Goodness-of-Fit Tests for Parametric Regression Models

Jianqing FAN and Li-Shan HUANG

Several new tests are proposed for examining the adequacy of a family of parametric models against large nonparametric alternatives. These tests formally check if the bias vector of residuals from parametric fits is negligible by using the adaptive Neyman test and other methods. The testing procedures formalize the traditional model diagnostic tools based on residual plots. We examine the rates of contiguous alternatives that can be detected consistently by the adaptive Neyman test. Applications of the procedures to the partially linear models are thoroughly discussed. Our simulation studies show that the new testing procedures are indeed powerful and omnibus. The power of the proposed tests is comparable to the F-test statistic even in the situations where the F test is known to be suitable and can be far more powerful than the F-test statistic in other situations. An application to testing linear models versus additive models is also discussed.

KEY WORDS: Adaptive Neyman test; Contiguous alternatives; Partial linear model; Power; Wavelet thresholding.

1. INTRODUCTION

Parametric linear models are frequently used to describe the association between a response variable and its predictors. The adequacy of such parametric fits often arises. Conventional methods rely on residual plots against fitted values or a covariate variable to detect if there are any systematic departures from zero in the residuals. One drawback of the conventional methods is that a systematic departure that is smaller than the noise level cannot be observed easily. Recently many articles have presented investigations of the use of nonparametric techniques for model diagnostics. Most of them focused on one-dimensional problems, including Dette (1999), Dette and Munk (1998), and Kim (2000). The book by Hart (1997) gave an extensive overview and useful references. Chapter 5 of Bowman and Azzalini (1997) outlined the work by Azzalini, Bowman, and Härdle (1989) and Azzalini and Bowman (1993). Although a lot of work has focused on the univariate case, there is relatively little work on the multiple regression setting. Hart (1997, section 9.3), considered two approaches: one is to regress residuals on a scalar function of the covariates and then apply one-dimensional goodness-of-fit tests; the second approach is to explicitly take into account the multivariate nature. The test by Härdle and Mammen (1993) was based on the L_2 error criterion in d-dimensions, Stute, González Mantoiga, and Presedo Quindimil (1998) investigated a marked empirical process based on the residuals and Aerts, Claeskens, and Hart (1999) constructed tests based on orthogonal series that involved choosing a nested model sequence in the bivariate regression.

In this article, a completely different approach is proposed and studied. The basic idea is that if a parametric family of models fits data adequately, then the residuals should have nearly zero bias. More formally, let $(\mathbf{x}_1, Y_1), \ldots, (\mathbf{x}_n, Y_n)$ be independent observations from a population,

$$Y = m(\mathbf{x}) + \varepsilon, \qquad \varepsilon \sim N(0, \sigma^2), \tag{1.1}$$

where $\mathbf{x}$ is a p-dimensional vector and $m(\cdot)$ is a smooth regression surface. Let $f(\cdot, \theta)$ be a given parametric family. The null hypothesis is

$$H_0 : m(\cdot) = f(\cdot, \theta) \qquad \text{for some } \theta. \tag{1.2}$$

Examples of $f(\cdot, \theta)$ include the widely used linear model $f(\mathbf{x}, \theta) = \mathbf{x}^T\theta$ and a logistic model $f(x, \theta) = \theta_0/(1 + \theta_1 \exp(\theta_2^T \mathbf{x}))$.

The alternative hypothesis is often vague. Depending on the situations of applications, one possible choice is the saturated nonparametric alternative

$$H_1 : m(\cdot) \neq f(\cdot, \theta) \qquad \text{for all } \theta, \tag{1.3}$$

and another possible choice is the partially linear models (see Green and Silverman 1994 and the references therein)

$$H_1 : m(\mathbf{x}) = f(x_1) + \mathbf{x_2}^T\beta_2, \qquad \text{with } f(\cdot) \text{ nonlinear}, \tag{1.4}$$

where x_1 and $\mathbf{x}_2$ are the covariates. Traditional model diagnostic techniques involve plotting residuals against each covariate to examine if there is any systematic departure, against the index sequence to check if there is any serial correlation, and against fitted values to detect possible heteroscedasticity, among others. This amounts to informally checking if biases are negligible in the presence of large stochastic noises. These techniques can be formalized as follows. Let $\hat{\theta}$ be an estimate under the null hypothesis and let $\hat{\varepsilon} = (\hat{\varepsilon}_1, \ldots, \hat{\varepsilon}_n)^T$ be the resulting residuals with $\hat{\varepsilon}_i = Y_i - f(\mathbf{x}_i, \hat{\theta})$. Usually, $\hat{\theta}$ converges to some vector θ_0. Then under model (1.1), conditioned on $\{\mathbf{x}_i\}_{i=1}^n$, $\hat{\varepsilon}$ is nearly independently and normally distributed with mean vector $\eta = (\eta_1, \ldots, \eta_n)^T$, where $\eta_i = m(\mathbf{x}_i) - f(\mathbf{x}_i, \theta_0)$. Namely, any finite-dimensional components

Jianqing Fan is Professor, Department of Statistics, Chinese University of Hong Kong and Professor, Department of Statistics, University of North Carolina, Chapel Hill, NC 27599-3260 (E-mail: *jfan@email.unc.edu*). Li-Shan Huang is Assistant Professor, Department of Biostatistics, University of Rochester Medical Center, Rochester, NY 14642 (E-mail: *lhuang@bst.rochester.edu*). Fan's research was partially supported by NSA 96-1-0015 and NSF grant DMS-98-0-3200. Huang's work was partially supported by NSF grant DMS-97-10084. The authors thank the referee and the associate editor for their constructive and detailed suggestions that led to significant improvement in the presentation of the article.

of $\hat{\varepsilon}$ are asymptotically independent and normally distributed with mean being a subvector of η. Thus, the problem becomes

$$H_0 : \eta = 0 \quad \text{versus} \quad H_1 : \eta \neq 0, \tag{1.5}$$

based on the observations $\hat{\varepsilon}$. Note that $\hat{\varepsilon}_i$ will not be independent in general after estimating θ, but the dependence is practically negligible, as will be demonstrated. Thus, the techniques for independent samples such as the adaptive Neyman test in Fan (1996) continue to apply. Plotting a covariate X_j against the residual $\hat{\varepsilon}$ aims at testing if the biases in $\hat{\varepsilon}$ along the direction X_j are negligible, namely whether the scatter plot $\{(X_j, \eta_j)\}$ is negligible in the presence of the noise.

In testing the significance of the high-dimensional problem (1.5), the conventional likelihood ratio statistic is not powerful due to noise accumulation in the n-dimensional space, as demonstrated in Fan (1996). An innovative idea proposed by Neyman (1937) is to test only the first m-dimensional subproblem if there is prior that most of nonzero elements lie on the first m dimension. To obtain such a qualitative prior, a Fourier transform is applied to the residuals in an attempt to compress nonzero signals into lower frequencies. In testing goodness of fit for a distribution function, Fan (1996) proposed a simple data-driven approach to choose m based on power considerations. This results in an adaptive Neyman test. See Rayner and Best (1989) for more discussions on the Neyman test. Some other recent work motivated by the Neyman test includes Eubank and Hart (1992), Eubank and LaRiccia (1992), Kim (2000), Inglot, Kallenberg, and Ledwina (1994), Ledwina (1994), Kuchibhatla and Hart (1996), and Lee and Hart (1998), among others. Also see Bickel and Ritov (1992) and Inglot and Ledwina (1996) for illuminating insights into nonparametric tests. In this article, we adopt the adaptive Neyman test proposed by Fan (1996) to the testing problem (1.5). It is worth noting that the test statistic studied by Kuchibhatla and Hart (1996) looks similar to that of Fan (1996), but they are decidedly different. See the second paragraph of Section 3 for further discussions. The adaptive Neyman test of Kuchibhatla and Hart (1996) also will be implemented to demonstrate the versatility of our proposed idea. We would note that the adaptive Neyman tests in Fan (1996) and Kuchibhatla and Hart (1996) were proposed for independent samples. Herein, the adaptive Neyman test of Fan (1996) will be justified to remain applicable for weakly dependent $\hat{\varepsilon}$ in the current regression setting; see Theorem 1.

The power of the adaptive Neyman test depends on the smoothness of $\{\eta_i\}_{i=1}^n$ as a function of i. Let us call this function $\eta(\cdot)$, namely $\eta(i/n) = \eta_i$. The smoother the function $\eta(\cdot)$, the more significant are the Fourier coefficients on the low frequencies and hence the more powerful the test will be; see Theorems 2 and 3. When $m(\cdot)$ is completely unknown such as in (1.3), there is no information on how to make the function $\eta(\cdot)$ smoother by ordering the residuals. Intuitively, the closer the two consecutive covariate vectors $\mathbf{x}_i$ and $\mathbf{x}_{i+1}$, the smaller the difference between $m(\mathbf{x}_i) - f(\mathbf{x}_i, \theta_0)$ and $m(\mathbf{x}_{i+1}) - f(\mathbf{x}_{i+1}, \theta_0)$, and hence the smoother the sequence $\{m(\mathbf{x}_i) - f(\mathbf{x}_i, \theta_0)\}$ indexed by i. Thus, a good proxy is to order the residuals $\{\varepsilon_i\}_{i=1}^n$ in a way that the corresponding $\mathbf{x}_i$'s are close as a sequence. Note that when there is only one predictor variable (i.e., $p = 1$), such ordering is straightforward. However, it is quite challenging to order a multivariate vector. A method based on the principal component analysis is described in Section 2.2.

