Data-Driven Smooth Tests When the Hypothesis Is Composite

Wilbert C. M. KALLENBERG and Teresa LEDWINA

In recent years several authors have recommended smooth tests for testing goodness of fit. However, the number of components in the smooth test statistic should be chosen well; otherwise, considerable loss of power may occur. Schwarz's selection rule provides one such good choice. Earlier results on simple null hypotheses are extended here to composite hypotheses, which tend to be of more practical interest. For general composite hypotheses, consistency of the data-driven smooth tests holds at essentially any alternative. Monte Carlo experiments on testing exponentiality and normality show that the data-driven version of Neyman's test compares well to other, even specialized, tests over a wide range of alternatives.

KEY WORDS: Goodness of fit; Monte Carlo study; Neyman's test; Schwarz's BIC criterion.

1. INTRODUCTION

In recent years there has been substantial interest in data-driven procedures for testing goodness of fit (see, e.g., Bickel and Ritov 1992; Eubank and Hart 1992; Eubank, Hart, and LaRiccia 1993; Eubank and LaRiccia 1992; Fan 1996; Inglot and Ledwina 1996; Kallenberg and Ledwina 1994, 1995a,b; Ledwina 1994). Although smooth tests are often recommended for this purpose (see, e.g., Milbrodt and Strasser 1990; Rayner and Best 1990), a wrong choice for the number k of components in the test statistic can result in a considerable loss of power. Therefore, a good procedure is needed for choosing a value for k that can be used in practice. Renewed research in the area of smooth tests shows that a deterministic procedure gives no simple answer (see Inglot, Kallenberg, and Ledwina 1994a). Hence Ledwina (1994) introduced a data-driven version of Neyman's test for testing uniformity. An important role in the 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.

To apply this procedure for testing *composite* hypotheses (e.g., for testing normality or exponentiality), Schwarz's selection rule must be extended. This is done in Section 2 by inserting an estimator of the parameters involved in the composite null hypothesis.

It turns out that the data-driven smooth test with this selection rule is *consistent* against essentially any alternative. A modification of the selection rule is also presented, which gives a consistent test but is easier to calculate.

To investigate the behavior of the tests for finite sample sizes, Monte Carlo experiments were conducted for testing *exponentiality* and *normality*. The results from these experiments are reported in Sections 4 and 5. We also analyze the power behavior of smooth tests for different fixed val-

ues of the number of components and study the behavior of Schwarz's selection rule under alternatives. The new tests work very well and are even competitive with well-known "special" tests, such as Gini's test for exponentiality and the Shapiro–Wilk's test for normality. A more detailed description of the conclusions is given at the end of Sections 4 and 5.

To apply the test in practice, critical values or p values can be simulated. One can also rely on a simple and accurate approximation of critical values and p values, which is presented in Section 3.

Schwarz's selection rule is not the only solution for choosing the dimension. However, a comparison of selection rules with other penalties in Section 6 gives sufficient support to recommend the use of Schwarz's rule for general applications. The article concludes with a real example in Section 7.

2. DATA-DRIVEN SMOOTH TESTS

2.1 Smooth Tests

Let $X_1, X_2, \ldots, X_n$ be iid random variables with density $f(x)$. We want to test the null hypothesis

$$H_0: f(x) \in \{f(x; \beta): \beta \in B\},$$

where $B \subset \mathbb{R}^q$ and $\{f(x; \beta): \beta \in B\}$ is a given family of densities. Of particular interest are the problems of testing for exponentiality wherein, for $\beta > 0$

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

and testing for normality where $f(x; \beta) = (\sqrt{2\pi}\sigma)^{-1} \exp\{-\frac{1}{2}(x-\mu)^2/\sigma^2\}$, for $\beta = (\mu, \sigma)$ with $\mu \in \mathbb{R}$ and $\sigma > 0$.

To test the null hypothesis, a sequence of exponential families is built, which extends from the null model using classes of alternatives that become larger and larger as the dimension of the exponential family grows. Testing H_0 within this exponential family then leads to a standard testing problem.

To be more precise, 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. Typical examples of $\phi_0, \phi_1, \ldots, \phi_k$ are the orthonormal Legendre polynomials on $[0, 1]$, corresponding to Neyman's smooth test and the cosine system, given by $\phi_j(x) = \sqrt{2}\cos(j\pi x)$, $j = 1, 2, \ldots$.

Now let $F(x; \beta)$ be the distribution function of X_i when β applies. For $k = 1, 2, \ldots$ define exponential families (with respect to θ) by their density

$$g(x; \theta, \beta) = \exp\{\theta \circ \phi[F(x; \beta)] - \psi_k(\theta)\}f(x; \beta), \quad (1)$$

where

$$\theta = (\theta_1, \ldots, \theta_k), \qquad \phi = (\phi_1, \ldots, \phi_k),$$

$$\psi_k(\theta) = \log \int_0^1 \exp\{\theta \circ \phi(y)\}\, dy,$$

and $\circ$ stands for the inner product in $\mathbb{R}^k$. In what follows, the Euclidean norm in $\mathbb{R}^k$ (or $\mathbb{R}^q$) is denoted by $\|\cdot\|$. When there is no confusion, the dimension k is sometimes suppressed in the notation.

Testing H_0 within the exponential family (1) means testing $H\colon \theta = 0$ against $\theta \neq 0$. An obvious test statistic for this testing problem is the score statistic (see Javitz 1975; Kopecky and Pierce 1979; Neyman 1980; Rayner and Best 1989; Thomas and Pierce 1979). Denoting by $\mathbf{I}$ the $k \times k$ identity matrix and the transpose of a matrix or vector by $'$, the score statistic is given by

$$W_k = n\mathbf{Y}'_n(\hat{\beta})\{\mathbf{I} + \mathbf{R}(\hat{\beta})\}\mathbf{Y}_n(\hat{\beta}),$$

where, writing E_β for the expected value under $f(x; \beta)$,

$$\mathbf{Y}_n(\beta) = (\bar{\phi}_1(\beta), \ldots, \bar{\phi}_k(\beta))'$$

$$= n^{-1}\sum_{i=1}^{n}(\phi_1[F(X_i; \beta)], \ldots, \phi_k[F(X_i; \beta)])',$$

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

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

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

and $\hat{\beta}$ is the maximum likelihood estimator (MLE) of β under H_0.

2.2 Selection Rules and Data-Driven Smooth Tests

To choose the "right" dimension, an extension of Schwarz's (1978) Bayesian information criterion (BIC) to

the composite hypothesis case is introduced. The likelihood of the independent random variables $X_1, \ldots, X_n$, each having density (1), is given by

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

In Schwarz's rule the maximized log-likelihood is penalized with $k/2\log n$ for dimension k. Therefore, when β is known, this yields

$$S(\beta) = \min\{k\colon 1 \leq k \leq d(n), L_k(\beta)$$
$$\geq L_j(\beta), j = 1, \ldots, d(n)\}, \quad (2)$$

where

$$L_k(\beta) = n\sup_{\theta \in \mathbb{R}^k}\{\theta \circ \mathbf{Y}_n(\beta) - \psi_k(\theta)\} - \frac{1}{2}k\log n.$$

The number $d(n)$ is the upper bound of the dimension of the exponential families under consideration. It turns out that $S(\beta)$ is stable for large $d(n)$. A more detailed discussion on the choice of $d(n)$ is given in Section 6.2.

