Title
Goodness-of-fit test for parametric models

Authors
Jianqing Fan
Lishan Huang

Publication Date
2011-10-25

Goodness-of-Fit Test for Parametric Regression Models

Jianqing Fan*
Department of Statistics
University of North Carolina
Chapel Hill, N.C. 27599-3260

Li-Shan Huang†
Department of Statistics
Florida State University
Tallahassee, F.L. 32306-4330

July 15, 1999

Abstract

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 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 are addressed.

Keywords: Adaptive Neyman test; contiguous alternatives; partial linear model; power; wavelet thresholding.

Abbreviated title: Regression goodness-of-fit

*Partially supported by NSF Grant NSA 96-1-0015 and NSF DMS-9803200.

†Partially supported by NSF Grant DMS-9710084.

1 Introduction

Parametric linear models are frequently used to describe the association between a response variable and its predictors, and the adequacy of 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 can not be observed easily. Recently there have been many papers investigating the use of nonparametric techniques for model diagnostics. Most of them focus on one-dimensional problems. The book by Hart (1997) gave an extensive overview and many useful references. Chapter 5 of Bowman and Azzalini (1997) outlined the work by Azzalini *et al.* (1989) and Azzalini and Bowman (1993). While a lot of work has focused on univariate case, there is relatively little work on multiple regression setting. In Section 9.3, Hart (1997) considered two approaches: one is to regress residuals on a scalar function of the covariates and then apply one-dimensional goodness-of-fit tests, and the second approach is to explicitly take into account of the multivariate nature. The test by Härdle and Mammen (1993) is based on the L_2 error criterion in d-dimensions, and Aerts *et al.* (1998) construct tests based on orthogonal series which involve choosing a nested model sequence in the (bivariate) multiple regression. While these ideas are useful, their implementations encounter the difficulty of so-called "curse of dimensionality".

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

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

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

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

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

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

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

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

$$H_1 : m(x) = f(x_1) + \mathbf{x}_2^T \beta_2, \text{with } f \text{ nonlinear}, \tag{1.4}$$

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 heteroscedacity, among others. This amounts to informally checking if biases are negligible in presence of large stochastic noises. These techniques can be formalized as follows. Let $\hat{\theta}$ be an estimate under the null hypothesis, and $\hat{\varepsilon} = (\hat{\varepsilon}_1, \cdots, \hat{\varepsilon}_n)^T$ be the resulting residuals with $\hat{\varepsilon}_i = Y_i - f(\mathbf{x}_i, \hat{\theta})$. Then under model (1.1), conditioned on $\{\mathbf{x}_i\}_{i=1}^n$ and $\hat{\theta}$,

$$\hat{\varepsilon} \overset{a}{\sim} N_n(\eta, \sigma^2 I_n), \tag{1.5}$$

where "$\overset{a}{\sim}$" means "distributed approximately", N_n denotes an n-dimensional multivariate normal distribution, and $\eta = (\eta_1, \cdots, \eta_n)^T$ with $\eta_i = m(\mathbf{x}_i) - f(\mathbf{x}_i, \hat{\theta})$. Thus, the problem (1.1) becomes

$$H_0 : \eta = 0 \qquad \text{versus} \qquad H_1 : \eta \neq 0, \tag{1.6}$$

based on the observations $\hat{\varepsilon}$. Plotting a covariate X_j against the residual $\hat{\varepsilon}$ intends to testing if the biases in $\hat{\varepsilon}$ along the direction X_j are negligible.

In testing the significance of the high-dimensional problem (1.6), the conventional likelihood ratio statistic is not powerful due to noise accumulation, as demonstrated in Fan (1996). An innovative idea was proposed by Neyman (1937) who tested only first m-dimensional subproblem if there is prior that most of nonzero elements lie on the first m-dimension. To obtain such a kind qualitative prior, a Fourier transform is applied to the residuals in an attempt to compress nonzero signals into lower frequencies. 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 (1992), Inglot *et al.* (1994), Ledwina (1994), Kuchibhatla and Hart (1996), Lee and Hart (1998), among others. See also Bickel and Ritov (1992) and Inglot and Ledwina (1996) for illuminating insights into nonparametric tests. 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. In the present paper we adopt the adaptive Neyman test proposed by Fan (1996) to the testing problem (1.6). The adaptive Neyman test proposed by Kuchibhatla and Hart (1996) will also be implemented to demonstrate the versatility of our proposed idea.

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 as η, namely $\eta(i/n) = \eta_i$. The smoother the function η, the more significantly the Fourier coefficients are on the low-frequencies and hence the more powerful the test will be. See Theorem 3. When $m(\cdot)$ is completely unknown such as those in (1.3), there is no information on how to make the function $\eta(\cdot)$ smoother by ordering the residuals. Intuitively, the closer $\mathbf{x}_i$ and $\mathbf{x}_{i+1}$, the smaller difference between $m(\mathbf{x}_i)$ and $m(\mathbf{x}_{i+1})$ and hence the smoother the function η. Thus, a good proxy is to order the covariates $\{\mathbf{x}_i\}_{i=1}^n$ in a way that they are close as a sequence. Note that when there is only one predictor variable, i.e. $p = 1$, the 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 regression surface 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. The parameters β_2 in model (1.4) can be estimated easily and the problem is then reduced to the one-dimensional nonparametric setting. A simple and effective new estimator for β_2 is proposed. See Section 2.4 for details.