When there is some information about the alternative hypothesis such as the partially linear model in (1.4), it is sensible to order the residuals according to the covariate x_1. Indeed, our simulation studies show that it improves a great deal the power of the generic ordering procedure outlined in the last paragraph improved a great deal. The parameters β_2 in model (1.4) can be estimated easily at the $n^{-1/2}$ rate via a difference based estimator, for example, that of Yatchew (1997). Then the problem is, heuristically, reduced to the one-dimensional nonparametric setting; see Section 2.4 for details.

Wavelet transform is another popular family of orthogonal transforms. It can be applied to the testing problem (1.5). For comparison purposes, wavelet thresholding tests are also included in our simulation studies. See Fan (1996) and Spokoiny (1996) for various discussions on the properties of the wavelet thresholding tests.

This article is organized as follows. In Section 2, we propose a few new tests for the testing problems (1.2) versus (1.3) and (1.2) versus (1.4). These tests include the adaptive Neyman test and the wavelet thresholding tests in Sections 2.1 and 2.3, respectively. Ordering of multivariate vectors is discussed in Section 2.2. Applications of the test statistics in the partially linear models are discussed in Section 2.4 and their extension to additive models is outlined briefly in Section 2.5. Section 2.6 outlines simple methods for estimating the residual variance σ^2 in the high-dimensional setting. In Section 3, we carry out a number of numerical studies to illustrate the power of our testing procedures. Technical proofs are relegated to the Appendix.

2. METHODS AND RESULTS

As mentioned in the Introduction, the hypothesis testing problem (1.2) can be reduced to the high-dimensional problem (1.5) based on the weakly dependent Gaussian random vector $\hat{\varepsilon}$. The adaptive Neyman test and the wavelet thresholding test proposed in Fan (1996) are adapted for the current regression setting. The case of a multiple linear model, $f(\mathbf{x}, \theta) = \mathbf{x}^T\theta$, is addressed here and its extension to a general parametric model can be treated similarly.

2.1 The Adaptive Neyman Test

Assume a parametric linear model under the null hypothesis,

$$Y = \mathbf{x}^T\theta + \varepsilon, \qquad \varepsilon \sim N(0, \sigma^2),$$

and assume $\hat{\varepsilon}$ is the resulting residual vector from a least-squares fit. Let $\hat{\varepsilon}^* = (\hat{\varepsilon}_1^*, \ldots, \hat{\varepsilon}_n^*)^T$ be the discrete Fourier transform of the residual vector $\hat{\varepsilon}$. More precisely, we define

$$\hat{\varepsilon}_{2j-1}^* = (2/n)^{1/2} \sum_{i=1}^n \cos(2\pi ij/n)\hat{\varepsilon}_i,$$

$$\hat{\varepsilon}_{2j}^* = (2/n)^{1/2} \sum_{i=1}^n \sin(2\pi ij/n)\hat{\varepsilon}_i, \qquad j = 1, \ldots, [n/2].$$

When n is odd, an additional term $\hat{\varepsilon}_n^* = (1/\sqrt{n/2})\sum_{i=1}^n \hat{\varepsilon}_i$ is needed. Note that for linear regression with an intercept, this term is simply zero and hence can be ignored. As mentioned in the Introduction, the purpose of the Fourier transform is to compress useful signals into low frequencies so that the power of the adaptive Neyman test can be enhanced. Let $\hat{\sigma}_1$ be a $n^{-1/2}$ consistent estimate of σ under both the null and the alternative hypotheses. Methods for constructing such an estimate are outlined in Section 2.6. The adaptive Neyman test statistic is defined as

$$T^*_{\mathrm{AN},1} = \max_{1\le m\le n} \frac{1}{\sqrt{2m\hat{\sigma}_1^4}} \sum_{i=1}^m \left(\hat{\varepsilon}_i^{*2} - \hat{\sigma}_1^2\right), \tag{2.1}$$

and the null hypothesis is rejected when $T^*_{\mathrm{AN},1}$ is large. See Fan (1996) for motivations and some optimal properties of the test statistic. Note that the factor $2\sigma^4$ in (2.1) is the variance of ε^2 with $\varepsilon \sim N(0,\sigma^2)$. Thus, the null distribution of the testing procedure depends on the normality assumption. For nonnormal noises, a reasonable method is to replace the factor $2\hat{\sigma}_1^4$ by a consistent estimate $\hat{\sigma}_2^2$ of $\mathrm{Var}(\varepsilon_i^2)$. This leads to the test statistic

$$T^*_{\mathrm{AN},2} = \max_{1\le m\le n} \frac{1}{\sqrt{m\hat{\sigma}_2^2}} \sum_{i=1}^m \left(\hat{\varepsilon}_i^{*2} - \hat{\sigma}_1^2\right). \tag{2.2}$$

The null distribution of $T^*_{\mathrm{AN},2}$ is expected to be more robust against the normality assumption. Following Fan (1996), we normalize the test statistics as

$$\begin{aligned} T_{\mathrm{AN},j} = {} & \sqrt{2\log\log n}\, T^*_{\mathrm{AN},j} \\ & - \{2\log\log n + 0.5\log\log\log n \\ & \quad - 0.5\log(4\pi)\} \qquad \text{for } j = 1,2. \end{aligned} \tag{2.3}$$

Under the null hypothesis (1.5) and the independent normal model with a known variance, it can be shown that

$$P(T_{\mathrm{AN},j} < x) \to \exp(-\exp(-x)) \qquad \text{as } n\to\infty. \tag{2.4}$$

See Darling and Erdős (1956) and Fan (1996). However, the approximation (2.4) is not so good. Let us call the exact null distribution of $T_{\mathrm{AN},1}$ under the independent normal model with a known variance as J_n. Based on a million simulations, the distribution of J_n was tabulated in Fan and Lin (1998). We have excerpted some of their results and they are presented in Table 1.

We now show that the approximation (2.4) holds for testing linear models where $\{\hat{\varepsilon}_i\}$ are weakly dependent. For convenience of technical proofs, we modify the range of maximization over m as $[1, n/(\log\log n)^4]$. This hardly affects our theoretical understanding of the test statistics and has little impact on practical implementations of the test statistics.

Theorem 1. Suppose that Conditions A1–A3 in the Appendix hold. Then, under the null hypothesis of (1.2) with $f(\mathbf{x},\theta) = \mathbf{x}^T\theta$, we have

$$P(T_{\mathrm{AN},j} < x) \to \exp(-\exp(-x)) \qquad \text{as } n\to\infty, \text{ for } j = 1,2.$$

The proof is given in the Appendix. As a consequence of Theorem 1, the critical region

$$T_{\mathrm{AN},j} > -\log\{-\log(1-\alpha)\} \qquad (j = 1,2) \tag{2.5}$$

has an asymptotic significance level α.

Assume that $\hat{\theta}$ converges to θ_0 as $n\to\infty$. Then, $f(\cdot,\theta_0)$ is generally the best parametric approximation to the underlying regression function $m(\cdot)$ among the candidate models $\{f(\cdot,\theta)\}$. Let $\eta_i = m(\mathbf{x}_i) - f(\mathbf{x}_i,\theta_0)$ and let η^* be the Fourier transform of the vector $\eta = (\eta_1,\ldots,\eta_n)^T$. Then we have the following power result.

Theorem 2. Under the assumptions in Theorem 1 and Condition A4 in the Appendix, if

$$(\log\log n)^{-1/2} \max_{1\le m\le n} m^{-1/2} \sum_{i=1}^m \eta_i^{*2} \to \infty, \tag{2.6}$$

then the critical regions (2.5) have an asymptotic power 1.

We now give an implication of condition (2.6). Note that by Parseval's identity,

$$m^{-1/2}\sum_{i=1}^m \eta_i^{*2} = m^{-1/2}\left\{\sum_{i=1}^n \eta_i^2 - \sum_{i=m+1}^n \eta_i^{*2}\right\}. \tag{2.7}$$

Suppose that the sequence $\{\eta_i\}$ is smooth so that

$$n^{-1}\sum_{i=m+1}^n \eta_i^{*2} = O(m^{-2s}) \tag{2.8}$$

for some $s > 0$. Then the maximum value of (2.7) over m is of order

$$\left\{n^{-1}\left(\sum_{i=1}^n \eta_i^2\right)^{4s+1}\right\}^{1/(4s)}.$$

For random designs with a density h,

$$\begin{aligned} \sum_{i=1}^n \eta_i^2 &= \sum_{i=1}^n \{m(\mathbf{x}_i) - f(\mathbf{x}_i,\theta_0)\}^2 \\ &\approx n\int \{m(\mathbf{x}) - f(\mathbf{x},\theta_0)\}^2 h(\mathbf{x})\,d\mathbf{x}. \end{aligned}$$

By Theorem 2, we have the following result.

Theorem 3. Under the conditions of Theorem 2, if the smoothness condition (2.8) holds, then the adaptive Neyman test has an asymptotic power 1 for the contiguous alternative with

$$\int \{m(\mathbf{x}) - f(\mathbf{x},\theta_0)\}^2 h(\mathbf{x})\; d\mathbf{x} = n^{-4s/(4s+1)}(\log\log n)^{2s/(4s+1)} c_n$$

for some sequence with $\liminf c_n = \infty$.

