Electronic Journal of Statistics
Vol. 12 (2018) 767–792
ISSN: 1935-7524
https://doi.org/10.1214/18-EJS1405

Hypothesis testing sure independence screening for nonparametric regression[*]

Adriano Zanin Zambom

e-mail: adriano.zambom@csun.edu

and

Michael G. Akritas

e-mail: mga@stat.psu.edu

Abstract: In this paper we develop a sure independence screening method based on hypothesis testing (HT-SIS) in a general nonparametric regression model. The ranking utility is based on a powerful test statistic for the hypothesis of predictive significance of each available covariate. The sure screening property of HT-SIS is established, demonstrating that all active predictors will be retained with high probability as the sample size increases. The threshold parameter is chosen in a theoretically justified manner based on the desired false positive selection rate. Simulation results suggest that the proposed method performs competitively against procedures found in the literature of screening for several models, and outperforms them in some scenarios. A real dataset of microarray gene expressions is analyzed.

Keywords and phrases: ANOVA, false discovery rate, lack-of-fit test, multiple testing, nonparametric regression.

Received January 2017.

1. Introduction

In recent years, fast advances in technology and data collection have facilitated the acquisition of high-dimensional data in several areas of research. The challenge arises when the number of predictors is larger than the sample size, which can be found for example in studies with genomic microarrays, high frequency functional MRI or imaging decoding. Several regularization methods can be used to perform variable selection in such situations, including the LASSO [17], the SCAD [7], the LARS [6], the elastic net [24] and the Dantzig selector [4]. Although these methods yield good results for high-dimensional data, when the number of predictors is ultra-high they may not perform well due to computational problems or statistical accuracy. In order to deal with these challenges, it becomes necessary to develop methods that reduce the dimensionality of the predictor space from an ultra-high scale to a relatively high scale.

[*] This is an original survey paper.

Fan and Lv [8] were pioneers in studying theoretical aspects for the idea of screening out unimportant predictors in a regression model. They introduced the concept of sure independence screening (SIS), that is, with probability tending to 1, a well chosen subset of the predictors will contain the true set of predictors that contribute to the underlying model. The theoretical properties of this procedure were obtained under the strong assumption of a linear model. However, if this assumption is not accurate, predictors with high predictive significance whose effects are nonlinear might not be detected.

In order to identify nonlinear effects in a regression model, Fan, Feng and Song (2011) [11] considered nonparametric independence screening (NIS) with an additive model, ranking the utility of the covariates with $Em_j^2(X_j)$, where $m_j = E(Y|X_j)$, the projection of Y onto X_j. For multi-index models Zhu, Li, Li and Zhu (2011) [23] used $E[xE\{I(Y < y)|x\}]$ as the population utility measure for a covariate, estimating it with the statistic $(1/n)\sum_j^n \left[(1/n)\sum_i^n X_i I(Y_i < Y_j)\right]^2$. Several other authors have recently developed methods for variations of linear and nonlinear models, see for instance [13], [9], [10], [12], [18], [21], [14] and [22]. However, little is found in the literature regarding screening for fully nonparametric regression models. Li, Zhong and Zhu (2012) [15] innovatively considered a model-free sparse regression whose active predictors are those which $F(Y|\mathbf{X})$ is functionally dependent on. In order to allow for arbitrary regression relationship, they used the distance correlation (DC-SIS) between each covariate and the response variable as the ranking for screening.

In this paper we propose a novel screening method that, differently from the focus of the procedures in the literature, is based on a test statistic for the hypothesis that each available predictor has predictive significance. The signal strength of active predictors is based on the variance of the marginal nonparametric regression function. We use a powerful nonparametric test proposed by Zambom and Akritas (2014) [20] to compute the marginal utility of each predictor. New asymptotic theory is developed in order to establish the rates of convergence of the test statistic with a new Berry Essen type bound for its distribution and exponential convergence rates for the variance estimator. The proposed method is performed under a very general heteroscedastic nonparametric regression model, which does not require strong assumptions such as linearity or additivity of the mean regression function. Moreover, due to the fact that the predictors are ranked using a test statistic, a meaningful choice of the threshold parameter can be made, a fundamental advantage over the ad-hoc approaches in other procedures in the literature.