Wavelet transform is another popular family of orthogonal transforms. It can be applied to the testing problem (1.6). For comparison purpose, 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.

The paper 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 include the adaptive

Neyman test and wavelet thresholding tests in Sections 2.1 and 2.3. 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. Section 2.5 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. A strategy for testing linear models against additive models is also discussed. 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 (1.6) based on the approximate Gaussian random vector $\hat{\varepsilon}$. Two methods, the adaptive Neyman test and the wavelet thresholding test, were proposed in Fan (1996) for testing (1.6). They can be adapted to the current regression setting.

2.1 The Adaptive Neyman test

Let $\hat{\varepsilon}^* = (\hat{\varepsilon}_1^*, \cdots, \hat{\varepsilon}_n^*)^T$ be the discrete Fourier transform of the residual vector $\hat{\varepsilon}$. 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 alternative hypotheses. Methods for constructing such an estimate will be outlined in Section 2.5. The adaptive Neyman test statistic is defined as

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

The null hypothesis in (1.2) is rejected when $T_{AN,1}^*$ 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 non-normal 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 following test statistic:

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

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

$$T_{AN,j} = \sqrt{2\log\log n}T^*_{AN,j} - \{2\log\log n + 0.5\log\log\log n - 0.5\log(4\pi)\}, \text{ for } j = 1, 2.$$

Under the null hypothesis and the exact normal model with a known variance, it can be shown that

$$P(T_{AN,j} < x) \to \exp(-\exp(-x)), \qquad \text{as } n \to \infty. \tag{2.3}$$

See Darling and Erdös (1956) and Fan (1996). However, the approximation (2.3) is not so good. Let us call the exact null distribution of $T_{AN,1}$ under the 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 excerpted some of their results in Table 1. These sample quantiles can be used for choosing the critical values of the adaptive Neyman tests $T_{AN,j}$.

Table 1: α upper quantile of the distribution $J_n^\dagger$

$\alpha\backslash n$	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

We now show that the approximation (2.3) holds for testing linear models. 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), we have*

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

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

$$T_{AN,j} > -\log\{-\log(1-\alpha)\} \quad (j = 1, 2) \tag{2.4}$$

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 $m(\cdot)$. Let $\eta_i = m(\mathbf{x}_i) - f(\mathbf{x}_i, \theta_0)$ and η^* be the Fourier transform of the vector $\eta = (\eta_1, \cdots, \eta_n)^T$. Then we have the following power result.

Theorem 2 *Under Conditions (A1) – (A4) in the appendix, if*

$$(\log\log n)^{-1/2} \max_{1\le m\le n} m^{-1/2} \sum_{i=1}^{m} \eta_i^{*2} \to \infty, \tag{2.5}$$

then the critical regions (2.4) have an asymptotic power one.

We now give an implication of condition (2.5). Note that by the Parseval's identity

$$m^{-1/2} \sum_{i=1}^{m} \eta_i^{*2} = m^{-1/2} \{ \sum_{i=1}^{n} \eta_i^2 - \sum_{i=m+1}^{n} \eta_i^{*2} \}. \tag{2.6}$$

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

$$n^{-1} \sum_{i=m+1}^{n} \eta_i^{*2} = O(m^{-2s}). \tag{2.7}$$

Then, the maximum value of (2.6) over m is of order

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

For random designs with a density f,

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

By theorem 2, we have the following result.

Theorem 3 *Under conditions of Theorem 2, if the smoothness condition (2.7) holds, then the adaptive Neyman test has an asymptotic power one for the contiguous alternative with*

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

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

The significance of the above theoretical result is its adaptivity. The adaptive Neyman test does not at all depend on the smoothness assumption s. Nevertheless, it can adaptively detect alternatives with rates $O(n^{-2s/(4s+1)} (\log\log n)^{2s/(4s+1)})$, 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.

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 m so that its large Fourier coefficients are concentrated on low frequencies.

When the alternative is the partial linear model (1.4), one 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, among other ordering schemes. For saturated nonparametric alternative (1.3), there is little useful information on how to order the sequence. A good strategy is to order the covariates $\{\mathbf{x}_i\}_{i=1}^n$ so that they are close to each other consecutively. 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 $\{m(\mathbf{x}_i) - f(\mathbf{x}_i, \theta_0)\}$. Therefore it is useful to order the covariates in such a way that the distance between two consecutive covariates are small. This problem can easily 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 i-th observation and then order the observations according to the rank of s_i. This is called "reduced ordering" in Barnett (1976).

What could be a reasonable score when $p > 1$? 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 $\{\mathbf{x}_i, i = 1, \cdots, n\}$. 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 $\mathbf{x}_i - \bar{\mathbf{x}} = z_{i,1}\xi_1 + \ldots + z_{i,p}\xi_p$, 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 as the "sample score of variation". See also Gnanadesikan and Ketterring (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 3.3 for further discussions.

In the case of two covariates, Aerts *et al.* (1998) construct a score test statistic and its power depends also on 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. One may naturally ask how other families of orthogonal transforms behave. For comparison purposes, we apply the wavelet hard-thresholding test of Fan (1996) to the current setting. Only brief outlines are given here.

