The Annals of Statistics
1997, Vol. 25, No. 3, 1222–1250

DATA DRIVEN SMOOTH TESTS FOR COMPOSITE HYPOTHESES

BY TADEUSZ INGLOT,[1] WILBERT C. M. KALLENBERG AND TERESA LEDWINA[1]

Technical University of Wrocław, University of Twente and Technical University of Wrocław

The classical problem of testing goodness-of-fit of a parametric family is reconsidered. A new test for this problem is proposed and investigated. The new test statistic is a combination of the smooth test statistic and Schwarz's selection rule. More precisely, as the sample size increases, an increasing family of exponential models describing departures from the null model is introduced and Schwarz's selection rule is presented to select among them. Schwarz's rule provides the "right" dimension given by the data, while the smooth test in the "right" dimension finishes the job. Theoretical properties of the selection rules are derived under null and alternative hypotheses. They imply consistency of data driven smooth tests for composite hypotheses at essentially any alternative.

1. Introduction. When testing for goodness-of-fit, alternative hypotheses are often vague and an omnibus test is welcome. By an omnibus test, we mean a test that is consistent against essentially all alternatives. In this paper we propose and investigate an omnibus test for testing composite goodness-of-fit hypotheses. To motivate our choice, let us start with some background.

Let $X_1, \dots, X_n$ be i.i.d. random variables with density $f(x)$. First consider the simple hypothesis H: $f(x) = f_0(x)$, where f_0 is some specified density. Among the most celebrated tests for H are the Kolmogorov–Smirnov test defined in 1933 and the Cramér–von Mises test proposed in 1928 by Cramér and corrected in 1936 by Smirnov. These test statistics are reported in most textbooks and a lot of work has been done on their empirical and asymptotic powers, efficiencies and other properties. As a result, nowadays there is strong evidence that, although the tests are omnibus, for moderate sample sizes only a few deviations from f_0 can be detected by these tests with substantial frequency. Simulation results confirming this observation can be found in Quesenberry and Miller (1977), Locke and Spurrier (1978), Miller and Quesenberry (1979), Eubank and LaRiccia (1992) and Kim (1992).

There are also very interesting asymptotic results on this phenomenon due to Neuhaus (1976) and Milbrodt and Strasser (1990). See also Janssen (1995) for some recent developments. Their results show how these tests distribute their asymptotic powers in the space of all alternatives. In particular, they show that there are only very few directions of deviations from f_0 for which the tests are of reasonable asymptotic power. These directions corre-

Received February 1996; revised November 1996.
[1]Research supported by Grant KBN 350 044.
AMS 1991 *subject classifications*. 62G20, 62G10.
Key words and phrases. Schwarz's BIC criterion, data driven procedure, goodness-of-fit, smooth test, Neyman's test.

spond to very smooth departures from the null density. There is only one direction with the highest asymptotic power that is possible and for "bad" directions the asymptotic power is close to the significance level. A conclusion is that these tests behave very much like a parametric test for a one-dimensional alternative and not like a well-balanced test for higher dimensional alternatives.

The above results caused renewed interest in Neyman's (1937) smooth test of fit and the whole class of smooth tests. For details see Rayner and Best (1989, 1990), Milbrodt and Strasser (1990), Eubank and LaRiccia (1992) and Kaigh (1992). While smooth tests are recommended, it turns out that a wrong choice of the number of components in the test statistic may give a considerable loss of power [cf. Inglot, Kallenberg and Ledwina (1994) and Kallenberg and Ledwina (1995a)]. Therefore, a good procedure for choosing k is very welcome. Some data driven versions of smooth tests have been recently proposed by Bickel and Ritov (1992), Eubank and LaRiccia (1992), Eubank, Hart and LaRiccia (1993), Ledwina (1994), Kallenberg and Ledwina (1995a) and Fan (1996). Extensive simulations presented in Ledwina (1994) and Kallenberg and Ledwina (1995a) show that the data driven Neyman's test proposed in Ledwina (1994) and extended in Kallenberg and Ledwina (1995a) compares very well to classical tests and other competitors.

An important role in this data driven smooth test is played by Schwarz's selection rule. It provides the "right" dimension (or, equivalently, number of components) for the smooth test. The selection rule may be seen as the first step, followed by the finishing touch of applying the smooth test in the selected dimension. Kallenberg and Ledwina (1995a) have shown that this test is consistent at essentially any alternative. Moreover, for a very large set of alternatives $\{f_n\}$ converging to the null density f_0, Inglot and Ledwina (1996) have shown that this data driven test is asymptotically as efficient as the most powerful Neyman–Pearson test of f_0 against f_n. So, the test has quantitatively and qualitatively better properties than the Kolmogorov–Smirnov and Cramér–von Mises tests, for example.

Next consider the composite null hypothesis

$$H_0\colon f(x) \in \{f(x;\beta),\ \beta \in B\}, \tag{1.1}$$

where $B \subset \mathbb{R}^q$ and $\{f(x;\beta),\ \beta \in B\}$ is a given family of densities (for instance, the family of normal or exponential densities). Again a lot of work has been done to construct and investigate Kolmogorov–Smirnov and Cramér–von Mises test statistics in this case. As is well known, when a nuisance parameter β is present, the situation is more complicated than in the case of testing a simple null hypothesis. The reason is that, in general, a natural counterpart of the empirical process, on which these statistics are based, is no longer distribution free or even asymptotically distribution free. For an exhaustive discussion of the problem and for proposed partial solutions, we refer to Durbin (1973), Neuhaus (1979), Khmaladze (1982) and D'Agostino and Stephens (1986). To solve the above problem two general solutions have been proposed also. The first one, proposed by Khmaladze (1981), relies on modifying the natural empirical process with estimated parameters to get a martingale converging weakly, under the null hypothesis, to a Wiener process. This makes it possible to construct some counterparts of the classical Kolmogorov–Smirnov and Cramér–von Mises statistics based on the new process. The second solution, given by Burke and Gombay (1988), consists in taking a single bootstrap sample to estimate β, which makes the Kolmogorov–Smirnov and Cramér–von Mises statistics, based on the related empirical process, asymptotically distribution free. The two above-mentioned solutions, mathematically very elegant, were proposed in principle to enable the use of classical solutions in a more complicated situation when the nuisance parameter is present. However, it is hardly expected that the power behavior of such counterparts of the classical tests will be more appealing than for the case of testing the simple hypothesis. In fact, the above-mentioned results of Neuhaus (1976) on the Cramér–von Mises test hold for the general case when nuisance parameters are present as well. Moreover, simulation studies by Angus (1982), Ascher (1990) and Gan and Koehler (1990) show that more specialized tests like Gini's test for exponentiality or Shapiro-Wilk's test for normality in most situations dominate both Kolmogorov–Smirnov and Cramér–von Mises tests constructed to verify those hypotheses. Therefore, as in the case of testing a simple hypothesis, it seems to be promising to consider data driven smooth tests. This is the subject of the present work.

To be more specific, we shall present now the basic version of the data driven smooth test considered in this paper. A crucial step in the construction of a smooth test is embedding the null density into a larger family of models. We took the family defined as follows.

Let $F(x;\beta)$ be the distribution function of X_i when β applies. For $k = 1, 2, \dots, d(n)$, $d(n) \to \infty$ as $n \to \infty$, define exponential families (w.r.t. θ) by their density

$$g_k(x;\theta,\beta) = \exp\{\theta \circ \phi[F(x;\beta)] - \psi_k(\theta)\} f(x;\beta), \tag{1.2}$$

where

$$\theta = (\theta_1, \dots, \theta_k)', \quad \phi = (\phi_1, \dots, \phi_k)', \quad \psi_k(\theta) = \log \int_0^1 \exp\{\theta \circ \phi(y)\}\, dy.$$

When there is no confusion, the dimension k is sometimes suppressed in the notation. The Euclidean norm in $\mathbb{R}^k$ (or $\mathbb{R}^q$) is denoted by $\|\cdot\|$, the transpose of a matrix or vector by $'$ and $\circ$ stands for the inner product in R^k. The functions $\phi_1, \phi_2, \dots$ and $\phi_0 \equiv 1$ form an orthonormal system in $L_2([0,1])$. The functions $\phi_1, \phi_2, \dots$ are assumed to be bounded but not necessarily uniformly bounded.

Assume for a moment that β is known and consider $g_k(x;\theta,\beta)$ for a fixed k. In this situation, an asymptotically optimal test for testing H_0 or, equivalently, $\theta = 0$, rejects the null hypothesis for large values of

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

This is the smooth test or, equivalently, the score test for $\theta = 0$ against $\theta \neq 0$ in the family (1.2). To select k, Ledwina (1994) proposed to apply Schwarz's rule. The idea behind this choice is clear. First we select a likely model, among $d(n)$ models given by (1.2), fitting the data at hand. Then we apply the optimal test for the fitted model. To define the Schwarz rule (in case β is known) set

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

with j depending on the context. The likelihood of the independent random variables $X_1, \dots, X_n$ each having density (1.2) is given by

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

Schwarz's (1978) Bayesian information criterion for choosing submodels corresponding to successive dimensions yields

$$S(\beta) = \min\{k; 1 \leq k \leq d(n),\ L_k(\beta) \geq L_j(\beta),\ j = 1, \dots, d(n)\}, \tag{1.3}$$

where

$$L_k(\beta) = n \sup_{\theta \in R^k} \{\theta \circ Y_n(\beta) - \psi_k(\theta)\} - \tfrac{1}{2} k \log n.$$

Although it is not mentioned in the notation, $S(\beta)$ depends of course on the upper bound $d(n)$ of the dimensions of the exponential families under consideration. So, the data driven smooth test statistic, when β is known, is $T_{S(\beta)}$. $T_{S(\beta)}$ shares the property of classical goodness-of-fit statistics being, under the null hypothesis, distribution free for each fixed n.

While some authors [cf. Eubank, Hart and LaRiccia (1993) and references mentioned there] include dimension 0 as a candidate dimension or even take the selection rule itself as test statistic, others [e.g., Bickel and Ritov (1992)] start from dimension 1. We prefer the latter approach as argued in Kallenberg and Ledwina (1995a), Section 2.

Our extension of the above solution to the case when β is unknown is based on a similar idea. For fixed k we use an asymptotically optimal test statistic for testing H_0 in the family $g_k(x; \theta, \beta)$. Such a statistic is Neyman's $C(\alpha)$ statistic or, equivalently, the score test for $\theta = 0$ against $\theta \neq 0$ in (1.2). This test statistic is also called (generalized) smooth or Neyman's smooth statistic. For details see Javitz (1975) or Thomas and Pierce (1979). This statistic is denoted throughout by $W_k(\tilde{\beta})$, where $\tilde{\beta}$ is an estimator of β. As in the case of testing a simple hypothesis, the choice of k is crucial for the power of $W_k(\tilde{\beta})$. This is clearly illustrated in Kallenberg and Ledwina (1997). Therefore, a careful selection of k is very important. In this paper we extend Schwarz's selection rule $S(\beta)$ to the present case by considering

$$S = S(\tilde{\beta}). \tag{1.4}$$

In many cases we take as estimator of β the maximum likelihood estimator $\hat{\beta}$ of β under $\theta = 0$, that is, the maximum likelihood estimator of β based on

$\prod_{i=1}^{n} f(x_i;\beta)$. Some other versions of the selection rule, which are easier to calculate, are discussed as well. We show, under mild regularity conditions, that under the null hypothesis $S(\tilde{\beta})$ selects $k=1$ with probability tending to 1 as $n\to\infty$.

