Test of Significance Based on Wavelet Thresholding and Neyman's Truncation *

Jianqing Fan

December 12, 1994

Traditional nonparametric tests, such as the Kolmogorov-Smirnov test and the Cramér-Von Mises test, are based on the empirical distribution functions. While these tests possess root-n consistency, they effectively only use information contained in the first few dimensions. This leads to low power in detecting fine features such as sharp and short aberrants as well as global features such as high frequency alternations. The drawback can be repaired via smoothing-based test statistics. In this paper, we propose two such kinds of test statistics based on the wavelet thresholding and the Neyman truncation. We provide extensive evidence to demonstrate that the proposed tests have high power in detecting sharp peaks and high frequency alternations, while maintain the same capability in detecting smooth alternative densities as the traditional tests. Similar conclusions can be made for two-sample nonparametric tests of distribution functions. In that case, the traditional linear rank tests such as the Wilcoxon (1938) test and the Fisher-Yates (1945) test have low power in detecting two nearby densities where one has local features or contains high frequency components, since these procedures are essentially testing the uniform distribution based on the sample mean of rank statistics. In contrast, the proposed tests use more fully the sampling information and have better ability in detecting subtle features.

KEY WORDS: Goodness-of-fit, adaptive Neyman test, wavelet thresholding, hard thresholding-parameter, soft thresholding-parameter, two-sample test.

SHORT TITLE: Thresholding and Truncation Tests.

*Jianqing Fan is Assistant Professor, Department of Statistics, University of North Carolina, Chapel Hill, NC 27599-3260. His work was partially supported by NSF Grants DMS-9203135, an NSF Postdoctoral Fellowship and a grant from the Research Council of UNC Chapel Hill. The work was completed while visiting the Institut de Statistique, Université Catholique de Louvain, Belgium. The author thanks the institute for its hospitality and support and thanks Professor Peter J. Bickel, an associate editor and two referees for helpful suggestions.

1 Introduction

Let $X_1, \cdots, X_n$ be an i.i.d sample from a certain distribution $F(x)$. Consider the following goodness-of-fit testing problem:

$$H_0 : F(x) = F_0(x) \quad \text{versus} \quad H_1 : F(x) \neq F_0(x), \tag{1.1}$$

where F_0 is a given distribution function. Traditional approaches to this problem are based on distribution functions and use certain distance between the empirical distribution $\hat{F}_n$ and F_0 as a test statistic. Witness of this includes the Kolmogorov-Smirnov test (KS-test)

$$\hat{T}_{KS} = n^{1/2} \sup_x |\hat{F}_n(x) - F_0(x)|,$$

and the Cramér-Von Mises test (CVM-test)

$$\hat{T}_{CVM} = n \int \{\hat{F}_n(x) - F_0(x)\}^2 dF_0(x).$$

While these testing procedures are popular due to their simplicity, they suffer from low power in detecting densities containing high frequency components or local features such as bumps. Indeed, any distribution-based (in contrast with smoothing-based) test statistic uses effectively only the information contained in the first few dimensions and shares similar drawbacks.

To elaborate the above points further, we first note that the testing problem (1.1) is equivalent to the problem:

$$H_0 : F(x) = \text{uniform} \quad \text{versus} \quad H_1 : F(x) \neq \text{uniform} \tag{1.2}$$

by using the transformation F_0. Consider the Fourier transform:

$$\theta_{2j-1} = \int_0^1 \cos(2\pi jx) dF(x), \quad \theta_{2j} = \int_0^1 \sin(2\pi jx) dF(x), j = 1, 2, \cdots.$$

Correspondingly, let $\hat{\theta}_j (j = 1, 2, \cdots)$ be the empirical Fourier coefficients. The testing problem is equivalent to the problem

$$H_0 : \theta_j = 0, j = 1, 2, \cdots \quad \text{versus} \quad H_1 : \text{at least one of } \theta_j \neq 0, \tag{1.3}$$

based on the observations $\hat{\theta}_j, j = 1, 2, \cdots$. The CVM-test can be expressed as

$$\hat{T}_{CVM} = \frac{n}{2\pi^2} \sum_{j=1}^{\infty} j^{-2} (\hat{\theta}_{2j-1}^2 + \hat{\theta}_{2j}^2). \tag{1.4}$$

The weight j^{-2} makes this testing procedure effectively use only the first few coefficients. This is evidently shown in Eubank and LaRiccia (1992) and in Section 3. Although the Kolmogorov-Smirnov test does not admit an expression as explicit as (1.4), the expression

$$\hat{F}_n(x) - x \approx \pi^{-1} \sum_{j=1}^{n/2} j^{-1} \{-\hat{\theta}_{2j} \cos(2\pi jx) + \hat{\theta}_{2j-1} \sin(2\pi jx)\}, 0 < x < 1$$

shows that any distribution-based test statistic greatly weighs down the contributions from high-frequency components. This explains why the traditional approaches have low power against the distributions with large energy at high frequencies. Throughout this paper, high frequency component and energy refer respectively to the component $\exp(i2\pi jx)$ with reasonably large j and the modulus of its Fourier coefficient.

As noted above, the testing problem (1.1) is equivalent to problem (1.3) using the data $\hat{\theta}_j, j = 1, \cdots, n$, since these Fourier coefficients are a sufficient statistic. Intuitively, under the null hypothesis or contiguous alternatives, the coefficients are asymptotically independent and normally distributed: $\hat{\theta}_j \sim N(\theta_j, n^{-1})$ for $j = 1, \cdots, N$, where $N/n \to 0$. This leads us to considering the Gaussian white noise model as follows. Let $X \sim N(\theta, n^{-1} I_N)$ be an N-dimensional normal random vector with I_N the identity matrix. We wish to test

$$H_0 : \theta = 0 \quad \text{versus} \quad H_1 : \theta \neq 0.$$

This simpler structure allows us to provide deeper insights into the nonparametric goodness-of-fit problem. It enables us to use other orthogonal transforms such as wavelet transforms (See Meyer 1990 and Daubechies 1992).

The above Gaussian white noise model has often been used as a canonical model to nonparametric density estimation and nonparametric regression. There is a large literature on this subject. See, for example, Pinsker (1980), Ibragimov and Hasminskii (1984), Donoho and Liu (1991) and Donoho and Johnstone (1994), among others. Although the contexts of these papers are completely different from ours, the scenario is the same: The Gaussian white noise model can provide deeper insights into its related statistical problems. Since many models are closely related to the Gaussian white noise model, the methods developed here are expected to have applications in nonparametric regression, nonparametric discriminant analysis and spectral analysis.

The low power of the traditional distribution-based tests can also be improved upon using density-based test statistics. Bickel and Rosenblatt (1973) proposed to use the L_2-distance between the estimated density $\hat{f}_n$ and its expected value $E\hat{f}_n$ under H_0 as a test statistic and derived its asymptotic distribution. Bowman (1992) confirms via simulations that gains can be made by using Bickel and Rosenblatt's procedure and argued in favor of the Bickel-Rosenblatt test (instead of the L_2-distance $\|\hat{f}_n - f_0\|$). An adaptive version of this test for multivariate normality is proposed in Bowman and Foster (1993). Eubank and LaRiccia (1992) use the Neyman (1937) test and show that the resulting procedure has good power against reasonably high frequencies. Bickel and Ritov (1992) give illuminating accounts for nonparametric tests. For related tests in regression setup, see Eubank and Hart (1992). Significant insights can be gained via a study of the Gaussian white noise model as presented in Sections 2 and 3, where we show via asymptotic studies and simulations that classical tests such as the Cramér-Von Mises test effectively only use information contained in the first few dimensions and the Neyman smooth test with the proposed rule for selecting the truncation parameter can detect alternatives containing high frequency components.

We begin by studying the testing problem in the n-dimensional Gaussian white noise model $X \sim N(\theta, I_n)$. Since the dimensionality is so high, testing all dimensions (i.e. test whether at least one of $\theta_j \neq 0$ for all $1 \leq j \leq n$) would accumulate stochastic errors and deteriorate the performance of resulting procedures. Instead, we propose two approaches to test for a certain subset: truncation and thresholding. If there is a *priori* evidence that most of the energy is concentrated at low frequencies (namely, large θ_j's are located at small j), then it is reasonable to use the Neyman test statistic $\sum_{j=1}^{m} X_j^2$, where m is the truncation parameter to be selected. A data driven version of this is given by (2.5). If there is no such a *priori*, the thresholding statistic $\sum_{j=1}^{n} X_j^2 I(|X_j| > \delta)$ is preferable, where δ is a thresholding parameter. We propose two methods of selecting the thresholding parameter: hard-thresholding and soft-thresholding parameter. The thresholding method is particularly appealing when most of the energy is concentrated on a few dimensions but with unknown locations. See Section 2 for more discussions. The performance of the truncation and thresholding procedures is demonstrated via a large scale simulations, which are designated to verify previous claims. See Section 3.

We then, in Section 4, apply the procedures in the Gaussian white noise model to the testing problem (1.2) via a binning technique. The binning technique links the nonparametric test problem (1.2) with the Gaussian white noise model. When the Fourier transform is used we apply our truncation approach since most of the energy of an unknown function is expected to concentrate at low frequencies. The thresholding techniques are used when one uses wavelet transforms. We demonstrate that for contiguous alternatives with local characters, the thresholding approach has more power than the truncation test and the latter is far more powerful than traditional tests. When testing contiguous alternatives with high frequency alternations, both the truncation and thresholding based approaches have good powers, but this is again not true for the traditional tests. When testing classical models such as mixture normals, we show that both truncation and thresholding methods work at least comparably with the traditional methods. In conclusion, the preferences are given to the wavelet based thresholding test and the Fourier transform based truncation approach.

Our approach can also be applied to the two-sample nonparametric test of distribution functions. Given two independent samples $X_1, \cdots, X_m \sim F_1$ and $Y_1, \cdots, Y_n \sim F_2$, we wish to test

$$H_0 : F_1 = F_2 \quad \text{versus} \quad H_1 : F_1 \neq F_2 \tag{1.5}$$

The maximum invariant statistics under monotonic transforms are the rank statistics $R_1, \cdots, R_m$ of the first sample $X_1, \cdots, X_m$ in the pooled sample $\{X_1, \cdots, X_m; Y_1, \cdots, Y_n\}$. Under the null hypothesis, $R_1, \cdots, R_m$ are uniformly distributed among $\{1, \cdots, m+n\}$ (outcomes of drawing without replacements). The traditional linear rank statistics measure the departure from the uniformity by checking whether the observed sample mean of the rank statistics deviated away from 0.5 can be explained by chance variation or not (or more generally, how far the sample mean $n^{-1}\sum a(R_i)$ is away from $Ea(U)$ with $U \sim$ uniform). Although these tests possess some optimalities such as the locally most powerful rank test in some given parametric family $\{P_\theta\}$, they do not exploit all the information contained in the data. As a consequence, the linear rank tests have very low power against alternatives other than the family $\{P_\theta\}$. Hájek and Šidák (1967) even hesitated to call such a kind of tests as nonparametric tests. Since testing the uniformity of $R_1, \cdots, R_m$ is analogous to (1.2), our approaches continue to apply with some modifications. Compared with the linear rank tests, the proposed nonparametric tests exploit much more fully the sampling information, and thus are more powerful, retaining high power against wider families of alternative hypotheses. See Section 5.