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 indices. 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), $J_0 = 3$ will be used, which keeps the wavelet coefficients in the first two resolution levels intact. Then T_H has an asymptotic normal distribution under H_0 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 further improved by using the following techniques due to Fan (1996). Under the Gaussian model and the null hypothesis, the maximum of $\hat{\varepsilon}_{W,i}$ is of order $\sqrt{2\log n}$. 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. However, under the alternative hypothesis, the latter is smaller than a_n and hence the procedure uses automatically a smaller thresholding parameter and hence enhances the 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 it improves a great deal on the power of the generic methods. Nevertheless, the power can still 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 $\hat{\varepsilon}_i = Y_i - \mathbf{x}_{i,2}^T\hat{\beta}_2$. Based on the induced data $\{(X_{i1}, \hat{\varepsilon}_i)\}$, the problem reduces to testing the one-dimensional linear regression model. We can therefore apply the adaptive Neyman and the other tests to the data $\{(X_{i1}, \hat{\varepsilon}_i)\}$.

There is a large literature on discussing efficient estimation of coefficients β_2. See, for example, Wahba (1984), Speckman (1988), Cuzick (1992), Green and Silverman (1994) and the references therein. The methods proposed in the literature involve choices of smoothing parameters. For our purpose, a root-n consistent estimator of β_2 suffices. Here we offer a simple alternative method to estimate β_2.

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

$$Y_{(i+1)} - Y_{(i)} = (\mathbf{x}_{(i+1),2} - \mathbf{x}_{(i),2})^T\beta_2 + e_i + O_P(n^{-1}), \quad i = 1, \cdots, n-1, \tag{2.8}$$

where $\{e_i\}$ are correlated stochastic errors with $e_i = \varepsilon_{(i+1)} - \varepsilon_{(i)}$. Thus, β_2 can be estimated from the linear model (2.8). To correct the bias in the above approximation, particularly for those $x_{(i),1}$ at tails (hence the spacing can be wide), we fit the following linear regression model in our implementation:

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

by using the ordinary least-squares method. The general least-squares method can also be used here to cope with the correlated data, but we opt not to use it just for the sake of simplicity.

2.5 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 deteriorate the power of the tests. A root-n consistent estimator can be constructed by using the residuals of a nonparametric kernel regression fit. But this theoretically satisfactory method encounters the difficulty of 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 *et al.* (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,

$$\hat{\varepsilon}_j^* \overset{a}{\sim} N(0, \sigma^2), \qquad \text{for large } j. \tag{2.9}$$

This gives us a simple and effective way to estimate σ^2 by using the sample variance of $\{\hat{\varepsilon}_i^*, i = I_n + 1, \ldots, n\}$:

$$\hat{\sigma}_1^2 = \frac{1}{n - I_n} \sum_{i=I_n+1}^{n} \hat{\varepsilon}_i^{*2} - \{\frac{1}{n - I_n} \sum_{i=I_n+1}^{n} \hat{\varepsilon}_i^*\}^2. \tag{2.10}$$

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 for 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 $\{\hat{\varepsilon}_i^{*2}, i = n/4 + 1, \ldots, n\}$:

$$\hat{\sigma_2}^2 = \frac{1}{n - I_n} \sum_{i=I_n+1}^{n} \hat{\varepsilon}_i^{*4} - \{\frac{1}{n - I_n} \sum_{i=I_n+1}^{n} \hat{\varepsilon}_i^{*2}\}^2. \tag{2.11}$$

In the case of multiple regression ($p > 1$), the above 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 (2.9) holds. Thus, the estimator (2.10) can involve with substantial bias. Indeed, it is difficult to find a practically satisfactory estimate for large p. This problem is beyond the scope of the paper. 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. An estimate $\hat{\sigma}_\ell$ can be obtained by fitting linear splines with four knots and the interaction terms between continuous predictors:

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

where $t_{i,j}$ denotes the $(20j)$-th percentile, $j = 1, 2, 3, 4$, of the i-th covariate, $i = 1, 2, 3$ (continuous covariates only). In this model, there are twenty 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 ($T_{AN,1}$ and $T_{AN,2}$) and the wavelet thresholding test T_H via simulations. The simulated examples consist of testing 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%, give or take 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 that in the multiple regression setting, repeated measurements are available at some design points. Another reason for using the known variances 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 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 only report the results of the wavelet thresholding test when the residual variances are given.

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

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

where p is the number of parameters fitted in the null model. Comparing with 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 smooth, and less powerful than the adaptive Neyman test when the alternative is not 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's 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:

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

The sample size is 64 and the residuals $\hat{\varepsilon}_i$'s 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 $T_{AN,1}$ and $T_{AN,2}$ 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

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

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 ideal F-test, while the wavelet thresholding test falls behind the other three tests. See Figures 1(a) & (b). In fact, the wavelet tests are not designated for testing this kind of very smooth alternatives. This example also is designated to show how large price the adaptive Neyman and the KH

tests have to pay in order to be more omnibus. Surprisingly, Figure 1 demonstrates that both procedures paid very little price.