The significance of the preceding theoretical result is its adaptivity. The adaptive Neyman test does not depend at all on the smoothness assumption (2.8). Nevertheless, it can adaptively detect alternatives with rates $O(n^{-2s/(4s+1)}(\log\log n)^{s/(4s+1)})$, which is the optimal rate of adaptive testing (see Spokoiny 1996). In this sense, the adaptive Neyman test is truly adaptive and is adaptively optimal.

Table 1. α Upper Quantile of the Distribution J_n

$\alpha \backslash n$	10	20	30	40	60	80	100	120	140	160	180	200
0.01	5.78	6.07	6.18	6.22	6.32	6.37	6.41	6.41	6.42	6.42	6.47	6.43
0.025	4.57	4.75	4.82	4.87	4.91	4.93	4.95	4.95	4.95	4.95	4.95	4.95
0.05	3.67	3.77	3.83	3.85	3.88	3.89	3.90	3.90	3.90	3.91	3.90	3.89
0.10	2.74	2.78	2.81	2.82	2.85	2.85	2.86	2.87	2.86	2.87	2.87	2.86

2.2 Ordering Multivariate Vector

The adaptive Neyman test statistics depend on the order of the residuals. Thus, we need to order the residuals first before using the adaptive Neyman test. By Theorem 2, the power of the adaptive Neyman tests depends on the smoothness of the sequence $\{m(\mathbf{x}_i) - f(\mathbf{x}_i, \theta_0)\}$ indexed by i. A good ordering should make the sequence $\{m(\mathbf{x}_i) - f(\mathbf{x}_i, \theta_0)\}_{i=1}^n$ as smooth as possible for the given function m so that its large Fourier coefficients are concentrated on low frequencies.

When the alternative is the partial linear model (1.4), we can order residuals according to the covariate X_1 so that the sequence $\{m(\mathbf{x}_i) - f(\mathbf{x}_i, \theta_0)\}_{i=1}^n$ is smooth for given m. For saturated nonparametric alternative (1.3), there is little useful information on how to order the sequence. As discussed in the Introduction, a good strategy is to order the covariates $\{\mathbf{x}_i\}_{i=1}^n$ so that they are close to each other consecutively. This problem easily can be done for the univariate case ($p = 1$). However, for the case $p > 1$, there are limited discussions on ordering multivariate observations. Barnett (1976) gave many useful ideas and suggestions. One possible approach is to first assign a score s_i to the ith observation and then order the observations according to the rank of s_i. Barnett (1976) called this "reduced ordering."

What could be a reasonable score $\{s_i\}$? We may consider this problem from the viewpoint of principal component (PC) analysis (see, e.g., Jolliffe 1986). Let $\mathbf{S}$ be the sample covariance matrix of the covariate vectors $\{\mathbf{x}_i, i = 1, \ldots, n\}$. Denote by $\lambda_1, \ldots, \lambda_p$ the ordered eigenvalues of $\mathbf{S}$ with corresponding eigenvectors $\xi_1, \ldots, \xi_p$. Then $z_{i,k} = \xi_k^T \mathbf{x}_i$ is the score for the ith observation on the kth sample PC, and λ_k can be interpreted as the sample variance of $\{z_{1,k}, \ldots, z_{n,k}\}$. Note that $\mathbf{x}_i - \bar{\mathbf{x}} = z_{i,1}\xi_1 + \cdots + z_{i,p}\xi_p$, where $\bar{\mathbf{x}}$ is the sample average. Thus a measure of variation of $\mathbf{x}_i$ may be formed by taking

$$s_i = \frac{1}{n-1} \sum_{k=1}^{p} \lambda_k z_{i,k}^2.$$

We call s_i the sample score of variation. Also see Gnanadesikan and Kettenring (1972) for an interpretation of s_i as a measure of the degree of influence for the ith observation on the orientation and scale of the PCs. A different ordering scheme was given in Fan (1997).

Ordering according to a certain covariate is another simple and viable method. It focuses particularly on testing if the departure from linearity in a particular covariate can be explained by chance. It formalizes the traditional scatterplot techniques for model diagnostics. The approach can be powerful when the alternative is the additive models or partially linear models. See Section 2.5 for further discussions.

In the case of two covariates, Aerts et al. (1999) constructed a score test statistic, and the power of which depends also on the ordering of a sequence of models. It seems that ordering is needed in the multiple regression setting, whether choosing an ordering of residuals or a model sequence.

2.3 Hard-Thresholding Test

The Fourier transform is used in the adaptive Neyman test. We may naturally ask how other families of orthogonal transforms behave. For comparison purposes, we apply the wavelet hard-thresholding test of Fan (1996) that is based on the "extremal phase" family of wavelets. See Daubechies (1992) for details on construction of the discrete wavelet transform with various wavelets. The Splus WaveThresh package includes many useful routines for wavelets.

Assume that the residuals are ordered. We first standardize the residuals as $\hat{\varepsilon}_{S,i} = \hat{\varepsilon}_i / \hat{\sigma}_3$, where $\hat{\sigma}_3$ is some estimate of σ. Let $\{\hat{\varepsilon}_{W,i}\}$ be the empirical wavelet transform of the standardized residuals, arranged in such a way that the coefficients at low resolution levels correspond to small indexes. The thresholding test statistic [equation (17) in Fan 1996] is defined as

$$T_H = \sum_{i=1}^{J_0} \hat{\varepsilon}_{W,i}^2 + \sum_{i=J_0+1}^{n} \hat{\varepsilon}_{W,i}^2 I(|\hat{\varepsilon}_{W,i}| > \delta)$$

for some given J_0 and $\delta = \sqrt{2\log(na_n)}$ with $a_n = \log^{-2}(n - J_0)$. As in Fan (1996), $J_0 = 3$ is used, which keeps the wavelet coefficients in the first two resolution levels intact. Then it can be shown analogously that T_H has an asymptotic normal distribution under the null hypothesis in Theorem 1, and the $(1-\alpha)$ critical region is given by

$$\sigma_H^{-1}(T_H - \mu_H) > \Phi^{-1}(1-\alpha),$$

where $\mu_H = J_0 + \sqrt{2/\pi} a_n^{-1} \delta(1 + \delta^{-2})$ and $\sigma_H^2 = 2J_0 + \sqrt{2/\pi} a_n^{-1} \delta^3 (1 + 3\delta^{-2})$. The power of the wavelet thresholding test can be improved further by using the following techniques owing to Fan (1996). Under the null hypothesis (1.5), the maximum of n independent Gaussian noises is on the order of $\sqrt{2\log n}$. Thus, replacing a_n in the thresholding parameter δ by $\min(4(\max_i \hat{\varepsilon}_{W,i})^{-4}, \log^{-2}(n - J_0))$ does not alter the asymptotic null distribution. Then, under the alternative hypothesis, the latter is smaller than a_n and hence the procedure uses a smaller thresholding parameter automatically, leading to better power of the wavelet thresholding test.

2.4 Linear Versus Partially Linear Models

Suppose that we are interested in testing the linear model against the partial linear model (1.4). The generic methods

outlined in Sections 2.1–2.3 continue to be applicable. In this setting, a more sensible ordering scheme is to use the order of X_1 instead of the score s_i given in Section 2.2. Our simulation studies show that this improves a great deal on the power of the generic methods. Nevertheless, the power can further be improved when the partially linear model structure is fully exploited.

The basic idea is to estimate the coefficient vector β_2 first and then to compute the partial residuals $z_i = Y_i - \mathbf{x}_{i,2}^T\hat{\beta}_2$. Based on the synthetic data $\{(x_{i1}, z_i)\}$, the problem reduces to testing the one-dimensional linear model, but with slightly dependent data. Theoretically, this dependence structure is asymptotically negligible, following the same proof as that of Theorem 1, as long as $\hat{\beta}_2$ is estimated at the root-n rate. We can therefore apply the adaptive Neyman and other tests to the data $\{(x_{i1}, z_i)\}$.

There is a large literature on efficient estimation of coefficients β_2. See, for example, Wahba (1984), Speckman (1988), Cuzick (1992), and Green and Silverman (1994). Most of the methods proposed in the literature involve choices of smoothing parameters. For our purpose, a root-n consistent estimator of β_2 suffices. This can be much more easily constructed than the semiparametric efficient estimator. The basic idea is as follows.

First the data are arranged according to the corresponding order of $\{X_1\}$, yielding the sorted data $\{(\mathbf{x}_{(i)}, Y_{(i)}), i = 1, \ldots, n\}$. Note that for a random sample, $x_{(i+1),1} - x_{(i),1}$ is $O_P(1/n)$. Thus, for the differentiable function f in (1.4), we have

$$Y_{(i+1)} - Y_{(i)} = (\mathbf{x}_{(i+1),2} - \mathbf{x}_{(i),2})^T\beta_2 + e_i + O_P(n^{-1}), \quad i = 1, \ldots, n-1, \quad (2.9)$$

where $\{e_i\}$ are correlated stochastic errors with $e_i = \varepsilon_{(i+1)} - \varepsilon_{(i)}$. Thus, β_2 can be estimated using the ordinary least-squares estimator from the approximate linear model (2.9). This idea was proposed by Yatchew (1997) (and was independently proposed by us around the same time with the following improvements). Yatchew (1997) showed that such an estimate β_2 is root-n consistent. To correct the bias in the preceding approximation at finite sample, particularly for those $x_{(i),1}$ at tails (hence the spacing can be wide), we fit the linear regression model

$$Y_{(i+1)} - Y_{(i)} = (x_{(i+1),1} - x_{(i),1})\beta_1 + (\mathbf{x}_{(i+1),2} - \mathbf{x}_{(i),2})^T\beta_2 + e_i$$

into our implementation by using the ordinary least-squares method. This improves the performance of the estimator based on the model (2.9).