The basic version of the data driven smooth test considered in this paper is $W_{S(\hat{\beta})}(\hat{\beta})$, denoted for short by $W_{S(\hat{\beta})}$. By the above, the test statistic $W_{S(\hat{\beta})}$ has as asymptotic null distribution a central chi-square distribution with 1 degree of freedom. So, similarly to the development by Khmaladze (1981, 1993) and Burke and Gombay (1988), the limiting distribution of the test statistic is the same for a large class of null distributions $F(x;\beta)$. However, contrary to their approach, much simpler tools suffice to get the result. Furthermore, while Khmaladze (1981, 1993) and Burke and Gombay (1988) restrict attention to asymptotically efficient estimators of β, we provide also a version of the score statistic and selection rule related to a $\sqrt{n}$-consistent estimator of β. Moreover, we show that rejecting H_0 for large values of $W_{S(\hat{\beta})}$ provides a test procedure which is consistent at essentially any alternative. Similar results are proved for some modified selection rules and the corresponding data driven tests. The main ingredients for deriving the asymptotic null distribution and for proving consistency are properties of the selection rule $S(\tilde{\beta})$ and its modifications. These results are of independent interest.

The main theme of this paper is to prove that the proposed tests have a simple asymptotic distribution under the null hypothesis (chi-square-one) and good power properties (consistency). As a counterpart it is interesting to check the validity of the proposed construction for finite sample sizes. Therefore, the method has been applied in Kallenberg and Ledwina (1995c) to testing exponentiality and normality. From the extensive simulation study reported in that paper it follows that the data driven smooth tests, in contrast to Kolmogorov–Smirnov and Cramér–von Mises tests, compare well for a wide range of alternatives with other, more specialized tests, such as Gini's test for exponentiality and Shapiro–Wilk's test for normality. Finally, it is worthwhile to emphasize that the solution presented here is based on general likelihood methods and hence can be extended to a wide class of other problems both univariate and multivariate. On the other hand, the solution is naturally related to sieve methods and can be extended to some other nonparametric problems as well.

The paper is organized as follows. In Section 2 the selection rules are formally defined, the assumptions are stated and the asymptotic null distribution and behavior of selection rules under alternatives is discussed. Section 3 presents smooth tests for composite hypotheses. In Section 4 the selection rules and the smooth test statistics are combined to give data driven smooth tests for composite hypotheses. Consistency at essentially any alternative is proved. The Appendix is mainly devoted to the proof of Theorem 2.1 and Theorem 3.1. This involves new results on exponential families with dimension growing with n and may be of independent interest.

Except for the basic questions addressed in the present paper, many other aspects may be interesting from an applicability point of view. We conclude this section by discussing some of them.

The first question concerns the choice of the orthonormal system on which the smooth test is based. Neyman's (1937) smooth test is based on the orthonormal Legendre polynomials. Typical statements for choosing the orthonormal system are "Neyman's test would have substantially greater power than the chi-squared test for smooth alternatives" (Barton, 1985) and "to detect alternatives of particular interest, an orthonormal basis should be selected that gives a compact representation of those alternatives" [Rayner and Best (1986, 1990)]. In Bogdan (1995), data driven versions of Pearson's chi-square test for uniformity are investigated and in Bogdan and Ledwina (1996) increasing log-spline families are studied. Such families have been extensively exploited in recent years in the context of nonparametric density estimation [cf., e.g., Stone (1990) and Barron and Sheu (1991)]. In Bogdan and Ledwina (1996) it is shown that the theoretical results of Kallenberg and Ledwina (1995a) can be extended to cover log-spline models. However, for moderate sample sizes there is no substantial gain of empirical power in comparison with the much simpler data driven Neyman's test based on the orthonormal Legendre polynomials. This agrees with our experience that the sensitivity w.r.t. the number of components k is much larger than the sensitivity w.r.t the (commonly used) orthonormal systems. On the one hand, as a rule, the orthonormal systems are complete and hence the systems as a whole are not really different, implying that the "ordering" within the system is the most important feature. On the other hand, a change in the ordering of a *given* orthonormal system may be considered as resulting in a "new" orthonormal system. So, the choice of the system and the ordering are intimately related. In this paper the interesting question of the (possibly data driven) choice of the orthonormal system is not further addressed, except for the discussion on the ordering within a chosen orthonormal system, which is given below.

Well-known examples of the orthonormal functions $\phi_0, \phi_1, \phi_2, \dots$ are the orthonormal Legendre polynomials on [0,1] and the cosine system, given by $\phi_j(x) = \sqrt{2}\cos(j\pi x)$. To these two systems we will refer explicitly in the next sections.

When referring to the orthonormal Legendre polynomials, it will be implicitly assumed that ϕ_j is of degree j. One might ask what happens if the "ordering" of a system is mixed up, for instance, for the Legendre polynomials in a way that ϕ_1 is of degree 3 and not 1. The question on "ordering" is also addressed in Eubank, Hart, Simpson and Stefanski (1995), Section 3. In fact, it means that one considers another orthonormal system. For the example of the Legendre polynomials, the "new" orthonormal system then starts with the Legendre polynomial of degree 3 instead of degree 1. It should be noted that due to the penalty in the selection rule, the ordering reflects what kind of deviations from the null hypothesis are of greatest concern.

As a rule, the results of this paper will go through for such mixings of the orthonormal system as well. One simply has to verify the conditions for the "new" system and except for extreme mixings (e.g., strongly dependent on n) the conditions will be satisfied and the results also hold for the "new" system. However, in the finite sample case the power may change.

When applying the orthonormal Legendre polynomials in testing the simple goodness-of-fit hypothesis of uniformity on [0,1], it seems natural to start with degree 1, corresponding to testing a shift of the mean. Presumably, this is of prime interest. If there is no significant difference in mean, with degree 2 one investigates the next interesting property: the variance. If there is also no significant difference in variance, with degree 3 the skewness is tested and so on. In this way well understood properties of the distribution are systematically investigated in a natural ordering of interest. Moreover, if one starts with the Legendre polynomial with degree 3, being $\sqrt{7}(20x^3-30x^2+12x-1)$, one starts testing a mixture of difference in mean, variance and skewness. Therefore, in the simple hypothesis case, the "degree ordering" starting with degree 1 seems (when aiming for an omnibus test) the most appropriate.

In the composite hypothesis case when testing, for instance, normality, one may propose to start with degree 3 instead of 1, because of location and scale invariance. However, it should be noted that in the composite hypothesis case we are not testing a shift in mean of X_i, but of $F(X_i;\beta)$, when dealing with degree 1 and similarly for degree 2. Nevertheless, if we consider a symmetric alternative, there will be no shift in the mean of $F(X_i;\beta)$. Therefore, in simulations we have experimented with selection starting at dimension 2. Indeed, for symmetric alternatives a higher power is obtained, but for skew alternatives some power is lost. Since the aim is to get a well performing *omnibus* test, we do not recommend starting at dimension 2 or 3.

One may also ask whether the problem of choosing the number of components k is replaced by the choice of $d(n)$. However, in contrast to the power of W_k, the power of W_S does not change for larger $d(n)$. For empirical evidence of this and a discussion of some other aspects, see Kallenberg and Ledwina (1997).

2. Selection rules.

2.1. *Definitions, assumptions, notation.* In this section we set out some of the notation, definitions and conditions to be used in subsequent sections. First we present an alternative to Schwarz's rule (1.3) for choosing k that we will also investigate.

Schwarz's rule $S(\beta)$ as given in (1.3) compares (penalized) maximized likelihoods. It turns out (cf. also Remark 2.5) 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 $\frac{1}{2}n\|Y_n(\beta)\|^2$. The following modification of Schwarz's rule is based on this fact and is easier to calculate:

$$
\begin{aligned}
S2=S2(\tilde\beta)=\min\bigl\{k\colon 1\le k\le d(n),\ & n\|Y_n(\tilde\beta)\|^2_{(k)}-k\log n \\
& \ge n\|Y_n(\tilde\beta)\|^2_{(j)}-j\log n,\ j=1,\dots,d(n)\bigr\},
\end{aligned}
\tag{2.1}
$$

where the index of the norm denotes the dimension.

Denote by P_β that X_i has density $f(x;\beta)$ and by E_β and var_β the corresponding expected value and variance, respectively. For the family $\{f(x;\beta)\colon \beta\in$

$B\}$ we will need the following *regularity* conditions. These conditions should hold on an open subset B_0 of B. (The "true" value of β is supposed to lie in B_0.)

(R1) For $t, u = 1, \dots, q$, $(\partial/\partial\beta_t)f(x;\beta)$ and $(\partial^2/\partial\beta_t\partial\beta_u)f(x;\beta)$ exist almost everywhere and are such that for each $\beta_0 \in B_0$ uniformly in a neighborhood of β_0,

$$|(\partial/\partial\beta_t)f(x;\beta)| \le H_t(x)$$

and

$$|(\partial^2/\partial\beta_t\partial\beta_u)f(x;\beta)| \le G_{tu}(x),$$

where

$$\int_{\mathbb{R}} H_t(x)\,dx < \infty \quad \text{and} \quad \int_{\mathbb{R}} G_{tu}(x)\,dx < \infty.$$

(R2) For $t, u = 1, \dots, q$, $(\partial/\partial\beta_t)\log f(x;\beta)$ and $(\partial^2/\partial\beta_t\partial\beta_u)\log f(x;\beta)$ exist almost everywhere and are such that the Fisher information matrix,

$$I_{\beta\beta} = E_\beta\left\{\left[\frac{\partial}{\partial\beta}\log f(X_1;\beta)\right]\left[\frac{\partial}{\partial\beta}\log f(X_1;\beta)\right]'\right\},$$

is finite, positive definite and continuous and, as $\delta \to 0$, we have

$$E_\beta\left\{\sup_{\{h:\ \|h\|\le\delta\}}\left\|\frac{\partial^2}{\partial\beta\partial\beta'}\log f(X_1;\beta+h) - \frac{\partial^2}{\partial\beta\partial\beta'}\log f(X_1;\beta)\right\|\right\} \to 0.$$

(R3) For each $\beta_0 \in B_0$ there exists $\eta = \eta(\beta_0) > 0$ with

$$\sup_{\|\beta-\beta_0\|<\eta}\ \sup_{x\in\mathbb{R}}\left|\frac{\partial^2}{\partial\beta_t\partial\beta_u}F(x;\beta)\right| < \infty, \qquad t, u = 1, \dots, q$$

and

$$\sup_{x\in\mathbb{R}}\left|\frac{\partial}{\partial\beta_t}F(x;\beta)\right|_{\beta=\beta_0} < \infty, \qquad t = 1, \dots, q.$$

(R4) There exist positive constants c_1, c_2, ρ_1 and n_1 such that the estimator $\tilde\beta$ of $\beta \in B_0$ satisfies

$$P_\beta(\sqrt{n}\|\tilde\beta - \beta\| \ge r) \le c_1\exp(-c_2 r^2)$$

for all $r = \rho\sqrt{\log n}$ with $0 < \rho \le \rho_1$ and $n \ge n_1$.

The next conditions concern the orthonormal *system* $\{\phi_j\}_{j=0}^\infty$.

(S1) $$\sup_{x\in[0,1]} |\phi'_j(x)| \le c_3 j^{m_1}$$

for each $j = 1, 2, \dots, d(n)$ and some $c_3 > 0$, $m_1 > 0$.

(S2) $$\sup_{x\in[0,1]} |\phi''_j(x)| \le c_4 j^{m_2}$$

for each $j = 1, 2, \dots, d(n)$ and some $c_4 > 0$, $m_2 > 0$.

Finally we have conditions on the *dimension* $d(n)$ of the exponential family.