Put Figure 1 about here

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

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

This example intends to examine how powerful each testing procedure is for detecting alternatives with different frequency components. The result is presented in Figures 1 (c) & (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 latter 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:

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

where $X_1 \sim N(0,1)$. Figures 1 (e) & (f) depict 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, while it is dominated by the adaptive Neyman and KH tests.

3.2 Testing linearity in partially linear models

In this section, we are testing 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 $X_1, \cdots, X_4$ are covariates. The covariates X_1, X_2 and X_3 are normally distributed with mean 0 and variance 1. Further, the correlation coefficients among these three random variables are 0.5. The covariate X_4 is binary, independent of X_1, X_2 and X_3, with

$$P(X_4 = 1) = 0.4 \quad \text{and} \quad P(X_4 = 0) = 0.6.$$

The techniques proposed in Section 2.3 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 linear model against the quadratic model

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

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

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

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. This is due partially to the fact that an overparametrized 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 in Section 2.4.

Put Figure 2 about here

Example 5. We simulate the response variable from

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

The adaptive Neyman, wavelet thresholding, and KH tests perform well when σ is known. The F-test gives low power for $\theta \geq 1.5$, 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 can not 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 in order to have a powerful test, σ should be estimated well under both null and 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 a generic method in Section 2.2. Thus, 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's and using $\hat{\sigma}_\ell$ ($c_\alpha = 3.54$).

To demonstrate the versatility of our proposal, we also apply 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 $\hat{\sigma}_\ell$ ($c_\alpha = 2.38$).

In Section 9.3, Hart (1997) proposes performing one-dimensional test a couple of times along the first couple principal components and adjust the significance level accordingly. For simplicity, we only use 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 conventional F-test against quadratic models are also included. The simulation results are reported in Figure 3.

Put Figure 3 about here

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 with the adaptive Neyman with the ideal ordering, while for Example 5, the F-test fails to detect high-frequency components in the model.

The above results are encouraging. Correctly ordering the covariate can produce a test that is nearly as powerful as knowing the alternative model (comparing Figure 3 with Figure 2). This in turn suggests that one can powerfully test the linear model versus the additive model

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

simply via ordering residuals according to each of the covariates. The Bonferroni adjustment can be used to combine the resulting p test statistics.

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. One just needs to order covariates 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 the appendix, we will establish the asymptotic distribution given in Theorem 1 and the asymptotic power expression given in Theorem 2. We first introduce some necessary notation 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

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

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

Condition A

(1) There exists a positive definite matrix A such that $n^{-1}X^TX \to A$.

(2) There exists an integer n_0 such that

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

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

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

(4) $\frac{1}{n}\sum_{i=1}^n m(\mathbf{x}_i)\mathbf{x}_i \to b$ 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 by

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

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 by $\mathbf{z} = \Gamma\varepsilon$ 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 i^{th} component 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-Schwartz inequality,

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

where $\hat{\beta}_j$ and β_j are the j^{th} component of the vectors $\hat{\beta}$ and β. 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

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

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

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

The last equality in (A.5) follows from Theorem 1 in Darlin and Erdös (1956) [see also (A.7) below]. By the Cauchy-Schwartz inequality, (A.4) and (A.5), we have

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

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

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

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 Darlin and Erdös (1956), we have

$$P[\sqrt{2\log\log n} T_n^* - \{2\log\log n + 0.5\log\log\log n - 0.5\log(4\pi)\} \le x] \to \exp(-\exp(-x)) \tag{A.7}$$

Hence,

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

and

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

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

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

Combination of this and (A.6) entails that

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

The conclusion for $T_{AN,1}$ follows from (A.7) and the conclusion for $T_{AN,2}$ 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 by $c_\alpha = -\log\{-\log(1-\alpha)\}$.

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

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

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

$$P\{T_{AN,j} > c_\alpha\} = P[T_{AN}^* > (\log\log n)^{1/2}\{1 + o(1)\} + o\{(\eta_n^*)^{1/2}\}]. \tag{A.8}$$

Then, by (A.8), we have

$$\begin{aligned} & P\{T_{AN,j} > c_\alpha\} \\ \geq\ & P\{(2m_0^*\sigma^4)^{-1/2}\sum_{i=1}^{m_0^*}(z_i^2-\sigma^2) > 2(\log\log n)^{1/2}\} \\ \geq\ & P\{(2m_0^*\sigma^4)^{-1/2}\sum_{i=1}^{m_0^*}(z_i^2-\eta_i^{*2}-\sigma^2) > 2(\log\log n)^{1/2}-\eta_n^*\}. \end{aligned} \tag{A.9}$$

Since the sequence of random variables have

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

the mean zero and standard deviation one, they are tight. It follows from (A.9) and the assumption (2.7) that the power of the critical regions in (2.4) tends to one.

References

Aerts, M., Claeskens, G. & Hart, J.D. (1998). Testing lack of fit in multiple regression. *Manuscript*.

Azzalini, A. & Bowman, A.N. (1993). On the use of nonparametric regression for checking linear relationships. *J. Roy. Statist. Soc. Ser.B* **55**, 549-557.

Azzalini, A., Bowman, A.N. & 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). *J. R. Statist. Soc.* A. **139**, 318-55.

Bickel, P.J. & Ritov, Y. (1992). Testing for goodness of fit: a new approach. In *Nonparametric Statistics and Related Topics*, Ed. A.K.Md.E. Saleh, pp.51-7. New York: North-Holland.