2.5 Linear Models Versus Additive Models

As will be demonstrated in Section 3.3, when the data are properly ordered, the proposed tests are quite powerful. To test linear versus partially linear models, we show in Section 3.3 that the adaptive Neyman tests based on ordered residuals according to the variable X_1 for model (1.4) are nearly as powerful as those tests in Section 2.4 when the structure of the partially linear model is taken into account. This encouraged us to extend the idea to testing a linear model against the additive model

$$Y = f_1(X_1) + f_2(X_2) + \cdots + f_p(X_p) + \varepsilon.$$

Let $\widehat{T}_j$ be the normalized test statistic (2.3) when the residuals are ordered according to variable X_j. Let $\widehat{T} = \max_{1\le j\le p}\widehat{T}_j$. Reject the null hypothesis when $\widehat{T} > J_n(\alpha/p)$, with $J_n(\alpha/p)$ the α/p upper quantile of the distribution J_n in Table 1. This is simply the Bonferroni adjustment applied to the combined test statistics.

2.6 Estimation of Residual Variance

The implementations of the adaptive Neyman test and other tests depend on a good estimate of the residual variance σ^2. This estimator should be good under both the null and alternative hypotheses, because an overestimate (caused by biases) of σ^2 under the alternative hypothesis will significantly deteriorate the power of the tests. A root-n consistent estimator can be constructed by using the residuals of a nonparametric kernel regression fit. However, this theoretically satisfactory method encounters the "curse of dimensionality" in practical implementations.

In the case of a single predictor ($p = 1$), there are many possible estimators for σ^2. A simple and useful technique is given in Hall, Kay, and Titterington (1990). An alternative estimator can be constructed based on the Fourier transform, which constitutes raw materials of the adaptive Neyman test. If the signals $\{\eta_i\}$ are smooth as a function of i, then the high frequency components are basically noise even under the alternative hypothesis, namely, the discrete Fourier transform $\hat{\varepsilon}_j^*$ for j large has a mean approximately equal to 0 and variance σ^2. This gives us a simple and effective way to estimate σ^2 by using the sample variance of $\{\hat{\varepsilon}_i^*, i = I_n + 1, \ldots, n\}$:

$$\hat{\sigma}_1^2 = \frac{1}{n - I_n}\sum_{i=I_n+1}^{n}\hat{\varepsilon}_i^{*2} - \left\{\frac{1}{n - I_n}\sum_{i=I_n+1}^{n}\hat{\varepsilon}_i^*\right\}^2. \quad (2.10)$$

for some given I_n ($= [n/4]$, say). Under some mild conditions on the smoothness of $\{\eta_i\}$, this estimator can be shown to be root-n consistent even under the alternative hypothesis. Hart (1997) suggested taking $I_n = m + 1$, where m is the number of Fourier coefficients chosen based on some data-driven criteria. Similarly, an estimate $\hat{\sigma}_2^2$ for $\mathrm{Var}(\varepsilon^2)$ can be formed by taking the sample variance of $\{\hat{\varepsilon}_i^{*2}, i = n/4 + 1, \ldots, n\}$:

$$\hat{\sigma}_2^2 = \frac{1}{n - I_n}\sum_{i=I_n+1}^{n}\hat{\varepsilon}_i^{*4} - \left\{\frac{1}{n - I_n}\sum_{i=I_n+1}^{n}\hat{\varepsilon}_i^{*2}\right\}^2. \quad (2.11)$$

In the case of multiple regression ($p > 1$), the foregoing method is still applicable. However, as indicated in Section 2.2, it is hard to order residuals in a way that the resulting residuals $\{\eta_i\}$ are smooth so that $E(\hat{\varepsilon}_j^*) = 0$ for large j. Thus, the estimator (2.10) can possess substantial bias. Indeed, it is difficult to find a practically satisfactory estimate for large p. This problem is beyond the scope of this article. Here we describe an ad hoc method that is implemented in our simulation studies. For simplicity of description, assume that there are three continuous and one discrete

covariates, X_1, X_2, X_3, and X_4, respectively. An estimate $\hat{\sigma}_\ell$ can be obtained by fitting linear splines with four knots and the interaction terms between continuous predictors,

$$\theta_0 + \sum_{i=1}^{4} \theta_i X_i + \sum_{i=1}^{3}\sum_{j=1}^{4} \theta_{i,j}(X_i - t_{i,j})_+ + \gamma_1 X_1 X_2 + \gamma_2 X_1 X_3 + \gamma_3 X_2 X_3,$$

where $t_{i,j}$ denotes the $(20j)$th percentile, $j = 1, 2, 3, 4$, of the ith covariate, $i = 1, 2, 3$ (continuous covariates only). In this model, there are 20 parameters and the residual variance is estimated by

$$\hat{\sigma}_\ell^2 = \text{ residual sum of squared errors}/(n - 20).$$

3. SIMULATIONS

We now study the power of the adaptive Neyman tests ($T_{\mathrm{AN},1}$ and $T_{\mathrm{AN},2}$) and the wavelet thresholding test T_H via simulations. The simulated examples consist of testing the goodness of fit of the linear models in the situations of one predictor, partially linear models, and multiple regression. The results are based on 400 simulations and the significance level is taken to be 5%. Thus, the power under the null hypothesis should be around 5%, with a Monte Carlo error of $\sqrt{0.05 * 0.95/400} \approx 1\%$ or so. For each testing procedure, we examine how well it performs when the noise levels are estimated and when the noise levels are known. The latter mimics the situations where the variance can be well estimated with little bias. An example of this is when repeated measurements are available at some design points. Another reason for using the known variance is that we intend to separate the performance of the adaptive Neyman test and the performance of the variance estimation. For the wavelet thresholding test, the asymptotic critical value 1.645 is used. The simulated critical values in table 1 of Fan and Lin (1998) are used for the adaptive Neyman test if σ is known; empirical critical values are taken when σ_1 and/or σ_2 are estimated. For simplicity of presentation, we report only the results of the wavelet thresholding test when the residual variances are given.

To demonstrate the versatility of our proposed testing scheme, we also include the test proposed by Kuchibhatla and Hart (1996) (abbreviated as the KH test),

$$S_n = \max_{1 \le m \le n-p} \frac{1}{m} \sum_{j=1}^{m} \frac{2n\hat{\varepsilon}_i^{*2}}{\sigma^2},$$

where p is the number of parameters fitted in the null model. Compared to the adaptive Neyman test, this test tends to select a smaller dimension m, namely the maximization in S_n is achieved at smaller m than that of the adaptive Neyman test. Therefore, the KH test will be somewhat more powerful than the adaptive Neyman test when the alternative is very smooth and will be less powerful than the adaptive Neyman test when the alternative is not as smooth. Again, the results of S_n are given in cases when variance is known and when variance is estimated. The same variance estimator as the adaptive Neyman tests is applied.

3.1 Goodness of Fit for Simple Linear Regression

In this section, we study the power of the adaptive Neyman test and other tests for univariate regression problems:

$$H_0 : m(x) = \alpha + \beta x \quad \text{versus} \quad H_1 : m(x) \neq \alpha + \beta x.$$

The sample size is 64 and the residuals $\hat{\varepsilon}_i$ are ordered by their corresponding covariate. The power of each test is evaluated at a sequence of alternatives given in Examples 1–3. The empirical critical values for $T_{\mathrm{AN},1}$ and $T_{\mathrm{AN},2}$ are 4.62 and 4.74, respectively, and 3.41 for the KH test when σ is estimated. For comparison purposes, we also include the parametric F test for the linear model against the quadratic regression $\beta_0 + \beta_1 x + \beta_2 x^2$.

Example 1. The covariate X_1 is sampled from uniform $(-2, 2)$, and the response variable is drawn from

$$Y = 1 + \theta X_1^2 + \varepsilon, \qquad \varepsilon \sim N(0, 1), \tag{3.1}$$

for each given value of θ. The power function for each test is evaluated under the alternative model (3.1) with a given θ. This is a quadratic regression model where the F test is derived. Nevertheless, the adaptive Neyman tests and the KH test perform close to the F test in this ideal setting, whereas the wavelet thresholding test falls behind the other three tests; see Figure 1, (a) and (b). In fact, the wavelet tests are not specifically designated for testing this kind of very smooth alternatives. This example is also designated to show how large a price the adaptive Neyman and the KH tests have to pay to be more omnibus. Surprisingly, Figure 1 demonstrates that both procedures paid very little price.