(D1)
$$\{d(n)V_{d(n)}\}^2 n^{-1}\log n \to 0 \quad \text{as } n\to\infty,$$
$$\text{where } V_k = \max_{1\le j\le k}\ \sup_{x\in[0,1]} \mid \phi_j(x)\mid .$$

(D2) $$d(n) = o(\{n/\log n\}^{(2m)^{-1}}) \quad \text{as } n\to\infty \text{ where } m=\max(m_1,m_2).$$

(D3) $$d(n) = o(n^c) \quad \text{as } n\to\infty$$

for some $c < c_2 b^{-2}$ if $\rho_1 b \ge 1$, and with $c = c_2\rho_1^2$, otherwise, where c_2 and ρ_1 are given by (R4) and

$$b = \left\{\sum_{t=1}^{q} \operatorname{var}_\beta \frac{\partial}{\partial\beta_t}\log f(X;\beta)\right\}^{1/2}.$$

It should be noted that some of the lemmas and theorems require only a subset of these assumptions.

If we take $\{\phi_j\}$ to be the orthonormal Legendre polynomials on [0, 1] we get [cf. Sansone (1959), page 190] $V_k = (2k+1)^{1/2}$ and hence (D1) reduces in this case to $\{d(n)\}^3 n^{-1}\log n \to 0$ as $n\to\infty$. Moreover, (S1) and (S2) are fulfilled with [cf. Sansone (1959) page 251] $m_1 = 5/2$ and $m_2 = 9/2$.

If $\{\phi_j\}$ is the cosine system we have $V_k = \sqrt{2}$ and hence (D1) reduces in this case to $\{d(n)\}^2 n^{-1}\log n \to 0$ as $n\to\infty$. Moreover, (S1) and (S2) are fulfilled with $m_1 = 1$ and $m_2 = 2$.

If the family $\{f(x;\beta)\colon \beta\in B\}$ is a location-scale family $\{\sigma^{-1}f_0(\sigma^{-1}(x-\mu))\colon \mu\in\mathbb{R},\ \sigma>0\}$, say, and if f_0 is continuously differentiable with $x^2 f_0'(x)$ and $xf_0(x)$ bounded, then (R3) holds. Consequently, (R3) holds for the normal family. Similarly, it can be shown that (R3) holds for many other well-known families as, for instance, the scale family of exponential distributions.

As a rule condition (R4) can be checked by application of standard moderate deviation theory. A particularly important example is for $\tilde\beta = (\bar X, \{n^{-1}\sum_{i=1}^n (X_i-\bar X)^2\}^{1/2})'$, which is often used in the location-scale case. We study this case in Appendix A.3 and show, for example, that in the normal case (R4) holds and (D4) reduces to $d(n) = o(n^c)$ for some $c < \frac{1}{6}$. Other estimators based on sample moments can be treated in a similar way.

Another important example is when $\{P_\beta\}$ is an exponential family with β the natural parameter in this family and $\tilde\beta = \hat\beta$, the maximum likelihood estimator. Condition (R4) then easily follows from Example 2.1 (continued) on page 502 of Kallenberg (1983).

For M-estimators, including sample quantiles, (R4) follows from the results in Jurečková, Kallenberg and Veraverbeke (1988). For L-statistics we refer to Section 5 of Inglot, Kallenberg and Ledwina (1992), where also many other references can be found.

Condition (R4) can also be checked by an application of Berry–Esséen theorems. Suppose that for some $c^* > 0$,

$$\sup_x \bigl| P_\beta\bigl(\sqrt{n}|\tilde\beta-\beta| \le x\bigr) - \Phi(x)\bigr| \le c^* n^{-1/2}$$

with Φ the standard normal distribution function. Then we have for all $r = \rho\sqrt{\log n}$ with $0 < \rho \le 1$,

$$P_\beta\big(\sqrt{n}|\tilde{\beta} - \beta| \ge r\big) \le \Phi(-r) + c^* n^{-1/2} \le \big\{\tfrac{1}{2} + c^*\big\}\exp\big(-\tfrac{1}{2}r^2\big)$$

and hence (R4) holds with $c_1 = \{\frac{1}{2} + c^*\}$, $c_2 = \frac{1}{2}$ and $\rho_1 = 1$.

If $\tilde{\beta}$ is of the form $T(\hat{F}_n)$ with T some functional and $\hat{F}_n$ the empirical distribution function, (R4) may occasionally be checked by application of the Dvoretzky–Kiefer–Wolfowitz (DKW) inequality. Suppose that $\tilde{\beta} - \beta = T(\hat{F}_n) - T(F_\beta)$ satisfies $\|T(\hat{F}_n) - T(F_\beta)\| \le \tilde{c}\|\hat{F}_n - F_\beta\|_\infty$ for some $\tilde{c} > 0$, where $\|\cdot\|_\infty$ is the sup-norm and $F_\beta(x) = F(x;\beta)$. Then, by the DKW inequality [cf. Massart (1990)], for all r,

$$P_\beta\big(\sqrt{n}\|\tilde{\beta} - \beta\| \ge r\big) \le P_\beta\big(\sqrt{n}\|\hat{F}_n - F_\beta\|_\infty \ge r/\tilde{c}\big) \le 2\exp(-2r^2/\tilde{c}^2)$$

and hence (R4) holds for any ρ_1, with $c_1 = 2$ and $c_2 = 2/\tilde{c}^2$.

2.2. *Null distribution of selection rules.* This subsection gives the asymptotic null distribution of the selection rules. It will be tacitly assumed that the "true" parameter β belongs to B_0 [cf. (R3), (R4)]. The probability measure P_β denotes that X_i has density $f(x;\beta)$.

The first theorem states how close $Y_n(\tilde{\beta})$ is to $Y_n(\beta)$.

THEOREM 2.1. *Assume* (R1)–(R4), (S1), (S2), (D2), (D3). *Then there exists* $\varepsilon > 0$ *such that*

$$\lim_{n\to\infty}\sum_{k=2}^{d(n)} P_\beta\big(\|Y_n(\tilde{\beta}) - Y_n(\beta)\| \ge (1-\varepsilon)\{(k-1)n^{-1}\log n\}^{1/2}\big) = 0. \tag{2.2}$$

The proof of Theorem 2.1 is given in the Appendix.

The next theorem shows that under H_0, Schwarz's selection rule and its modification asymptotically concentrate on dimension 1. For empirical evidence of this, see Kallenberg and Ledwina (1997).

THEOREM 2.2. *Assume* (R1)–(R4), (S1), (S2), (D1), (D2), (D3); *then*

$$\lim_{n\to\infty} P_\beta(S(\tilde{\beta}) \ge 2) = 0, \tag{2.3}$$

$$\lim_{n\to\infty} P_\beta(S2(\tilde{\beta}) \ge 2) = 0. \tag{2.4}$$

Before proving Theorem 2.2, we present the following lemma, which is given in Inglot and Ledwina (1996) (cf. Theorem 7.4). Let $u_k = k^{1/2}V_k$.

LEMMA 2.3. *For every* $k \ge 1$, $0 < \varepsilon < \min\{1, \frac{2}{3}u_k^2\}$ *and* $a \le (2-\varepsilon)\varepsilon^2/(16u_k^2)$ *we have*

$$\Big\{x \in \mathbb{R}^k\colon \sup_{\theta\in\mathbb{R}^k}[\theta\circ x - \psi_k(\theta)] \ge a\Big\} \subset \big\{x\colon \|x\|^2 \ge (2-\varepsilon)a\big\}.$$

PROOF OF THEOREM 2.2. Take any $k \in \{2, \dots, d(n)\}$. Let $a_k = (k-1) \times n^{-1} \log n$. Then we get

$$\{S(\tilde{\beta}) = k\} \subset \Big\{\sup_{\theta \in \mathbb{R}^k} (\theta \circ Y_n(\tilde{\beta}) - \psi_k(\theta)) \geq \tfrac{1}{2} a_k\Big\}.$$

In view of (D1), we have for any $0 < \varepsilon < 1$, $2 \leq k \leq d(n)$ and sufficiently large n,

$$\tfrac{1}{2} a_k \leq (2 - \varepsilon)\varepsilon^2 / (16 u_k^2)$$

and hence by Lemma 2.3,

$$\{S(\tilde{\beta}) = k\} \subset \{\|Y_n(\tilde{\beta})\| \geq [(1 - \tfrac{1}{2}\varepsilon) a_k]^{1/2}\}. \tag{2.5}$$

By Theorem 2.1 there exists $0 < \varepsilon_0 < 2 - \sqrt{2}$ such that

$$\lim_{n\to\infty} \sum_{k=2}^{d(n)} P_\beta\big(\|Y_n(\tilde{\beta}) - Y_n(\beta)\| \geq (1 - \varepsilon_0) a_k^{1/2}\big) = 0. \tag{2.6}$$

Take $0 < \varepsilon < 1$ in (2.5) such that $(1 - \frac{1}{2}\varepsilon)^{1/2} = 1 - \frac{1}{2}\varepsilon_0$. Combining (2.5) and (2.6) then yields

$$\begin{aligned} P_\beta(S(\tilde{\beta}) \geq 2) &= \sum_{k=2}^{d(n)} P_\beta(S(\tilde{\beta}) = k) \\ &\leq \sum_{k=2}^{d(n)} P_\beta\big(\|Y_n(\tilde{\beta})\| \geq \big(1 - \tfrac{1}{2}\varepsilon_0\big) a_k^{1/2}\big) \\ &\leq \sum_{k=2}^{d(n)} \big[P_\beta\big(\|Y_n(\beta)\| \geq \tfrac{1}{2}\, \varepsilon_0 a_k^{1/2}\big) \\ &\qquad + P_\beta\big(\|Y_n(\tilde{\beta}) - Y_n(\beta)\| \geq (1 - \varepsilon_0) a_k^{1/2}\big)\big] \\ &= \sum_{k=2}^{d(n)} P_\beta\big(\|Y_n(\beta)\| \geq \tfrac{1}{2}\varepsilon_0 a_k^{1/2}\big) + o(1). \end{aligned} \tag{2.7}$$

Next we apply formula (2) of Prohorov (1973) with, in the notation of that paper, $\rho = \frac{1}{2}\, \varepsilon_0 \{n a_k\}^{1/2}$, $m = k, \lambda = 1$, $a = \frac{1}{2}\, \varepsilon_0 k^{1/2} a_k^{1/2} V_k$, yielding [cf. also the proof of Theorem 3.2 in Kallenberg and Ledwina (1995a)]

$$\sum_{k=2}^{d(n)} P_\beta\big(\|Y_n(\beta)\| \geq \tfrac{1}{2}\varepsilon_0 a_k^{1/2}\big) \leq c_5 \sum_{k=2}^{d(n)} \exp\big\{-\tfrac{1}{9}\, \varepsilon_0^2 (k-1) \log n\big\} \tag{2.8}$$

for some positive constant c_5. The right-hand side of (2.8) tends to 0 as $n \to \infty$. This completes the proof of (2.3). Since

$$P_\beta(S2(\tilde{\beta}) \geq 2) \leq \sum_{k=2}^{d(n)} P_\beta\big(\|Y_n(\tilde{\beta})\|^2 \geq a_k\big),$$

it follows immediately from the previous proof [cf. (2.6), (2.7) and (2.8)] that (2.4) holds as well. □

REMARK 2.4. Since $\int_0^1 \phi_k^2(x)\,dx = 1$, we have $V_k \geq 1$ and hence $\frac{2}{3}u_k^2 \geq \frac{2}{3}k \geq 1$ if $k \geq 2$, and thus $\min\{1, \frac{2}{3}u_k^2\} = 1$ for $k \geq 2$ in Lemma 2.3.