The extension to the situation where β is an unknown nuisance parameter is obtained by inserting $\hat{\beta}$ in formula (2) to obtain

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

The maximized log-likelihood (which is in fact the log-likelihood ratio statistic for testing $H\colon \theta = 0$ against $\theta \neq 0$ when β is known) is locally equivalent to $\frac{1}{2}n\|\mathbf{Y}_n(\beta)\|^2$. This leads to the following modification $S2$ (with 2 from squared norm), which is much easier to calculate:

$$S2 = S2(\hat{\beta}) = \min\{k\colon 1 \leq k \leq d(n), n\|\mathbf{Y}_n(\hat{\beta})\|^2_{(k)}$$
$$- k\log n \geq n\|\mathbf{Y}_n(\hat{\beta})\|^2_{(j)} - j\log n, j = 1, \ldots, d(n)\},$$

where the index of the norm denotes the dimension.

The two *data-driven* smooth test statistics are now defined by

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

and

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

The null hypothesis is rejected for large values of W_S and W_{S2}.

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

Theoretical support of the tests W_S and W_{S2} is given by their *consistency*. Consider an alternative P. Under this alternative, $\hat{\beta}$ will as a rule converge to some element of B. This element will be called β. For instance, for testing exponentiality $\hat{\beta} = \bar{X}$, which converges under P to $E_P X$; therefore, in that case under the alternative P, β will refer to $E_P X$.

Table 1. 5% and 10% Critical Values of W_S and W_{S2} for Testing Exponentiality and Normality

						d(n)					
α		1	2	3	4	5	6	7	8	9	10
						Testing exponentiality					
.05	W_S	3.592	4.445	4.790	4.960	5.020	5.063	5.086	5.086	5.086	5.086
	W_{S2}	3.592	4.492	4.800	4.891	4.935	4.947	4.947	4.947	4.947	4.947
.10	W_S	2.562	2.783	2.886	2.937	2.955	2.959	2.967	2.967	2.967	2.967
	W_{S2}	2.562	2.781	2.864	2.905	2.921	2.937	2.937	2.937	2.937	2.937
						Testing normality					
.05	W_S	3.817	3.820	3.825	3.847	3.854	3.854	3.854	3.855	3.855	3.855
	W_{S2}	3.817	3.820	3.825	3.836	3.840	3.840	3.840	3.840	3.840	3.840
.10	W_S	2.725	2.726	2.728	2.745	2.750	2.757	2.757	2.757	2.757	2.757
	W_{S2}	2.725	2.726	2.728	2.735	2.734	2.734	2.734	2.734	2.734	2.734

NOTE: Each case is based on 10,000 samples; $n = 50$.

We consider alternatives P to the family $\{f(x; \beta): \beta \in B\}$ for which there exists (for the β associated with $\hat{\beta}$ and P) $K(\beta)$ such that

$$E_P \phi_1(F(X; \beta)) = \cdots = E_P \phi_{K(\beta)-1}(F(X; \beta))$$

$$= 0, E_P \phi_{K(\beta)}(F(X, \beta)) \neq 0. \quad (3)$$

Note that if (3) does not hold, then $E_P \phi_j(F(X; \beta)) = 0$ for all j, and thus essentially any alternative of interest satisfies (3).

Theorem 2.1. Under the conditions given in the Appendix, the tests based on W_S and W_{S2} are consistent against any alternative satisfying (3).

Consequently, the following corollary.

Corollary 2.1. For testing exponentiality or normality the test based on W_S and W_{S2} with $\{\phi_j\}$, the orthonormal Legendre polynomials on [0, 1] are consistent against any alternative P having finite second moment and satisfying (3) if $d(n) = o(\{n/\log n\}^{1/9})$. The same result holds if $\{\phi_j\}$ is the cosine system, in which case $d(n) = o(\{n/\log n\}^{1/4})$ suffices for testing exponentiality and for each $\varepsilon > 0$, $d(n) = o(n^{(1/6)-\varepsilon})$ suffices for testing normality.

The proof of Theorem 2.1 was provided by Inglot et al. (1994b). (A more general case is treated in Inglot et al. 1997.) The proof of Corollary 2.1 is provided in the Appendix.

3. CRITICAL VALUES

The null hypothesis corresponds to $\theta = 0$ in (1) for any dimension. Because the penalty in the selection rules increases substantially with the dimension, it may be expected that under H_0, the lowest dimension is prevalent. Indeed, it can be shown that the selection rules S and $S2$ converge under H_0 in probability to 1 (Inglot et al. 1994b). To show how well this limiting theorem describes the finite-sample

Table 2. Second-Order Approximation of $P(W_S \leq x)$ for Testing Exponentiality With x the 5% Critical Value From Table 1

x	4.445	4.790	4.960	5.020	5.063	5.086
Appr.	.942	.947	.949	.950	.950	.950

behavior, the null distribution of the selection rules was simulated for testing exponentiality and normality. In all of the simulations we take smooth tests and selection rules, with the ϕ_j's being the orthonormal Legendre polynomials on [0, 1]. The simulations are performed in a similar way to that described by Ledwina (1994). Every Monte Carlo experiment reported here was repeated 10,000 times. Hence the standard deviation of the simulated powers and probabilities concerning the selection rules does not exceed $(40,000)^{-1/2} = .005$.

Let $n = 50$ and $d(50) = 10$. Then for testing exponentiality, we found the empirical relative frequencies of choosing $s = 1, 2,$ and 3 to be .96, .03, and .01. Similarly, for testing normality, these probabilities were .99, .00, and .01.

The same holds for $S2$. Thus it appears that the selection rules do indeed concentrate on dimension 1, the concentration in the case of testing normality being even stronger than that for testing exponentiality.

Under classical regularity conditions, we have under H_0 that

$$\sqrt{n} \mathbf{Y}_n(\hat{\beta}) \xrightarrow{D} N(0, \mathbf{I} - \mathbf{A}_\beta), \quad \text{with} \quad \mathbf{A}_\beta = \mathbf{I}'_\beta \mathbf{I}_{\beta\beta}^{-1} \mathbf{I}_\beta.$$

Because

$$\mathbf{A}_\beta = \mathbf{I}'_\beta \mathbf{I}_{\beta\beta}^{-1} (\mathbf{I}_{\beta\beta} - \mathbf{I}_\beta \mathbf{I}'_\beta)(\mathbf{I}_{\beta\beta} - \mathbf{I}_\beta \mathbf{I}'_\beta)^{-1} \mathbf{I}_\beta$$

$$= \mathbf{R}(\beta) - \mathbf{A}_\beta \mathbf{R}(\beta),$$

we get

$$(\mathbf{I} - \mathbf{A}_\beta)(\mathbf{I} + \mathbf{R}(\beta)) = \mathbf{I} + \mathbf{R}(\beta) - \mathbf{A}_\beta - \mathbf{A}_\beta \mathbf{R}(\beta) = \mathbf{I}.$$

It now follows that under H_0, $W_k \xrightarrow{D} \chi_k^2$, where χ_k^2 denotes a random variable with a chi-squared distribution with k df. Hence, due to the concentration of S and $S2$ on dimension 1, we obtain $W_S \xrightarrow{D} \chi_1^2$ and $W_{S2} \xrightarrow{D} \chi_1^2$. Therefore, one might expect the critical values to be close to those of a chi-squared random variable with 1 df. A summary of some simulated critical values for testing exponentiality and normality with $n = 50$ is given in Table 1.