The remaining of the paper is as follows. In Section 2 we present the nonparametric regression model and preliminary asymptotic properties of the test statistic. The screening method HT-SIS and its sure independence properties are examined in Section 3. Section 4 describes a procedure to select the threshold parameter in order to maintain a desired false positive rate. Section 5 presents a comparison of the performance of HT-SIS, the parametric SIS and the model-free DC-SIS and finally a microarray dataset is analyzed in Section 6.

2. The model and preliminary results

Let Y denote the response variable, $\mathbf{X} = (X_1, \ldots, X_d)$ the vector of available predictors, and with some abuse of notation, let X_{ki} be the i-th observation of the k-th covariate. Assume that the data come from the heteroscedastic nonparametric regression model

$$Y = m(\mathbf{X}) + \sigma(\mathbf{X})\epsilon, \tag{1}$$

where ϵ is the independent error with $E(\epsilon) = 0$ and constant variance (w.l.o.g. assume $\text{Var}(\epsilon) = 1$), uncorrelated with $\mathbf{X}$. When the dimension d of the vector of covariates is high, it is often assumed that the regression model is sparse, in the sense that, there is a unique subset of indices I_0 such that the regression function $m(\cdot)$ is influenced only by those predictors whose indices are in I_0. Hence, we define $I_0 \subseteq \{1, \ldots, d\}$ such that the true underlying model is

$$Y = m(\mathbf{X}_{I_0}) + \sigma(\mathbf{X})\epsilon.$$

Note that the variance function $\sigma(\cdot)$ is not restricted to the set of predictors in I_0, for we are only interested in selecting predictors that have predictive significance, that is, those that contribute to the underlying mean regression function.

There are several procedures in the literature for testing whether a covariate has no predictive value. The most common idea is to test for a constant conditional expectation of the response given the covariate. The majority of the literature proposes tests which assume homoscedasticity and hence become liberal under heteroscedasticity. Thus, a covariate with no predictive value stands a good chance of being selected as a predictor if the variance function, or even other aspects of the conditional distribution of the response, are not constant with respect to the covariate. Based on a sample of n iid observations from model (1), we propose ranking the utility of the covariates using, marginally, the test statistic introduced by Zambom and Akritas (2014) [20]. We now briefly recall the test statistic and its main properties. For the marginal regression model $Y = m_k(X_k) + \sigma_k(X_k)\epsilon_k$, consider the null hypothesis

$$H_0^k : m_k(x_k) = C_k, k = 1, \ldots, d, \tag{2}$$

for a constant C_k. Let $(Y_i, X_{ki}), i = 1, \ldots, n$ represent data from a high-dimensional one-way ANOVA design with Y_i being the observation at “level” X_{ki}. Because of the ANOVA requirement of more than one observation per cell, each cell is augmented with neighboring observations in the following way. Consider that X_{ki} is arranged in order of magnitude. Define the augmented cell X_{ki} to consist of Y_i and the Y_j’s corresponding to the $(p-1)/2$ X_{kj}’s on either side of X_{ki}, for a fixed odd constant p. Then, the set of indices j composing the augmented cell X_{ki} can be written as

$$W_i^k = \left\{ \ell : |\hat{F}_{X_k}(X_{k\ell}) - \hat{F}_{X_k}(X_{ki})| \le \frac{p-1}{2n} \right\}, \tag{3}$$

where $\hat{F}_{X_k}$ is the empirical distribution function of X_k, so that W_i^k defines the augmented cell corresponding to X_{ki}. The test statistic for the hypothesis in (2) is based on the high-dimensional one-way ANOVA type test statistic

$$T_k = MST_k - MSE_k = \frac{p}{n-1}\sum_{i=1}^{n}(Y_{i\cdot} - Y_{\cdot\cdot})^2 - \frac{1}{np-n}\sum_{i=1}^{n}\sum_{j\in W_i^k}(Y_j - Y_{i\cdot})^2, \quad (4)$$

