	G		W_S		W_{S1}		W_{S2}	
	$n=20$	$n=50$	$n=20$	$n=50$	$n=20$	$n=50$	$n=20$	$n=50$
Agnus (1982)								
χ^2_1	66	93	69	96	72	96	75	97
χ^2_3	29	58	24	64	17	51	25	62
χ^2_4	62	95	53	97	41	93	55	97
$LN(0;0.8)$	45	74	40	75	43	76	39	74
$LN(0;1)$	19	22	29	52	29	62	25	47
$LN(0;1.2)$	29	46	49	89	37	81	45	84
Weibull(1;0.8)	34	59	30	59	33	56	35	60
Weibull(1;1.2)	21	43	17	45	10	34	16	44
Weibull(1;1.5)	64	97	50	97	40	93	52	96
Beta(1;2)	41	81	30	79	23	71	29	78
Uniform(0;2)	82	100	74	100	61	99	68	100
Shifted exp.(0.2;1)	35	68	56	98	31	83	46	91
Shifted exp.(0.2;0.7)	23	45	35	80	20	58	28	66
Pareto(1;0.2)	57	74	99	100	95	100	96	100
Pareto(0.8;0.01)	72	94	97	100	94	100	94	100
Shifted Pareto	56	86	49	85	51	84	50	85

like Gini's test. Also in comparison with other tests for exponentiality in literature, the power of W_S and W_{S2} behaves well. In many cases taken from Gan and Koehler (1990), when power of W_S is 100, classical goodness-of-fit tests have definitely smaller powers (cf. Tab. 4 in Gan and Koehler (1990)). As is seen from Table III and Table IV, for $n = 50$ W_S and W_{S2} often have higher power than Gini's test with great differences in Shifted exp.(0.2;1), Beta(0.5;1), Beta(0.5;0.5), $SB(0.5333;0.5)$, $LN(0;1)$, $LN(0;1.2)$, Shifted exp.(0.2;0.7), Pareto (1;0.2). Data driven smooth tests improve considerably (and much faster than Gini's test) from $n = 20$ to $n = 50$.

Next we consider the null hypothesis of normality, corresponding to (writing $\beta = (\mu, \sigma)$)

$$f(x; \beta) = (\sqrt{2\pi}\sigma)^{-1} \exp\left\{ -\frac{1}{2}(x-\mu)^2/\sigma^2 \right\}$$

and

$$\hat{\beta} = \left(\bar{X}, \left\{ n^{-1} \sum_{i=1}^{n} (X_i - \bar{X})^2 \right\}^{\frac{1}{2}} \right) \text{ with } \bar{X} = n^{-1} \sum_{i=1}^{n} X_i.$$

Here we use the often recommended Shapiro-Wilk test. According to Bowman (1992) the Shapiro-Wilk test in particular sets a very high standard as an omnibus test of normality. The corresponding test statistic is denoted by W. Table V shows the results for a variety of symmetrical and skew alternatives for $n=20$, 50. A more extensive version of Table V is given in Kallenberg and Ledwina (1994).

As in Pearson *et al.* (1977) both symmetrical and skew alternatives in Table V are ordered according to increasing kurtosis. It turns out that except for "picked" symmetrical cases as the first 3 cases, which are close to the null hypothesis (in e.g., Kullback-Leibler distance), W_S performs reasonable in the symmetrical case and works well for skew alternatives. That W_S performs less for symmetrical alternatives is not

TABLE V Estimated powers (in%) of W, W_S, W_{S1} and W_{S2} testing normality; $\alpha=0.05$, $n=20$, 50; each case is based on 10000 samples; $d(20)=6$ for S and 12 for $S1$, $S2$

alt	W		W_S		W_{S1}		W_{S2}	
	$n=20$	$n=50$	$n=20$	$n=50$	$n=20$	$n=50$	$n=20$	$n=50$
Pearson *et al.* (1977) Symmetric alternatives								
SB(0;0.5)	44	99	36	67	34	93	26	55
TU(1.5)	26	92	20	34	19	74	14	26
TU(0.7)	12	62	11	13	9	45	8	9
Logistic(1)	12	13	10	13	13	21	11	12
TU(10)	82	99	82	99	87	100	85	99
SC(0.05;3)	19	31	17	25	19	38	16	24
SC(0.2;5)	71	95	65	92	74	98	65	92
SC(0.05;5)	36	62	33	55	37	66	32	55
SC(0.05;7)	45	74	42	70	46	77	42	70
SU(0;1)	43	68	36	61	47	81	38	61
Skew alternatives								
SB(0.533;0.5)	73	100	63	96	47	95	55	92
SB(1;1)	31	81	29	72	17	57	29	71
LC(0.2;3)	31	60	27	68	19	52	28	69
Weibull(2)	15	41	15	41	10	29	16	41
LC(0.1;3)	25	50	24	58	21	51	26	58
χ^2_{10}	25	57	23	61	18	48	26	62
LC(0.05;3)	18	32	17	33	18	37	17	34
LC(0.1;5)	76	98	72	97	72	98	73	97
SU(−1;2)	22	37	19	43	20	40	21	42
χ^2_4	53	95	51	94	38	86	52	93
LC(0.05;5)	55	85	48	79	54	87	49	78
LC(0.05;7)	65	92	64	91	65	92	63	90
SU(1;1)	73	96	73	98	68	97	73	98
LN(0;1)	94	100	91	100	85	100	92	100