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

Cuzick, J. (1992). Semiparametric additive regression. *J. R. Statist. Soc.* B **54**, 831-43.

Darling, D.A. & Erdös, P. (1956). A limit theorem for the maximum of normalized sums of independent random variables. *Duke Math. J.* **23**, 143-55.

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

Eubank, R.L. & LaRiccia, V.M. (1992). Asymptotic comparison of Cramér-von Mises and nonparametric function estimation techniques for testing goodness-of-fit. *Ann. Statist.* **20**, 2071-86.

Fan, J. (1996). Test of significance based on wavelet thresholding and Neyman's truncation. *J. Am. Statist. Assoc.* **91**, 674-88.

Fan, J. (1997). Comments on Polynomial splines and their tensor products in extended linear modeling. *Ann. Statist.* **25**, 1425–32.

Fan, J. & Lin, S.K. (1998). Test of significance when data are curves. *J. Am. Statist. Assoc.*, to appear.

Gnannadesikan, R. & Kettenring, J.R. (1972). Robust estimates, residuals, and outlier detection with multiresponse data. *Biometrics* **28**, 81-124.

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

Härdle, W. & Mammen, E. (1993). Comparing nonparametric versus parametric regression fits. *Ann. Statist.* **21**, 1926–47.

Hall, P., Kay, J.W. & Titterington, D.M. (1990). Asymptotically optimal difference-based estimation of variance in nonparametric regression *Biometrika* **77**, 521-8.

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

Inglot, T., Kallenberg, W.C.M. & Ledwina, T. (1994). Power approximations to and power comparison of smooth goodness-of-fit tests. *Scand. J. Statist.* **21**, 131-45.

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

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

Kim, J. (1992). Testing goodness-of-fit via order selection criteria. *PhD Thesis.* Department of Statistics, Texas A & M University.

Kuchibhatla, M. & Hart, J.D. (1996). Smoothing-based lack-of-fit tests: variations on a theme. *Jour. Nonpara. Statist.* **7**, 1-22.

Lee, G. & Hart, J.D. (1998). An L_2 error test with order selection and thresholding. *Statist. Probab. Lett.* **39**, 61-72.

Ledwina, T. (1994). Data-driven version of Neyman's smooth test of fit. *J. Am. Statist. Assoc.* **89**, 1000-05.

Neyman, J. (1937). Smooth test for goodness of fit. *Skandinavisk Aktuarietidskrift* **20**, 149-99.

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

Speckman, P. (1988). Kernel smoothing in partial linear models. *J. R. Statist. Soc.* B **50**, 413–36.

Spokoiny, V.G. (1996). Adaptive hypothesis testing using wavelets. *Ann. Statist.* **24**, 2477-98.

Wahba, G. (1984). Partial spline models for the semiparametric estimation of functions of several variables. In *Analyses for Time Series, Japan-U.S. Joint Sem.*, pp.319–29. Tokyo: Institute of Statistical Mathematics.

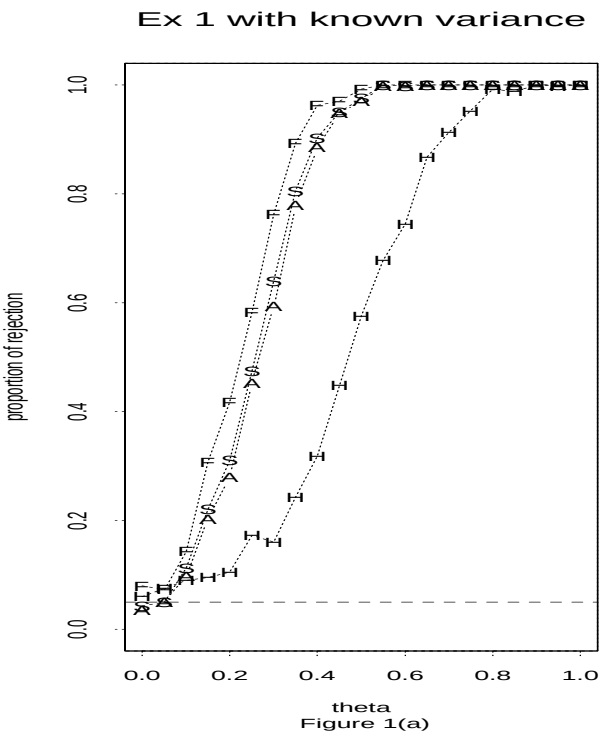

Figure 1(a)