The tests proposed in this paper do not assume specific alternatives. However, many classical tests are designated for certain specific family of alternatives. See, for example, Hájek and Šidák (1967), Hollander and Wolfe (1973), D'Agostino and Stephens (1986) and Rayner and Best (1989). Comparing the proposed approaches with these classical tests under their designated alternatives would be an interesting future topic.

2 Testing Multivariate normal mean in high dimension

Let $X \sim N(\theta, I_n)$ be an n-dimensional normal random vector. We wish to test

$$H_0 : \theta = 0 \qquad \text{versus} \qquad H_1 : \theta \neq 0. \tag{2.1}$$

Given a specific alternative $\theta = \theta_0$, the Neyman-Person fundamental theory states that the best testing procedure for the problem

$$H_0 : \theta = 0 \qquad \text{versus} \qquad H_1 : \theta = \theta_0$$

is to reject H_0 when $\theta_0^T X > \|\theta_0\| \Phi^{-1}(1-\alpha)$, where α is the significant level, Φ is the standard normal distribution function, and $\|\cdot\|$ is the L_2-norm. The power of this optimal test is $1 - \Phi(z_{1-\alpha} - \|\theta_0\|)$ with $z_{1-\alpha} = \Phi^{-1}(1-\alpha)$. However, this optimal test depends on the specific alternative. A naive mimicking of this testing procedure is to estimate θ_0 by X and uses the testing procedure $\|X\|^2$. The resulting test suffers from very low power at the alternative θ_0. The power of such a test is approximately

$$1 - \Phi\left(\frac{z_{1-\alpha} - \|\theta_0\|^2/\sqrt{2n}}{\sqrt{1 + 2\|\theta_0\|^2/n}}\right) \approx 1 - \Phi\left(z_{1-\alpha} - \|\theta_0\|^2/\sqrt{2n}\right) \tag{2.2}$$

provided that $\|\theta_0\|^2 = o(n)$ and it tends to α even though $\|\theta_0\| \to \infty$ with $\|\theta_0\|^2 = o(\sqrt{n})$.

2.1 Adaptive Neyman's test

The expression (2.2) suggests that it is not a good idea to test on each coefficient. Neyman (1937) proposed to test the first m-dimensional subproblem, leading to the test statistic $\sum_{j=1}^m X_j^2$.

(Admittedly, Neyman's motivation is different from ours; see also Rayner and Best (1989) for some historical notes). If the large coefficients of θ_0 concentrate on the first m dimensions, this test would have a reasonable performance. For such a case, we have an option of selecting m. For problem (1.2), Eubank and LaRiccia (1992) propose to select m, via a cross-validation criterion in curve estimation setting.

Here we offer a simple and powerful procedure to select m based on the power consideration. The power of the Neyman test is approximately equal to (as $m \to \infty$, c.f. (2.2))

$$1 - \Phi\Big(z_{1-\alpha} - (2m)^{-1/2} \sum_{j=1}^{m} \theta_{j0}^2\Big), \tag{2.3}$$

provided that $\sum_{j=1}^m \theta_{j0}^2 = o(m^{-1})$, where θ_{j0} is the j^{th} component of θ_0. Thus, if the alternative hypothesis is given, an ideal choice of m is the one that maximizes $(2m)^{-1/2}\sum_{j=1}^m \theta_{j0}^2$. This suggests us to select

$$\hat{m} = \mathrm{argmax}_{m:1\le m\le n}\Big\{m^{-1/2}\sum_{j=1}^{m}(X_j^2 - 1)\Big\}. \tag{2.4}$$

Observe that the quantity $m^{-1/2}\sum_{j=1}^m (X_j^2-1)$ is an unbiased estimator of $m^{-1/2}\sum_{j=1}^m \theta_{j0}^2$. This leads to the adaptive Neyman test for problem (2.1) as follows: Reject H_0 when

$$T_{AN}^* = (\sqrt{2\hat{m}})^{-1}\sum_{j=1}^{\hat{m}}(X_j^2 - 1) = \max_{1\le m\le n}\Big\{(\sqrt{2m})^{-1}\sum_{j=1}^{m}(X_j^2-1)\Big\}$$

is too large or when

$$T_{AN} = \sqrt{2\log\log n}\,T_{AN}^* - \{2\log\log n + 0.5\log\log\log n - 0.5\log(4\pi)\} \tag{2.5}$$

is too large.

Theorem 2.1. The asymptotic distribution under H_0 is

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

This result is a corollary of Darlin and Erdös (1956). As a consequence of Theorem 2.1, the critical region

$$T_{AN} > -\log(-\log(1-\alpha))$$

has asymptotic significant level α.

We would also like to remark that although we use the Gaussian white noise model as our motivation, the adaptive Neyman procedure continues to apply to non Gaussian case. One needs only to replace $\sqrt{2}$ in the definition of $\hat{T}_{AN}^*$ by $\sqrt{\mathrm{var}(X_j^2)}$ (or its consistent estimate). Theorem 2.1 continues to hold.

Theorem 2.2. The adaptive Neyman test has the power at the alternative θ_0 at least

$$P_{\theta_0}\Big\{(2m_0)^{-1/2}\sum_{j=1}^{m_0}(X_j^2 - 1 - \theta_{j0}^2) \ge \sqrt{2\log\log n}(1+o(1)) - \max_{1\le m\le n}(2m)^{-1/2}\sum_{j=1}^{m}\theta_{j0}^2\Big\},$$

where $m_0 = \mathrm{argmax}_{1 \le m \le n} (2m)^{-1/2} \sum_{j=1}^m \theta_{j0}^2$. In particular, when $\max_{1 \le m \le n} (2m)^{-1/2} \sum_{j=1}^m \theta_{j0}^2 >> (\log\log n)^{1/2}$, the adaptive Neyman test has asymptotic power 1.

The power of the ideal Neyman test $\sum_{j=1}^{m_0} X_j^2$, which depends on the specific alternative θ_0, is $P_{\theta_0}\{(2m_0)^{-1/2} \sum_{j=1}^{m_0} (X_j^2 - 1 - \theta_{j0}^2) \ge c_{m_0,\alpha} - \max_{1 \le m \le n} (2m)^{-1/2} \sum_{j=1}^m \theta_{j0}^2\}$ for a critical value $c_{m_0,\alpha}$ that is bounded. With the simple rule (2.4), the adaptive Neyman test performs at least as well as the ideal Neyman test within a factor of $\sqrt{2 \log\log n}$.

The CVM-test $\hat{T}_{CVM} = \sum_{j=1}^n X_j^2 / j^2$ significantly weighs down the contributions from high frequencies. In particular, regarding $\theta_{10}, \cdots, \theta_{n0}$ as a realization of n i.i.d. random variables with a distribution $F_n(\theta)$ (namely, $F_n(\theta)$ is the prior distribution of θ_{j0}), we have for almost surely all realizations that the asymptotic power of the adaptive Neyman test is 1 while the asymptotic power of the CVM-test is α, provided that $\int \theta^2 dF_n(\theta) = o(1)$ and $\int \theta^2 dF_n(\theta) >> (n/\log\log n)^{-1/2}$. In other words, for such a kind of alternatives, the CVM-test has almost no power. In general, the inequalities

$$E \max_{1 \le m \le n} (2m)^{-1/2} \sum_{j=1}^m \theta_{j0}^2 \ge (n/2)^{1/2} E\theta_{j0}^2 >> E \sum_{j=1}^n \theta_{j0}^2 / j^2 \approx 1.64 E\theta_{j0}^2$$

are also an excellent evidence of supporting the adaptive Neyman test.

2.2 Hard and Soft-thresholding tests

The Neyman test does not perform well against the alternative hypothesis whose energy concentrates at very high frequencies: i.e. large values of θ_{j0} appear at large indices. An improved version of Neyman's test is to select a subset $S = \{j : |\theta_{j0}| > a\}$ and to base the test on

$$\sum_{j \in S} X_j^2 = \sum_{j=1}^n X_j^2 I(|\theta_{j0}| > a) \sim \chi_m^2 \Big(\sum_{j \in S} \theta_j^2 \Big),$$

where $m = \#(S)$. This test requires an "Oracle" to select a subset S which contains the first m largest coefficients. If there is such an oracle, then the power of the testing procedure at the point θ_0 is approximately (as $m \to \infty$ and $\sum_{j \in S} \theta_{0j}^2 = o(m)$, c.f. (2.2))

$$1 - \Phi\Big(z_{1-\alpha} - (2m)^{-1/2} \sum_{j \in S} \theta_{0j}^2 \Big), \tag{2.6}$$

which is no smaller than that of (2.3), i.e. the power of the ideal Neyman's test. The oracle test will choose m to maximize $\sqrt{m^{-1}} \sum_{j \in S} \theta_{0j}^2$. Let S_o denote the resulting best subset.

In practice, we do not have such an oracle. Selecting the best subset is cumbersome and accumulates stochastic errors in the selection process. A simple procedure of selecting a good subset without data mining is the subset $\hat{S} = \{j : |X_j| > \delta\}$ for some thresholding constant $\delta \equiv \delta_n$, leading to the testing statistic

$$\hat{T}_H^* = \sum_{j=1}^n X_j^2 I(|X_j| > \delta).$$

In order not to downgrade the noise accumulation in the subset selection process, δ must be of the form: $\delta = \sqrt{2\log n}(1+o(1))$. This choice of δ suppresses most of the noise, since $\sqrt{2\log n}$ is the maximum of n independent Gaussian white noises. On the other hand, δ can not be too large to filter out all of the important coefficients. This suggests us to take

$$\delta = \sqrt{2\log(na_n)}, \quad a_n = c(\log n)^{-d} \tag{2.7}$$

for some positive constants c and d. With this choice of the thresholding parameter, we have the following testing procedure: Reject if

$$\hat{T}_H = \sigma_{n,H}^{-1}\left(\hat{T}_H^* - \mu_{n,H}\right) > \Phi^{-1}(1-\alpha), \tag{2.8}$$

where

$$\mu_{n,H} = \sqrt{2/\pi}a_n^{-1}\delta(1+\delta^{-2}), \qquad \sigma_{n,H}^2 = \sqrt{2/\pi}a_n^{-1}\delta^3(1+3\delta^{-2}).$$

Note that we have normalized the test statistic $\hat{T}_H^*$ by its asymptotic mean and variance, which are expanded to the second order. Following Donoho and Johnstone (1994), we call this testing procedure as "hard-thresholding" test. The "soft-thresholding" test is

$$\hat{T}_S = \sigma_{n,S}^{-1}\Big(\sum_{j=1}^{n}\psi_\delta^2(X_j) - \mu_{n,S}\Big) > \Phi^{-1}(1-\alpha), \tag{2.9}$$

where with t_+ denoting the positive part of t,

$$\psi_\delta(x) = \operatorname{sgn}(x)(|x|-\delta)_+$$

is the soft-thresholding function, introduced by Bickel (1983), and

$$\mu_{n,S} = 2\sqrt{2/\pi}a_n^{-1}\delta^{-3}(1-4\delta^{-2}), \quad \sigma_{n,S}^2 = 2\sqrt{2/\pi}a_n^{-1}\delta^{-1}$$

Asymptotically, both the hard thresholding and soft-thresholding test have significance level α, as evidenced by

Theorem 2.3. Under the null hypothesis, $\hat{T}_H \xrightarrow{\mathcal{L}} N(0,1)$, provided that $a_n \log^{1/2} n \to 0$. The same result holds for the soft-thresholding test $\hat{T}_S$, if $a_n \log^{5/2} n \to 0$.