where $Y_{i\cdot} = (1/p)\sum_{j\in W_i^k} Y_j$ and $Y_{\cdot\cdot} = (1/np)\sum_{i=1}^{n}\sum_{j\in W_i^k} Y_j$. Note that T_k can be written in a quadratic form as $T_k = \mathbf{Y}_{W^k}^T A \mathbf{Y}_{W^k}$ where $\mathbf{Y}_{W^k}$ is the vector of $(n-p+1)p$ augmented observations

$$\mathbf{Y}_{W^k} = (Y_i, i\in W_1^k, \ldots, Y_i, i \in W_n^k)^T \quad (5)$$

in the high-dimensional one-way ANOVA and the matrix A is

$$A = \frac{np-1}{n(n-1)p(p-1)} \oplus_{i=1}^{n} \underline{\mathrm{J}}_p - \frac{1}{n(n-1)p}\underline{\mathrm{J}}_{np} - \frac{1}{n(p-1)}\underline{\mathrm{I}}_{np}, \quad (6)$$

where $\underline{\mathrm{I}}_r$ is an identity matrix of dimension r, $\underline{\mathrm{J}}_r$ is a rxr matrix of 1's and $\oplus$ is the Kronecker sum or direct sum.

Remark 1. Simulations suggest that the choice of the window size p has little influence on the performance of the test, as long as it is not too small or too large. Choosing $p < 5$ tends to make the test procedure liberal, while a large value of p has the opposite effect. In simulations we used $p = 11$. A way to gain confidence in the choice of p in any practical situation is to run the test after randomly permuting the observed response variables among the covariate values, in order to induce the validity of the null hypothesis.

To obtain insight on the properties of T_k for ranking the utility of X_k in the nonparametric regression, we recall the following theorem

Theorem 1. *(Zambom and Akritas, 2014 [20]) Assume that $\sigma_k^2(x_k)$ is Lipschitz continuous, $\sup_x \sigma_k^2(x_k) < \infty$, the marginal density f_{X_k} of X_k is uniformly continuous and bounded away from 0 and $E(\epsilon_k^4) < \infty$. Then under H_0 in (2), the asymptotic distribution of the test statistic in (4) is given by*

$$n^{1/2}(MST_k - MSE_k) \xrightarrow{d} N(0, v_k),$$

where $v_k = [2p(2p-1)\tau_k^2]/[3(p-1)]$ and $\tau_k = \int \left[\sigma_k^2(x_k)\right]^2 f_{X_k}(x_k)dx_k$.

In order to estimate v_k, assume that the response values $Y_i, i = 1, \ldots, n$ are sorted according to X_k, in other words, assume that Y_i is the observation corresponding to $X_{k(i)}$, where $X_{k(i)}$ is the i-th largest observation of the sample $X_{k1}, \ldots, X_{kn}$. Then, a consistent estimator of v_k (see Lemmas 2 and 3) is

$$\hat{v}_k = \frac{2p(2p-1)}{3(p-1)}\frac{1}{4(n-3)}\sum_{j=2}^{n-2}(Y_j - Y_{j-1})^2(Y_{j+2} - Y_{j+1})^2. \quad (7)$$

Note that both MST_k and MSE_k are averages and converge to constants. Under the null hypothesis (2), both converge to the same constant. Under local alternatives, Zambom and Akritas [20] showed that the asymptotic distribution of the test statistic is Normal with mean given by $p\mathrm{Var}(m_k(X_k))$. The hypothesis is hence rejected for large values of the test statistic, so that it is expected that T_k is a useful statistic to rank the utility of each predictor.

3. The screening procedure and main results

The Hypothesis Testing Nonparametric Independence Screening (HT-SIS) procedure consists of selecting a superset of indices $\hat{I}$ that contains the index set I_0 with probability increasing to one as the sample size increases. The challenge addressed in screening is to deal with the situation where the number of predictors d greatly exceeds the sample size n. Define the superset $\hat{I}$ as

$$\hat{I} = \left\{ k : \frac{T_k}{\sqrt{\hat{v}_k}} \geq cpn^{-\alpha}, 1 \leq k \leq d \right\}, \tag{8}$$