Although the results are satisfactory for testing normality, for testing exponentiality, substantial differences occur between the simulated and asymptotic critical values. In view of the higher concentration on dimension 1 for S and

Table 3. Estimated $P(S(\hat{\beta}) = s)$ and $P(S2(\hat{\beta}) = s)$ Under Alternatives (in %) When Testing Exponentiality

		s						
Alternative		1	2	3	4	5	6	7–10
e1	$P(S(\hat{\beta}) = s)$	56	40	2	2			
	$P(S2(\hat{\beta}) = s)$	62	36	2				
e2	$P(S(\hat{\beta}) = s)$	58	14	24	2	2		
	$P(S2(\hat{\beta}) = s)$	64	16	19	1			
e3	$P(S(\hat{\beta}) = s)$	69	26	2	2			
	$P(S2(\hat{\beta}) = s)$	64	27	2	4	1	1	2
e4	$P(S(\hat{\beta}) = s)$	16	72	6	5	1	1	
	$P(S2(\hat{\beta}) = s)$	22	64	14	1			
e5	$P(S(\hat{\beta}) = s)$	27	41	14	12	3	2	1
	$P(S2(\hat{\beta}) = s)$	43	49	7	1			
e6	$P(S(\hat{\beta}) = s)$	71	18	7	3	1		1
	$P(S2(\hat{\beta}) = s)$	62	27	4	3	1	1	3

NOTE: Each case is based on 10,000 samples; $n = 50$, $d(50) = 10$.

$S2$ in the case of testing normality, this is not surprising. Although in the case of testing exponentiality, S and $S2$ also tend to concentrate on dimension 1, there is still a difference, because as a rule when S or $S2$ is large, this implies a large value of W_S or W_{S2} as well.

To get p values (or critical values), one may perform simulations. However, it is also possible to apply an approximation, which is much more accurate than the chi-squared approximation but is still simple. We may write

$$P(W_S \leq x) = P(W_1 \leq x, S = 1)$$
$$+ P(W_2 \leq x, S = 2) + P(W_S \leq x, S \geq 3).$$

Under H_0, $P(S \geq 3)$ is negligible. The event $\{S = 1\}$ means that dimension 1 "beats" dimension 2, 3, and so on. The most important event is that dimension 1 "beats" dimension 2, and this is approximately equal to $\{n(\bar{\phi}_2(\hat{\beta}))^2 \leq \log n\}$. Similarly, for $\{S = 2\}$, the most important part corresponds to dimension 2 "beating" dimension 1. Therefore, under H_0, we get the following approximation:

$$P(W_S \leq x) \approx P(W_1 \leq x, n(\bar{\phi}_2(\hat{\beta}))^2 \leq \log n)$$
$$+ P(W_2 \leq x, n(\bar{\phi}_2(\hat{\beta}))^2 > \log n).$$

Except for very large x, $n(\bar{\phi}_2(\hat{\beta}))^2 > \log n$ implies $W_2 > x$, and hence we omit the second term in the approximation. Taking the limit distribution of $\sqrt{n}(\bar{\phi}_1(\hat{\beta}), \bar{\phi}_2(\hat{\beta}))$, writing $\mathbf{A}$ for $\mathbf{A}_{\hat{\beta}}$ (note that for location/scale families $\mathbf{A}$ does not depend on $\hat{\beta}$), and defining a, b, c, and ρ by

$$\mathbf{I} - \mathbf{A} = \begin{pmatrix} a & b \\ b & c \end{pmatrix}, \qquad \rho = b/\sqrt{ac},$$

we get as an approximation [Bickel and Doksum 1977, (1.4.18), p. 25]

$$P(W_S \leq x) \approx P(\{\sqrt{1 - \rho^2}U_1 + \rho U_2\}^2 \leq x, cU_2^2 \leq \log n),$$

where U_1 and U_2 are independent and $N(0, 1)$ distributed.

A further look at the integration region, noting that a $N(0, 1)$-distributed random variable has almost no probability mass above $2 \log n$, suggests the following very simple

approximation:

$$P(W_S \leq x) \approx 2\Phi(\sqrt{x}) - 1$$
$$- 2[1 - \Phi\{(c^{-1} \log n)^{1/2}\}][\Phi(b(x)) - \Phi(a(x))],$$

with

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

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

and Φ denoting the standard normal distribution function. In case of the exponential distribution we get $c = 31/36$ and $\rho = -\sqrt{15/31}$. Inserting for x the simulated critical values of Table 1 (with $d(n) \geq 2$), we get the results shown in Table 2. As seen in Table 2, the approximation appears to work quite well in the exponential case. For testing normality, we get $\rho = 0$ and $c = 1 - 15/(2\pi^2) = .2401$. For instance, if $n = 50$, then $(c^{-1} \log n)^{1/2} = 4.0366$, and hence $P(cU_2^2 \geq \log n) = 2\{1 - \Phi(4.0366)\} = .00005$, implying that the approximation above is very close to the chi-squared-one approximation. This explains why in the normal case, the simulated critical values are close to the chi-squared-one critical values (cf. Table 1).

Both examples (testing exponentiality and normality) show that the foregoing approximation gives p values in a simple and accurate way and can also serve to find critical values. The approximation can be used for W_{S2} as well.

4. POWER FOR TESTING EXPONENTIALITY

Let $X_1, X_2, \ldots$ be iid random variables each distributed according to some probability distribution P. For testing exponentiality, $\hat{\beta} = \bar{X}$, which converges under P to $E_P X$ and thus, under alternatives, $\beta = E_P X$. We consider P to be an alternative if there exists $K(\beta) = K(E_P X)$ such that

$$E_P \phi_1(F(X; \beta)) = \cdots = E_P \phi_{K(\beta)-1}(F(X; \beta))$$
$$= 0, E_P \phi_{K(\beta)}(F(X; \beta)) \neq 0, \quad (4)$$

with $F(x; \beta) = 1 - \exp(-\beta^{-1}x)$, the distribution function of the exponential distribution.

While under H_0 the selection rules concentrate on dimension 1, under alternatives we have