a big surprise in view of Kopecky and Pierce's (1979, p. 397) statement that W_1 provides little or no protection against nearby symmetric alternatives.

As in testing exponentiality W_{S2} is close to W_S, while W_{S1} performs better for symmetric alternatives, but is less stable for skew alternatives. Further comparison with Table 8 of Pearson *et al.* (1977) shows that for most of the skew alternatives considered there, W_S dominates moment-based tests. Extensive simulation study of some new and classical tests (such as Anderson-Darling, Cramér-von Mises, Kolmogorov, Watson) is given in Gan and Koehler (1990). Comparison of Table V with their Table 3 shows that W_S compares well to classical and new tests introduced in this paper both for symmetrical and skew alternatives. A similar conclusion can be drawn from studying Table III and Table IV in Baringhaus *et al.* (1989), where W is compared, among others, to the test based on the empirical characteristic function, introduced by Epps and Pulley (1983).

The next table concerns alternatives from the Johnson-system of distributions (cf. also Baringhaus *et al.* (1989)). Parameters of the Johnson-system are taken from Pearson and Hartley (1972).

It is seen that W, W_S and W_{S2} are close to each other with W_S and W_{S2} slightly better for larger b_2 (kurtosis) and not too large $\sqrt{b_1}$

TABLE VI Estimated powers (in%) of W, W_S, W_{S1} and W_{S2} testing normality; $\alpha = 0.05$, $n = 50$; each case is based on 10000 samples; $d(50) = 10$ for S and 12 for $S1$, $S2$; alternatives of the Johnson-system with varying skewness $\sqrt{b_1}$ and kurtosis b_2 cf. Baringhaus *et al.* (1989)

		$b_2 = 2.6$					$b_2 = 3.0$		
$\sqrt{b_1}$	W	W_S	W_{S1}	W_{S2}	$\sqrt{b_1}$	W	W_S	W_{S1}	W_{S2}
0	6	4	4	3	0	5	5	5	5
0.2	9	7	6	7	0.2	–	–	–	–
0.4	27	22	16	23	0.4	15	18	12	19
0.6	76	59	47	58	0.6	43	44	30	44
0.8	100	95	91	94	0.8	90	82	69	81
1.0	100	100	99	100	1.0	100	99	98	99
		$b_2 = 3.4$					$b_2 = 3.8$		
0	7	7	9	7	0	9	10	15	10
0.2	8	10	10	10	0.2	10	12	15	12
0.4	14	17	14	17	0.4	15	18	18	18
0.6	30	36	24	36	0.6	25	31	23	31
0.8	69	67	52	67	0.8	51	58	42	57
1.0	99	95	89	95	1.0	90	87	77	87

(skewness), while W performs better for $b_2 = 2.6$. If b_2 is large and $\sqrt{b_1}$ small W_{S1} is the better one, but in other cases W_{S1} performs less.

The last table concerns stable alternatives.

Clearly, W, W_S and W_{S2} are competitive, while W_{S1} is slightly better. Power is not varying much with b.

The simulations confirm the existing consensus of opinion that data driven methods are advantageous when the underlying density is heavily skewed or long-tailed. For details see Bowman and Foster (1993). Moreover, the simulations show that the general construction of data driven smooth tests proposed in this paper works well when applied to some standard situations. Especially for moderately sample size ($n = 50$) it gives results comparable to those provided by recognized tests proposed in special situations. Needless to say that hitherto existing efforts of extending e.g., Shapiro-Wilk's statistic to testing exponentiality were ineffective. The solution presented here is based on general likelihood methods and hence it can be extended to a very wide class of problems.

One may ask whether the problem of choosing the number of components k is replaced by the choice of $d(n)$. This is certainly not the case, since in contrast to the power of W_k, the power of W_S, W_{S1} and W_{S2} is stable for larger $d(n)$, cf. Kallenberg and Ledwina (1997).