The proof of this theorem is given in the appendix. Remark that the condition on a_n is also a necessary condition for the asymptotic normality. This can easily be seen from (2.8) and the fact that a necessary condition for a sequence of nonnegative random variables to be asymptotically normal is the ratio of the means of the random variables to their standard deviations must tend to ∞.

Although $\hat{T}_S$ is asymptotically normal, we have experienced slow convergence. See also the condition on a_n in Theorem 2.3. Since computing P-values, which relies on the asymptotic normality, is also very important in practice, we will not pursue further on $\hat{T}_S$, but focus only on $\hat{T}_H$

The constant a_n can possibly be selected to maximize the power under a fixed alternative or be estimated from data. This is beyond the scope of this study. For concreteness, we will take $a_n = (\log n)^{-2}$. For $n = 25, 50, 100$ and 200, the values of the thresholding parameter δ_n are respectively 1.37, 1.54, 1.76 and 1.98. This selection is the 'largest' a_n that makes asymptotic normality plausible: A smaller choice of a_n would also retain the asymptotic normality, but passes too much noise in the thresholding process.

Under H_0, $X_{\max} = \max_{1\le i\le n} X_i \approx (2\log n)^{1/2}$, and it is expected to be larger under the alternative hypothesis. Thus, replacing $\log n$ in the definition of a_n by $X_{\max}^2/2$ would enhance the power of the test without increasing asymptotic type I error. This leads to the following hard-thresholding parameter:

$$\delta_H = \sqrt{2\log(n\hat{a}_n)}, \quad \hat{a}_n = \min(4(X_{\max})^{-4}, \log^{-2} n). \tag{2.10}$$

This selection enhances a great deal of the power when a lot of energy is concentrated at one dimension. See for instance Example 3 in Section 3. A soft-thresholding parameter is defined by

$$\delta_S = \sqrt{2\log(n\hat{a}_n)}, \quad \hat{a}_n = \min(\{\log(\sum_{i=1}^{n} X_i^2)\}^{-2}, \log^{-2} n). \tag{2.11}$$

This soft-thresholding parameter increases the power of the test since it is than under H_0 than under H_1.

The power of the thresholding test is hard to compute under the alternative hypothesis since the random variables are no longer i.i.d and the test statistic is not necessarily asymptotic normal. We have the following lower bound:

Theorem 2.4. The power of the hard thresholding test $\hat{T}_H$ is at least

$$P_{\theta_0}\Big\{\sum_{i\in S_0}(X_i^2 - 1) \ge m_0\delta^2 + \sigma_{n,H}(\Phi^{-1}(1-\alpha) - Z_n)\Big\},$$

provided that $m_0 = o(na_n^{1/2})$, where S_0 is the "Oracle" best subset, $m_0 = \#(S_0)$ and Z_n is a sequence of random variables converging to the standard normal random variable and independent of $X_i's, i \in S_0$. Moreover, when $\sum_{j=1}^n \theta_j^2 I(|\theta_j| > \sqrt{2\log n}) >> \sigma_{n,H}$, the asymptotic power tends to one.

Should we know the "Oracle" best subset S_0, we would use the oracle test with the critical region $(2m_0)^{-1/2}\sum_{i\in S_0}(X_i^2 - 1) > C_{m_0,\alpha}$, where $C_{m_0,\alpha}$ is the critical value and is bounded since the random variables $\{(2m_0)^{-1/2}\sum_{i\in S_0}(X_i^2-1)\}$ are tight under H_0. Thus, the power of the oracle test is

$$P_{\theta_0}\Big\{(2m_0)^{-1/2}\sum_{i\in S_0}(X_i^2 - 1) \ge C_{m,\alpha}\Big\}.$$

Comparing it with the thresholding test, we paid at most the price of the order $O(\sqrt{m_0}\delta^2 + \sigma_{n,H})$. When m_0 is of logarithmic order, so is this price. In other words, the thresholding test performs well when the energy of θ_0 is concentrated in a few dimensions. In that case, we mimic the performance of the "Oracle" test within a factor of the logarithmic order.

Sometimes, one might have vague knowledge that $|\theta_j|$ $(j = 1, \cdots, J_0)$ can be reasonably large. This is particularly the case when one does a wavelet transform: The wavelet coefficients at low resolution level can be reasonably large. This knowledge can be incorporated into the thresholding estimator. In this case, the hard-thresholding estimator is defined by

$$\hat{T}_H^* = \sum_{j=1}^{J_0} X_j^2 + \sum_{j=J_0+1}^{n} X_j^2 I(|X_j| \geq \delta), \tag{2.12}$$

where $a_n = \log^{-2}(n - J_0)$ and $\delta = \sqrt{2\log\{(n - J_0)a_n\}}$. The normalization constants in (2.8) are modified as

$$\mu_{n,H} = J_0 + \sqrt{2/\pi}a_n^{-1}\delta(1 + \delta^{-2}), \qquad \sigma_{n,H}^2 = 2J_0 + \sqrt{2/\pi}a_n^{-1}\delta^3(1 + 3\delta^{-2}). \tag{2.13}$$

In this paper, we use $J_0 = 3$ and obtain the results which are not qualitatively different from those with $J_0 = 0$ as reported in an earlier version of this paper (Fan, 1994).

3 Simulations

3.1 Densities under null hypothesis

The asymptotic distributions of the proposed test statistics are derived under the null hypothesis. But, we do not know when the asymptotic theory takes in effect. One possible way to obtain the finite-sample distributions of these test statistics are via simulations. As an illustration, we study these finite-sample distributions for $n = 50, 200, 800$ via 1600 simulations. For each testing procedure, the distributions of the 1600 simulated test statistics under the null hypothesis are presented using a kernel density estimate. More precisely, let $Z_1, \cdots, Z_N$ be a sequence of independently identically distributed random variables with a common density f. Then, its kernel density estimator is

$$\hat{f}_h(x) = N^{-1}\sum_{i=1}^{N} h^{-1}K\{(Z_i - x)/h\},$$

where K is a symmetric density function, called a kernel function, and h is the bandwidth, controlling the size of the local neighborhood. See for example Silverman (1986). Here, we use the Gaussian kernel function $K(x) = (2\pi)^{-1/2}\exp(-x^2/2)$ and Silverman's bandwidth selection rule: $h = 1.06 s_N N^{-1/5}$, where s_N is the standard deviation of $Z_1, \cdots, Z_N$.

For each testing procedure $\hat{T}$, we calculated $Z_1 = \hat{T}_1, \cdots, \hat{Z}_N = \hat{T}_N$ from the 1600 simulations. The results are presented in Figure 1.

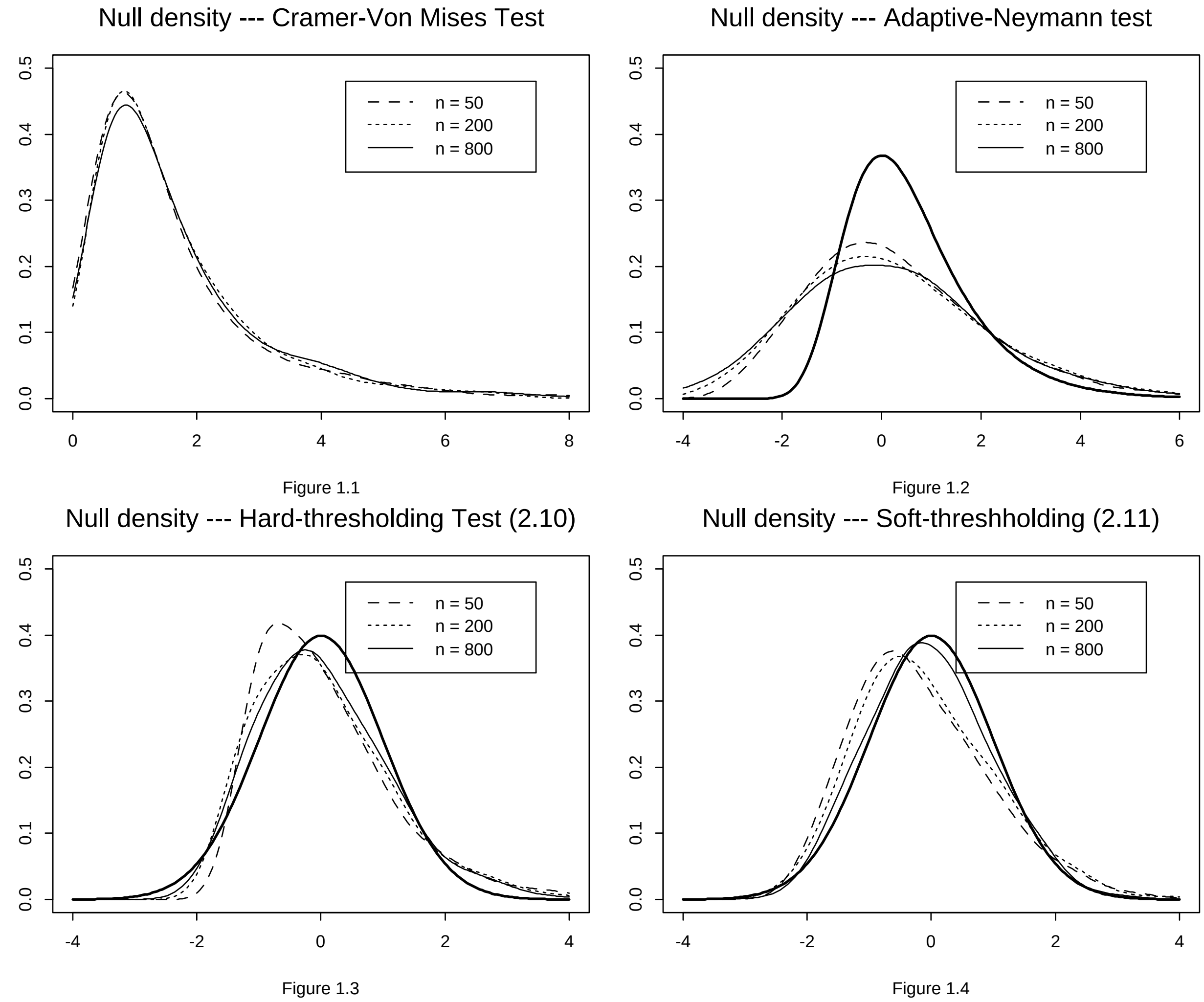

Figure 1: *The estimated densities for the test statistics under the null hypothesis for* $n = 50, 200, 800$*, based on 1600 simulations. Solid thick lines are the asymptotic densities.*

3.2 Power comparisons

The following simulations are designed to compare the efficiency of the three proposed testing procedures, introduced in Section 2, with the Cramér-Von Mises test-statistic. For each example, we took $\alpha = 0.05$. All results reported here are based on 1600 simulations so that the standard error is less than $0.5/\sqrt{1600} = 1.25\%$. For each example, we generate a multivariate normal $X \sim N(\theta, 1)$ of dimensions $n = 50$ and $n = 200$, respectively, and present the powers of each test at a sequence of alternatives indexed by k:

$$H_k : \theta_{k+j} = \eta_j, \text{ for } j = 0, \cdots, \ell_0 - 1; \qquad \theta_j = 0 \text{ otherwise}, \tag{3.1}$$

where η_j is a given sequence. This is designated to examine how much the power would decrease as k increases, i.e. when the energy is shifted towards high frequencies.