Example 2. Let X_1 be $N(0, 1)$ and

$$Y = 1 + \cos(\theta X_1 \pi) + \varepsilon, \qquad \varepsilon \sim N(0, 1). \tag{3.2}$$

This example examines how powerful each testing procedure is for detecting alternatives with different frequency components. The result is presented in Figure 1, (c) and (d). Clearly, the adaptive Neyman and KH tests outperform the F test when σ is given, and they lose some power when σ is estimated. The loss is due to the excessive biases in the estimation of σ when θ is large. This problem can be resolved by setting a larger value of I_n in the high-frequency cases. The wavelet thresholding test performs nicely too. As anticipated, the adaptive Neyman test is more powerful than the KH test for detecting a high-frequency component.

Example 3. We now evaluate the power of each testing procedure at a sequence of logistic regression models,

$$Y = \frac{10}{1.0 + \theta \exp(-2X_1)} + \varepsilon, \qquad \varepsilon \sim N(0, 1). \tag{3.3}$$

where $X_1 \sim N(0, 1)$. Figure 1, (e) and (f), depicts the results. The adaptive Neyman and KH tests far outperform the F test, which is clearly not omnibus. The wavelet thresholding test is also better than the F test, whereas it is dominated by the adaptive Neyman and KH tests.

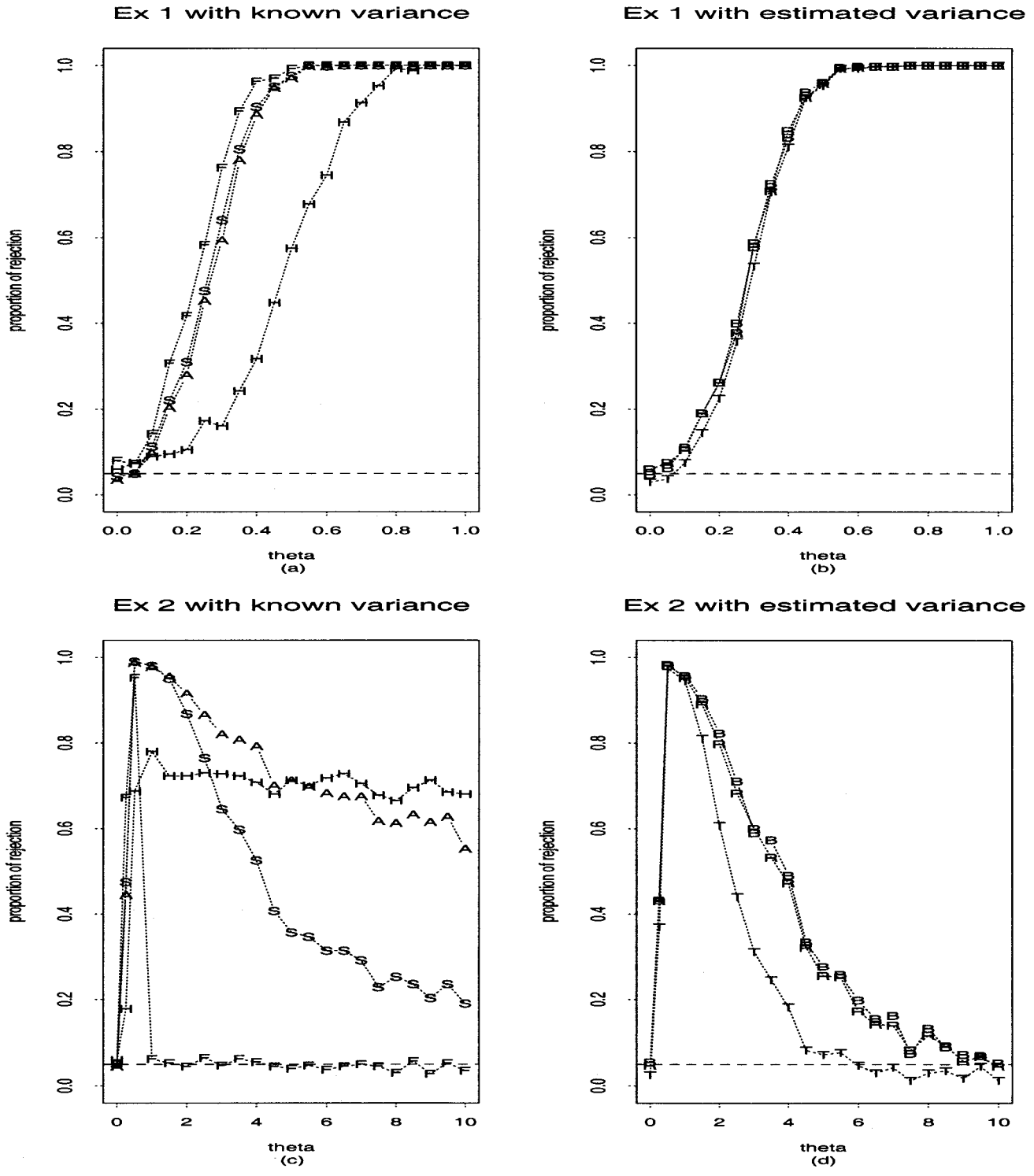

Figure 1. Power of the Adaptive Neyman Tests, the KH Test, the Wavelet Thresholding Test, and the Parametric F Test for Examples 1–3 With $n = 64$. Key: A, the adaptive Neyman test with known variance; B, the adaptive Neyman test with estimated variance (2.10); F, the parametric F-test statistic; H, the wavelet thresholding test with known variance; R, the robust version of the adaptive Neyman test with estimated variance (2.10) and (2.11); S, the KH test with known variance; T, the KH test with estimated variance (2.10).

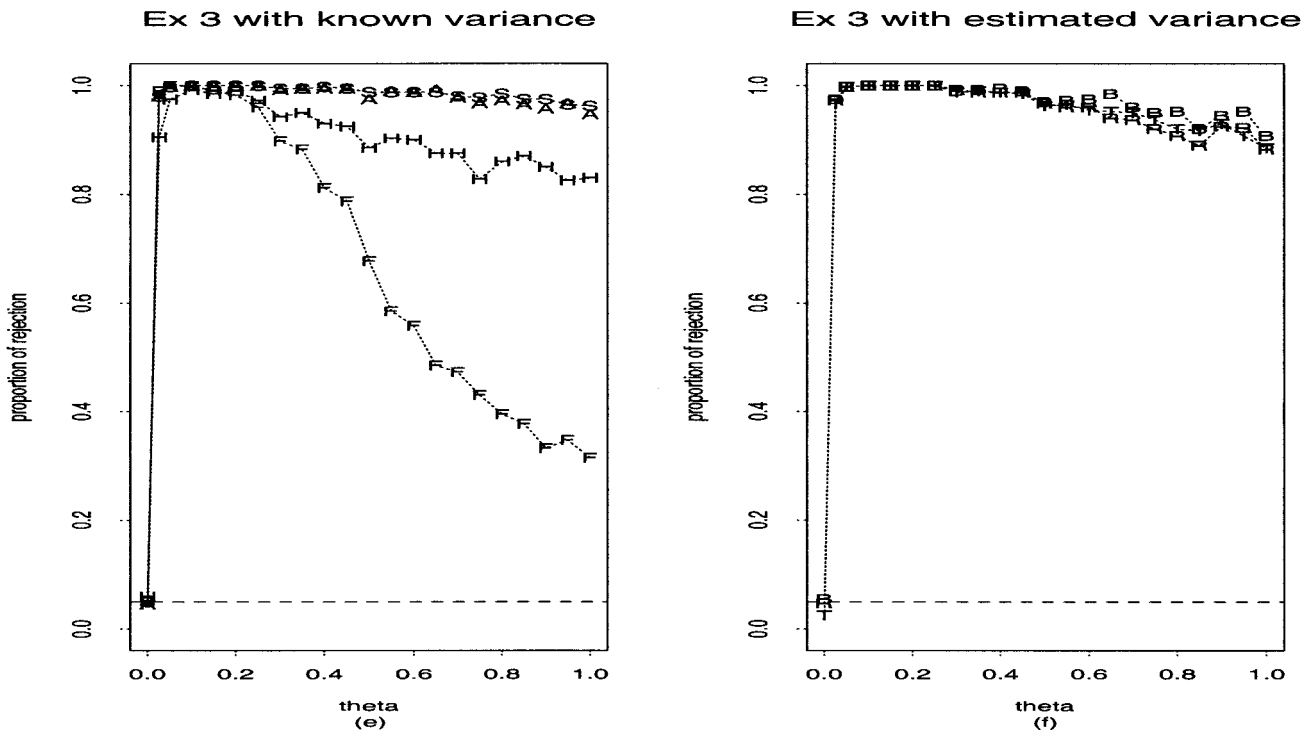

Figure 1. Continued

3.2 Testing Linearity in Partially Linear Models

In this section, we test the linear model versus a partial linear model

$$m(X) = \theta_0 + \theta_1 X_1 + f(X_2) + \theta_3 X_3 + \theta_4 x_4,$$

where $X_1, \ldots, X_4$ are covariates. The covariates X_1, X_2, and X_3 are normally distributed with mean 0 and variance 1. Furthermore, the correlation coefficients among these three random variables are .5. The covariate X_4 is binary, independent of X_1, X_2, and X_3, with

$$P(X_4 = 1) = .4 \quad \text{and} \quad P(X_4 = 0) = .6.$$

The techniques proposed in Section 2.4 are employed here for the adaptive Neyman and KH tests, and the same simulated critical values for $n = 64$ as in the case of the simple linear model are used. For comparison, we include the parametric F test for testing the linear model against the quadratic model

$$\beta_0 + \sum_{i=1}^{4} \beta_i X_i + \sum_{1 \le i \le j \le 3} \beta_{i,j} X_i X_j.$$

We evaluate the power of each testing procedure at two particular partially linear models as follows.