$$\lim_{n \to \infty} P(S(\hat{\beta}) \geq K(\beta)) = 1$$

and

$$\lim_{n \to \infty} P(S2(\hat{\beta}) \geq K(\beta)) = 1$$

(see Inglot et al. 1994b). To illustrate this result, we present some Monte Carlo results. For testing exponentiality, we present the following alternatives:
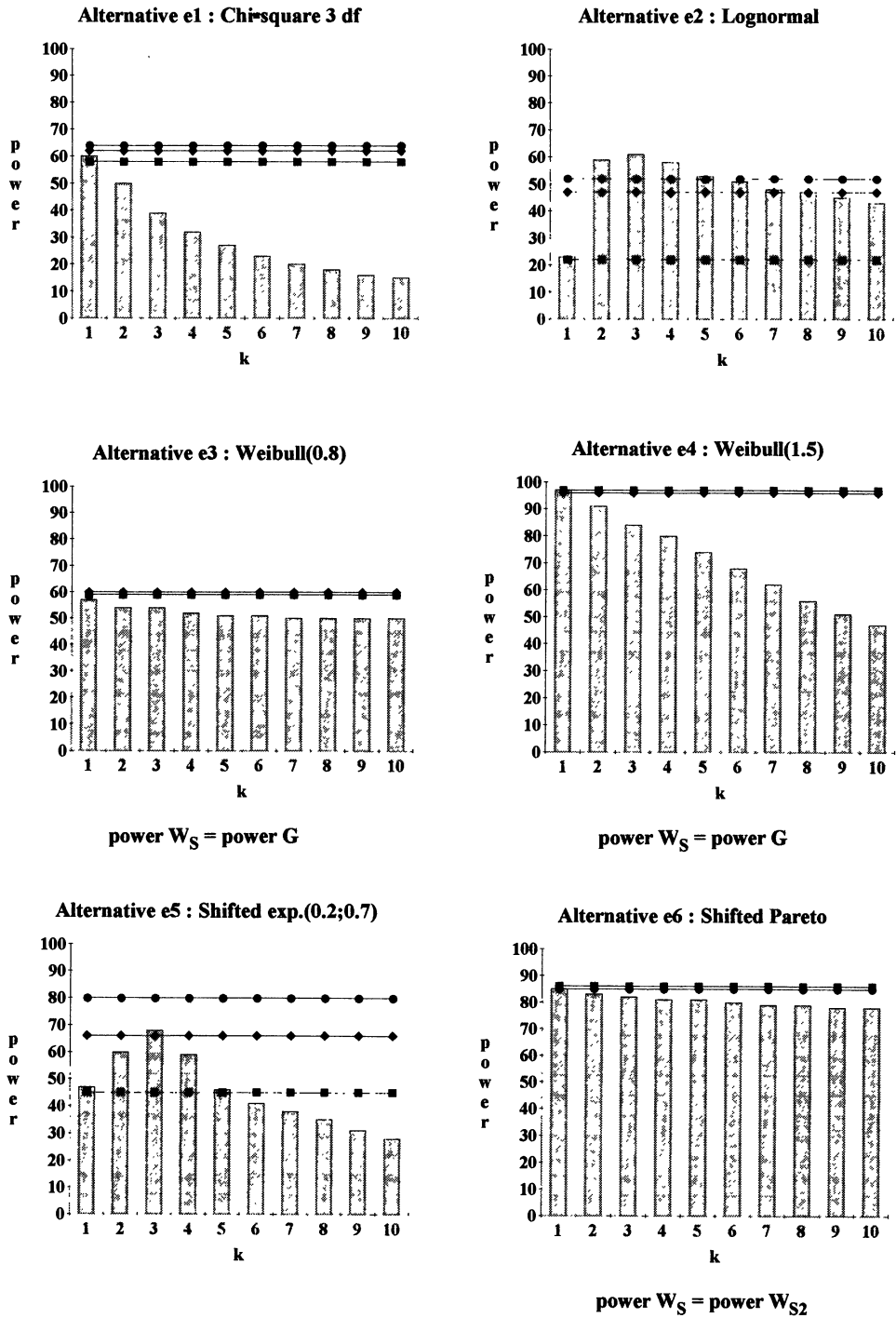
Alternative e1 : Chi-square 3 df

Alternative e2 : Lognormal

Alternative e3 : Weibull(0.8)

power W_S = power G

Alternative e4 : Weibull(1.5)

power W_S = power G

Alternative e5 : Shifted exp.(0.2;0.7)

Alternative e6 : Shifted Pareto

power W_S = power W_{S2}

Figure 1. Estimated Powers (in %) of W_k (bars), W_S (— • —), W_{S2} (— ♦ —), G (— ■ —) Testing Exponentiality; $\alpha = .10$, $n = 50$; $d(50) = 10$. Each case is based on 10,000 samples.

Notation	Alternative	Density	
e1	χ_3^2	$\{2^{3/2}\Gamma(\frac{3}{2})\}^{-1}x^{1/2}\exp(-\frac{1}{2}x)$,	$x > 0$
e2	Lognormal	$(x\sqrt{2\pi})^{-1}\exp\{-\frac{1}{2}(\log x)^2\}$,	$x > 0$
e3	Weibull (.8)	$.8x^{-.2}\exp(-x^{.8})$,	$x > 0$
e4	Weibull (1.5)	$\frac{3}{2}x^{1/2}\exp(-x^{3/2})$,	$x > 0$
e5	Shifted exponential (.2; .7)	$.7\exp\{-(x-.2).7\}$,	$x > .2$
e6	Shifted Pareto	$2(1+x)^{-3}$,	$x > 0$

Table 3 shows the behavior of the selection rules under alternatives when testing exponentiality against these alternatives.

It is seen that the behavior under alternatives may be quite different from that under H_0 (see Sec. 3). Indeed, according to the theory, there is less concentration on dimension 1, and the theoretical similarity between $S(\hat{\beta})$ and $S2(\hat{\beta})$ is reflected in the simulation results.

Theoretical support of the tests W_S and W_{S2} is given by their *consistency*; see Theorem 2.1 and Corollary 2.1.

For power comparison in the finite-sample case, we consider $W_k (k = 1, \ldots, 10)$, W_S, W_{S2}, and G, Gini's test. This last 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

Table 4. Estimated $P(S(\hat{\beta}) = s)$ and $P(S2(\hat{\beta}) = s)$ Under Alternatives (in %) When Testing Normality

Alternative		s									
		1	2	3	4	5	6	7	8	9	10
n1	$P(S(\hat{\beta}) = s)$	30			3	1	58	6	2		
	$P(S2(\hat{\beta}) = s)$	40			4	1	33	18	4		1
n2	$P(S(\hat{\beta}) = s)$	97	1	1	1						
	$P(S2(\hat{\beta}) = s)$	97	1	1	1						
n3	$P(S(\hat{\beta}) = s)$	79		14	2	3					
	$P(S2(\hat{\beta}) = s)$	82		14	2	1					
n4	$P(S(\hat{\beta}) = s)$	45		44	7	1	1		1		
	$P(S2(\hat{\beta}) = s)$	47		47	6	1					
n5	$P(S(\hat{\beta}) = s)$	87	2	9	1	1					
	$P(S2(\hat{\beta}) = s)$	88	2	9	1						
n6	$P(S(\hat{\beta}) = s)$	74		19	3	4	1				
	$P(S2(\hat{\beta}) = s)$	76		21	2	1					

NOTE: Each case is based on 10,000 samples; $n = 50$; $d(50) = 10$.

by LaRiccia (1991).

The conclusions from the simulation results presented in Figure 1 and from other simulation results that we have performed (see, e.g., Kallenberg and Ledwina 1994) are as follows:

1. A wrong choice of the number k of components in W_k may give a considerable loss of power.

2. The power of W_S and W_{S2} is high and stable, often as high or even higher than the "best" of W_k.

3. The tests W_S and W_{S2}, although based on general ideas, can compare favorably even with "special" tests for exponentiality, like Gini's test. W_S and W_{S2} often have higher power than Gini's test and in many cases much higher.

4. In most cases, W_S and W_{S2} have similar powers.

5. POWER FOR TESTING NORMALITY

Again we consider alternatives of the form (4), now with $\beta = (E_P X, \{var_P X\}^{1/2})$, corresponding to the MLE $\hat{\beta}$ under normality and with $F(x; \beta) = \Phi(x; \beta) = \Phi((x - E_P X)/\{var_P X\}^{1/2})$. The same theoretical results as for testing exponentiality apply here.

We consider the following alternatives. Here Z denotes a $N(0, 1)$ random variable and φ denotes its density.

Notation	Alternative	Density/Definition
n1	Normal mixture	$.33\{2\varphi(2x)\} + .33\{2\varphi(2[x - 2])\}$ $+.34\{2\varphi(2[x - 4])\}$, $-\infty < x < \infty$
n2	Logistic	$e^x(1 + e^x)^{-2}$, $-\infty < x < \infty$
n3	$LC(0.1; 3)$	$.1\varphi(x - 3) + .9\varphi(x)$, $-\infty < x < \infty$
n4	.65uniform + .35beta(10, 20)	$.65 + .35x^9(1 - x)^{19}\{B(10, 20)\}^{-1}$, $0 < x < 1$
n5	$SU(-1; 2)$	$Z = -1 + 2\sinh^{-1}(X)$, $-\infty < X < \infty$
n6	$SB(1.580, 1.357)$	$Z = 1.580 + 1.357\log\{X(1 - X)\}$, $0 < X < 1$

Note that SB is the alternative from the Johnson system with skewness $\sqrt{b_1} = .8$ and kurtosis $b_2 = 3.4$.