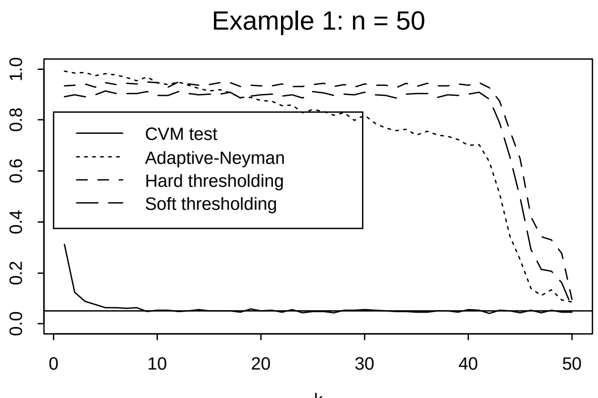

Figure 2.1

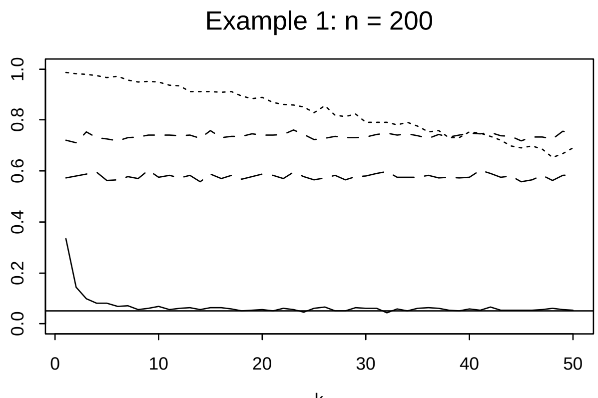

Figure 2.2

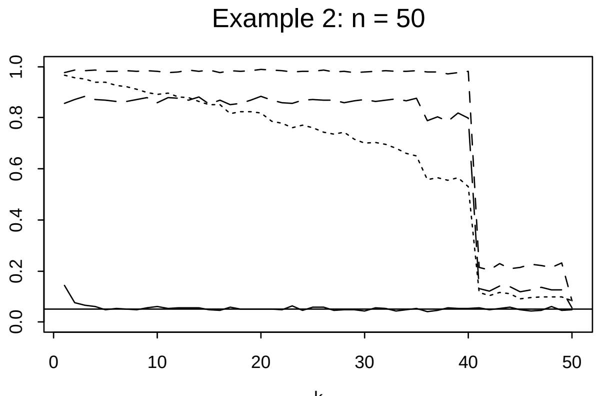

Figure 2.3

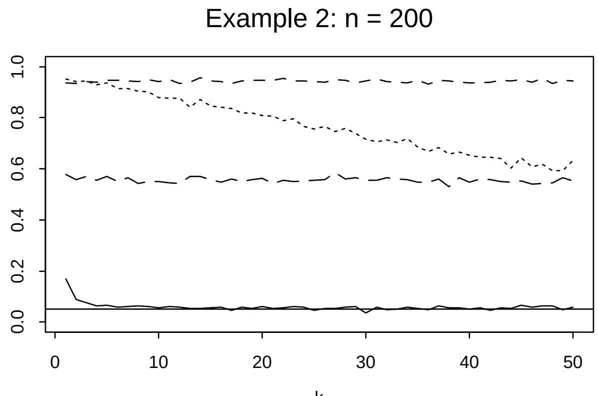

Figure 2.4

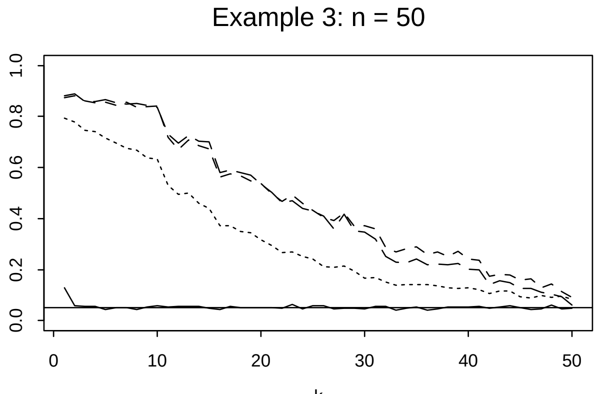

Figure 2.5

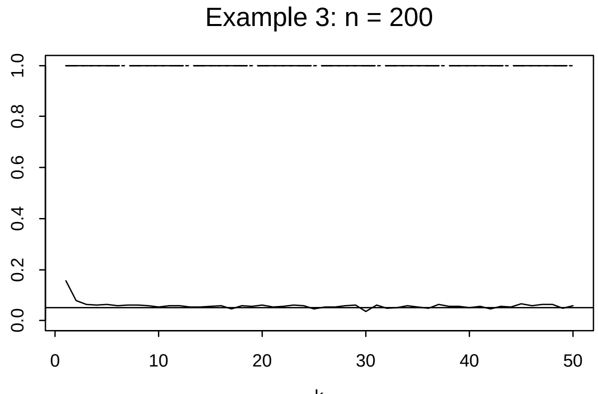

Figure 2.6

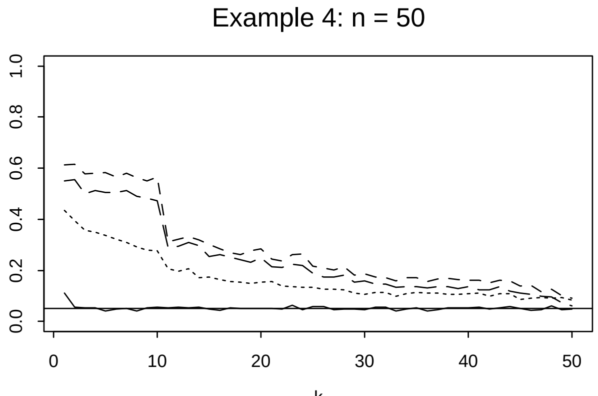

Figure 2.7

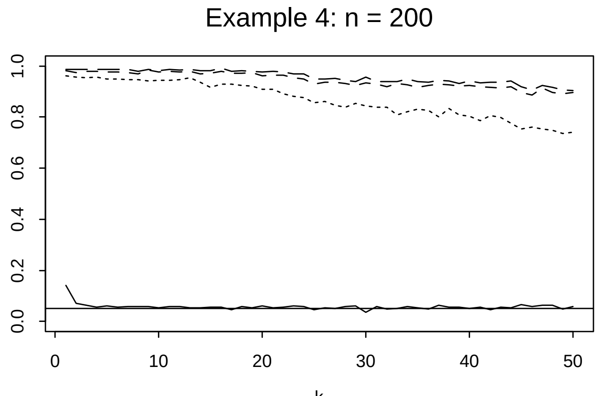

Figure 2.8

Figure 2: *The plot of the power functions of each testing procedure against a sequence of alternatives H_k given by (3.1) for each simulated example. The larger the value of k, the more energy is concentrated at high frequencies. The horizontal line indicates 5%-significant level. Solid curve – the power of the CVM test. Dashed curves — from the shortest to the longest dash: The power of the adaptive Neyman test, the hard- and soft-thresholding tests.*

We now describe the values of $\eta' s$.

Example 1: We generated $\ell_0 = 10$ numbers of η_j from uniform$(0, 3)$ and kept them fixed across simulations. These values of η's were

$$0.63, 2.11, 1.61, 0.48, 1.44, 2.72, 2.10, 2.56, 2.33,$$

corresponding to $\eta_0, \cdots, \eta_9$ in (3.1). This example is designed to see the performance of each testing procedure when the energy is scattered at low frequencies.

Example 2: We took $\ell_0 = 16$, $\eta_0 = 2, \eta_{10} = 5, \eta_{15} = 2$ and the rest η_j are zero. The experiment is designed to see how each testing procedure performs when the energy is concentrated on a few unknown locations.

Example 3: We took $\ell_0 = n$ and generated n standard normal random numbers as the vector η_j. This example shows how each testing procedure performs under 'normal' situations.

Example 4: Once again, we took $\ell_0 = n$, and for the even indexed η's, we assigned their values by the standard normal random numbers, and for odd indexed η's, we assigned their values by zero. This corresponds to the situation when the density is symmetric so that its Fourier coefficients are zero for odd indices. For such a 'half-black' object, we want to study how each procedure performs.

The results of simulations are depicted in Figure 2, where we plot the power of each test against the index k of the alternatives.

We first remark that the Cramér-von Mises test has almost no power when $k > 3$. This means that the CVM-test uses mainly information contained in the first 3 dimensions — the rest has little effect.

When we have a *priori* information that most of the large coefficients are concentrated on the first few dimensions, the adaptive-Neyman test should perform better than the thresholding procedures. This is demonstrated in Figures 2.1 and 2.2 when k is small. However, as k increases, the power of the adaptive-Neyman test decreases. The adaptive-Neyman test performs surprisingly well even when k is reasonably large. In other words, this procedure has good power for testing the situations when the energy is concentrated at reasonably high frequencies.

When we have no prior knowledge about the energy concentration, we would use the thresholding procedures. These tests do not use any *priori* information and hence are robust against the locations of the large observations. They are particularly appealing when the object is nearly black, i.e. when the energy is concentrated on a few unknown locations, such as in Example 2. The power of the thresholding tests is nearly constant against k.

We have observed some decreases of the powers for the thresholding tests. This is not due to the testing procedures themselves, but rather due to the changes of models. For example for $n = 50$ and $k > 40$, most of the $\theta_j' s$ are zero in Examples 1 & 2 and hence the problem now becomes harder for any test. This is why there is a sudden power drop in Figures 2.1 and 2.3. There is no such an effect when $n = 200$. We would also like to remark that the powers in Figure 2.2 are smaller than those in Figure 2.1 for the thresholding tests. This is because for $n = 50$, we need less effort to search some effective dimensions (e.g. search 3 largest coefficients among 50 dimensions versus among 200 dimensions) and hence accumulate less stochastic errors.

4 Applications to Goodness-of-fit

As evidenced in the previous section, both the adaptive Neyman test and the thresholding tests are genuine nonparametric — they effectively test many dimensions. When there is a *priori* evidence that most of energy is scattered at low frequencies, we would use the adaptive Neyman test; and the thresholding tests, otherwise. When the Fourier transform is used for the goodness-of-fit, we would expect that the energy is scattered at low frequencies and hence employ the adaptive Neyman test. We will still call this procedure as the adaptive Neyman test. However, when we use the wavelet transforms, we would not have such a *priori* and hence would employ the thresholding tests.