where c and α are threshold parameters defined in condition C8 below and p is the window size defined in (3). Note that in Section 4 we set $cpn^{-\alpha} = \lambda_n$ and provide a method for choosing λ_n. In order to establish the sure screening properties of HT-SIS, consider the following conditions. For any $1 \leq i, j \leq n$ and some $s > 0$

$$C1 : \sup_d \max_{1\leq k\leq d} E(\exp\{s\sigma_k^2(X_{ki})\epsilon_{ki}^2\}) < \infty$$

$$C2 : \sup_d \max_{1\leq k\leq d} E(\exp\{sm_k^2(X_{ki})\}) < \infty$$

$$C3 : \sup_d \max_{1\leq k\leq d} E(\exp\{s\sigma_k(X_{ki})\epsilon_{ki}\sigma_k(X_{kj})\epsilon_{kj}\}) < \infty$$

$$C4 : \sup_d \max_{1\leq k\leq d} E(\exp\{sm_k(X_{ki})m_k(X_{kj})\}) < \infty$$

$$C5 : \sup_d \max_{1\leq k\leq d} E(\exp\{s\sigma_k(X_{ki})\epsilon_{ki}m_k(X_{ki})\}) < \infty$$

$$C6 : m_k(\cdot) \text{ and } \sigma_k(\cdot) \text{ are Lipschitz continuous for } k = 1, \ldots, d$$

$$C7 : f_{X_k}(\cdot), k = 1, \ldots, d, \text{ are bounded away from } 0.$$

where f_{X_k} is the density of X_k, with support in $\mathcal{X}_k$. Conditions C1-C7 are necessary for the derivation of Theorem 2 and supporting Lemmas 1 - 3. Conditions C1-C5 are similar to condition C1 in Li, Zhong and Zhu (2012) [15], which require finite expected values of exponential functions of $\sigma_k(X_k)\epsilon_k$ and $m_k(X_k)$. These conditions follow if $\sigma_k^2(\cdot)$ and $m_k(\cdot)$ are bounded uniformly. Conditions C6 and C7 are usual conditions in nonparametric regression (see for example Fan, Feng and Song, 2011 [11]), where C7 for example follows for distributions with compact support.

In all theoretical results that follow, the constants in the $O(\cdot)$ notation may depend, as indicated, on the expected value of functions of $\sigma_k(X_k)$ and $m_k(X_k)$, and hence also on f_{X_k}. We denote these constants by $C_\sigma, C_{m\sigma}$. Their exacts expressions are suppressed for ease of notation. Note that these constants, although sometimes with the same subscript, may take different values at each appearance. In the following lemma, we establish the rate at which the test statistic T_k converges in probability to its expected value.

Lemma 1. *Under conditions C1-C5, for any $0 < \gamma < 1/2 - \alpha$, there exists constants $c_1 > 0$ and $c_2 > 0$ such that*

$$P(\max_{1\leq k\leq d} |T_k - E(T_k)| \geq cn^{-\alpha}) \leq O(d[\exp(-c_1 n^{1-2(\gamma+\alpha)}) + nC_{m\sigma}\exp(-c_2 n^{\gamma})]).$$

Note that for an active predictor $X_k, k \in I_0$, we expect the value of T_k not to be too small, or at least larger than most of those of inactive predictors. For the sure independence screening property of HT-SIS, we require the following condition

$$C8 : \min_{k\in I_0} \operatorname{Var}(m_k(x)) \geq 2cn^{-\alpha},$$

for some constant c and $0 \leq \alpha < 1/2$. Condition C8 is similar to condition 3 of Fan and Lv (2008) [8] where it is assumed that the true correlation between the predictor and the response is above a certain threshold. In the present case, we assume that the signal strength, measured by the variance of $m_k(\cdot)$, is not too small, however, intuitively, it is 0 if the relationship of X_k and Y is constant.

In Lemmas 2 and 3 we explore the rate of convergence of $\hat{v}_k$, used to standardize the proposed raking utility T_k (see Theorem 1). Note that Lemma 3 establishes the consistency of $\hat{v}_k$ as n goes to infinity. Using these lemmas and in connection with Lemma 1, we can show the sure screening property of HT-SIS, which is stated in Theorem 2.