REMARK 2.5. As a counterpart of Lemma 2.3 one may also prove that for every $k \geq 1$, $0 < \varepsilon \leq 1$ and $a \leq \varepsilon^2(2+\varepsilon)^{-3}u_k^{-2}$ it holds that

$$\Big\{x \in \mathbb{R}^k\colon \sup_{\theta\in\mathbb{R}^k}[\theta \circ x - \psi_k(\theta)] \geq a\Big\} \supset \{x\colon \|x\|^2 \geq (2+\varepsilon)a\};$$

compare Inglot and Ledwina (1996), Theorem 7.3. This shows that the likelihood ratio statistic for testing H: $\theta = 0$ against A: $\theta \neq 0$ in the model (1.2) with β known; that is,

$$n \sup_{\theta\in\mathbb{R}^k}\big[\theta \circ Y_n(\beta) - \psi_k(\theta)\big],$$

is locally equivalent to $\frac{1}{2}n\|Y_n(\beta)\|^2$, not only in fixed dimension (as is well known), but also as the dimension tends to infinity with n.

2.3. *Selection rules under alternatives.* Here we consider the behavior of the selection rules under alternatives. So, $X_1, X_2, \ldots$ are i.i.d. r.v.'s each distributed according to P. Under alternatives, the meaning of β is at first sight less clear. We only have an estimator $\tilde{\beta}$. Under the alternative distribution P, $\tilde{\beta}$ will as a rule converge to some element of B_0. This element will then be called β [cf. (2.10)]. So, the "artificial" parameter β under P is determined by the estimator and the alternative distribution P. For instance, if we estimate a location parameter by the sample mean, under the alternative P the parameter β will be the expectation of X_i under P. However, if we estimate the location parameter by the sample median, the parameter β should be read as the median of X_i under P.

We shall consider P to be an alternative to the family $\{f(x;\beta)\colon \beta \in B\}$ if there exists (for the β associated with $\tilde{\beta}$ and P) $K(\beta)$ such that

$$\begin{gathered} E_P\phi_1(F(X;\beta)) = \cdots = E_P\phi_{K(\beta)-1}(F(X;\beta)) = 0, \\ E_P\phi_{K(\beta)}(F(X;\beta)) \neq 0. \end{gathered} \tag{2.9}$$

Note that if (2.9) does not hold, $E_P\phi_j(F(X;\beta)) = 0$ for all j and therefore essentially any alternative of interest satisfies (2.9).

While under H_0 the selection rules concentrate on dimension 1 (cf. Theorem 2.2), under alternatives, higher dimensions also play a role.

THEOREM 2.6. *Assume that* (2.9) *holds and let* β *be associated with* $\tilde{\beta}$ *and* P *in the sense that*

$$\|\tilde{\beta} - \beta\| \to_P 0. \tag{2.10}$$

Further assume that there exists $\eta > 0$ *such that*

$$\sup_{\|\gamma-\beta\|<\eta} \sup_{x\in\mathbb{R}} \left| \frac{\partial}{\partial\gamma_t} F(x;\gamma) \right| < \infty, \qquad t = 1, \dots, q$$

and that

$$\sup_{x\in[0,1]} \left|\phi'_j(x)\right| < \infty, \qquad j = 1, \dots, K(\beta).$$

Then

$$\lim_{n\to\infty} P(S(\tilde\beta) \geq K(\beta)) = 1,$$
$$\lim_{n\to\infty} P(S2(\tilde\beta) \geq K(\beta)) = 1.$$

PROOF. Consider a fixed $k \in \{1, \dots, K(\beta)\}$. Since

$$\bar\phi_k(\tilde\beta) - \bar\phi_k(\beta) = O(\|\tilde\beta - \beta\|)$$

and, by the law of large numbers,

$$\bar\phi_k(\beta) \to_P E_P \phi_k(F(X;\beta)),$$

it follows that

$$\bar\phi_k(\tilde\beta) \to_P E_P \phi_k(F(X;\beta)). \tag{2.11}$$

By continuity of the function

$$x \to \sup_{\theta\in\mathbb{R}^k} [\theta \circ x - \psi(\theta)]$$

at $x = 0$, it follows that for each $k \in \{1, \dots, K(\beta) - 1\}$,

$$\sup_{\theta\in\mathbb{R}^k} [\theta \circ Y_n(\tilde\beta) - \psi_k(\theta)] \to_P 0. \tag{2.12}$$

Since

$$\frac{d}{dt} \psi_{K(\beta)}(0, \dots, 0, t)\Big|_{t=0} = E_\beta \phi_{K(\beta)}(F(X;\beta)) = 0,$$

it holds that for every $a \neq 0$,

$$\sup_{t\in\mathbb{R}} [ta - \psi_{K(\beta)}(0, \dots, 0, t)] > 0.$$

Further we have

$$\sup_{\theta\in\mathbb{R}^{K(\beta)}} [\theta \circ Y_n(\tilde\beta) - \psi_{K(\beta)}(\theta)] \geq \sup_{t\in\mathbb{R}} [t\bar\phi_{K(\beta)}(\tilde\beta) - \psi_{K(\beta)}(0, \dots, 0, t)]. \tag{2.13}$$

By the continuity of the function

$$x \to \sup_{t\in\mathbb{R}} [tx - \psi_{K(\beta)}(0, \dots, 0, t)]$$

we therefore get

$$
\begin{aligned}
(2.14)\qquad & \sup_{t\in\mathbb{R}}\big[t\bar{\phi}_{K(\beta)}(\tilde{\beta}) - \psi_{K(\beta)}(0,\dots,0,t)\big] \\
& \quad \to_P \sup_{t\in\mathbb{R}}\big[tE_P\phi_{K(\beta)}(F(X;\beta)) - \psi_{K(\beta)}(0,\dots,0,t)\big] > 0.
\end{aligned}
$$

Combination of (2.12), (2.13) and (2.14) yields, for each $k \in \{1,\dots,K(\beta)-1\}$,

$$
\begin{aligned}
P(S(\tilde{\beta}) = k) \le P\Big(&\sup_{\theta\in\mathbb{R}^k}\{\theta \circ Y_n(\tilde{\beta}) - \psi_k(\theta)\} \ge -\tfrac{1}{2}\{K(\beta)-k\}n^{-1}\log n \\
&+ \sup_{\theta\in\mathbb{R}^{K(\beta)}}\{\theta \circ Y_n(\tilde{\beta}) - \psi_{K(\beta)}(\theta)\}\Big) \to 0
\end{aligned}
$$

as $n \to \infty$ and therefore

$$P(S(\tilde{\beta}) \ge K(\beta)) \to 1.$$

Since for $k \in \{1,\dots,K(\beta)-1\}$,

$$\|Y_n(\tilde{\beta})\|^2_{(k)} \to_P 0$$

and

$$\|Y_n(\tilde{\beta})\|^2_{(K(\beta))} \to_P \big\{E_P\phi_{K(\beta)}(F(X;\beta))\big\}^2 > 0,$$

it easily follows that

$$P(S2(\tilde{\beta}) \ge K(\beta)) \to 1. \qquad \square$$

For illustration of Theorem 2.6 by Monte Carlo results we refer to Kallenberg and Ledwina (1997). Indeed, the behavior under alternatives is quite different from that under H_0 with less concentration on dimension 1.

REMARK 2.7. The natural setting for $d(n)$ is that $d(n) \to \infty$ as $n \to \infty$. Nevertheless, all results in the paper also hold if $d(n)$ is bounded, provided that, in case of results on alternatives, $\liminf_{n\to\infty} d(n) \ge K(\beta)$.

3. Smooth tests for composite hypotheses. Our test statistic is a combination of the selection rule and the score test within the family (1.2). Having introduced the selection rule in Section 2, we now present the score test. The most popular version of the score test statistic for testing $\theta = 0$ against $\theta \neq 0$ in (1.2) is defined as follows. Let I be the $k \times k$ identity matrix. Further define

$$
\begin{aligned}
& I_\beta = \left\{-E_\beta \frac{\partial}{\partial\beta_t}\phi_j[F(X;\beta)]\right\}_{t=1,\dots,q,\ j=1,\dots,k}, \\
(3.1)\qquad & I_{\beta\beta} = \left\{-E_\beta \frac{\partial^2}{\partial\beta_t\partial\beta_u}\log f(X;\beta)\right\}_{t=1,\dots,q,\ u=1,\dots,q}, \\
& R(\beta) = I'_\beta(I_{\beta\beta} - I_\beta I'_\beta)^{-1}I_\beta.
\end{aligned}
$$

Applying $\hat{\beta}$, the maximum likelihood estimator of β under H: $\theta = 0$, the score statistic is given by

$$W_k = nY'_n(\hat{\beta})\{I + R(\hat{\beta})\}Y_n(\hat{\beta}). \tag{3.2}$$

A more general version, which gives more flexibility in choosing an estimator of β, is defined in the following way. Let $\tilde{\beta}$ be a $\sqrt{n}$-consistent estimator of β (we will assume this condition for $\tilde{\beta}$ in the rest of the paper), then the score statistic is given by [cf. Cox and Hinkley (1974), page 324, Thomas and Pierce (1979), page 443]

$$W_k(\tilde{\beta}) = n\tilde{Y}'_n(\tilde{\beta})\{I + R(\tilde{\beta})\}\tilde{Y}_n(\tilde{\beta}) \tag{3.3}$$

with

$$\tilde{Y}_n(\beta) = Y_n(\beta) - I'_\beta I^{-1}_{\beta\beta} C_n(\beta), \tag{3.4}$$

where

$$C_n(\beta) = n^{-1}\sum_{i=1}^{n}\left(\frac{\partial}{\partial\beta_1}\log f(X_i;\beta), \dots, \frac{\partial}{\partial\beta_q}\log f(X_i;\beta)\right)'. \tag{3.5}$$

Note that $C_n(\hat{\beta}) = 0$ and indeed in that case the extra term vanishes; compare (3.2).

In the case that the ϕ_j's are polynomials, Thomas and Pierce (1979) have formulated a set of assumptions which guarantee that W_k, given by (3.2), has an asymptotic chi-square distribution. A general theory on the asymptotic distribution of score test statistics when the maximum likelihood estimator is used is presented in Sen and Singer (1993), Chapter 5. The assumptions they impose are standard Cramér-type regularity conditions, sufficient for asymptotic normality of maximum likelihood estimators, as developed by Le Cam (1956) and Hájek (1972). Checking those assumptions for the family (1.2), conditions can be formulated sufficient to ensure that W_k, given in (3.2), is asymptotically chi-squared distributed. Hall and Mathiason (1990) introduced an extension of the score statistic which they have called Neyman–Rao, or effective score statistic. The modification allows both $\sqrt{n}$-consistent estimators of β and a nonsingular consistent estimate of $R(\beta)$. They also give a set of assumptions sufficient to derive the asymptotic distribution of their statistic.

The asymptotic distribution of $W_k(\tilde{\beta})$ with ϕ_j's being the Legendre polynomials has been considered in Javitz (1975). The general case is treated in Theorem 3.1. Its proof is given in the Appendix. By χ^2_k we denote a r.v. with a central chi-square distribution with k degrees of freedom.

THEOREM 3.1. *Assume* (R1)–(R3) *and* (S1), (S2). *Then*

$$W_k(\tilde{\beta}) \to_{P_\beta} \chi^2_k.$$

REMARK 3.2. Theorem 3.1 shows that $W_k(\tilde{\beta})$ is asymptotically distribution free, in the sense that the limiting distribution of it under H_0 does not depend in any way on F and β. For location-scale families also the finite sample null distribution of W_k does not depend on β (cf. Section 4).