4.1 Connections with multivariate normal model

Let $X_1, \ldots, X_n$ be independent identically distributed random variables on $[0, 1]$ with density f and we wish to test problem (1.2). To use the adaptive Neyman test and the wavelet thresholding tests, we first bin the data as follows. Partition the interval $[0, 1]$ into n_{bin} subintervals $\{I_j, j = 1, \ldots, n_{\text{bin}}\}$. Let N_j be the number of data $\{X_i, i = 1, \ldots, n\}$ falling in the subinterval I_j. Do the square-root transform to stabilize the variance:

$$Y_j = 2\Big\{N_j^{1/2} - (n/n_{\text{bin}})^{1/2}\Big\}. \tag{4.1}$$

Let x_j be the center of I_j and $p_j = \int_{I_j} f(x)dx \approx f(x_j)/n_{\text{bin}}$. Then,

$$N_j \sim \text{Binomial}(n, p_j) \approx N\Big(f(x_j)\frac{n}{n_{\text{bin}}}, f(x_j)\frac{n}{n_{\text{bin}}}\Big).$$

Thus asymptotically,

$$Y_j \sim N\Big(2\{\sqrt{f(x_j)} - 1\}(n/n_{\text{bin}})^{1/2}, 1\Big)$$

and the sequence $\{Y_j\}$ is independent, namely, the vector $Y \sim N(\theta, I_{n_{\text{bin}}})$. Now compress the information via an orthogonal transform Γ and let the resulting vector be $X_\Gamma = \Gamma Y$. We would expect that X_Γ has the asymptotic distribution $N(\Gamma\theta, I_{n_{\text{bin}}})$. Thus, we reduce the nonparametric goodness-of-fit problem to the Gaussian white noise model (2.1), based on the observed vector X_Γ. Formally, we have the following results, which show that the theory developed in the Gaussian white noise model continues to apply.

Theorem 4.1. Let $f(x) > 0$ be bounded and continuous. Suppose that $n_{\text{bin}} \to \infty$ and $n_{\text{bin}} n^{-1} \log n \to 0$. Then by changing a probability space if necessary, we have the following stochastic representation:

$$Y_j = \theta_j + \varepsilon_j + O_p(n_{\text{bin}}^{-1/2} + (n^{-1} n_{\text{bin}} \log n)^{1/2}),$$

uniformly in j, where $\{\varepsilon_j\}$ is a sequence of i.i.d. Gaussian white noise and $\theta_j = 2(\sqrt{p_j/n_{\text{bin}}} - 1)(n/n_{\text{bin}})^{1/2}$. Indeed, the following stronger result holds: For any $r > 0$,

$$E \max_{1 \le j \le n_{\text{bin}}} |Y_j - \theta_j - \varepsilon_j|^r = O\left(n_{\text{bin}}^{-r/2} + (n^{-1} n_{\text{bin}} \log n)^{r/2}\right),$$

provided that $n_{\text{bin}} \to \infty$ and $n_{\text{bin}}^{1+\gamma}/n \to 0$ for some $\gamma > 0$.

Theorem 4.2. Suppose that with at most finitely many possible exceptions, rows of the orthogonal matrix Γ are orthogonal to the vector $(1, \cdots, 1)$. Let T_{AN} be the adaptive Neyman test (2.5) using the data X_Γ. Then, under the null hypothesis of (1.2),

$$P(T_{AN} \le x) \to \exp(-\exp(-x)), \text{ as } n \to \infty, \tag{4.2}$$

provided that $n_{\text{bin}} \to \infty$ and $n_{\text{bin}}^2 \log^2 n/n \to 0$.

Theorem 4.3. Let $\hat{T}_H$ be the thresholding test statistic (2.8) using the data X_Γ. Then, under conditions of Theorem 4.2, $\hat{T}_H \xrightarrow{\mathcal{L}} N(0, 1)$, if $a_n \log^{1/2} n_{\text{bin}} \to 0$ and $n^{-1} n_{\text{bin}}^{2+\gamma} \to 0$ for some $\gamma > 0$.

Two popular bases of the orthogonal transforms are used in the literature. One is the Fourier transform Γ_F and the other is the wavelet transform Γ_W, which is introduced to statistics by Donoho (1992) and Donoho and Johnstone (1994). Both of them satisfy the conditions of Theorems 4.2 and 4.3. Since the Fourier transform is very familiar, we omit the details. Let us call the resulting procedure (2.5) with the Fourier transformed data X_Γ ($\Gamma = \Gamma_F$) as the adaptive Neyman test.

Wavelet transform is an orthogonal transform which has capacity of localizing information in both time and frequency domains. After doing the wavelet transform, most of useful signals are expected to be concentrated only on a few dimensions, the rest are just noise. Thus, it is very appropriate to apply the thresholding approach. Daubechies (1992) provides two useful families of wavelet transforms: the 'extremal phase family' and 'the least asymmetric family' both indexed by the number of vanishing moments. Mallat's pyramid algorithm makes the wavelet transforms very fast to compute ($O(n)$-algorithm) and easy to program. This algorithm requires the vector to be transformed of length 2^k for some integer k. We omit the details here, but see Daubechies (1992). For concreteness, we take n_{bin} to be the largest dyadic integer 2^k such that it is smaller than n and use Daubechies' 'extremal phase family' with number of vanishing moments 1.

Once the data Y is transformed via the wavelet transform, we now apply the thresholding procedure (2.8) with hard-thresholding and soft-thresholding parameters (2.10) and (2.11). For brevity, we call the resulting procedures respectively the hard- and soft-thresholding test. Following Donoho and Johnstone (1994), we took the primary resolution that is equal to 2, leaving the first 3 wavelet coefficients unthresholded. Namely, we used the test statistic (2.12) with $J_0 = 3$.

It is shown by Donoho and Johnstone (1994) and Fan *et al* (1993) that when the alternative density has local features such as bumps and discontinuities, there are only a few very large values of wavelet coefficients. Therefore, according to Theorem 2.4, the wavelet thresholding methods have good power in detecting this kind of alternatives.

4.2 Power comparisons

We use four examples to compare the proposed procedures with the KS-test, the CVM-test, the Anderson-Darling (1952) test and the Shapiro-Wilk (1965) test. The families of the distributions are described in Figure 3. The comparison also provides insights to the strength of the wavelet thresholding procedures and the adaptive Neyman test. For each example, we simulated 1600 times for the sample sizes of $n = 50$ and $n = 200$.

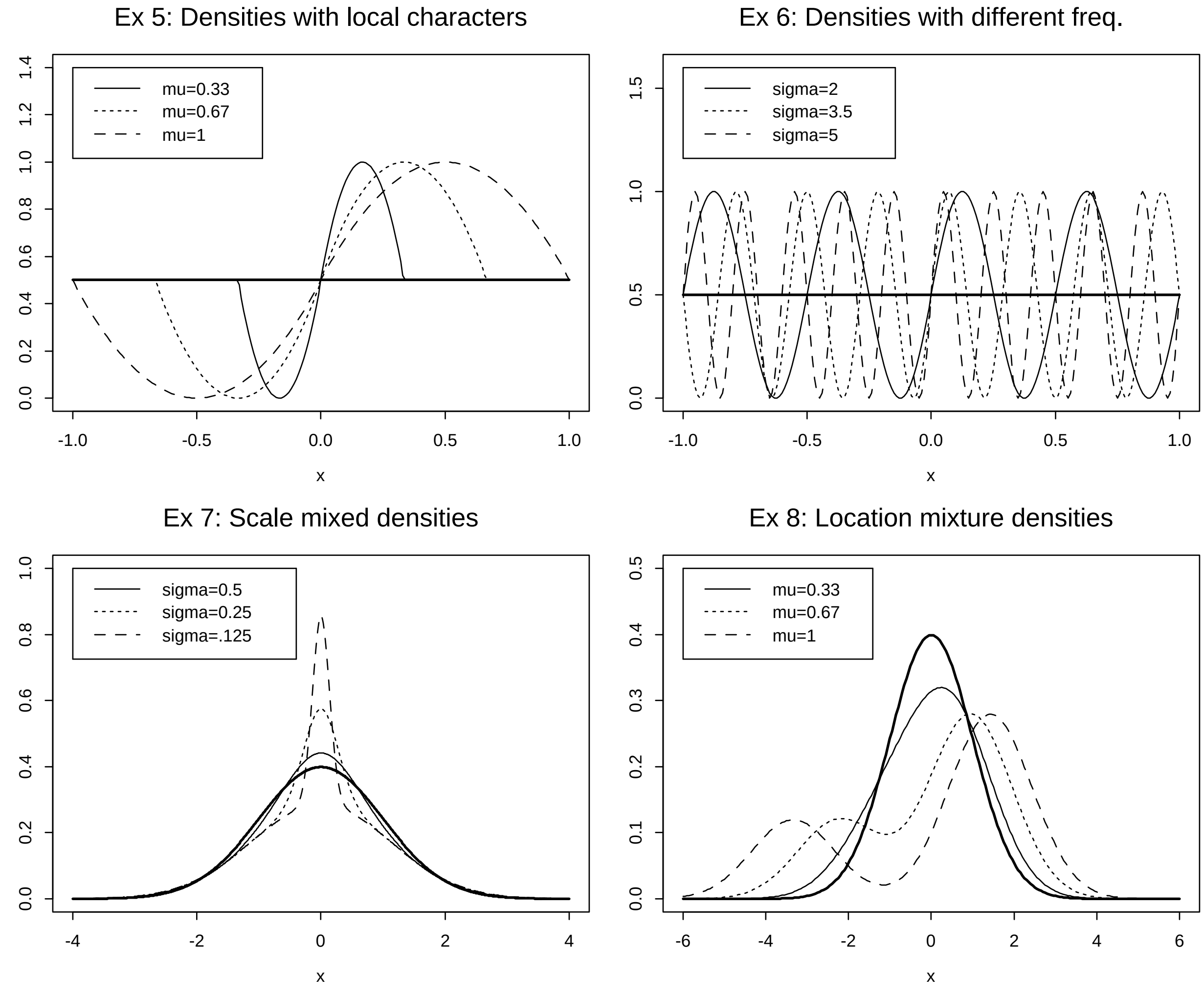

Figure 3: *Families of testing densities for Examples 5 – 8.*

Example 5: Consider the testing problem

$$H_0 : F = \text{uniform}(-1, 1) \quad \text{versus} \quad H_1 : F = F_\mu, \text{ with density}$$
$$F'_\mu(x) = \frac{1}{2} + 2x(\mu - |x|)/\mu^2 I(|x| < \mu), \quad (0 \leq \mu \leq 1).$$

When $\mu = 0$, the alternative hypothesis is the same as the null hypothesis. We deliberately include this special alternative so that we can examine whether the type I error for each procedure is 5%

or not. The task of this example is to demonstrate that the wavelet thresholding methods have better ability to detect the local feature around 0. The strength of the local feature depends on the value of μ – the larger, the stronger. Clearly, the wavelet thresholding tests have a high power in detecting the local character, while the KS-test and CVM-test have very poor power. See Figures 4.1 and 4.2.

Example 6: Instead of detecting local features, we now test the global features with different frequencies:

$$\begin{aligned}&H_0 : F = \text{uniform}(-1, 1) \quad \text{versus} \quad H_1 : F = F_\sigma, \text{ with density}\\ &F_\sigma'(x) = \tfrac{1}{2}\{1 + \sin(2\pi\sigma x)\}, \quad (0.5 \le \sigma \le 5).\end{aligned}$$

The results are depicted in Figures 3.3 and 3.4. Again, both the wavelet thresholding tests and the adaptive Neyman test have high power, while the KS-test and CVM-test suffer from low power at the high frequency alternatives. For sample size $n = 200$, the former tests made no type II error, but the latter made a very large proportion of type II error.