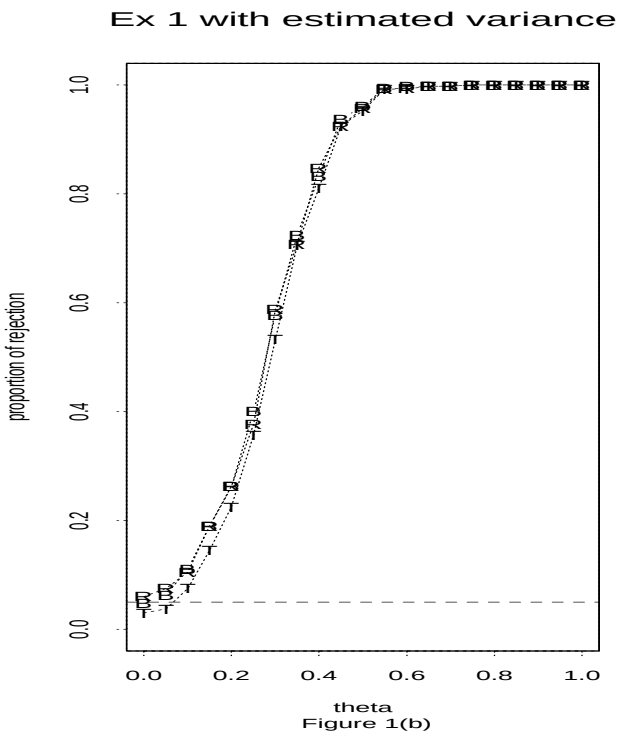

Figure 1(b)

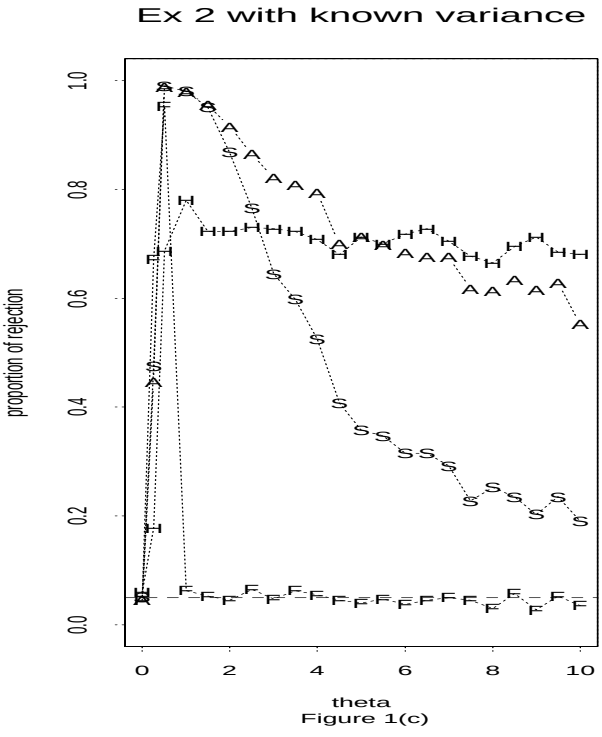

Figure 1(c)

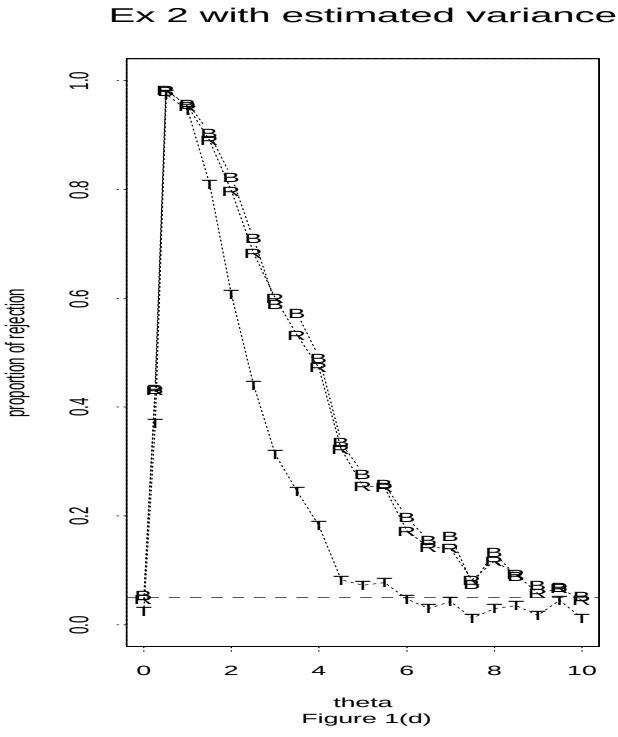

Figure 1(d)