Table 4 shows the behavior of the selection rules under alternatives when testing normality. Again the behavior under alternatives is different from that under H_0 (see Sec. 3) with less concentration on dimension 1. The similarity between $S(\hat{\beta})$ and $S2(\hat{\beta})$ is also clearly exhibited.

For power comparison in the finite-sample case, we use the prominent and often recommended Shapiro–Wilk's test. The corresponding test statistic is denoted by W.

The conclusions from the simulation results presented in Figure 2 and from other simulation results that we have performed (see, e.g., Kallenberg and Ledwina 1994) are as follows:

1. A wrong choice of the number k of components in W_k may give a considerable loss of power.

2. For skewed alternatives, the power of W_S and W_{S2} is high and stable, as a rule as high as the "best" of W_k.

3. The tests W_S and W_{S2}, although based on general ideas, can compare to skewed alternatives even with "special" tests for normality, like the Shapiro–Wilk's test, and can dominate moment-based tests.

4. The tests W_S and W_{S2} perform reasonably well in the symmetrical case. They compare well both to symmetrical and skewed alternatives to classical tests (such as Anderson–Darling, Cramér–von Mises, Kolmogorov, and Watson) and some new tests such as those given by Gan and Koehler (1990). It is not a big surprise that W_S and W_{S2} perform better for skewed alternatives than for symmetrical alternatives. It is well known that W_1 provides little or no protection against nearby symmetric alternatives (see Kopecky and Pierce 1979, p. 397). This is easily understood by noting that $E_P\phi_1\{\Phi(X; \beta)\} = 0$ for symmetric alternatives P.

5. In most cases W_S and W_{S2} have comparable empirical powers.

In view of the consistency and the simulation results, we feel that the conclusion of Rayner and Best (1990, p. 9)— "don't use those other methods—use a smooth test!"—may be slightly sharpened to "use a data-driven smooth test."

6. MODIFICATIONS

6.1 Changing the Penalty

In the selection rules, the penalty $k/2 \log n$ is used for dimension k. An obvious question is what happens for other penalty functions. The theoretical results, briefly described here, can be extended to other penalties as well. It should be noted that different penalties have an effect on p values and critical values. One should either simulate them or adapt the approximation of Section 3 to the situation at hand.

Of course, taking Schwarz's BIC-rule is not the only solution for defining a selection rule. By taking smaller weights, the power will be smaller for smooth and larger for highly oscillating alternatives. For illustration, we consider for $n = 50$ again the alternatives $e1, \ldots, e6$ when testing exponentiality and $n1, \ldots, n6$ when testing normality (Fig. 3). The different penalties are $(k/2)\log\log n = 1.4(k/2)$ (Hannan and Quin 1979), $2.5(k/2)$, $(k/2)\log n = 3.9(k/2)$ (which is Schwarz's rule), and $k/2(2\log n) = 7.8(k/2)$ (Haughton,
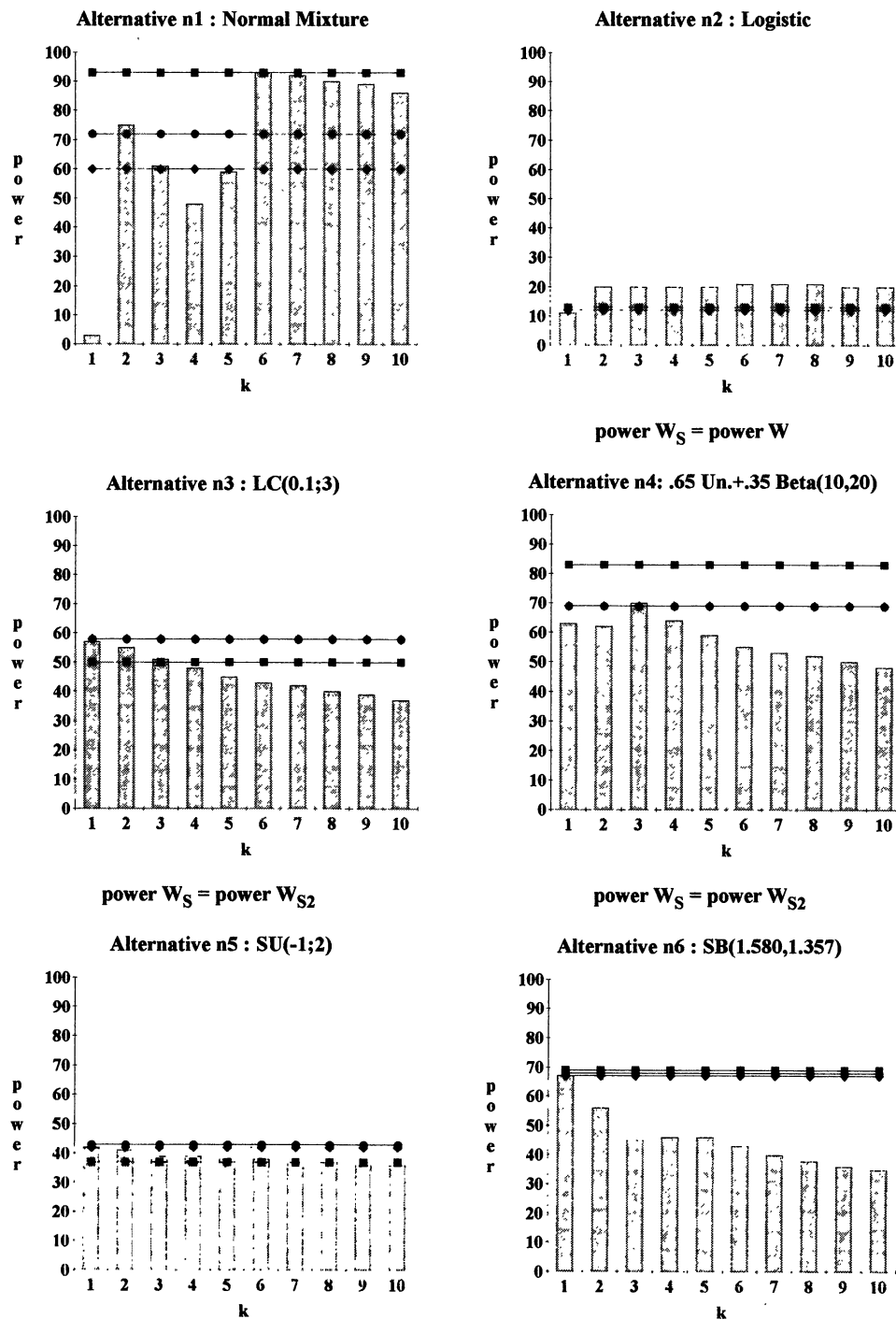
Figure 2. Estimated Powers (in %) of W_k (bar), W_S (—•—), W_{S2} (—♦—), W (—■—) Testing Normality; $\alpha = .05$, $n = 50$; $d(50) = 10$. Each case is based on 10,000 samples.

Haughton, and Izenman 1990). For every choice, corresponding critical values are simulated. The selection rule with penalty c instead of $\log n$ is called Sc.

We conclude from these and other simulations we have performed, that, although other penalties can be considered, there is sufficient support to adopt the Schwarz rule in general.

6.2 Choice of d(n)

One may ask whether the problem of choosing the number of components k is replaced by the choice of $d(n)$. Such a situation, where the power is not stable as a function of $d(n)$, happens to some order selection criteria (Bogdan 1995). This is certainly not the case if Schwarz's rule is used. In contrast to the power of W_k (see Figs. 1 and 2), the power of W_S does not change for larger $d(n)$. This is clearly illustrated in Figure 4.