REMARK 3.3. It is not always obvious that the smooth test statistic for the composite hypothesis has a chi-square limiting distribution, even in the location-scale case. This is shown by Boulerice and Ducharme (1995) when considering smooth test statistics proposed by Rayner and Best (1986). Theorem 3.1 shows that $W_k(\tilde{\beta})$ does not suffer from this problem.

4. Data driven smooth tests for composite hypotheses. In view of the good performance of the data driven smooth test in the simple hypothesis case, it is natural to investigate data driven versions of the smooth tests for composite hypotheses, being far more important in applications.

The data driven smooth test statistics for testing the composite hypothesis H_0 are given by

$$W_{S(\tilde{\beta})}(\tilde{\beta}), \qquad W_{S2(\tilde{\beta})}(\tilde{\beta}) \tag{4.1}$$

with $W_k(\tilde{\beta})$ given in (3.3), $S(\tilde{\beta})$ in (1.4), (1.3) and $S2(\tilde{\beta})$ given in (2.1). The null hypothesis is rejected for large values of the test statistic.

The asymptotic null distribution of the test statistics is given in the following theorem.

THEOREM 4.1. *Under the conditions of Theorem* 2.2 *we have*

$$W_{S(\tilde{\beta})}(\tilde{\beta}) \to_{P_\beta} \chi_1^2, \qquad W_{S2(\tilde{\beta})}(\tilde{\beta}) \to_{P_\beta} \chi_1^2.$$

PROOF. Since for $T = S(\tilde{\beta})$ or $S2(\tilde{\beta})$,

$$\begin{aligned} P_\beta(W_T(\tilde{\beta}) \le x) = P_\beta(W_1(\tilde{\beta}) \le x) - P_\beta(W_1(\tilde{\beta}) \le x,\ T \ge 2) \\ + P_\beta(W_T(\tilde{\beta}) \le x,\ T \ge 2) \end{aligned}$$

and, by Theorem 3.1,

$$W_1(\tilde{\beta}) \to_{P_\beta} \chi_1^2,$$

the result follows immediately from Theorem 2.2. □

Note that $W_{S(\tilde{\beta})}(\tilde{\beta})$ and $W_{S2(\tilde{\beta})}(\tilde{\beta})$ are asymptotically distribution free, in the sense that the limiting null distribution neither depends on F nor on β. For location-scale families, the finite sample null-distribution does not depend on β (cf. the end of this section).

Although the selection rules under H_0 concentrate on dimension 1, the implied chi-square distribution with one degree of freedom does not work very well as an approximation to establish accurate critical values [cf. Kallenberg and Ledwina (1997)]. 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 can be proposed for the composite null hypothesis, yielding a simple and accurate approximation of the critical values. For example, let $\{f(x;\beta): \beta \in B\}$ be a location-scale family and let $\hat\beta$ be the maximum likelihood estimator of β in this family. Under classical regularity conditions we have under H_0 that $\sqrt{n}(\bar\phi_1(\hat\beta), \bar\phi_2(\hat\beta))$ converges to a two-dimensional normal distribution with correlation coefficient r, say. Under H_0, $P_\beta(S \geq 3)$ and $P_\beta(W_2 \leq x,\ S = 2)$ are negligible. The most important part of $\{S = 1\}$ is that dimension 1 "beats" dimension 2 which is approximately equal to $\{n(\bar\phi_2(\hat\beta))^2 \leq \log n\}$. Hence [cf. (1.4.18) on page 25 of Bickel and Doksum (1977)],

$$P_\beta(W_S \leq x) \approx P_\beta\big(W_1 \leq x,\ n(\bar\phi_2(\hat\beta))^2 \leq \log n\big)$$
$$\approx Pr\big(\{\sqrt{1-r^2}\,U_1 + rU_2\}^2 \leq x,\ U_2^2 \leq v \log n\big),$$

where U_1 and U_2 are independent and $N(0,1)$-distributed and v^{-1} equals the limiting variance of $\sqrt{n}\bar\phi_2(\hat\beta)$. An even more simple approximation in terms of the standard normal distribution function is given by

$$P_\beta(W_S \leq x) \approx 2\Phi(\sqrt{x}) - 1 - 2\big[1 - \Phi\{(v \log n)^{1/2}\}\big]\big[\Phi(b(x)) - \Phi(a(x))\big]$$

with

$$b(x) = \left[\sqrt{x} - \frac{r}{2}\{\sqrt{v \log n} + \sqrt{2 \log n}\}\right]\Big/\sqrt{1-r^2},$$
$$a(x) = \left[-\sqrt{x} - \frac{r}{2}\{\sqrt{v \log n} + \sqrt{2 \log n}\}\right]\Big/\sqrt{1-r^2}.$$

For more details we refer to Kallenberg and Ledwina (1997).

Before proving consistency we consider the distribution of the test statistics under alternatives. First we take as estimator $\hat\beta$, the maximum likelihood estimator of β under H_0.

THEOREM 4.2. *Under the conditions of Theorem* 2.6 *we have*

$$W_{S(\hat\beta)} \to_P \infty, \qquad W_{S2(\hat\beta)} \to_P \infty.$$

PROOF. Since $R(\beta)$ is nonnegative definite, we have

$$W_k \geq n\|Y_n(\hat\beta)\|^2 \tag{4.2}$$

and hence for any $k \geq K(\beta)$,

$$W_k \geq n|\bar\phi_{K(\beta)}(\hat\beta)|^2. \tag{4.3}$$

Therefore, for $T = S(\hat\beta)$ or $S2(\hat\beta)$ and for any $x \in \mathbb{R}$,

$$P(W_T \leq x) = P(W_T \leq x,\ T \geq K(\beta)) + P(W_T \leq x,\ T \leq K(\beta) - 1)$$
$$\leq P(n|\bar\phi_{K(\beta)}(\hat\beta)|^2 \leq x) + P(T \leq K(\beta) - 1).$$

In view of (2.11) with $\tilde{\beta} = \hat{\beta}$, we have

$$\bar{\phi}_{K(\beta)}(\hat{\beta}) \to_P E_P \phi_{K(\beta)}(F(X;\beta)) \neq 0. \tag{4.4}$$

By Theorem 2.6 it holds that

$$P(T \leq K(\beta) - 1) \to 0$$

and hence, for any $x \in \mathbb{R}$,

$$\lim_{n\to\infty} P(W_T \leq x) = 0. \qquad \square$$

A combination of Theorems 4.1 and 4.2 yields Theorem 4.3.

THEOREM 4.3. *Under the conditions of Theorems* 4.1 *and* 4.2, *the tests based on* $W_{S(\hat{\beta})}$ *and* $W_{S2(\hat{\beta})}$ *are consistent against any alternative of the form* (2.9), *that is, against essentially any alternative of interest.*

COROLLARY 4.4. *For testing normality or exponentiality, the test based on* $W_{S(\hat{\beta})}$ *with* $\{\phi_j\}$ *the orthonormal Legendre polynomials on* $[0,1]$ *is consistent against any alternative of the form* (2.9) *with finite second moment if* $d(n) = o(\{n/\log n\}^{1/9})$.

The same result holds if $\{\phi_j\}$ *is the cosine system, in which case for each* $\varepsilon > 0$, $d(n) = o(n^{(1/6)-\varepsilon})$ *suffices for testing normality and* $d(n) = o(\{n/\log n\}^{1/4})$ *for testing exponentiality.*

An extensive simulation study of the power of tests based on $W_{S(\hat{\beta})}$ and $W_{S2(\hat{\beta})}$ with $\{\phi_j\}$ the Legendre polynomials is presented in Kallenberg and Ledwina (1995c). It turns out that these data driven versions of Neyman's test compare well for a wide range of alternatives with other, possibly more specialized, commonly used tests. The data driven smooth tests are competitive with well-known tests such as Shapiro–Wilk's test in case of normality and Gini's test for exponentiality.

Theorems 4.2, 4.3 and Corollary 4.4 can easily be extended to other estimators than $\hat{\beta}$ if we, apart from (2.9), assume that

$$E_P \phi_j(F(X;\beta)) - \left\{ I'_\beta I^{-1}_{\beta\beta} E_P \frac{\partial}{\partial\beta} \log f(X;\beta) \right\}_j \neq 0, \tag{4.5}$$

for some $j \leq K(\beta)$, where $\{\dots\}_j$ denotes the jth element of the vector between braces. Condition (4.5) is used in the adjusted (4.4); compare also (4.2) and (4.3).

In case of a location-scale family $\{f(x;\beta) \colon \beta \in B\}$ we write $\beta = (\mu, \sigma)'$, $f(x;\beta) = \sigma^{-1} f_0((x-\mu)/\sigma)$ and $F(x;\beta) = F_0((x-\mu)/\sigma)$. Now $R(\beta)$ defined in (3.1) does not depend on β. The statistics $S(\hat{\beta})$, $S2(\hat{\beta})$, $W_{S(\hat{\beta})}$ and $W_{S2(\hat{\beta})}$ all depend on $X_1, \dots, X_n$ by means of

$$\frac{X_i - \hat{\mu}}{\hat{\sigma}}, \qquad i = 1, \dots, n,$$

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

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

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 $\{f(x;\beta)\colon \beta \in B\}$ the finite sample null distributions of $S(\hat{\beta})$, $S2(\hat{\beta})$, $W_{S(\hat{\beta})}$ and $W_{S2(\hat{\beta})}$ do not depend on β. Moreover, if the alternative also belongs to a location-scale family, the distribution of $S(\hat{\beta})$, $S2(\hat{\beta})$, $W_{S(\hat{\beta})}$ and $W_{S2(\hat{\beta})}$ do not depend on the location-scale parameter of that family. Of course, the preceding remarks also apply to other estimators of β, provided they are also location-scale invariant. The same remark applies to location families and to scale families.

We close this section by refering to Section 7 of Kallenberg and Ledwina (1997) for some modifications. Another modification, not mentioned there, is to replace $n\|Y_n(\tilde{\beta})\|^2$ in $S2$ by $W_k(\tilde{\beta})$, thus taking into account that an estimator is plugged in. Since simulation (with $\tilde{\beta} = \hat{\beta}$) of this modified selection rule gives no better results, we do not work it out here.

APPENDIX

A.1. *Proof of Theorem* 2.1. We start with some additional notation:

$$U_{tj} = n^{-1/2}\sum_{i=1}^{n}\left\{\frac{\partial}{\partial\beta_t}\phi_j(F(X_i;\beta)) - E_\beta\frac{\partial}{\partial\beta_t}\phi_j(F(X;\beta))\right\},$$

$$R_{1j} = (\tilde{\beta} - \beta)'(n^{-1/2}U_j) \quad \text{with } U_j = (U_{1j}, \dots, U_{qj})',$$

$$R_{2j} = \frac{1}{2}n^{-1}\sum_{i=1}^{n}(\tilde{\beta} - \beta)'\frac{\partial^2}{\partial\beta\partial\beta'}\phi_j(F(X_i;\beta))\Big|_{\beta=\xi}(\tilde{\beta} - \beta),$$

$$Z_j = (\tilde{\beta} - \beta)'E_\beta\frac{\partial}{\partial\beta}\phi_j(F(X;\beta)),$$

$$a_k = (k-1)n^{-1}\log n,$$

$$v_{tj} = \operatorname{var}_\beta\frac{\partial}{\partial\beta_t}\phi_j(F(X;\beta)),$$

$$y_{nj} = \{\zeta^2(1-\varepsilon)(k-1)(4kq\rho_1^2)^{-1}n\}^{1/2},$$

where ε, ζ and ξ are defined below; compare Lemma A.1 and its proof.

Throughout Section A.1 *it is always assumed that the regularity conditions* (R1)–(R4) *hold*.