TABLE VII Estimated powers (in%) of W, W_S, W_{S1} and W_{S2} testing normality; $\alpha = 0.05$, $n = 50$; each case is based on 10000 samples; $d(50) = 10$ for S and 12 for $S1$, $S2$; stable alternatives $S(a, b)$, cf. Baringhaus *et al.* (1989)

		$a = 1.2$					$a = 1.4$		
b	W	W_S	W_{S1}	W_{S2}	b	W	W_S	W_{S1}	W_{S2}
0	96	95	98	94	0	86	82	91	82
0.2	96	95	98	94	0.2	85	83	91	82
0.4	96	95	98	95	0.4	86	83	92	83
0.6	96	95	98	95	0.6	87	86	92	85
0.8	97	97	98	96	0.8	90	90	92	89
1.0	100	100	99	100	1.0	96	97	95	97
		$a = 1.6$					$a = 1.8$		
0	66	61	74	60	0	36	33	43	32
0.2	66	61	74	61	0.2	37	33	43	33
0.4	68	64	75	63	0.4	38	35	44	34
0.6	70	69	76	68	0.6	40	38	44	37
0.8	75	75	78	75	0.8	42	42	45	42
1.0	81	84	81	84	1.0	46	47	47	47

Obviously, taking the original Schwarz's rule to choose the number of components k is not the only possible solution. Other consistent variants of Schwarz's solution as those proposed by Hannan and Quinn (1979), Nishii (1984), Bozdogan (1987), Haughton (1988) or Haughton *et al.* (1990) can be taken into account. Our experience is that by taking a heavier penalty than in Schwarz's rule the power will be larger for smooth alternatives and smaller for highly oscillating ones. For a lighter penalty the situation is reversed. This is demonstrated in Kallenberg and Ledwina (1997). Based on simulations we have performed, we find that Schwarz's rule is a nice compromise.

We end with application of the method to a real problem. Consider the Mississippi River Data, presented on p. 16 of Rayner and Best (1989). Taking $d(50)=10$ the dimension chosen by Schwarz's rule equals 1. The value of W_S for these data is 3.356 with corresponding p-value 0.06. The doubt on normality agrees well with the discussion on p. 18 in Rayner and Best (1989).

Acknowledgement

The research of Teresa Ledwina was supported by Grant KBN 350 044. The authors thank the referee for careful reading of the manuscript and for comments.

References

Angus, J. E. (1982) Goodness-of-fit tests for exponentiality based on a loss-of-memory type functional equation. *J. Statist. Plann. Inference*, **6**, 241–251.

Ascher, S. (1990) A survey of tests for exponentiality. *Comm. Statist. Theory Methods*, **19**, 1811–1825.

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

Baringhaus, L. and Henze, N. (1992) An adaptive omnibus test for exponentiality *Comm. Statist. Theory Methods*, **21**, 969–978.

Bickel, P. J. and Ritov, Y. (1992) Testing for goodness of fit: A new approach. In: *Nonparametric Statistics and Related Topics*. A. K. Md. E. Saleh (ed.). North-Holland, Amsterdam.

Bogdan, M. (1995) Data driven versions of Pearson's chi-square test for uniformity. *J. Statist. Comput. Simul.*, **52**, 217–237.

Bowman, A. W. (1992) Density-based tests for goodness of fit. *J. Statist. Comput. Simul.*, **40**, 1–13.

Bowman, A. W. and Foster, P. J. (1993) Adaptive smoothing and density-based tests of multivariate normality. *J. Amer. Statist. Assoc.*, **88**, 529–537.

Bozdogan, H. (1987) Model selection and Akaike's Information Criterion (AIC): the general theory and its analytical extensions. *Psychometrika*, **52**, 345–370.

Chambers, J. M., Mallows, C. L. and Stuck, B. W. (1976) A method for simulating stable random variables. *J. Amer. Statist. Assoc.*, **71**, 340–344.

Ebrahimi, N., Habibullah, M. and Soofi, E. S. (1992) Testing exponentiality based on Kullback-Leibler information. *J. Roy. Statist. Soc. Ser., B* **54**, 739–748.

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

Eubank, R. L. and Hart, J. D. (1992) Testing goodness-of-fit in regression via order selection criteria. *Ann. Statist.*, **20**, 1412–1425.

Eubank, R. L., Hart, J. D. and La Riccia, V. N. (1993) Testing goodness of fit via nonparametric function estimation techniques. *Comm. Statist. Theory Methods*, **22**, 3327–3354.

Eubank, R. L. and LaRiccia, V. N. (1992) Aymptotic comparison of Cramér-von Mises and nonparametric function estimation techniques for testing goodness-of-fit. *Ann. Statist.*, **20**, 2071–2086.