When using W_S, to ensure convergence of the iterative procedure searching for the MLEs in the exponential family, we recommend $d(n) = 6$, 7, and 10 for $n = 20$, 30 and 50, respectively. (For details, see section 3.1 in Ledwina 1994.) When using W_{S2}, no restrictions on $d(n)$ are needed. However, in view of the stability of the power, there is no need to take very large $d(n)$.
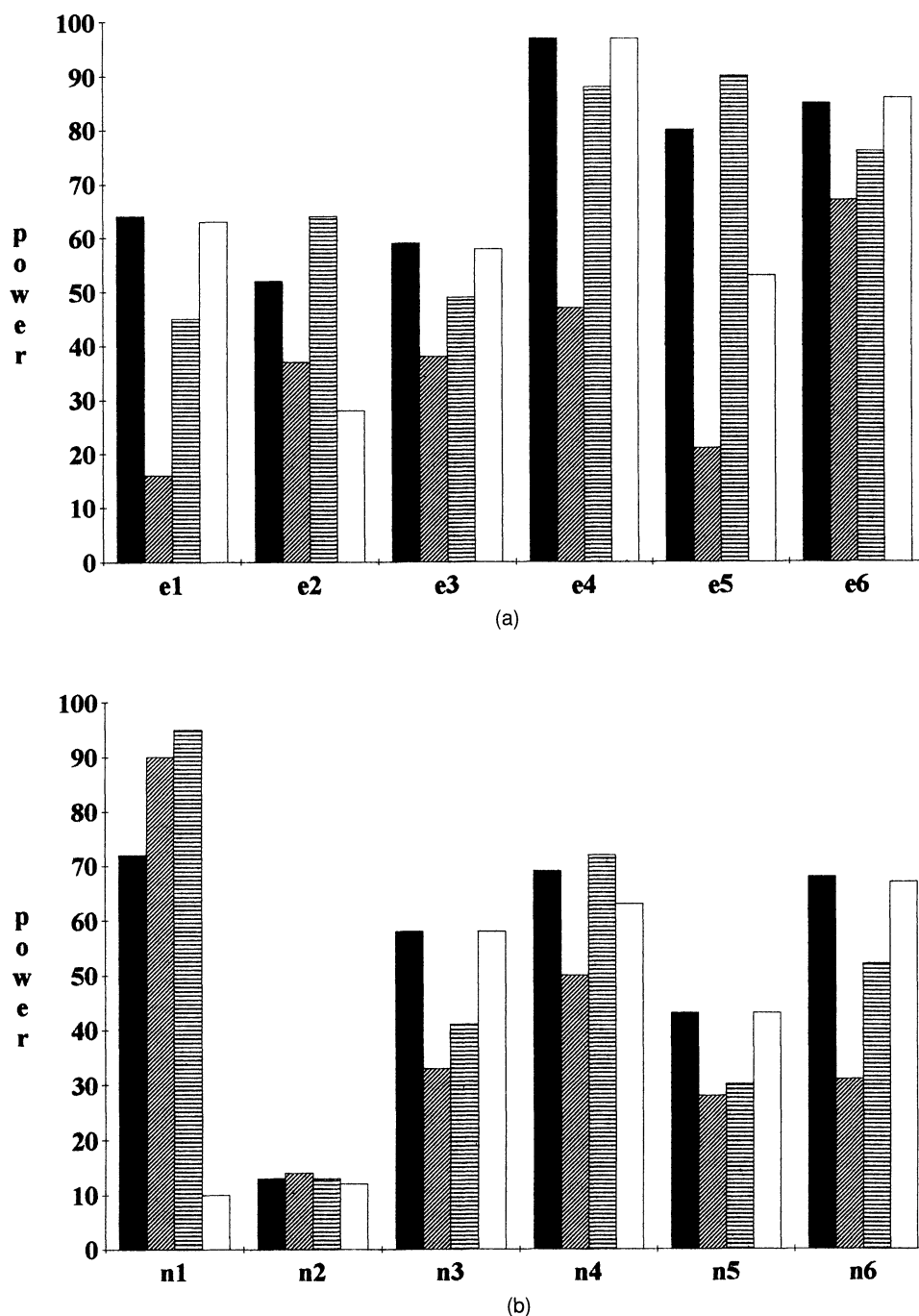
Figure 3. Estimated Powers (in %) of W_{Sc} for Different Penalties c; (a) for Testing Exponentiality $\alpha = .10$, $n = 50$; (b) for Testing Normality $\alpha = .05$, $n = 50$; $d(50) = 10$. Each case is based on 10,000 samples. ■, $c = log(n) = 3.9$; ▨, $c = log\ log(n) = 1.4$; ☰, $c = 2.5$; □, $c = 2\ log(n) = 7.8$.

6.3 Starting with Dimension 2

When testing normality against symmetrical alternatives, the first component $\bar{\phi}_1(\hat{\beta})$ will be close to 0, because $E_P\phi_1(\Phi(X_i;\beta)) = 0$. Therefore, one might suggest starting the selection with dimension 2 in that case. Indeed, for symmetric alternatives a higher power is obtained, but for skew alternatives some power is lost.

For $n = 50$ it turns out that W_S with S starting from 2, is slightly better than taking the penalty c equal to 2.5.