The proof of Theorem 2.1 consists of several lemmas. If $d(n)$ is bounded, Theorem 2.1 is easily obtained. We therefore assume w.l.o.g. $d(n) \to \infty$ as $n \to \infty$.

Let $\beta \in B_0$ and $\eta = \eta(\beta)$ as given in (R3).

LEMMA A.1. *Assume* (S1) *and* (S2). *For each* $0 < \varepsilon < 1$ *and* $0 < \zeta < 1$, *we have on the set* $\|\tilde{\beta} - \beta\| < \eta(\beta)$,

$$\left\{\|Y_n(\tilde{\beta}) - Y_n(\beta)\| \geq \sqrt{(1-\varepsilon)a_k}\right\} \subset \left\{\left(\sum_{j=1}^{k} Z_j^2\right)^{1/2} \geq (1-\zeta)\sqrt{(1-\varepsilon)a_k}\right\}$$
$$\cup \left\{\left(\sum_{j=1}^{k} R_{1j}^2\right)^{1/2} \geq \tfrac{1}{2}\zeta\sqrt{(1-\varepsilon)a_k}\right\}$$
$$\cup \left\{\left(\sum_{j=1}^{k} R_{2j}^2\right)^{1/2} \geq \tfrac{1}{2}\zeta\sqrt{(1-\varepsilon)a_k}\right\}.$$

PROOF. By the Taylor expansion we get on the set $\|\tilde{\beta} - \beta\| < \eta(\beta)$, for some ξ between $\tilde{\beta}$ and β,

$$\phi_j(F(X_i;\tilde{\beta})) - \phi_j(F(X_i;\beta)) = (\tilde{\beta}-\beta)'\frac{\partial}{\partial\beta}\phi_j(F(X_i;\beta))$$
$$+ \frac{1}{2}(\tilde{\beta}-\beta)'\frac{\partial^2}{\partial\beta\partial\beta'}\phi_j(F(X_i;\beta))\Big|_{\beta=\xi}(\tilde{\beta}-\beta).$$

Hence

$$\bar{\phi}_j(\tilde{\beta}) - \bar{\phi}_j(\beta) = Z_j + R_{1j} + R_{2j} \tag{A.1}$$

and the result follows from the triangle inequality. □

LEMMA A.2. *Assume* (S1), (S2), (D2). *For each* $0 < \varepsilon < 1$ *and* $0 < \zeta < 1$ *we have*, *for sufficiently large* n,

$$\sum_{k=2}^{d(n)} P_\beta\left(\left\{\sum_{j=1}^{k} R_{2j}^2\right\}^{1/2} \geq \tfrac{1}{2}\zeta\sqrt{(1-\varepsilon)a_k}\right) \leq c_1 d(n) n^{-c_2\rho_1^2}.$$

PROOF. In view of (R3), (S1) and (S2), there exists a constant $c_6 > 0$ such that on the set $\|\tilde{\beta} - \beta\| < \eta(\beta)$,

$$|R_{2j}| \leq c_6\|\tilde{\beta}-\beta\|^2 j^m \tag{A.2}$$

with $m = \max(m_1, m_2)$. Application of (R4) and (D2) yields, for sufficiently large n,

$$\sum_{k=2}^{d(n)} P_\beta\left(\left\{\sum_{j=1}^{k} R_{2j}^2\right\}^{1/2} \geq \tfrac{1}{2}\zeta\sqrt{(1-\varepsilon)a_k}\right)$$
$$\leq \sum_{k=2}^{d(n)} P_\beta\left(c_6\|\tilde{\beta}-\beta\|^2 k^{m+1/2} \geq \tfrac{1}{2}\zeta\sqrt{(1-\varepsilon)a_k}\right)$$

$$= \sum_{k=2}^{d(n)} P_\beta\Big(\sqrt{n}\|\tilde{\beta}-\beta\| \geq \Big\{nc_6^{-1}k^{-m-1/2}\tfrac{1}{2}\zeta\sqrt{(1-\varepsilon)a_k}\Big\}^{1/2}\Big)$$

$$\leq \sum_{k=2}^{d(n)} P_\beta\big(\sqrt{n}\|\tilde{\beta}-\beta\| \geq \rho_1\sqrt{\log n}\big)$$

$$\leq d(n)c_1 n^{-c_2\rho_1^2}. \qquad \square$$

LEMMA A.3. *Assume* (S1). *For each* $t = 1, \ldots, q$,

$$\sup_j \left| E_\beta \frac{\partial}{\partial \beta_t} \phi_j(F(X;\beta))\right|^2 \leq \operatorname{var}_\beta \frac{\partial}{\partial \beta_t} \log f(X;\beta).$$

PROOF. All expectations and (co)variances in this proof are with respect to P_β. Since for $j \geq 1$ by orthonormality,

$$E\phi_j(F(X;\beta)) = 0,$$

we get

$$\begin{aligned}
\operatorname{cov}\Big(\phi_j(F(X;\beta)), \frac{\partial}{\partial\beta_t}\log f(X;\beta)\Big) &= E\Big\{\phi_j(F(X;\beta))\frac{\partial}{\partial\beta_t}\log f(X;\beta)\Big\}\\
&= \int \phi_j(F(x;\beta))\frac{\partial}{\partial\beta_t} f(x;\beta)\,dx\\
&= \int \frac{\partial}{\partial\beta_t}\{\phi_j(F(x;\beta))f(x;\beta)\}\,dx\\
&\quad - E\Big\{\frac{\partial}{\partial\beta_t}\phi_j(F(X;\beta))\Big\}.
\end{aligned}$$

By the dominated convergence theorem [cf. (R1), (R3), (S1)], we have

$$\int \frac{\partial}{\partial\beta_t}\{\phi_j(F(x;\beta))f(x;\beta)\}\,dx = \frac{\partial}{\partial\beta_t} E\phi_j(F(X;\beta)) = 0$$

and hence

$$\begin{aligned}
\left|E\frac{\partial}{\partial\beta_t}\phi_j(F(X;\beta))\right|^2 &= \left|\operatorname{cov}\Big(\phi_j(F(X;\beta)), \frac{\partial}{\partial\beta_t}\log f(X;\beta)\Big)\right|^2\\
&\leq \operatorname{var}\phi_j(F(X;\beta))\operatorname{var}\frac{\partial}{\partial\beta_t}\log f(X;\beta)\\
&= \operatorname{var}\frac{\partial}{\partial\beta_t}\log f(X;\beta)
\end{aligned}$$

and the result follows. $\square$

LEMMA A.4. *Assume* (S1), (S2) *and* (D2). *For each* $0 < \varepsilon < 1$ *and* $0 < \zeta < 1$ *we have, for sufficiently large* n,

$$\sum_{k=2}^{d(n)} P_\beta\left(\left\{\sum_{j=1}^{k} R_{1j}^2\right\}^{1/2} \geq \tfrac{1}{2}\zeta\sqrt{(1-\varepsilon)a_k}\right) \leq c_1\left[\{d(n)\}^{-1} + d(n)n^{-c_2\rho_1^2}\right].$$

PROOF. By (R3) and (S1) we get, for some constant $c_7 > 0$,

$$\sup_{x\in\mathbb{R}}\left|\frac{\partial}{\partial\beta_t}\phi_j(F(x;\beta))\right| = \sup_{x\in\mathbb{R}}\left|\phi'_j(F(x;\beta))\frac{\partial}{\partial\beta_t}F(x;\beta)\right| \leq c_7\, j^{m_1}.$$

Application of Kolmogorov's exponential inequality [cf., e.g., Shorack and Wellner (1986), page 855] yields

$$P_\beta(U_{tj}^2 v_{tj}^{-1} > \lambda^2) < 2\exp\left\{-\tfrac{1}{2}\lambda h(\lambda)\right\}$$

with

$$h(\lambda) = \begin{cases} \lambda\left\{1 - \frac{1}{2}\lambda(nv_{tj})^{-1/2}c_7 j^{m_1}\right\}, & \text{if } \lambda \leq c_7^{-1} j^{-m_1}(nv_{tj})^{1/2}, \\ \frac{1}{2}(nv_{tj})^{1/2}c_7^{-1} j^{-m_1}, & \text{if } \lambda \geq c_7^{-1} j^{-m_1}(nv_{tj})^{1/2}. \end{cases}$$

In view of (R4) we obtain

$$\begin{aligned}
&\sum_{k=2}^{d(n)} P_\beta\left(\left\{\sum_{j=1}^{k} R_{1j}^2\right\}^{1/2} \geq \tfrac{1}{2}\,\zeta\sqrt{(1-\varepsilon)a_k}\right) \\
&\quad= \sum_{k=2}^{d(n)} P_\beta\left(\sum_{j=1}^{k} R_{1j}^2 \geq \tfrac{1}{4}\,\zeta^2(1-\varepsilon)a_k\right) \\
&\quad\leq \sum_{k=2}^{d(n)} P_\beta\left(n^{-1}\|\tilde\beta-\beta\|^2 \sum_{j=1}^{k}\sum_{t=1}^{q} U_{tj}^2 \geq \tfrac{1}{4}\,\zeta^2(1-\varepsilon)a_k\right) \\
&\quad\leq \sum_{k=2}^{d(n)} P_\beta\left(\sqrt{n}\|\tilde\beta-\beta\| \geq \rho_1\sqrt{\log n}\right) + \sum_{k=2}^{d(n)} P_\beta\left(\sum_{j=1}^{k}\sum_{t=1}^{q} U_{tj}^2 \geq kqy_{nj}^2\right) \qquad \text{(A.3)} \\
&\quad\leq d(n)c_1 n^{-c_2\rho_1^2} + \sum_{k=2}^{d(n)}\sum_{j=1}^{k}\sum_{t=1}^{q} P_\beta(U_{tj}^2 \geq y_{nj}^2) \\
&\quad\leq c_1 d(n)n^{-c_2\rho_1^2} + \sum_{k=2}^{d(n)}\sum_{j=1}^{k}\sum_{t=1}^{q} 2\exp\left\{-\tfrac{1}{2}v_{tj}^{-1/2}y_{nj}h(v_{tj}^{-1/2}y_{nj})\right\}.
\end{aligned}$$

If $\lambda \leq c_7^{-1} j^{-m_1}(nv_{tj})^{1/2}$, we have $h(\lambda) \geq \frac{1}{2}\lambda$. Therefore, if $v_{tj}^{-1/2}y_{nj} \leq c_7^{-1} j^{-m_1}(nv_{tj})^{1/2}$, we get, in view of (D2), for some positive constants c_8 and c_9 and sufficiently large n,

$$\begin{aligned}
\tfrac{1}{2}v_{tj}^{-1/2}y_{nj}h(v_{tj}^{-1/2}y_{nj}) &\geq \tfrac{1}{4}v_{tj}^{-1}y_{nj}^2 \geq c_8 n(v_{tj})^{-1} \geq c_9 nj^{-2m_1} \geq c_9 n\{d(n)\}^{-2m_1} \\
&\geq 4\log d(n).
\end{aligned}$$

Again in view of (D2), if $v_{tj}^{-1/2} y_{nj} \geq c_7^{-1} j^{-m_1}(nv_{tj})^{1/2}$, we have, by definition of h, for some positive constants c_{10} and c_{11} and sufficiently large n,

$$\begin{aligned}\tfrac{1}{2} v_{tj}^{-1/2} y_{nj} h(v_{tj}^{-1/2} y_{nj}) &\geq \tfrac{1}{4} y_{nj} n^{1/2} c_7^{-1} j^{-m_1} \geq c_{10} n j^{-m_1} \\ &\geq c_{10} n \{d(n)\}^{-m_1} \geq c_{11} n^{1/2} \geq 4 \log d(n).\end{aligned}$$