Example 4. The dependent variable is generated from the quadratic regression model

$$Y = X_1 + \theta X_2^2 + 2X_4 + \varepsilon, \qquad \varepsilon \sim N(0, 1), \tag{3.4}$$

for each given θ. Although the true function involves a quadratic of X_2, the adaptive Neyman and KH tests do slightly better than the F test, as shown in Figure 2, (a) and (b). This is due partially to the fact that an overparameterized full quadratic model is used in the F test. Again, the wavelet thresholding test does not perform as well for this kind of smooth alternatives. Note that this model is very similar to that in Example 1 and the power of the adaptive Neyman and KH tests behaves analogously to that in Example 1. This verifies our claim that the parametric components are effectively estimated by using our new estimator for β_2 in Section 2.4.

Example 5. We simulate the response variable from

$$Y = X_1 + \cos(\theta X_2 \pi) + 2X_4 + \varepsilon, \qquad \varepsilon \sim N(0, \sigma^2). \tag{3.5}$$

The results are depicted in Figure 2, (c) and (d). The adaptive Neyman, wavelet thresholding, and KH tests perform well when σ is known. The F test gives low power for $\theta \ge 1.5$, which indicates that it is not omnibus. Similar remarks to those given at the end of Example 4 apply. Note that when θ is large, the residual variance cannot be estimated well by any saturated nonparametric methods. This is why the power of each test is reduced so dramatically when σ is estimated by using our nonparametric estimate of σ. This also demonstrates that to have a powerful test, σ should be estimated well under both the null and the alternative hypotheses.

3.3 Testing for a Multiple Linear Model

Two models identical to those in Examples 4 and 5 are used to investigate the empirical power for testing linearity when there is no knowledge about the alternative, namely the alternative hypothesis is given by (1.3). The sample size 128 is used. If we know that the nonlinearity is likely to occur in the X_2 direction, we would order the residuals according to X_2 instead of using the generic method in Section 2.2. Thus,

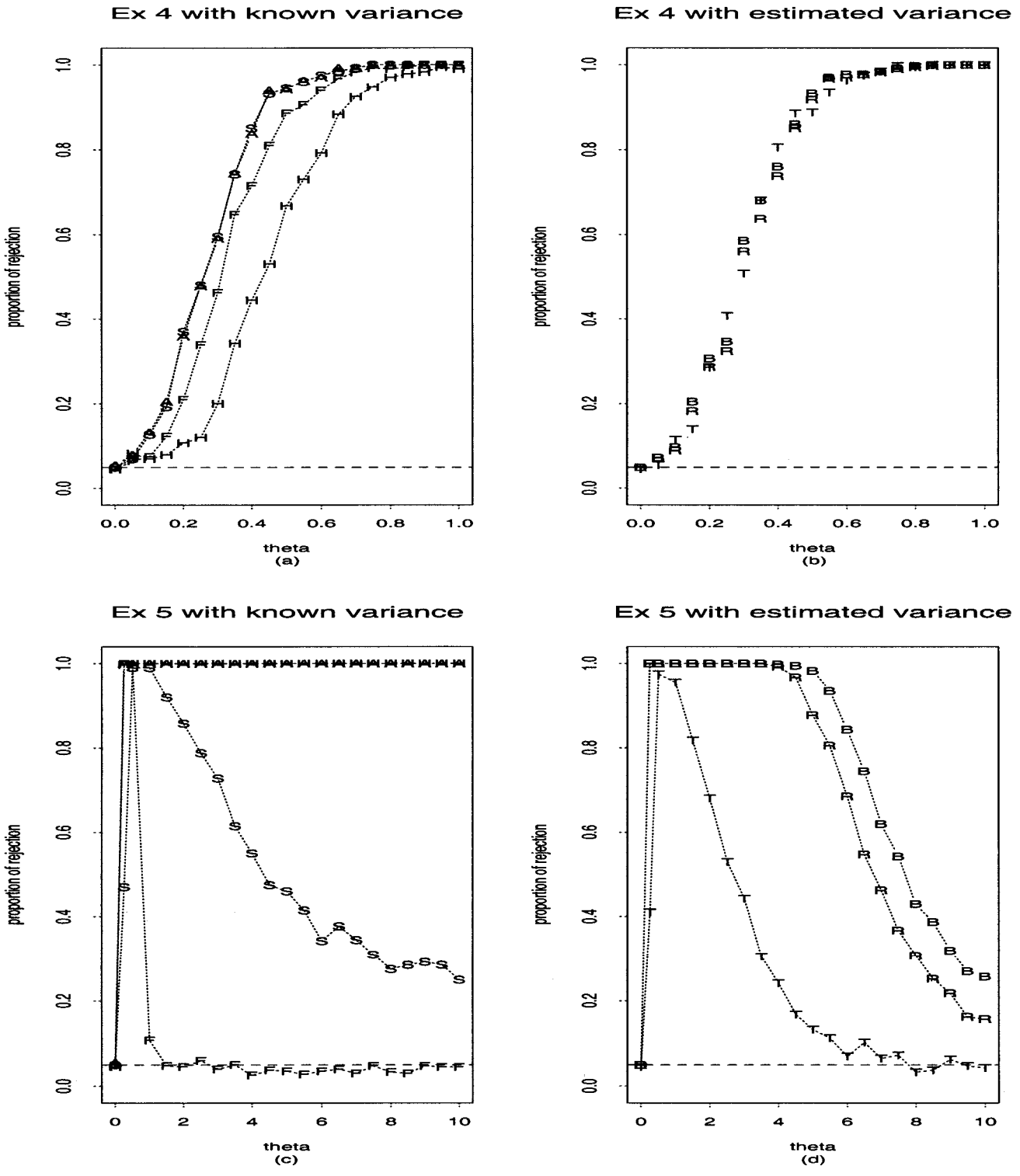

Figure 2. Power of the Adaptive Neyman Tests, the KH test, the Wavelet Thresholding Test, and the Parametric F Test for Examples 4 and 5 With Partial Linear Models as Alternative and n = 64. Key: A, the adaptive Neyman test with known variance; B, the adaptive Neyman test with estimated variance (2.10); F, the parametric F-test statistic; H, the wavelet thresholding test with known variance; R, the robust version of adaptive Neyman test with estimated variance (2.10) and (2.11); S, the KH test with known variance; T, the KH test with estimated variance (2.10).

for the adaptive Neyman test, we study the following four versions, depending on our knowledge of the model:

1. ordering according to X_2 and using the known variance ($c_\alpha = 3.17$, where c_α is the 5% empirical critical value)
2. ordering by X_2 and using $\hat{\sigma}_\ell$ ($c_\alpha = 3.61$)
3. ordering according to s_i in Section 2.2 and using the known variance ($c_\alpha = 3.06$)
4. ordering by s_i and using $\hat{\sigma}_\ell$ ($c_\alpha = 3.54$)

To demonstrate the versatility of our proposal, we also applied the KH test in the following ways:

1. ordering according to X_2 and using the known variance ($c_\alpha = 2.33$)
2. ordering by X_2 and using $\hat{\sigma}_\ell$ ($c_\alpha = 2.34$)
3. ordering according to the first principal component and using the known variance ($c_\alpha = 2.54$)
4. ordering by the first principal component and using $\hat{\sigma}_\ell$ ($c_\alpha = 2.38$)

Hart (1997, section 9.3) proposed performing a one-dimensional test a couple of times along the first couple of principal components and adjusting the significance level accordingly. For simplicity, we use only one principal direction in our implementation [see versions (3) and (4) of the KH test]. The results of the wavelet thresholding test with the ideal ordering (by X_2) and known variance are reported. The results of the conventional F test against the quadratic models are also included. The simulation results are reported in Figure 3.

First of all, we note that ordering according to X_2 is more powerful than ordering by the generic method or by the first principal direction. When ordering by X_2, the nonparametric procedures with estimated variance perform closely to the corresponding procedures with a known variance, except for the KH test at the alternative models (3.5) with high frequency. This in turn suggests that our generic variance estimator performs reasonably well. The F test in Example 4 performs comparably to the adaptive Neyman with ideal ordering, whereas for Example 5, the F test fails to detect high-frequency components in the model. The results are encouraging. Correctly ordering the covariate can produce a test that is nearly as powerful as knowing the alternative model.

4. SUMMARY AND CONCLUSION

The adaptive Neyman test is a powerful omnibus test. It is powerful against a wide class of alternatives. Indeed, as shown in Theorem 3, the adaptive Neyman test adapts automatically to a large class of functions with unknown degrees of smoothness. However, its power in the multiple regression setting depends on how the residuals are ordered. When the residuals are properly ordered, it can be very powerful as demonstrated in Section 3.3. This observation can be very useful for testing the linear model against the additive model. We just need to order the data according to the most sensible covariate. Estimation of residual variance also plays a critical role for the adaptive Neyman test. With a proper estimate of the residual variance, the adaptive Neyman test can be nearly as powerful as the case where the variance is known.