7. A NUMERICAL EXAMPLE

Thomas and Pierce (1979, sec. 5) and Rayner and Best (1989, pp. 15, 16) discussed a survey in which polychlorinated biphenyl (PCB) concentrations for 65 anacapa birds were tested for normality. (The data are from Risebrough 1972.) The values of $W_1,\dots,W_{10}$ are 4.09, 8.19, 8.60, 8.80, 9.26, 12.65, 14.01, 14.35, 16.00, and 16.13. Taking $d(65) = 10$, the dimension chosen by Schwarz's rule equals 1, and hence W_S equals 4.09. The p value of W_S equals .04. The same results are obtained for W_{S2}.
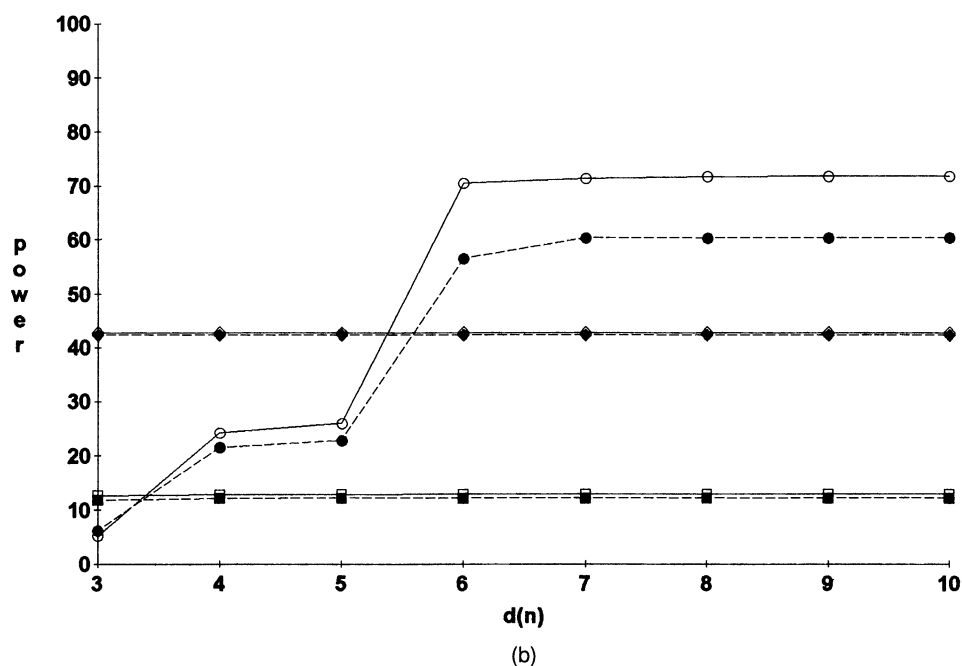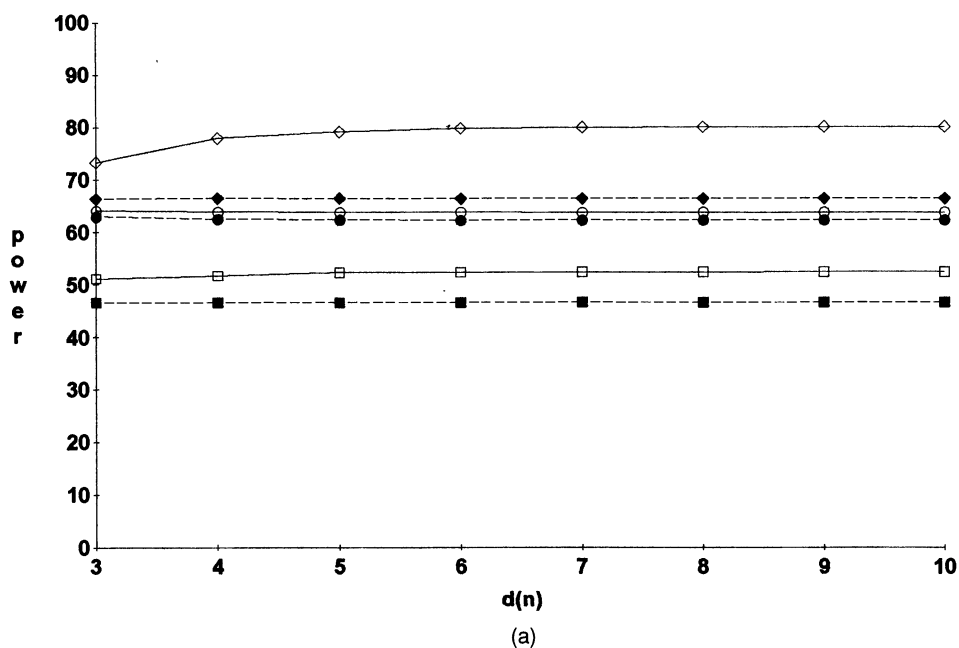
(a)

(b)

Figure 4. Estimated Powers (in %) of W_S and W_{S2} for $d(n) = 3 \ldots 10$; Each Case is Based on 10,000 Samples; ———— , W_S; – – –, W_{S2}. (a) For testing exponentiality $\alpha = .10$, $n = 50$; —o —, —●—, χ_3^2; —□—, - -■- -, Log normal; —◇—, - -♦- -, shifted exponential. (b) For testing normality $\alpha = .05$, $n = 50$; —o —, - -●- -, Normal mixture; —□—, - -■- -, Logistic; —◇—, - -♦- -, SU(−1; 2).

Thomas and Pierce (1979, sec. 5) gave simply $W_1, \ldots, W_4$ without making a choice, although in general they preferred W_2. The p value for W_2 is .017. Further, they present a histogram with an arbitrarily selected number of equal-width classes.

Rayner and Best (1989, p. 15) used the score test (related to some exponential family) with four nonzero components according to their recommendation on p. 84. The resulting p value was about .05. To give information about the "distance" from normality, they proposed using the Gram–Charlier type A density estimate based on the four com-

ponents (see also our Fig. 5). Moreover, they summarized the data by presenting also a histogram with equal width classes, where the number of classes is arbitrarily chosen, but differs from the one of Thomas and Pierce (1979). In comparison to the work of Thomas and Pierce (1979) and Rayner and Best (1989), we are using the same score test as Thomas and Pierce but with the number of components chosen by Schwarz's rule. We then summarize the data in Figure 5 by providing a likely model (among $d(65) = 10$ possible models given by (1)) fitted to the data in the dimension given by Schwarz's rule.
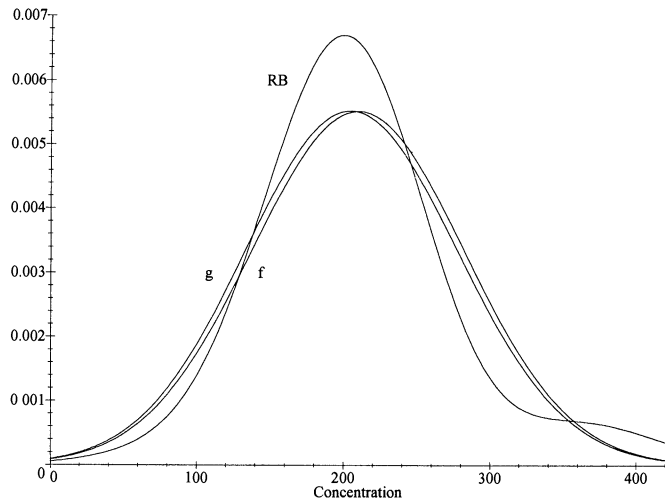
Fig. 5. Density Estimates for PCB Data; Estimated Null Density (f), Estimated Alternative Density (1) (g) and Estimated Gram–Charlier Type A density of Rayner and Best (RB).

APPENDIX: CONDITIONS AND PROOF

Conditions of Theorem 2.1

The conditions are divided in several parts: regularity conditions for the null family, conditions concerning the orthonormal system, and conditions on the upper bound $d(n)$ for the dimension of the exponential family.

Denote by P_β that X_i density $f(x; \beta)$ and by E_β and var_β the corresponding expected value and variance. For the family $\{f(x; \beta): \beta \in B\}$ and the alternative P, we set the following *regularity* conditions. The conditions (r1)–(r4) should hold on an open subset B_0 of B. (The "true" value of β under H_0 is supposed to lie in B_0.)

(r1). For $t, u = 1, \ldots, q$, $(\partial/\partial\beta_t)f(x; \beta)$ 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

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

and

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

where

$$\int_{\mathbb{R}} H_t(x)\, dx < \infty$$

and

$$\int_{\mathbb{R}} G_{tu}(x)\, dx < \infty:$$

(r2). For $t, u = 1, \ldots, q$, $(\partial/\partial\beta_t)\log f(x; \beta)$ and $(\partial^2/\partial\beta_t\partial\beta_u)\log f(x; \beta)$ exist almost everywhere and are such that the Fisher information matrix,

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

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

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

(r3). For each $\beta_0 \in B_0$, there exists $\eta = \eta(\beta_0) > 0$ with

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

and

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

(r4). There exist positive constants c_1, c_2, ρ_1, and n_1 such that

$$P_\beta(\sqrt{n}\|\hat{\beta} - \beta\| \geq r) \leq c_1 \exp(-c_2 r^2)$$

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

(r5). There exists $\beta \in B$ such that

$$\|\hat{\beta} - \beta\| \to 0$$

under the alternative P.

(r6). For β from (r5), there exists $\eta > 0$ such that

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

The next conditions concern the orthonormal system.

(s1). $\sup_{x\in[0,1]} |\phi_j'(x)| \leq c_3 j^{m_1}$ for each $j = 1, 2, \ldots, d(n)$ and some $c_3 > 0$, $m_1 > 0$.

(s2). $\sup_{x\in[0,1]} |\phi_j''(x)| \leq c_4 j^{m_2}$ for each $j = 1, 2, \ldots, d(n)$ and some $c_4 > 0$, $m_2 > 0$.