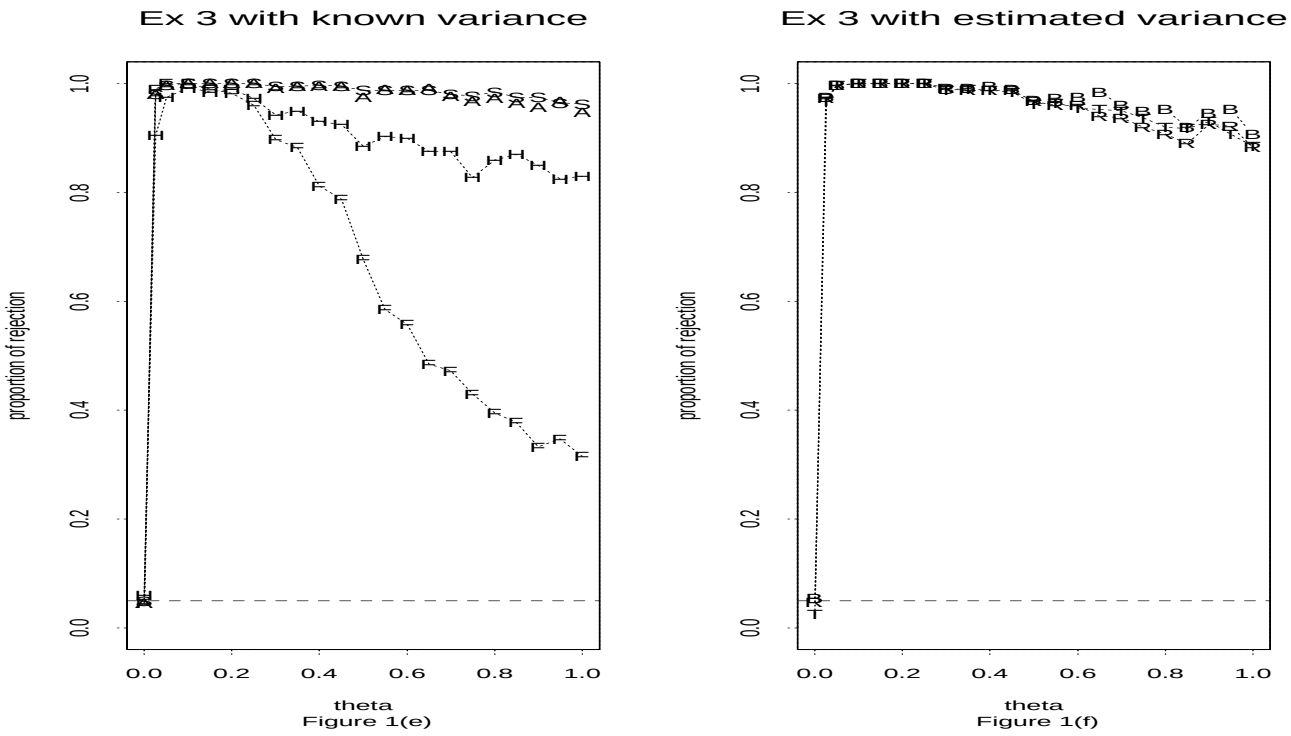

Figure 1: Power of the adaptive Neyman tests, the KH test, the wavelet thresholding test and parametric F-test for Examples 1-3. A – The adaptive Neyman test with known variance; B – The adaptive Neyman test with estimated variance (2.10); F – The parametric F-test statistic; H – The wavelet thresholding 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).