Therefore, in any case, we have, for sufficiently large n,

$$\begin{aligned}&\sum_{k=2}^{d(n)} \sum_{j=1}^{k} \sum_{t=1}^{q} 2 \exp\{-\tfrac{1}{2} v_{tj}^{-1/2} y_{nj} h(v_{tj}^{-1/2} y_{nj})\} \\ &\qquad \leq 2q\{d(n)\}^2 \exp\{-4 \log d(n)\} \leq c_1 \{d(n)\}^{-1},\end{aligned}$$

which together with (A.3) completes the proof. □

LEMMA A.5. *Assume* (S1). *Let* K *be fixed. For each* $0 < \varepsilon < 1$ *and* $0 < \zeta < 1$, *we have*

$$\lim_{n\to\infty} \sum_{k=2}^{K} P_\beta\left(\left\{\sum_{j=1}^{k} Z_j^2\right\}^{1/2} \geq (1-\zeta)\sqrt{(1-\varepsilon)a_k}\right) = 0.$$

PROOF. By definition of Z_j and in view of Lemma A.3, we get

$$Z_j^2 \leq \|\tilde{\beta} - \beta\|^2 \|E_\beta \frac{\partial}{\partial \beta} \phi_j(F(X;\beta))\|^2 \leq \|\tilde{\beta} - \beta\|^2 b^2$$

and hence, for each $2 \leq k \leq K$,

$$\begin{aligned}&P_\beta\left(\left\{\sum_{j=1}^{k} Z_j^2\right\}^{1/2} \geq (1-\zeta)\sqrt{(1-\varepsilon)a_k}\right) \\ &\qquad \leq P_\beta\left(\sqrt{n}\|\tilde{\beta} - \beta\| \geq \sqrt{n} k^{-1/2} b^{-1} (1-\zeta)\sqrt{(1-\varepsilon)a_k}\right),\end{aligned} \tag{A.4}$$

which tends to zero if $n \to \infty$ in view of (R4). □

LEMMA A.6. *Assume* (S1). *Let* $\rho_1 b \geq 1$. *For any* $c < c_2 b^{-2}$ *there exist* $0 < \varepsilon < 1$, $0 < \zeta < 1$ *and* K *such that, for all* $n \geq n_1$ *with* n_1 *from* (R4),

$$\sum_{k=K}^{d(n)} P_\beta\left(\left\{\sum_{j=1}^{k} Z_j^2\right\}^{1/2} \geq (1-\zeta)\sqrt{(1-\varepsilon)a_k}\right) \leq c_1 d(n) n^{-c}. \tag{A.5}$$

Let $\rho_1 b < 1$. *There exist* $0 < \varepsilon < 1$, $0 < \zeta < 1$ *and* K *such that, for all* $n \geq n_1$,

$$\sum_{k=K}^{d(n)} P_\beta\left(\left\{\sum_{j=1}^{k} Z_j^2\right\}^{1/2} \geq (1-\zeta)\sqrt{(1-\varepsilon)a_k}\right) \leq c_1 d(n) n^{-c_2 \rho_1^2}. \tag{A.6}$$

PROOF. Let $\rho_1 b \geq 1$ and let $c < c_2 b^{-2}$. Then there exist $0 < \varepsilon < 1$, $0 < \zeta < 1$ and K such that

$$c_2 b^{-2}(1-\zeta)^2(1-\varepsilon)K^{-1}(K-1) > c.$$

Since $(1-\zeta)\sqrt{1-\varepsilon}\{(K-1)/K\}^{1/2}b^{-1} \leq \rho_1$, we have by (A.4) and (R4), for all $k \geq K$ and $n \geq n_1$,

$$P_\beta\left(\left\{\sum_{j=1}^{k} Z_j^2\right\}^{1/2} \geq (1-\zeta)\sqrt{(1-\varepsilon)a_k}\right)$$
$$\leq c_1 \exp\{-c_2 b^{-2}(1-\zeta)^2(1-\varepsilon)k^{-1}(k-1)\log n\} \leq c_1 \exp(-c \log n)$$

and (A.5) easily follows.

Now let $\rho_1 b < 1$. Then there exist $0 < \varepsilon < 1$, $0 < \zeta < 1$ and K such that for all $k \geq K$, $\sqrt{n}k^{-1/2}b^{-1}(1-\zeta)\sqrt{(1-\varepsilon)a_k} \geq \rho_1\sqrt{\log n}$ and hence, again by (A.4) and (R4), for all $n \geq n_1$,

$$P_\beta\left(\left\{\sum_{j=1}^{k} Z_j^2\right\}^{1/2} \geq (1-\zeta)\sqrt{(1-\varepsilon)a_k}\right) \leq P_\beta\big(\sqrt{n}\|\tilde{\beta}-\beta\| \geq \rho_1\sqrt{\log n}\big)$$
$$\leq c_1 n^{-c_2\rho_1^2}$$

and (A.6) easily follows. □

PROOF OF THEOREM 2.1. Let $\rho_1 b \geq 1$. In view of (D3) there exists $c < c_2 b^{-2}$ such that $d(n) = o(n^c)$. Choose $0 < \varepsilon < 1$, $0 < \zeta < 1$ and K such that (A.5) holds. In view of Lemmas A.1, A.2, A.4, A.6 we get, for sufficiently large n,

$$\begin{aligned}
&\sum_{k=2}^{d(n)} P_\beta\big(\|Y_n(\tilde{\beta}) - Y_n(\beta)\| \geq (1-\varepsilon)\{(k-1)n^{-1}\log n\}^{1/2}\big)\\
&\quad\leq d(n)P_\beta\big(\|\tilde{\beta}-\beta\| \geq \eta(\beta)\big)\\
&\qquad+\sum_{k=2}^{K} P_\beta\left(\left\{\sum_{j=1}^{k} Z_j^2\right\}^{1/2} \geq (1-\zeta)\sqrt{(1-\varepsilon)a_k}\right)\\
&\qquad+c_1 d(n)n^{-c} + 2c_1 d(n)n^{-c_2\rho_1^2} + c_1\{d(n)\}^{-1}.
\end{aligned}\tag{A.7}$$

The first term on the right-hand side of (A.7) tends to 0 by (R4) and (D3), the second term by Lemma A.5, the third and fourth terms by (D3) (note that $c_2\rho_1^2 \geq c_2 b^{-2} > c$), while the last term tends to 0 by our assumption $d(n) \to \infty$, which we could make w.l.o.g.

The case $\rho_1 b < 1$ is similar. This completes the proof of Theorem 2.1. □

A.2. *Proof of Theorem* 3.1. Let $\beta \in B_0$ and $\eta(\beta)$ be as given in (R3). In view of the $\sqrt{n}$-consistency of $\tilde{\beta}$, we have

$$P_\beta(\|\tilde{\beta}-\beta\| > \eta(\beta)) \to 0$$

and therefore in the following we restrict attention to the set $\|\tilde{\beta}-\beta\| < \eta(\beta)$.

As in the proof of Lemma A.1 [cf. (A.1)], we write

$$\bar{\phi}_j(\tilde{\beta}) = \bar{\phi}_j(\beta) + Z_j + R_{1j} + R_{2j}.$$

Since $\sqrt{n}(\tilde{\beta} - \beta) = O_{P_\beta}(1)$ and since $n^{-1/2}U_j = o_{P_\beta}(1)$ by the law of large numbers, we get

$$\sqrt{n}R_{1j} = \sqrt{n}(\tilde{\beta} - \beta)' n^{-1/2} U_j \to_{P_\beta} 0.$$

As in the proof of Lemma A.2 [cf. (A.2)], we have

$$\sqrt{n}|R_{2j}| \le \sqrt{n} c_6 \|\tilde{\beta} - \beta\|^2 j^m$$

and hence, again by the $\sqrt{n}$-consistency of $\tilde{\beta}$,

$$\sqrt{n}R_{2j} \to_{P_\beta} 0.$$

Further note that by definition [cf. (3.1)],

$$(Z_1, \dots, Z_k)' = -(\tilde{\beta} - \beta)' I_\beta$$

and therefore

$$\sqrt{n}Y_n(\tilde{\beta}) = \sqrt{n}Y_n(\beta) - \sqrt{n}(\tilde{\beta} - \beta)' I_\beta + o_{P_\beta}(1). \tag{A.8}$$

By the Taylor expansion we get, for some ξ between $\tilde{\beta}$ and β,

$$\begin{aligned}\frac{\partial}{\partial \beta_t} \log f(X_i; \tilde{\beta}) &= \frac{\partial}{\partial \beta_t} \log f(X_i; \beta) + (\tilde{\beta} - \beta)' E_\beta \left(\frac{\partial^2}{\partial \beta \partial \beta_t} \log f(X_i; \beta) \right) \\ &\quad + (\tilde{\beta} - \beta)' \left\{ \frac{\partial^2}{\partial \beta \partial \beta_t} \log f(X_i; \xi) - E_\beta \left(\frac{\partial^2}{\partial \beta \partial \beta_t} \log f(X_i; \beta) \right) \right\}.\end{aligned}$$

By (R2),

$$n^{-1} \sum_{i=1}^n \left\{ \frac{\partial^2}{\partial \beta_u \partial \beta_t} \log f(X_i; \xi) - E_\beta \frac{\partial^2}{\partial \beta_u \partial \beta_t} \log f(X_i; \beta) \right\} \to_{P_\beta} 0$$

and therefore [cf. (3.1) and (3.5)],

$$C_n(\tilde{\beta}) = C_n(\beta) - \{ I_{\beta\beta} + o_{P_\beta}(1) \} (\tilde{\beta} - \beta). \tag{A.9}$$

In view of (R1), (R3), (S1) and the continuity of ϕ_j', the continuity of I_β is easily obtained by the dominated convergence theorem. The continuity of I_β and $I_{\beta\beta}$ [cf. (R2)], and the fact that $\tilde{\beta}$ is $\sqrt{n}$-consistent and that

$$\sqrt{n}C_n(\beta) = O_{P_\beta}(1),$$

imply, together with (A.9),

$$\sqrt{n} I'_{\tilde{\beta}} I^{-1}_{\tilde{\beta}\tilde{\beta}} C_n(\tilde{\beta}) = \sqrt{n} I'_\beta I^{-1}_{\beta\beta} C_n(\beta) - \sqrt{n} I'_\beta (\tilde{\beta} - \beta) + o_{P_\beta}(1)$$

and hence [cf. (3.4) and (A.8)],

$$\sqrt{n}\tilde{Y}_n(\tilde{\beta}) = \sqrt{n}Y_n(\beta) - \sqrt{n} I'_\beta I^{-1}_{\beta\beta} C_n(\beta) + o_{P_\beta}(1).$$

Also by the continuity of I_β, $I_{\beta\beta}$ and the convergence of $\tilde\beta$ to β, we have [cf. (3.1)]

$$R(\tilde\beta) = R(\beta) + o_{P_\beta}(1).$$

The multivariate central limit theorem now gives

$$\sqrt{n}\tilde Y_n(\tilde\beta) \to N(0; I - A_\beta) \quad \text{with } A_\beta = I'_\beta I^{-1}_{\beta\beta} I_\beta.$$

Since

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

we get

$$(I - A_\beta)(I + R(\beta)) = I + R(\beta) - A_\beta - A_\beta R(\beta) = I.$$

The proof is now easily completed by invoking the definition of $W_k(\tilde\beta)$ [cf. (3.3)]. □

A.3. *Condition* (R4) *for* $\tilde\beta = (\bar X, \{n^{-1}\sum_{i=1}^n (X_i - \bar X)^2\}^{1/2})'$. Let $\rho > 0$ and $r = \rho\{\log n\}^{1/2}$. All expectations and variances in this subsection are with respect to P_β. Writing $Y_i = (X_i - EX)/\sqrt{\operatorname{var} X}$, we have

$$\text{(A.10)}\qquad \begin{aligned} &P_\beta\big(\sqrt{n}\|\tilde\beta - \beta\| \ge r\big)\\ &\quad = P_\beta\left(\left\|\left(\bar Y, \left\{n^{-1}\sum_{i=1}^n (Y_i - \bar Y)^2\right\}^{1/2}\right)' - (0,1)'\right\|\sqrt{n} \ge r/\sqrt{\operatorname{var} X}\right),\end{aligned}$$

where $\beta = (EX, \sqrt{\operatorname{var} X})'$ by $\sqrt{n}$-consistency of $\tilde\beta$. Note that we do not restrict to location/scale families.