Examples 7. We now consider more classical type of testing problems:

$$H_0 : F = N(0, 1) \quad \text{versus} \quad F = 0.8N\left(0, \frac{1}{0.8 + 0.2\sigma^2}\right) + 0.2N\left(0, \frac{\sigma^2}{0.8 + 0.2\sigma^2}\right), \quad (1/8 \le \sigma \le 1).$$

Note that the mean and standard deviation under both H_0 and H_1 are respectively zero and one. When $\sigma = 1$, the alternative hypothesis is the same as the null hypothesis so that Figures 4.5 and 4.6 also report type I error. Even in such a normal mixture model, gains can be made by using our approaches. In particular, the adaptive Neyman test has high efficiency against the smooth alternative.

Example 8. Instead of considering the scale mixture, we examine the mean mixture model:

$$H_0 : F = N(0, 1) \quad \text{versus} \quad H_1 : F = 0.7N(\mu/0.7, 1) + 0.3N(-\mu/0.3, 1) \quad (0 \le \mu \le 1).$$

When $\mu = 0$, H_1 is the same as H_0 so that all tests should have power 0.05. Even for this type of alternatives, our approaches work comparably with the traditional approaches and once again the adaptive Neyman test is most powerful against the smooth alternatives.

In summary, the adaptive Neyman and the wavelet thresholding tests have good power in detecting subtle local changes and high frequency alternations, while maintain the strength of the classical tests such as the Kolmogorov-Smirnov and Cramér-Von Mises test.

The scenario does not end here. The idea in Section 2 has many other applications. We consider the applications to the nonparametric two-sample test of distribution functions in the next section.

5 Applications to two-sample test of distribution function

Consider the two independent random samples $X_1, \cdots, X_m \sim F_1$ and $Y_1, \cdots, Y_n \sim F_2$. We wish to test

$$H_0 : F_1 = F_2 \quad \text{versus} \quad H_1 : F_1 \ne F_2. \tag{5.1}$$

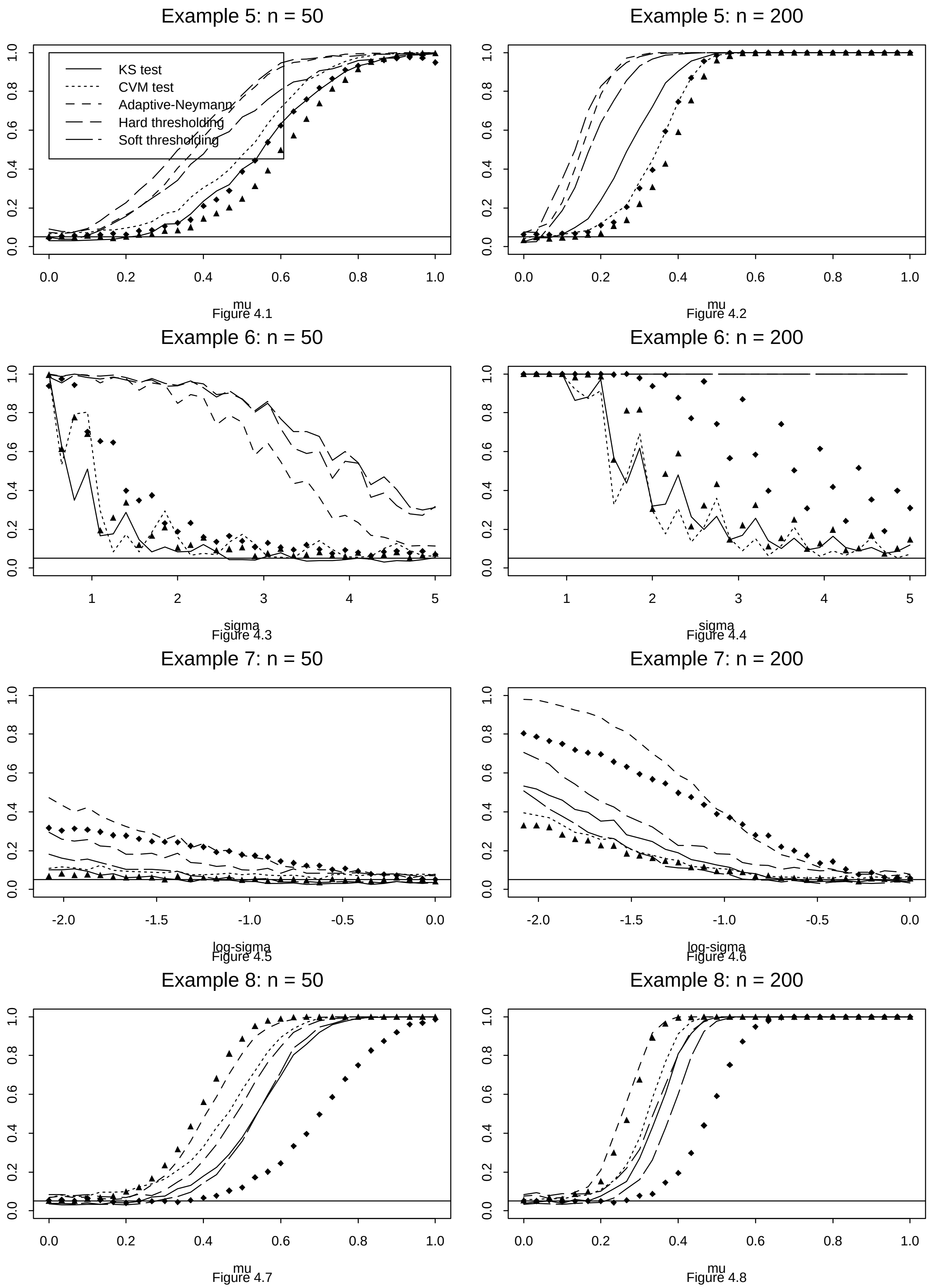

Figure 4: *The plot of power functions of each testing procedure against a sequence of alternatives given in Examples 5 – 8.* Δ *– power of the Anderson and Darling (1952) test.* $\Diamond$ *– power of Shapiro-Wilk (1965) test. Solid curve – power of the Kolmogorov-Smirnov test. Dashed curves — from the shortest to the longest dash: The power of the CVM-test, the adaptive Neyman test, the wavelet hard- and soft-thresholding tests. The horizontal line indicates 5%-significant level.*

Since the forms of F_1 and F_2 under H_0 are not specified, the order statistics of the observed data do not provide much information in this testing problem. The rank statistics $R_1, \cdots, R_m$ of the data $X_1, \cdots, X_m$ among the pooled sample $\{X_1, \cdots, X_m; Y_1, \cdots, Y_n\}$ contain most of the useful information. Indeed, they are the maximal invariant statistics under monotonic transformations. Under H_0, $R_1, \cdots, R_m$ are distributed as drawing m items without replacements from the box with items $\{1, \cdots, m+n\}$, or loosely speaking, uniformly distributed in $\{1, \cdots, m+n\}$. Testing the uniformity is discussed in the last section, except now that the rank statistics are dependent.

Let $X_i = R_i/(m+n+1)$ $(i = 1, \cdots, m)$, which are nearly uniform(0,1) distributed random variables. Temporarily ignoring the dependency of the data $\{X_i\}$, we can apply the methodology in Section 4.1. Let Y_j as in (4.1). Do data compression by using an appropriate orthonormal transform, namely, let $X_\Gamma = \Gamma Y$. To account for the dependent structure of x_j, we divide X_Γ by the standard deviation correction factor

$$c_{m,n} = \Big(\frac{N-m}{N-1}\Big)^{1/2} = \Big(\frac{n}{m+n-1}\Big)^{1/2},$$

where $N = m+n$, and denote the resulting vector $Z_\Gamma = X_\Gamma / c_{m,n}$. Now, apply the adaptive Neyman test to Z_Γ when $\Gamma = \Gamma_F$ (Fourier transform), and the wavelet thresholding procedures to Z_Γ when $\Gamma = \Gamma_W$ (Wavelet transform). We have not yet known the asymptotic distribution of these test statistics, so we use the simulated quantiles as critical values for the adaptive Neyman-test, and the wavelet thresholding tests.

The power of our testing procedures is compared with that of the Wilcoxon rank sum test and the Fisher-Yates normal scores test. These two tests are respectively the locally most powerful rank test against the location family of logistic type and of normal type. For details of these two test statistics, see Hájek and Šidák (1967) and Hollander and Wolfe (1973).

We simulated $m = n = 50$ and $m = n = 200$ random samples from both the null (generating X-samples) and alternative (generating Y-samples) hypotheses in Examples 5 – 8 and computed the results for the five testing procedures: The Wilcoxon rank sum, the Fisher-Yates normal scores, the adaptive Neyman, the wavelet thresholding with hard and soft thresholding parameters. The results of simulations are summarized in Figure 5, based on 1600 simulations. The type I error of each test is also reported in Figures 5.1, 5.3 and 5.4. To save space, we only presented the results for $m = n = 200$. Clearly, all testing procedures have significant level approximately 5% (see Figures 5.1 and 5.4 with $\mu = 0$ and Figure 5.3 with $\sigma = 1$) and the proposed nonparametric tests are far more powerful than the traditional ones.

6 Concluding Remarks

We have proposed the adaptive Neyman test and the wavelet thresholding tests based on a study of the Gaussian white noise model. These tests mimic the performance of the ideal Neyman test