APPENDIX

In this Appendix, we establish the asymptotic distribution given in Theorem 1 and the asymptotic power expression given in Theorem 2. We first introduce some necessary notation that is for the linear model. We use the notation β instead of θ to denote the unknown parameters under the null hypothesis. Under the null hypothesis, we assume that

$$Y_i = \mathbf{x}_i^T \beta + \varepsilon, \qquad \varepsilon \sim N(0, \sigma^2),$$

where β is a p-dimensional unknown vector. Let $\mathbf{X} = (x_{ij})$ be the design matrix of the linear model. Let $x_1^*, \ldots, x_p^*$ be, respectively, the discrete Fourier transforms of the first, ..., the pth column of the design matrix $\mathbf{X}$. We impose the following technical conditions.

Conditions

A1. There exists a positive definite matrix A such that $n^{-1}\mathbf{x}^T\mathbf{x} \to A$.

A2. There exists an integer n_0 such that

$$\{n/(\log\log n)^4\}^{-1} \sum_{i=n_0}^{n/(\log\log n)^4} x_{ij}^{*2} = O(1), \qquad j = 1, \ldots, p,$$

where x_{ij}^* is the j^{th} element of the vector x_i^*.

A3. $\hat{\sigma}_1^2 = \sigma^2 + O_P(n^{-1/2})$ and $\hat{\sigma}_2^2 = 2\sigma^4 + O_P\{(\log n)^{-1}\}$.

A4. $\frac{1}{n}\sum_{i=1}^n m(\mathbf{x}_i)\mathbf{x}_i \to b$ for some vector b.

Condition A1 is a standard condition for the least-squares estimator $\hat{\beta}$ to be root-n consistent. It holds almost surely for the designs that are generated from a random sample of a population with a finite second moment. By Parseval's identity,

$$\frac{1}{n}\sum_{i=1}^n x_{ij}^{*2} = \frac{1}{n}\sum_{i=1}^n x_{ij}^2 = O(1).$$

Thus, Condition A2 is a very mild condition. It holds almost surely for the designs that are generated from a random sample. Condition A4 implies that under the alternative hypothesis, the least-squares estimator $\hat{\beta}$ converges in mean square error to $\beta_0 = A^{-1}b$.

Proof of Theorem 1

Let Γ be the $n \times n$ orthonormal matrix generated by the discrete Fourier transform. Denote

$$\mathbf{X}^* = \Gamma\mathbf{X} \quad \text{and} \quad \mathbf{Y} = (Y_1, \ldots, Y_n)^T.$$

Under the null hypothesis, our model can be written as

$$\mathbf{Y} = \mathbf{X}\beta + \varepsilon. \tag{A.1}$$

Then the least-squares estimate is given by $\hat{\beta} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}$. Denote $\mathbf{z} = \Gamma\varepsilon$ and $\mathbf{o} = \mathbf{X}^*(\hat{\beta} - \beta)$. Then,

$$\hat{\varepsilon}^* = \Gamma\varepsilon - \Gamma\mathbf{X}(\hat{\beta} - \beta) = \mathbf{z} - \mathbf{o} \quad \text{and} \quad \mathbf{z} \sim N(0, \sigma^2 I_n).$$

Let z_i and o_i be the ith components of the vectors $\mathbf{z}$ and $\mathbf{o}$, respectively. Then

$$\sum_{i=1}^m \hat{\varepsilon}_i^{*2} = \sum_{i=1}^m (z_i^2 - 2z_i o_i + o_i^2). \tag{A.2}$$

We first evaluate the small order term o_i^2. By the Cauchy–Schwarz inequality,

$$o_i^2 = \left[\sum_{j=1}^p x_{ij}^*\{\hat{\beta}_j - \beta_j\}\right]^2 \le \|\hat{\beta} - \beta\|^2 \sum_{j=1}^p x_{ij}^{*2}, \tag{A.3}$$

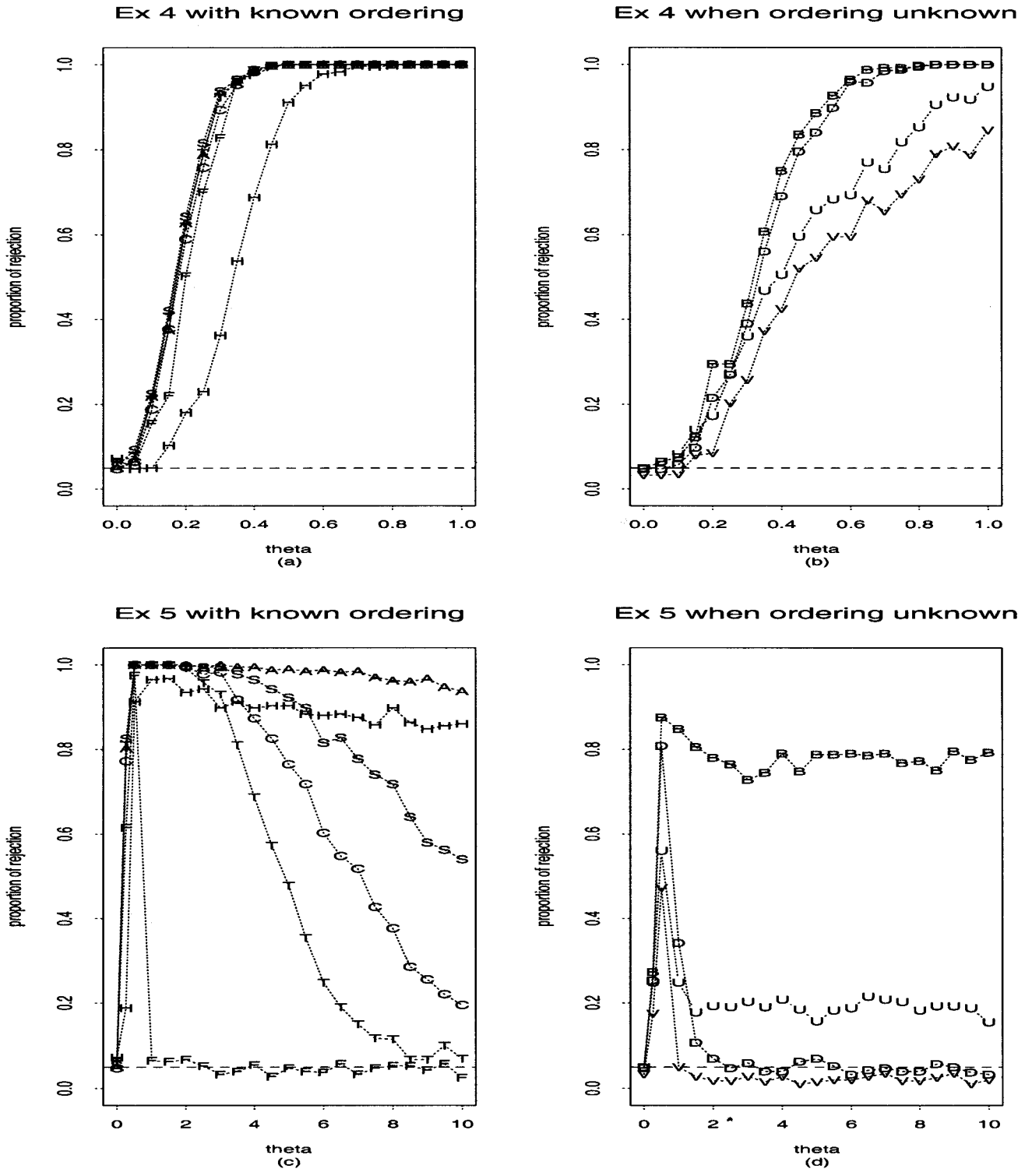

Figure 3. Power of the Adaptive Neyman Tests, the KH Test, the Wavelet Thresholding Test, and the Parametric F Test for Examples 4 and 5: When the Alternative Is Fully Nonparametric and $n = 128$. Key: A, the adaptive Neyman test (ANT) ordered according to X_2 and using the known variance; B, the ANT ordered according to s_i and using the known variance; C, the ANT test ordered by X_2 and using $\hat{\sigma}_\ell$; D, the ANT test ordered by s_i and using $\hat{\sigma}_\ell$; F, the parametric F-test statistic; H, the wavelet thresholding test ordered by X_2 and using the known variance; S, the KH test with known variance and ordered by X_2; T, the KH test with estimated variance $\hat{\sigma}_\ell$ and ordered by X_2; U, the KH test with known variance and ordered by first PC; V, the KH test with estimated variance $\hat{\sigma}_\ell$ and ordered by the first PC.

where $\hat{\beta}_j$ and β_j are the jth components of the vectors $\hat{\beta}$ and β, respectively. By the linear model (A.1) and Condition A1, we have

$$E(\hat{\beta}-\beta)(\hat{\beta}-\beta)^T = \sigma^2(\mathbf{X}^T\mathbf{X})^{-1} = n^{-1}A^{-1}\{\sigma^2+o(1)\}.$$

This and Condition A2 together yield

$$\sum_{i=n_0}^{m} o_i^2 \le \|\hat{\beta}-\beta\|^2 \sum_{j=1}^{p}\sum_{i=n_0}^{n/(\log\log n)^4} x_{ij}^{*2} = O_P\{(\log\log n)^{-4}\}. \quad (A.4)$$