Finally, we have conditions on the dimension $d(n)$ of the exponential family:

(d1). $\{d(n)V_{d(n)}\}^2 n^{-1}\log n \to 0$ as $n \to \infty$, where $V_k = \max_{1\leq j\leq k} \sup_{x\in[0,1]} |\phi_j(x)|$.

(d2). $d(n) = o(\{n/\log n\}^{(2m)^{-1}})$ as $n \to \infty$, where $m = \max(m_1, m_2)$.

(d3). $d(n) = o(n^c)$ as $n \to \infty$ for some $c < c_2 b^{-2}$ if $\rho_1 b \geq 1$, and with $c = c_2 \rho_1^2$ otherwise, where c_2 and ρ_1 are given by (r4) and

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

Proof of Corollary 2.1

Conditions (r1)–(r3) and (r6) can be easily checked by direct calculations. Condition (r5) follows from the law of large numbers. Condition (r4) is obtained by application of standard moderate deviation theory. Both in the exponential case and in the normal case, it holds with $c_2 = 1/2(\mathrm{var}_\beta X)^{-1}$, whereas ρ_1 may be taken as large as one wants. We omit the details (see, e.g. Kallenberg 1983 ex. 2.1, p. 502). Hence (d3) reads as $d(n) = o(n^c)$ for some $c < 1/2$ in the exponential case and as $d(n) = o(n^c)$ for some $c < 1/6$ in the normal case. Therefore, (d3) is satisfied under the conditions on $d(n)$ stated in the corollary.

Further, for the orthonormal Legendre polynomials on [0, 1] (s1) and (s2) are fulfilled with (Sansone 1959, p. 251) $m_1 = 5/2$ and $m_2 = 9/2$. For the cosine system, (s1) and (s2) are fulfilled with $m_1 = 1$ and $m_2 = 2$.

Because for the orthonormal Legendre polynomials on [0, 1], we get $V_k = (2k + 1)^{1/2}$ (Sansone 1959, p. 190), (d1) reduces in this case to $\{d(n)\}^3 n^{-1}\log n \to 0$ as $n \to \infty$, and hence (d1) is satisfied. For 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 $n \to \infty$, implying that also for the cosine system, (d1) is satisfied. Finally, (d2) exactly equals the condition on $d(n)$ in the corollary.

Because all conditions of Theorem 2.1 are satisfied, the theorem may be applied, thus completing the proof of Corollary 2.1.

[Received March 1995. Revised December 1996.]

REFERENCES

Ascher, S. (1990), "A Survey of Tests for Exponentiality," *Communications in Statistics, Part A—Theory and Methods*, 19, 1811–1825.

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

Bickel, P. J., and Ritov, Y. (1992), "Testing for Goodness of Fit: A New Approach," in *Nonparametric Statistics and Related Topics*, ed. A.K.Md.E. Saleh, Amsterdam: North-Holland, pp. 51–57.

Bogdan, M. (1995), "Data-Driven Versions of Pearson's Chi-Square Test for Uniformity," *Journal of Statistical Computation and Simulation*, 52, 217–237.

Eubank, R. L., and Hart, J. D. (1992), "Testing Goodness-of-Fit in Regression via Order Selection Criteria," *The Annals of Statistics*, 20, 1412–1425.

Eubank, R. L., Hart, J. D., and LaRiccia, V. N. (1993), "Testing Goodness of Fit via Nonparametric Function Estimation Techniques," *Communications in Statistics, Part A—Theory and Methods*, 22, 3327–3354.

Eubank, R. L., and LaRiccia, V. N. (1992), "Asymptotic Comparison of Cramér–von Mises and Nonparametric Function Estimation Techniques for Testing Goodness-of-Fit," *The Annals of Statistics*, 20, 2071–2086.

Fan, J. (1996), "Tests of Significance Based on Wavelet Thresholding and Neyman's Truncation," *Journal of the American Statistical Association*, 91, 674–688.

Gail, M. H., and Gastwirth, J. L. (1978), "A Scale-Free Goodness-of-Fit Test for the Exponential Distribution Based on the Gini Statistic," *Journal of the Royal Statistical Society*, 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," *Journal of the Royal Statistical Society*, Ser. B, 41, 190–195.

Haughton, D., Haughton, J., and Izenman, A. (1990), "Information Criteria and Harmonic Models in Time Series Analysis," *Journal of Statistical Computation and Simulation*, 35, 187–207.

Inglot, T., Kallenberg, W. C. M., and Ledwina, T. (1994a), "Power Approximations to and Power Comparison of Certain Goodness-of-Fit Tests," *Scandinavian Journal of Statistics*, 21, 131–145.

—— (1994b), "On Selection Rules With an Application to Goodness-of-Fit for Composite Hypotheses," Memorandum 1242, University of Twente, Faculty of Applied Mathematics.

—— (in press), "Data-Driven Smooth Tests for Composite Hypotheses," submitted to *The Annals of Statistics*, 25.

Inglot, T., and Ledwina, T. (1996), "Asymptotic Optimality of Data-Driven Neyman's Tests for Uniformity," *The Annals of Statistics*, 24, 1982–2019.

Javitz, H. S. (1975), "Generalized Smooth Tests of Goodness of Fit, Independence, and Equality of Distributions," unpublished Ph.D. dissertation, University of California, Berkeley, Dept. of Statistics.

Kallenberg, W. C. M. (1983), "On Moderate Deviation Theory in Estimation," *The Annals of Statistics*, 11, 498–504.

Kallenberg, W. C. M., and Ledwina, T. (1994), "Data-Driven Smooth Tests for Composite Hypotheses," Memorandum 1232, University of Twente, Faculty of Applied Mathematics.

—— (1995a), "Consistency and Monte Carlo Simulation of a Data-Driven Version of Smooth Goodness-of-fit Tests," *The Annals of Statistics*, 23, 1594–1608.

—— (1995b), "On Data-Driven Neyman's Tests," *Probability and Mathematical Statistics*, 15, 409–426.

Kopecky, K. J., and Pierce, D. A. (1979), "Efficiency of Smooth Goodness-of-Fit Tests," *Journal of the American Statistical Association*, 74, 393–397.

LaRiccia, V. N. (1991), "Smooth Goodness-of-Fit Tests: A Quantile Function Approach," *Journal of the American Statistical Association*, 86, 427–431.

Ledwina, T. (1994), "Data-Driven Version of Neyman's Smooth Test of Fit," *Journal of the American Statistical Association*, 89, 1000–1005.

Milbrodt, H., and Strasser, H. (1990), "On the Asymptotic Power of the Two-Sided Kolmogorov–Smirnov Test," *Journal of Statistical Planning and Inference*, 26, 1–23.

Neyman, J. (1980), "Some Memorable Incidents in Probabilistic/Statistical Studies," in *Asymptotic Theory of Statistical Tests and Estimation*, ed. I. M. Chakravarti, New York: Academic Press, pp. 1–32.

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

—— (1990), "Smooth Tests of Goodness of Fit: An Overview," *International Statistical Review*, 58, 9–17.

Risebrough, R. W. (1972), "Effects of Environmental Pollutants Upon Animals Other Than Man," in *Proceedings of the Sixth Berkeley Symposium on Mathematical Statistics and Probability*, Vol. VI, Berkeley: University of California Press, pp. 443–463.

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

Schwarz, G. (1978), "Estimating the Dimension of a Model," *The Annals of Statistics*, 6, 461–464.

Thomas, D. R., and Pierce, D. A. (1979), "Neyman's Smooth Goodness-of-Fit Test When the Hypothesis is Composite," *Journal of the American Statistical Association*, 74, 441–445.