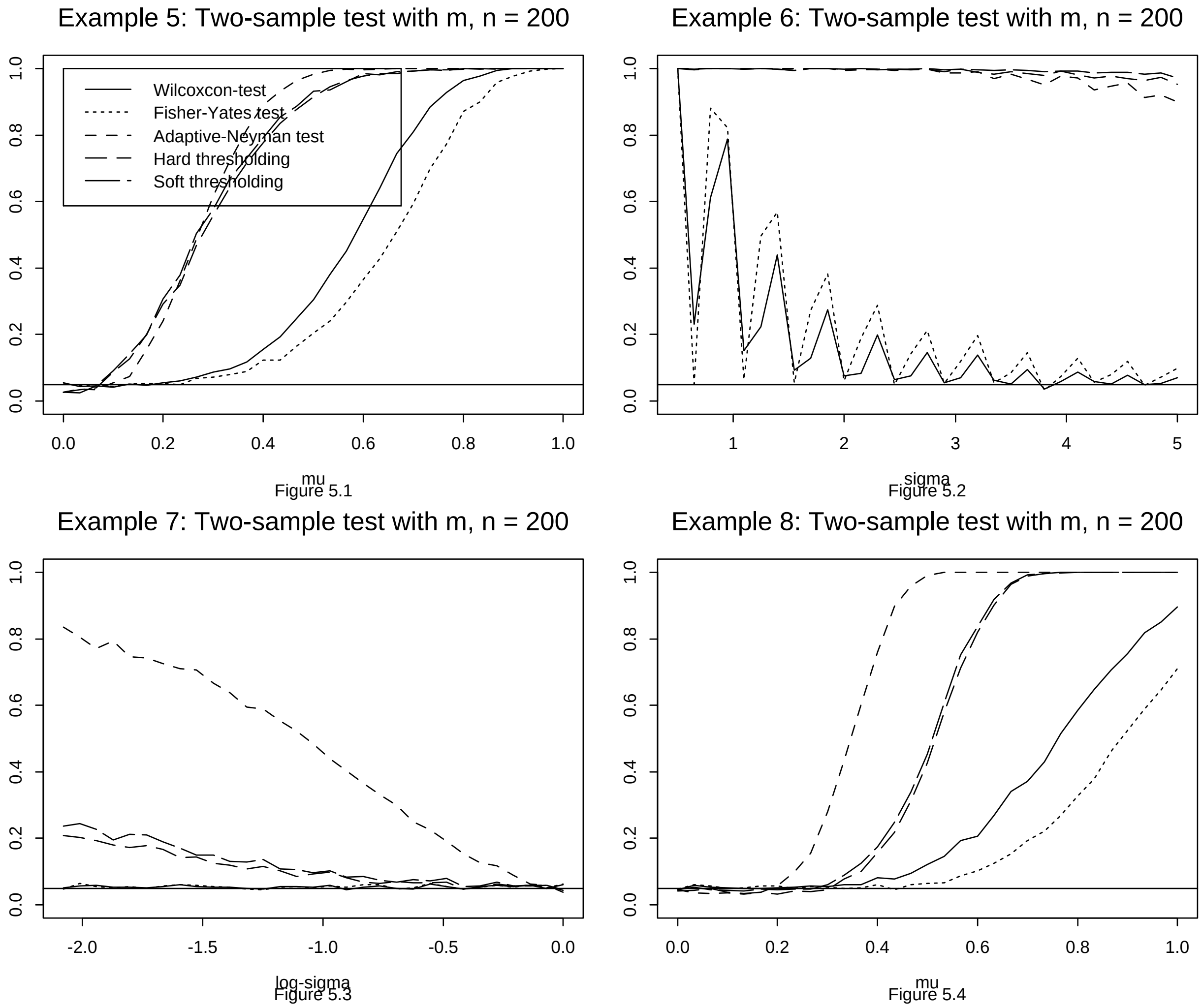

Figure 5: *The plot of power of each testing procedure against a sequence of alternatives given in each example. Solid curve – power of the Wilcoxon rank-sum test. Dashed curves — from shortest dash to longest dash: The power of the Fisher-Yates test, the adaptive Neyman test, the wavelet hard- and soft-thresholding tests. The horizontal line indicates 5%-significant level*

and the wavelet oracle test within a logarithmic factor. Our theory and simulations evidently show that the proposed tests use more fully the sampling information and therefore are more powerful than the traditional distribution-based or linear rank based test statistics in detecting subtle local features and high frequency alternations. They are also more robust against various kinds of alternatives. Our simulation experience indicates that the wavelet thresholding tests are somewhat more powerful than the adaptive Neyman test in detecting subtle local features, while the latter has somewhat better ability in detecting smoothed alternatives.

The asymptotic distributions of the proposed testing procedures for the two-sample problem (5.1) have not derived yet. In practical implementations, the finite-sample distributions are needed. With modern computers, the critical values can easily be simulated and tabulated. Our simulations show that even with 1600 simulations which took a few seconds to compute in a SUN Sparc 10 station, the critical values are practically accurate enough.

In practice, one can encounter the situation that F_0 involves a few unknown parameters. The proposed testing procedures continue to apply to this situation with the unknown parameters being suitably estimated. However, it remains open whether the resulting procedures have the desired power and the given significant level. Further investigation is needed.

Appendix — Proofs

Proof of Theorem 2.2. Let $c_\alpha = -\log(-\log(1-\alpha))$. Observe that

$$P_{\theta_0}\{T_{AN} > c_\alpha\} = P_{\theta_0}\{T_{AN}^* > \sqrt{2\log\log n}(1+o(1))\}.$$

Let $\eta_n = (2m_0)^{-1/2}\sum_{j=1}^{m_0}\theta_{j0}^2$. Then, since $T_{AN}^* \geq (2m_0)^{-1/2}\sum_{j=1}^{m_0}(X_j^2-1)$, we have that

$$P_{\theta_0}\{T_{AN} > c_\alpha\} \geq P_{\theta_0}\{(2m_0)^{-1/2}\sum_{j=1}^{m_0}(X_j^2 - 1 - \theta_{j0}^2) \geq \sqrt{2\log\log n}(1+o(1)) - \eta_n\}.$$

Since the sequence

$$\left\{(2m + \sum_{j=1}^{m}\theta_{j0}^2)^{-1/2}\sum_{j=1}^{m}(X_j^2 - 1 - \theta_{j0}^2), m = 1, 2, \cdots, n\right\}$$

is tight, when $\eta_n >> (\log\log n)^{1/2}$, the power tends to 1. This proves the second conclusion.

Proof of Theorem 2.3. We first calculate the first two moments of the hard-thresholding test statistic. Let $Y_{n,j} = X_j^2 I(|X_j| > \delta)$. By using the integration by parts, under H_0,

$$\begin{aligned} EY_{n,j}^k &= \sqrt{2/\pi}\int_\delta^\infty x^{2k}\exp(-x^2/2)dx \\ &= \sqrt{2/\pi}\{\delta^{2k-1}\exp(-\delta^2/2) + (2k-1)\int_\delta^\infty x^{2k-2}\exp(-x^2/2)dx\} \\ &= \sqrt{2/\pi}\delta^{2k-1}\{1 + (2k-1)\delta^{-2} + O(\delta^{-4})\}\exp(-\delta^2/2). \end{aligned}$$

Therefore,

$$E(\sum_{j=1}^n Y_{n,j}) = \sqrt{2/\pi}a_n^{-1}\delta\{1 + \delta^{-2} + O(\delta^{-4})\}$$

and

$$\mathrm{var}(\sum_{j=1}^n Y_{n,j}) = \sqrt{2/\pi}a_n^{-1}\delta^3\{1 + 3\delta^{-2} + O(\delta^{-4})\}.$$

Using the above moment calculation, it is easy to show that the Lyapounov's condition holds:

$$\begin{aligned} \sum_{j=1}^n E(Y_{n,j} - EY_{n,j})^4 &= O\{n\delta^7\exp(-\delta^2/2)\} \\ = O(a_n^{-1}\log^{7/2} n) &= o\{(a_n^{-1}\log^{3/2} n)^2\}. \end{aligned}$$

This establishes the asymptotic normality for $\hat{T}_H$. For the soft-thresholding test $\hat{T}_S$, by using the integration by parts, we have

$$E\psi_\delta^2(X_j) = \sqrt{2/\pi}2\delta^{-3}\exp(-\delta^2/2)\{1 - 4\delta^{-2} + O(\delta^{-4})\}$$

and

$$\begin{aligned} E\psi_\delta^4(X_j) &= \sqrt{2/\pi}\Big\{(\delta^2+3)\int_\delta^\infty (x-\delta)^2\exp(-x^2/2)dx \\ &\qquad -2\int_\delta^\infty (x-\delta)\exp(-x^2/2)dx\Big\} \\ &= 2\sqrt{2/\pi}\delta^{-1}\exp(-\delta^2/2)\{1+o(1)\}. \end{aligned}$$

We can similarly show that the Lyapounov's condition holds. This proves Theorem 2.3.

Proof of Theorem 2.4. Let $z_{1-\alpha} = \Phi^{-1}(1-\alpha)$. By differentiation with respect to θ_{j0} and using the integration by parts, we can show that the power function $P_{\theta_0}\{\hat{T}_H > z_{1-\alpha}\}$ is an increasing function of θ_{j0}. Thus,

$$P_{\theta_0}\Big\{\hat{T}_H > z_{1-\alpha}\Big\} \geq P_{\theta_0}\Big\{\sigma_{n,H}^{-1}\sum_{i\in S_0} X_i^2 I(|X_i| > \delta) + Z_n > z_{1-\alpha}\Big\}, \tag{A.1}$$

where with $Y_1, \cdots, Y_{n-m_0}$ being the i.i.d. $N(0,1)$ random variables, independent of $\{X_i, i \in S_0\}$,

$$Z_n = \sigma_{n,H}^{-1}\Big(\sum_{i=1}^{n-m_0} Y_i^2 I(|Y_i| > \delta) - \mu_{n,H}\Big).$$

By Theorem 2.3, $Z_n \to N(0,1)$. Since $X_j^2 I(|X_j| \geq \delta) \geq X_j^2 - \delta^2$, it follows from (A.1) that

$$P_{\theta_0}\Big\{\hat{T}_H > z_{1-\alpha}\Big\} \geq P_{\theta_0}\Big\{\sum_{i\in S_0}(X_i^2 - 1) > m_0(\delta^2 - 1) + \sigma_{n,H}(z_{1-\alpha} - Z_n)\Big\}.$$

This proves the first conclusion. To prove the second conclusion, we denote $X \sim N(\theta, 1)$. Then, it can be shown that when $\theta > \sqrt{2\log n}$,

$$EX^2 I(|X| > \delta) = \theta^2 + 1 + o(1), \quad EX^4 I(|X| > \delta) = \theta^4 + 6\theta^2 + O(1).$$

Thus, when $\theta > \sqrt{2\log n}$,

$$\mathrm{var}\{X^2 I(|X| > \delta)\} = 4\theta^2(1 + o(1)).$$

Let $U_n = \sum_{j=1}^n X_j^2 I(|X_j| > \delta, |\theta_j| > \sqrt{2\log n})$. By (A.1), we have

$$P_{\theta_0}\Big\{\hat{T}_H > z_{1-\alpha}\Big\} \geq P_{\theta_0}\left\{(U_n - EU_n)/\sqrt{\mathrm{var}(U_n)} > \{\sigma_{n,H}(z_{1-\alpha} - Z_n) - EU_n\}/\sqrt{\mathrm{var}(U_n)}\right\}. \tag{A.2}$$

Since the sequence $\{(U_n - EU_n)/\sqrt{\mathrm{var}(U_n)}\}$ is tight and

$$\{\sigma_{n,H}(z_{1-\alpha} - Z_n) - EU_n\}/\sqrt{\mathrm{var}(U_n)} \longrightarrow -\infty,$$

probability in (A.2) tends to 1. This proves the second conclusion.