Assume $E|X_i|^{4+\varepsilon} < \infty$ for some $\varepsilon > 0$, implying $E|Y_i^2 - 1|^{2+(1/2)\varepsilon} < \infty$. Denoting by Φ the standard normal distribution function we have [cf., e.g., Petrov (1975), page 252],

$$P_\beta\left(\sum_{i=1}^n (Y_i^2 - 1)\{\operatorname{var}(Y_i^2)\}^{-1/2} \ge x\sqrt{n}\right) = \{1 - \Phi(x)\}(1 + o(1))$$

for $0 \le x \le \{(\varepsilon/4)\log n\}^{1/2}$. Since $r = \rho\{\log n\}^{1/2}$, we get, for $\rho \le \{(\varepsilon/2)\operatorname{var}(X^2)\}^{1/2}$,

$$\text{(A.11)}\qquad P_\beta\left(\sum_{i=1}^n (Y_i^2 - 1) \ge r\sqrt{n}/\sqrt{2\operatorname{var} X}\right) \le \tfrac{1}{4}c_1\exp(-c_2 r^2)$$

for some positive constants c_1, c_2 and n large enough. Similarly, for $\rho \le [(\varepsilon/2)\min\{\operatorname{var} X, \operatorname{var}(X^2)\}]^{1/2} = \rho_1$, say,

$$\text{(A.12)}\qquad P_\beta\left(\sum_{i=1}^n (Y_i^2 - 1) \le -r\sqrt{n}/\sqrt{2\operatorname{var} X}\right) \le \tfrac{1}{4}c_1\exp(-c_2 r^2)$$

and

$$P_\beta\left(\left|\sum_{i=1}^{n} Y_i\right| \geq r\sqrt{n}/\sqrt{2\operatorname{var} X}\right) \leq \tfrac{1}{2}c_1 \exp(-c_2 r^2). \tag{A.13}$$

It is seen that for sufficiently large n and for all $0 < \rho \leq \rho_1$,

$$\left|\sum_{i=1}^{n}(Y_i^2 - 1)\right| < r\sqrt{n}/\sqrt{2\operatorname{var} X} \quad \text{and} \quad \left|\sum_{i=1}^{n} Y_i\right| < r\sqrt{n}/\sqrt{2\operatorname{var} X}$$

imply

$$\left\|\left(\bar{Y}, \left\{n^{-1}\sum_{i=1}^{n}(Y_i - \bar{Y})^2\right\}^{1/2}\right)' - (0,1)'\right\|\sqrt{n} < r/\sqrt{\operatorname{var} X}.$$

In combination with (A.10)–(A.13), the result is obtained.

The larger c_2, the larger the region for $d(n)$ in (D3). Therefore, one might look for the "best" c_2 assuming existence of a certain absolute moment of X_i. The following remark gives such optimal c_2. It requires a slightly more delicate proof than above.

REMARK A.1. Let λ be the largest characteristic root of the matrix

$$\begin{pmatrix} 1 & \operatorname{cov}(Y_i, \frac{1}{2}Y_i^2) \\ \operatorname{cov}(Y_i, \frac{1}{2}Y_i^2) & \operatorname{var}(\frac{1}{2}Y_i^2) \end{pmatrix}.$$

Assume that, for some $\varepsilon > 0$, a moment of order $4 + \varepsilon + 2\rho_1^2(\operatorname{var} X)^{-1}\lambda^{-1}$ exists for X_i. Using formula (101) on page 343 of Rubin and Sethuraman (1965), it can be shown that (R4) holds for any $c_2 < \frac{1}{2}(\operatorname{var} X)^{-1}\lambda^{-1}$ and ρ_1 from the above moment condition. If all absolute moments of X_i exist, ρ_1 may be taken as large as one wants and hence (D3) states $d(n) = o(n^c)$ for some $c < \frac{1}{2}(\operatorname{var} X)^{-1}\lambda^{-1}b^{-2}$. In particular, if X is normally distributed, $\lambda = 1$ and hence (D3) reduces to $d(n) = o(n^c)$ for some $c < \frac{1}{6}$.

Acknowledgment. The authors thank the referees for their constructive remarks.

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.

BARRON, A. R. and SHEU, C. (1991). Approximation of density functions by sequences of exponential families. *Ann. Statist.* **19** 1347–1369.

BARTON, D. E. (1985). Neyman's and other smooth goodness-of-fit tests. In *Encyclopedia of Statistical Sciences* **6** 230–232. Wiley, New York.

BICKEL, P. J. and DOKSUM, K. A. (1977). *Mathematical Statistics, Basic Ideas and Selected Topics.* Holden-Day, San Francisco.

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.) 51–57. North-Holland, Amsterdam.

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

BOGDAN, M. and LEDWINA, T. (1996). Testing uniformity via log-spline modeling. *Statistics* **28** 131–157.

BOULERICE, B. and DUCHARME, G. R. (1995). A note on smooth tests of goodness of fit for location-scale families. *Biometrika* **82** 437–438.

BURKE, M. D. and GOMBAY, E. (1988). On goodness-of-fit and the bootstrap. *Statist. Probab. Lett.* **6** 287–293.

COX, D. R. and HINKLEY, D. V. (1974). *Theoretical Statistics*. Chapman and Hall, London.

D'AGOSTINO, R. B. and STEPHENS, M. A. (1986). *Goodness-of-Fit Techniques*. Dekker, New York.

DURBIN, J. (1973). *Distribution Theory for Tests Based on the Sample Distribution Function*. SIAM, Philadelphia.

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

EUBANK, R. L., HART, J. D., SIMPSON, D. G. and STEFANSKI, L. A. (1995). Testing for additivity in nonparametric regression. *Ann. Statist.* **23** 1896–1920.

EUBANK, R. L. and LARICCIA, V. N. (1992). Asymptotic 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** 674–688.

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

HÁJEK, J. (1972). Local asymptotic minimax and admissibility in estimation. *Proc. Sixth Berkeley Symp. Math. Statist. Probab.* **1** 175–194. Univ. California Press, Berkeley.

HALL, W. J. and MATHIASON, D. J. (1990). On large-sample estimation and testing in parametric models. *Internat. Statist. Rev.* **58** 77–97.

INGLOT, T., KALLENBERG, W. C. M. and LEDWINA, T. (1992). Strong moderate deviation theorems. *Ann. Probab.* **20** 987–1003.

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

INGLOT, T. and LEDWINA, T. (1996). Asymptotic optimality of data driven Neyman's tests. *Ann. Statist.* **24** 1982–2019.

JANSSEN, A. (1995). Principal component decomposition of non-parametric tests. *Probab. Theory Related Fields* **101** 193–209.

JAVITZ, H. S. (1975). Generalized smooth tests of goodness of fit, independence, and equality of distributions. Ph.D. dissertation, Univ. California, Berkeley.

JUREČKOVÁ, J., KALLENBERG, W. C. M. and VERAVERBEKE, N. (1988). Moderate and Cramér-type large deviation theorems for M-estimators. *Statist. Probab. Lett.* **6** 191–199.

KAIGH, W. D. (1992). EDF and EQF orthogonal component decompositions and tests of uniformity. *J. Nonparametr. Statist.* **1** 313–334.

KALLENBERG, W. C. M. (1983). On moderate deviation theory in estimation. *Ann. Statist.* **11** 498–514.

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. (1995c). Data driven smooth tests for composite hypotheses: comparison of powers. Unpublished manuscript.

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

KHAMALADZE, E. V. (1981). Martingale approach to the theory of goodness of fit tests. *Theory Probab. Appl.* **26** 240–257.

KHAMALADZE, E. V. (1982). Some applications of the theory of martingales to statistics. *Russian Math. Surveys* **37** 215–237.

KHAMALADZE, E. V. (1993). Goodness of fit problem and scanning innovation martingales. *Ann. Statist.* **21** 798–829.

KIM, J. (1992). Testing goodness-of-fit via order selection criteria. Ph.D. dissertation, Texas A&M Univ.

LE CAM, L. (1956). On the asymptotic theory of estimation and testing hypotheses. *Proc. Third Berkeley Symp. Math. Statist. Probab.* **1** 129–156. Univ. California Press, Berkeley.

LEDWINA, T. (1994). Data-driven version of Neyman's smooth test of fit. *J. Amer. Statist. Assoc.* **89** 1000–1005.

LOCKE, C. and SPURRIER, J. D. (1978). On tests of uniformity. *Comm. Statist. Theory Methods* **7** 241–258.

MASSART, P. (1990). The tight constant in the Dvoretzky–Kiefer–Wolfowitz inequality. *Ann Probab.* **18** 1269–1283.

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

MILLER, F. L. and QUESENBERRY, C. P. (1979). Power studies of tests for uniformity II. *Comm. Statist. Simulation Comput.* **8** 271–290.

NEUHAUS, G. (1976). Asymptotic power properties of the Cramér–von Mises test under contiguous alternatives. *J. Multivariate Anal.* **6** 95–110.

NEUHAUS, G. (1979). Asymptotic theory of goodness of fit when parameters are present: a survey. *Math. Operationsforsch. Statist. Ser. Statist.* **10** 479–494.

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

PETROV, V. V. (1975). *Sums of Independent Random Variables*. Springer, New York.

PROHOROV, A. V. (1973). On sums of random vectors. *Theory Probab. Appl.* **18** 186–188.

QUESENBERRY, C. P. and MILLER, F. L. (1977). Power studies of some tests for uniformity. *J. Statist. Comput. Simulation* **5** 169–191.

RAYNER, J. C. W. and BEST, D. J. (1986). Neyman-type smooth tests for location-scale families. *Biometrika* **73** 437–446.

RAYNER, J. C. W. and BEST, D. J. (1989). *Smooth Tests of Goodness of Fit.* Oxford Univ. 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.

RUBIN, H. and SETHURAMAN, J. (1965). Probabilities of moderate deviations. *Sankhyā Ser. A* **27** 325–346.

SANSONE, G. (1959). *Orthogonal Functions*. Interscience, New York.

SEN, P. K. and SINGER, J. M. (1993). *Large Sample Methods in Statistics*. Chapman and Hall, New York.

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

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

STONE, C. J. (1990). Large sample inference for log-spline models. *Ann. Statist.* **18** 717–741.

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.

T. INGLOT
T. LEDWINA
INSTITUTE OF MATHEMATICS
TECHNICAL UNIVERSITY OF WROCŁAW
50-370 WROCŁAW
WYBRZEŻE WYSPIAŃSKIEGO 27
POLAND

W. C. M. KALLENBERG
FACULTY OF APPLIED MATHEMATICS
UNIVERSITY OF TWENTE
P.O. BOX 217
7500 AE ENSCHEDE
THE NETHERLANDS
E-MAIL: w.c.m.kallenberg@math.utwente.nl