Lemma 2. *Let $\hat{v}_k$ in (7) be the estimator of v_k. Under conditions C6 and C7 we have that*

$$E(\hat{v}_k) = v_k + O\left(\frac{C_{\sigma_k}}{c_{f_k} n}\right),$$

where C_{σ_k} is the Lipschitz constant for $\sigma_k(\cdot)$, and $c_{f_k} = \inf_{x\in\mathcal{X}_k} f_k(x)$.

Lemma 3. *Let $\hat{v}_k$ in (7) be the estimator of v_k. Under conditions C1 and C7, there exists constants $c_1 > 0$ and $c_2 > 0$ such that*

$$P(\max_{1\leq k\leq d} |\sqrt{\hat{v}_k} - \sqrt{v_k}| \geq cn^{-\alpha}) \leq O(d[\exp(-c_1 n^{1-2(\gamma+\alpha)}) + nC_\sigma \exp(-c_2 n^{\gamma})]).$$

Theorem 2. *Under conditions C1-C8, for $0 < \gamma + \alpha < 1/2$, there exists constants $c_1 > 0$ and $c_2 > 0$ such that for any $\varepsilon > 0$*

$$P(I_0 \subseteq \hat{I}) \geq 1 - O\left(d_0\left[\exp\left(-c_1 n^{1-2(\gamma+\alpha)}\right) + nC_{m\sigma}\exp(-c_2 n^{\gamma})\right]\right) - \varepsilon,$$

where d_0 is the cardinality of I_0.

Because the true model is assumed to be sparse, where only a small number d_0 of the predictors have predictive significance, Theorem 2 demonstrates that all significant predictors will be retained with high probability. Note that the theorem holds even when the number of covariates in the model is allowed to increase with the sample size n at an exponential rate.

Remark 2. All screening methods face challenges such as failing to identify important predictors that are marginally independent but maybe jointly correlated with the response or selecting spurious variables, that is, selecting unimportant predictors that are correlated with important predictors. An iterative version of HT-SIS, similar to the iterative versions of SIS, DC-SIS or NIS can easily be implemented in order to alleviate such issues. The asymptotic properties of the iterative versions of these methods is an interesting topic for further analysis.

4. The choice of the threshold parameter

Since Fan and Lv [8] introduced the notion of sure screening, several research studies have explored the theoretical and asymptotic properties of screening methods. However, in the majority of papers, the choice of the threshold parameter is not carefully addressed. Instead of setting a threshold for the ranking utility, most methods fix the maximum number of predictors to be kept after the screening procedure, for instance $n/log(n)$ or even $n-1$. These choices are ad-hoc and provide no meaningful interpretation, but do address the practical objective of ending up with fewer predictor than the sample size.

A characteristic only found in variable selection procedures based on test statistics is the possibility to control the False Positive Rate or the False Discovery Rate [1]. This idea was used by Zhao and Li (2012) [21] for the case of screening in linear Cox models based on a test statistic for the coefficients β_j's. For the linear regression model where the number of covariates is allowed to grow with n, Bunea, Wegkamp and Auguste (2006) [3] proposed a variable selection method based on FDR and showed that it is consistent in selecting the set of significant predictors. The p-values of these test statistics can be used to guarantee that the expected false positive rate will be below a chosen level. In this section we establish theoretical support for the choice of the threshold parameter when applying HT-SIS based on FDR. The asymptotic normality of the test statistic T_k provides a direct choice of the threshold parameter in connection with the cumulative distribution function.

Recall that I_0 is estimated by $\hat{I}$ in (8). Write $\hat{I}$ as

$$\hat{I} = \left\{ k : \frac{T_k}{\sqrt{\hat{v}_k}} \geq \lambda_n, 1 \leq k \leq d \right\},$$

so that λ_n is the threshold parameter to be chosen. If the true model I_0 has size $|I_0| = d_0$, the expected false positive rate is

$$E\left(\frac{|\hat{I} \cap I_0^c|}{|I_0^c|}\right) = \frac{1}{d-d_0} \sum_{k \in I_0^c} P\left(\frac{T_k}{\sqrt{\hat{v}_k}} \geq \lambda_n\right). \tag{9}$$

By Theorem 1 and the consistency of the estimator $\hat{v}_k$, $n^{1/2}T_k/\sqrt{\hat{v}_k}$ has an asymptotic standard Normal distribution, and the expected false positive rate is controlled at $(1-\Phi(n^{1/2}\lambda_n))$, where Φ is the cumulative function of a standard Normal.