Proof of Theorem 4.1. Let $\hat{F}_n$ be the empirical cumulative distribution function, and $t_j = j/n_{\text{bin}}$. Then $N_j = n\{\hat{F}_n(t_j) - \hat{F}_n(t_{j-1})\}$. By the "Hungarian Embedding" of Komlòs, Major and Tusnády (1975), there exists a version of Brownian bridge B_n such that for any $r > 0$,

$$E\|n(\hat{F}_n - F) - n^{1/2}B_n(F)\|_\infty^r = O(\log^r n).$$

Write $Z_n = O_r(a_n)$ if $E|X_n|^r = O(a_n^r)$. Then, we have

$$N_j = n\{F(t_j) - F(t_{j-1})\} + n^{1/2}\{B_n(F(t_j)) - B_n(F(t_{j-1}))\} + O_r(\log n).$$

Let $W_n(t)$ be the Brownian motion with $W_n(0) = 0$. By changing a probability space if needed, $B_n(t) = W_n(t) - tW_n(1)$. Write $p_j = F(t_j) - F(t_{j-1})$ and $\varepsilon_j = p_j^{-1/2}\{W_n(F(t_j)) - W_n(F(t_{j-1}))\}$. Then,

$$N_j = np_j + (np_j)^{1/2}\{\varepsilon_j - p_j^{1/2}W_n(1)\} + O_r(\log n).$$

By the independent increments and the covariance structure of the Brownian motion, $\varepsilon_1, \cdots, \varepsilon_{n_{\text{bin}}}$ are i.i.d. $N(0,1)$. Thus, $\max_{1\le j\le n_{\text{bin}}} \varepsilon_j = O_r(\log^{1/2} n_{\text{bin}}) = O_r(\log^{1/2} n)$. Using this and Taylor's expansion upto the second order, we obtain

$$\begin{aligned} Y_j &= 2[\{np_j + (np_j)^{1/2}\{\varepsilon_j - p_j^{1/2}W_n(1)\} + O_P(\log n)\}^{1/2} - (n/n_{\text{bin}})^{1/2}] \\ &= \theta_j + \varepsilon_j - p_j^{1/2}W_n(1) + O_P\{(n^{-1}n_{\text{bin}}\log n)^{1/2}\}. \end{aligned} \tag{A.3}$$

To prove (A.3) in a stronger mode, we need the following truncation arguments. Write

$$N_j \equiv np_j\{1 + (np_j)^{-1/2}A_j + (np_j)^{-1}b_j\},$$

where $A_j = \varepsilon_j - p_j^{1/2}W_n(1)$ and $\max_j b_j = O_r(\log n)$. Then,

$$\begin{aligned} &E\max_j |Y_j - \theta_j - A_j|^r = E\max_j |Y_j - \theta_j - A_j|^r \\ &\times[I\{|(np_j)^{-1/2}A_j + (np_j)^{-1}b_j| \le 0.5\} + I\{|(np_j)^{-1/2}A_j + (np_j)^{-1}b_j| > 0.5\}]. \end{aligned} \tag{A.4}$$

Using Taylor's expansion upto the second order, the first term is bounded by

$$O\left(E\max_j|(np_j)^{-1/2}b_j + (np_j)^{1/2}\{(np_j)^{-1/2}A_j + (np_j)^{-1}b_j\}^2|^r\right) = O\left((n^{-1}n_{\text{bin}}\log n)^{r/2}\right).$$

By the Cauchy-Schwartz inequality, the second term of (A.4) is bounded by

$$\left(E\max_j |Y_j - \theta_j - A_j|^{2r}\right)^{1/2} \left(P\{\max_j(np_j)^{-1/2}|A_j| > 0.25\} + P\{\max_j(np_j)^{-1}|b_j| > 0.25\}\right)^{1/2}.$$

Using the moment method to bound the tail probabilities, the second factor is of order $O((n_{\text{bin}}/n)^k)$ for any $k > 0$. Therefore, by choosing k that is large enough, the second term in (A.4) is also negligible. The result follows directly from (A.3) by observing that $p_j = O(n_{\text{bin}}^{-1})$.

Proof of Theorem 4.2. First of all, by (A.3) and the fact that $p_j = n_{\text{bin}}^{-1}$, we have

$$Y_j = \varepsilon_j - n_{\text{bin}}^{-1/2} W(1) + O_P\{(n^{-1} n_{\text{bin}} \log n)^{1/2}\}. \tag{A.5}$$

Note that for the orthogonal matrix $\Gamma = (\gamma_{i,j})$, by the assumption of Theorem 4.2, $\sum_{j=1}^{n_{\text{bin}}} \gamma_{i,j} = 0$, for almost all i. This together with (A.5) lead to

$$X_\Gamma = \Gamma\varepsilon + \Gamma V.$$

where $\max_j V_j = O_P\{(n^{-1} n_{\text{bin}} \log n)^{1/2}\}$. Thus, $\|\Gamma V\|_2 = \|V\|_2 = O_P\{(n^{-1} n_{\text{bin}}^2 \log n)^{1/2}\}$.

Let $(X_1, \cdots, X_{n_{\text{bin}}})^T = \Gamma\varepsilon \sim N(0, I_{n_{\text{bin}}})$. Then, by the definition of the adaptive Neyman test, we have

$$T_{AN}^* = \max_{1 \le m \le n_{\text{bin}}} \{(2m)^{-1/2} \sum_{j=1}^{m} (X_j^2 - 1)\} + O_P(\|\Gamma V\|_2)$$

Therefore, by the result of Darlin and Erdös (1956), we obtain the desired result.

Proof of Theorem 4.3. Following the notation in the Proof of Theorem 4.2, by the Cauchy-Schwartz inequality, $\sum_{j=1}^{n_{\text{bin}}} |\gamma_{i,j}| \le n_{\text{bin}}^{1/2}$. Using this and Theorem 4.1, we have $X_{\Gamma,j} = X_j + T_j, j = 1, \cdots, n_{\text{bin}}$ with $\|T\|_\infty \equiv \max_j T_j = O_r\{(n^{-1} n_{\text{bin}}^2 \log n)^{1/2}\}$. Therefore with $\alpha_n = \log^{-1} n$, we have

$$P(|T_j| > \alpha_n) \le P(\|T\|_\infty > \alpha_n) \le \log^r n E\|T\|_\infty^r = O(n^{-a})$$

for any $a > 0$ by choosing a sufficiently large r. Observe that

$$\begin{aligned} \hat{T}_H^* &= \sum_{1}^{n_{\text{bin}}} (X_j + T_j)^2 I(|X_j + T_j| > \delta) \\ &= \sum_{1}^{n_{\text{bin}}} X_j^2 I(|X_j + T_j| > \delta) + O_P\{\|T\|_\infty^2 \sum_{1}^{n_{\text{bin}}} I(|X_j + T_j| > \delta)\}. \end{aligned} \tag{A.6}$$

The second term is of order

$$O_P(n^{-1} n_{\text{bin}} \log n \{\sum_{j=1}^{n} [P(|X_j| \ge \delta - \alpha_n) + P(|T_j| \ge \alpha_n)]\}) = o_P(n^{-1} n_{\text{bin}}^{2+\gamma/2}) = o_P(n^{-\gamma/2}),$$

which is negligible. The first term can be expressed as

$$\sum_{1}^{n_{\text{bin}}} X_j^2 I(|X_j + T_j| > \delta, |T_j| > \alpha_n) + \sum_{1}^{n_{\text{bin}}} X_j^2 I(|X_j + T_j| > \delta, |T_j| \le \alpha_n). \tag{A.7}$$

The first term in (A.7) is bounded by

$$\max_j X_j^2 \sum_{1}^{n_{\text{bin}}} I(|T_j| > \alpha_n) = O_P\{n_{\text{bin}} \log n \max_j P(|T_j| \ge \alpha_n)\} = O(n^{-a}), \tag{A.8}$$

for any a. Therefore, we need only to consider the main term

$$S_n \equiv \sum_{1}^{n_{\text{bin}}} X_j^2 I(|X_j + T_j| > \delta, |T_j| < \alpha_n)$$

This main term is bounded from above and below by

$$\sum_{1}^{n_{\text{bin}}} X_j^2 I\{|X_j| > \delta + \alpha_n\} I(\max_j |T_j| < \alpha_n) \leq S_n \leq \sum_{1}^{n_{\text{bin}}} X_j^2 I\{|X_j| > \delta - \alpha_n). \qquad \text{(A.9)}$$

By the same argument as (A.8), we have

$$\begin{aligned} S_n &\geq \sum_{1}^{n_{\text{bin}}} X_j^2 I\{|X_j| > \delta + \alpha_n\} - n_{\text{bin}} \max_j X_j^2 I(\max_j |T_j| > \alpha_n) \\ &= \sum_{1}^{n_{\text{bin}}} X_j^2 I\{|X_j| > \delta + \alpha_n\} - O_P(n^{-a}), \end{aligned}$$

for any $a > 0$. Since both sides of (A.9) have the same limiting distribution, we obtain the desired result.

REFERENCES

Anderson, T.W. and Darling, D.A. (1952), "Asymptotic theory of certain goodness-of-fit criteria based on stochastic processes," *The Annals of Mathematical Statistics*, 23, 193-212.

Bickel, P.J. (1983), "Minimax estimation of a normal mean subject to doing well at a point," *Recent Advances in Statistics* (M.H. Rizvi, J.S. Rustagi, and D. Siegmund, eds.), Academic Press, New York, 511 – 528.

Bickel, P.J. and Rosenblatt, M. (1973), "On some global measures of the deviation of density function estimates," *The Annals of Statistics*, 1, 1071 – 1095.

Bickel, P.J. and Ritov, Y. (1992), "Testing for goodness of fit: a new approach," *Nonparametric Statistics and Related Topics* (A.K.Md.E. Saleh, ed.), Elsevier Science Publishers B.V., 51-57.

Bowman, A.W. (1992), "Density based tests for goodness-of-fit," *Journal of Statistical Computation and Simulation* 40,1-13.

Bowman, A.W. and Foster, P.J. (1993), "Adaptive smoothing and density based tests of multivariate normality," *Journal of the American Statistics Association*, 88, 529-37.

D'Agostino, R.B. and Stephens, M.A. (1986), "Goodness-of-Fit Techniques," Marcel Dekker, New York.

Darling, D. A. and P. Erdös (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*, SIAM, Philadelphia.

Donoho, D.L. (1992), "Nonlinear solution of linear inverse problems by wavelet-vaguelette decomposition," *Technical report 403*, Department of Statistics, Stanford University.

Donoho, D.L. and Johnstone, I.M. (1994), "Ideal spatial adaptation by wavelet shrinkage," *Biometrika*, 81, 425 - 455.

Donoho, D.L. and Liu, R.C. (1991), "Geometrizing rate of convergence II," *The Annals of Statistics.*, 19, 633 – 667.

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.N. (1992), "Asymptotic comparison of Cramér-von Mises and nonparametric function estimation techniques for testing goodness-of-fit," *The Annals of Statistics*, 20, 2071-2086.

Fan, J. (1994), Test of significance based on wavelet thresholding and Neyman's truncation (1994), Discussion paper # 9315, Institut de Statistique, Université Catholique de Louvain, Belgium.

Fan, J., P. Hall, M. Martin and P. Patil (1993), Adaptation to high spatial inhomogeneity based on wavelets and on local linear smoothing, *Institute of Statistics Mimeo Series # 2307*, University of North Carolina at Chapel Hill.

Fisher, R.A. and Yates, F. (1938), *Statistical tables for biological, agricultural and medical research*, Oliver & Boyd, Edinburgh-London.

Hájek, J. and Šidák, Z. (1967), *Theory of Rank Tests*, Academic Press, New York.

Hollander, M. and Wolfe, D.A. (1973), *Nonparametric statistical methods*, Wiley, New York.

Ibragimov, I.A. and Hasminksii, R.Z. (1984), "On nonparametric estimation of a linear functional in a Gaussian white noise," *Theory Probab. Appl.*, **29**, 19 – 32.

Komlós, J., P. Major and Tusnády, G. (1975), "An approximation of partial sums of independent RV's, and the sample DF. I," *Z. Wahrscheinlichkeitsheorie verw. Gebiete* **32**, 111-131.

Mallat, S. (1989), "A theory for multiresolution signal decomposition: The wavelet representation," *IEEE Trans. Pattern Anal. and Machine Intell.*, **11**, 674 – 693.

Meyer, Y. (1990), *Ondelettes*, Hermann, Paris.

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

Pinsker, M.S. (1980), "Optimal filtering of square integrable signals in Gaussian white noise," *Problems Inform. Transmission*, **16**, 52 – 68.

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

Shapiro, S. S. and Wilk, M.B. (1965), "An analysis of variance test for normality," *Biometrika*, 52, 591-611.

Silverman, B. W. (1986), *Density Estimation for Statistics and Data Analysis*, Chapman and Hall, London.

Wilcoxon, F. (1945), "Individual comparisons by ranking methods," *Biometrics Bulletin*, 1, 80-83.