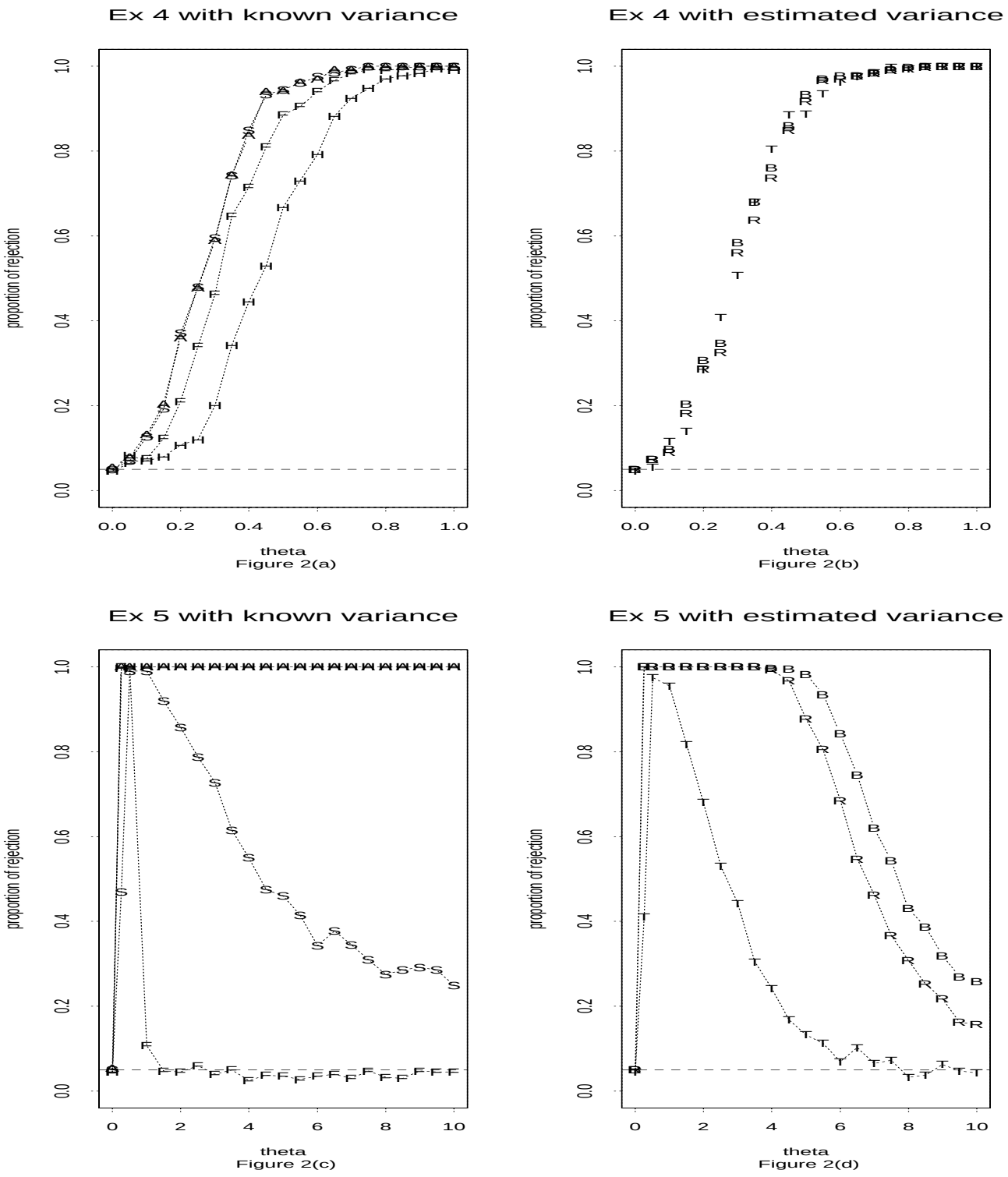

Figure 2: Power of the adaptive Neyman tests, the KH test, the wavelet thresholding test and parametric F-test for Examples 4-5 with partial linear models as alternative. A – The adaptive Neyman test with known variance; B – The adaptive Neyman test with estimated variance (2.10);. F – The parametric 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).

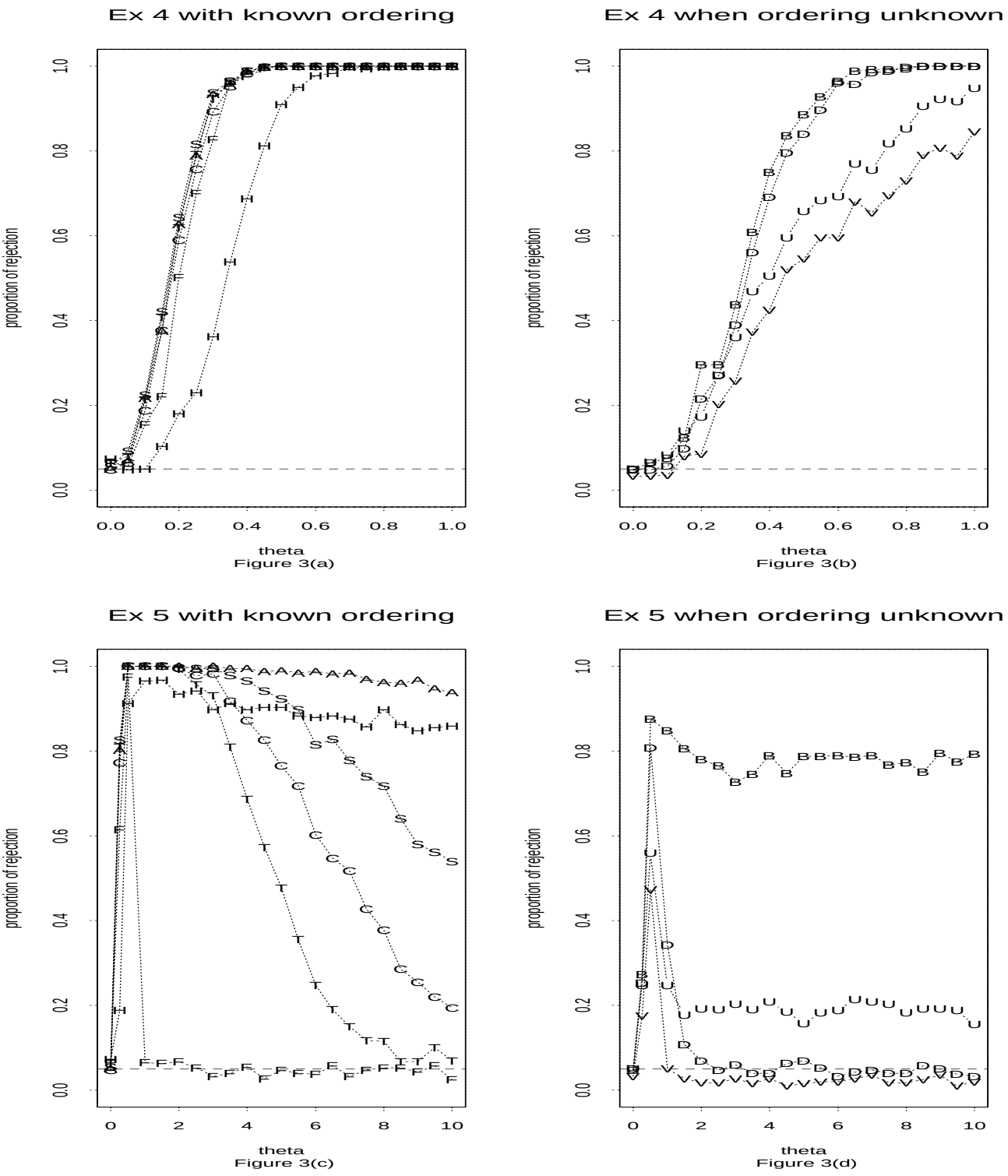

Figure 3: Power of the adaptive Neyman tests, the KH test, the wavelet thresholding test and parametric F-test for Examples 4-5, when the alternative is fully nonparametric. A – the adaptive Neyman test (ANT) ordered according to X_2 and using the known variance; B – the ANT ordered according to s_i 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's 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 first PC.