Fan, J. (1996) Test of significance based on wavelet thresholding and Neyman's truncation. *J. Amer. Statist. Assoc.*, **91**, 647–688.

Gail, M. H. and Gastwirth, J. L. (1978) A scale-free goodness-of-fit test for the exponential distribution based on the Gini statistic. *J. Roy. Statist. Soc. Ser., B* **40**, 350–357.

Gan, F. F. and Koehler, K. J. (1990) Goodness-of-fit tests based on P-P probability plots. *Technometrics*, **32**, 289–303.

Hannan, E. J. and Quinn, B. G. (1979) The determination of the order of an autoregression. *J. Roy. Statist. Soc. Ser., B* **41**, 190–195.

Haughton, D., Haughton, J. and Izenman, A. J. (1990) Information criteria and harmonic models in time series analysis. *J. Statist. Comput. Simulation*, **35**, 187–207.

Haughton, D. (1988) On the choice of a model to fit data from an exponential family. *Ann. Statist.*, **16**, 342–355.

Inglot, T., Kallenberg, W. C. M. and Ledwina, T. (1994a) Power approximations to and power comparison of certain goodness-of-fit tests. *Scand. J. Statist.*, **21**, 131–145.

Inglot, T., Kallenberg, W. C. M. and Ledwina, T. (1994b) *On selection rules, with an application to goodness-of-fit for composite hypotheses*. Memorandum 1242, University of Twente.

Inglot, T. and Ledwina, T. (1996) Asymptotic optimality of data driven Neyman's tests for uniformity. *Ann. Statist.*, **24**, 1982–2019.

Javitz, H. S. (1975) Generalized smooth tests of goodness of fit, independence, and equality of distributions. Unpublished thesis, University of California, Berkeley.

Kallenberg, W. C. M. and Ledwina, T. (1994) *Data driven smooth tests for composite hypotheses*. Memorandum 1232, University of Twente.

Kallenberg, W. C. M. and Ledwina, T. (1995a) Consistency and Monte Carlo simulation of a data driven version of smooth goodness-of-fit tests. *Ann. Statist.*, **23**, 1594–1608.

Kallenberg, W. C. M. and Ledwina, T. (1995b) On data driven Neyman's tests. *Probab. Math. Statist.*, **15**, 409–426.

Kallenberg, W. C. M. and Ledwina, T. (1997) Data driven smooth tests when the hypothesis is composite. *J. Amer. Statist. Assoc.*, to appear.

Kopecky, K. J. and Pierce, D. A. (1979) Efficiency of smooth goodness-of-fit tests. *J. Amer. Statist. Assoc.*, **74**, 393–397.

LaRiccia, V. N. (1991) Smooth goodness-of-fit tests: A quantile function approach. *J. Amer. Statist. Assoc.*, **86**, 427–431.

Le Cam, L. and Lehmann, E. L. (1974) J. Neyman – On the occasion of his 80th birthday, *Ann. Statist.*, **2**(3), vii–xiii.

Ledwina, T. (1994) Data driven version of the Neyman smooth test of fit. *J. Amer. Statist. Assoc.*, **89**, 1000–1005.

Milbrodt, H. and Strasser, H. (1990) On the asymptotic power of the two-sided Kolmogorov-Smirnov test. *J. Statist. Plann. Inference*, **26**, 1–23.

Neyman, J. (1937) 'Smooth test' for goodness of fit. *Skand. Aktuarietidskr.*, **20**, 149–199.

Neyman, J. (1980) Some memorable incidents in probabilistic/statistical studies. In: *Asymptotic Theory of Statistical Tests and Estimation* (I. M. Chakravarti, ed.) 1–32. Academic, New York.

Nishii, R. (1984) Asymptotic properties of criteria for selection of variables in multiple regression. *Ann. Statist.*, **12**, 758–765.

Pearson, E. S., D'Agostino, R. B. and Bowman, K. O. (1977) Tests for departure from normality: Comparison of powers. *Biometrika*, **64**, 231–246.

Pearson, E. S. and Hartley, H. O. (1972) *Biometrika Tables for Statisticians Vol.* II. Cambridge University Press.

Rayner, J. C. W. and Best, D. J. (1989) *Smooth Tests of Goodness of Fit.* Oxford University Press, New York.

Rayner, J. C. W. and Best, D. J. (1990) Smooth tests of goodness of fit: An overview. *Internat. Statist. Rev.*, **58**, 9–17.

Schwarz, G. (1978) Estimating the dimension of a model. *Ann. Statist.*, **6**, 461–464.

Thomas, D. R. and Pierce, D. A. (1979) Neyman's smooth goodness-of-fit test when the hypothesis is composite. *J. Amer. Statist. Assoc.*, **74**, 441–445.