We now deal with the second term in (A.2). Observe that

$$m^{-1}\sum_{i=1}^{m} z_i^2 \le \sigma^2 + \max_{1\le m\le n} m^{-1}\sum_{i=1}^{m}(z_i^2-\sigma^2) = o_P(\log\log n). \quad (A.5)$$

The last equality in (A.5) follows from (A.8). By the Cauchy–Schwarz inequality, (A.4), and (A.5), we have

$$\left|m^{-1/2}\sum_{i=n_0}^{m} o_i z_i\right| \le \left\{\sum_{i=n_0}^{m} o_i^2\right\}^{1/2}\left\{m^{-1}\sum_{i=n_0}^{m} z_i^2\right\}^{1/2} = o_P\{(\log\log n)^{-3/2}\}.$$

This together with (A.2) and (A.4) entails

$$m^{-1/2}\sum_{i=n_0}^{m}\hat{\varepsilon}_i^{2*} = m^{-1/2}\sum_{i=n_0}^{m} z_i^2 + o_P\{(\log\log n)^{-3/2}\}. \quad (A.6)$$

Let

$$T_n^* = \max_{1\le m\le n}(2m\sigma^4)^{-1/2}\sum_{i=1}^{m}(z_i^2-\sigma^2).$$

Then, by theorem 1 of Darling and Erdös (1956), we have

$$P[\sqrt{2\log\log n}\,T_n^* - \{2\log\log n + .5\log\log\log n - .5\log(4\pi)\} \le x] \to \exp(-\exp(-x)) \quad (A.7)$$

Hence,

$$T_n^* = \{2\log\log n\}^{1/2}\{1+o_P(1)\} \quad (A.8)$$

and

$$T_{\log n}^* = \{2\log\log\log n\}^{1/2}\{1+o_P(1)\}.$$

This implies that the maximum of T_n^* cannot be achieved at $m < \log n$. Thus, by (2.1),

$$T_{\mathrm{AN},1}^* = \max_{1\le m\le n/(\log\log n)^4}\frac{1}{\sqrt{2m\sigma^4}}\sum_{i=1}^{m}(\hat{\varepsilon}_i^{*2}-\sigma^2) + o_p\left\{(\log\log n)^{-3/2}\right\}.$$

Combination of this and (A.6) entails

$$T_{\mathrm{AN},1}^* = T_n^* + O_p\{(\log\log n)^{-3/2}\}.$$

The conclusion for $T_{\mathrm{AN},1}$ follows from (A.7) and the conclusion for $T_{\mathrm{AN},2}$ follows from the same arguments.

Proof of Theorem 2

Let m_0^* be the index such that

$$\eta_n^* = \max_{1\le m\le n}(2\sigma^4 m)^{-1/2}\sum_{i=1}^{m}\eta_i^{*2}$$

is achieved. Denote $c_\alpha = -\log\{-\log(1-\alpha)\}$.

Using arguments similar to but more tedious than those in the proof of Theorem 1, we can show that under the alternative hypothesis,

$$T_{\mathrm{AN,j}}^* = T_{\mathrm{AN}}^* + o_p\{(\log\log n)^{-3/2} + (\log\log n)^{-3/2}(\eta_n^*)^{1/2}\},$$

where $\mathbf{z}$ in T_{AN}^* is distributed as $N(\eta^*, \sigma^2 I_n)$. Therefore, the power can be expressed as

$$P\{T_{\mathrm{AN},j} > c_\alpha\} = P[T_{\mathrm{AN}}^* > (\log\log n)^{1/2} \times \{1+o(1)\} + o\{(\eta_n^*)^{1/2}\}]. \quad (A.9)$$

Then, by (A.9), we have

$$\begin{aligned} P\{T_{\mathrm{AN},j} > c_\alpha\} &\ge P\left\{(2m_0^*\sigma^4)^{-1/2}\sum_{i=1}^{m_0^*}(z_i^2-\sigma^2) > 2(\log\log n)^{1/2}\right\} \\ &\ge P\left\{(2m_0^*\sigma^4)^{-1/2}\sum_{i=1}^{m_0^*}(z_i^2-\eta_i^{*2}-\sigma^2) \right. \\ &\qquad \left. > 2(\log\log n)^{1/2} - \eta_n^*\right\}. \end{aligned} \quad (A.10)$$

The sequence of random variables

$$\left\{\left(2m\sigma^4 + 4\sigma^2\sum_{i=1}^{m}\eta_i^2\right)^{-1/2}\sum_{i=1}^{m}(z_i^2-\eta_i^{*2}-\sigma^2)\right\}$$

is tight, because they have the mean zero and standard deviation 1. It follows from (A.10) and the assumption (2.8) that the power of the critical regions in (2.5) tends to 1.

[Received April 1999. Revised June 2000.]

REFERENCES

Aerts, M., Claeskens, G., and Hart, J. D. (1999), "Testing Lack of Fit in Multiple Regression," *Journal of the American Statistical Association*, 94, 869–879.

Azzalini, A., and Bowman, A. N. (1993), "On the Use of Nonparametric Regression for Checking Linear Relationships," *Journal of the Royal Statistical Society*, Ser. B, 55, 549–557.

Azzalini, A., Bowman, A. N., and Härdle, W. (1989), "On the Use of Nonparametric Regression for Model Checking," *Biometrika*, 76, 1–11.

Barnett, V. (1976), "The Ordering of Multivariate Data" (with Discussion), *Journal of the Royal Statistical Society*, Ser. A, 139, 318–355.

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, New York: North-Holland, pp. 51–57.

Bowman, A. W., and Azzalini, A. (1997), *Applied Smoothing Techniques for Data Analysis*, London: Oxford University Press.

Cuzick, J. (1992), "Semiparametric Additive Regression," *Journal of the Royal Statistical Society*, Ser. B, 54, 831–843.

Darling, D. A., and Erdős, P. (1956), "A Limit Theorem for the Maximum of Normalized Sums of Independent Random Variables," *Duke Mathematical Journal*, 23, 143–155.

Daubechies, I. (1992), *Ten Lectures on Wavelets*, Philadelphia: SIAM.

Dette, H. (1999), "A Consistent Test for the Functional Form of a Regression Based on a Difference of Variance Estimators," *The Annals of Statistics*, 27, 1012–1040.

Dette, H., and Munk, A. (1998), "Validation of Linear Regression Models," *The Annals of Statistics*, 26, 778–800.

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., and LaRiccia, V. M. (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), "Test of Significance Based on Wavelet Thresholding and Neyman's Truncation," *Journal of the American Statistical Association*, 91, 674–688.

———(1997), Comments on "Polynomial Splines and Their Tensor Products in Extended Linear Modeling," *The Annals of Statistics*, 25, 1425–1432.

Fan, J., and Lin, S. K. (1998), Test of Significance When Data Are Curves," *Journal of the American Statistical Association*, 93, 1007–1021.

Gnanadesikan, R., and Kettenring, J. R. (1972), "Robust Estimates, Residuals, and Outlier Detection with Multiresponse Data," *Biometrics*, 28, 81–124.

Green, P. J., and Silverman, B. W. (1994), *Nonparametric Regression and Generalized Linear Models: A Roughness Penalty Approach*, London: Chapman and Hall.

Härdle, W., and Mammen, E. (1993), "Comparing Nonparametric Versus Parametric Regression Fits," *The Annals of Statistics*, 21, 1926–1947.

Hall, P., Kay, J. W., and Titterington, D. M. (1990), "Asymptotically Optimal Difference-Based Estimation of Variance in Nonparametric Regression," *Biometrika*, 77, 521–528.

Hart, J. D. (1997), *Nonparametric Smoothing and Lack-of-Fit Tests*, New York: Springer-Verlag.

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

Inglot, T., and Ledwina, T. (1996), "Asymptotic Optimality of Data-Driven Neyman Tests for Uniformity," *The Annals of Statistics*, 24, 1982–2019.

Jolliffe, I. T. (1986), *Principal Component Analysis*, New York: Springer-Verlag.

Kim, J. (2000), "An Order Selection Criteria for Testing Goodness of Fit," *Journal of the American Statistical Association*, 95, 829–835.

Kuchibhatla, M., and Hart, J. D. (1996), "Smoothing-Based Lack-of-Fit Tests: Variations on a Theme," *Journal of Nonparametric Statistics*, 7, 1–22.

Lee, G., and Hart, J. D. (1998), "An L_2 Error Test with Order Selection and Thresholding," *Statistics and Probability Letters*, 39, 61–72.

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

Neyman, J. (1937), "Smooth Test for Goodness of Fit," *Skandinavisk Aktuarietidskrift*, 20, 149–199.

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

Speckman, P. (1988), "Kernel Smoothing in Partial Linear Models," *Journal of the Royal Statistical Society*, Ser. B, 50, 413–436.

Spokoiny, V. G. (1996), "Adaptive Hypothesis Testing Using Wavelets," *The Annals of Statistics*, 24, 2477–2498.

Stute, W., González Manteiga, W., and Presedo Quindimil, M. (1998), "Bootstrap Approximations in Model Checks for Regression," *Journal of the American Statistical Association*, 93, 141–149.

Wahba, G. (1984), "Partial Spline Models for the Semiparametric Estimation of Functions of Several Variables," in *Analyses for Time Series*, Tokyo: Institute of Statistical Mathematics, pp. 319–329.

Yatchew, A. (1997), "An Elementary Estimator of the Partial Linear Model," *Economics Letters*, 57, 135–143.