In order to have the false positive rate (# false positives)/(#negatives) decrease when the sample size increases, fix the number of false positives r we are willing to tolerate in the screening procedure. Then the false positive rate $r/(d-d_0)$ decreases with the sample size since d is allowed to increase with n (see the rate in Theorem 3). Now by conservatively setting the expected false positive rate as $(1-\Phi(n^{1/2}\lambda_n)) = r/d < r/(d-d_0)$, we obtain $\lambda_n = n^{-1/2}\Phi^{-1}(1-r/d)$. A similar idea was also used in Zhao and Li (2012). Theorem 3 establishes the bounds for the expected false positive rate of the proposed screening method using the Berry-Essen-type bound for T_k derived in Lemma 5.

Lemma 4. *Under assumption C1, for $k \in I_0^c$ we have*

$$n^{1/2}[\mathbf{Y}_{W_k}^T A\mathbf{Y}_{W_k} - (\mathbf{Y}_{W_k} - C_kI_N)^T A_d(\mathbf{Y}_{W_k} - C_kI_N)] = O_p\left(\frac{\mu_{\sigma_k}}{n^{1/2}}\right),$$

where $\mu_{\sigma_k} = E(\sigma_k^2(X_k))$, and A_d is the block diagonal matrix $A_d = diag\{B_1, \dots, B_n\}$, with $B_i = (\underline{J}_p - \underline{I}_p)/(n(p-1))$.

Lemma 5. *Under conditions C6-C7, for $k \in I_0^c$ we have*

$$\sup_x |P(n^{1/2}(\mathbf{Y}_{W_k} - C_kI_N)^T A_d(\mathbf{Y}_{W_k} - C_kI_N)/\sqrt{v_k} \leq x) - \Phi(x)| \leq C_\sigma n^{-3/10}.$$

Theorem 3. *Under conditions C1-C8, for the choice of threshold parameter $\lambda_n = n^{-1/2}\Phi^{-1}(1-r/d)$ and $log(d) = O(n^{1-2\alpha})$, then there exists a constant $c > 0$ such that*

$$E\left(\frac{|\hat{I} \cap I_0^c|}{|I_0^c|}\right) \leq \frac{r}{d} + C_\sigma n^{-3/10},$$

while the sure independence property (Theorem 2) holds.

Theorem 3 establishes that the false positive rate is maintained close to the nominal level chosen r/d, while retaining all active predictors with high probability. The rate at which the number of predictors d is allowed to increase with the sample size is comparable to those of Fan and Lv (2008) [8], where $log(d) = O(n^\xi)$, for some $\xi > 0$ (Condition 1).

Note that the False Discovery Rate is defined as the expected value of $|\hat{I} \cap I_0^c|/|\hat{I}|$. Moreover, $|\hat{I} \cap I_0^c|/|\hat{I}|$ can be written as the product of the false positive rate $|\hat{I} \cap I_0^c|/|I_0^c|$ and $|I_0^c|/|\hat{I}|$. Because $|I_0^c|/|\hat{I}| < d/|\hat{I}|$, the False Discovery Rate can be controlled at $r/|\hat{I}|$ conditionally on $|\hat{I}|$, as long as the false positive rate is controlled at r/d.

5. Simulation study

In this section we analyze the performance of HT-SIS with simulation studies for 7 different models. For comparison purposes, the well known Sure Independence Screening (SIS) [8], the Distance Correlation Sure Independence Screening (DC-SIS) [15] and the Nonparametric Independence Screening (NIS) [11] are also evaluated. All results were obtained in R (www.r-project.org), using packages *NonpModelCheck*, *SIS* and *energy* for HT-SIS, SIS and DC-SIS respectively.

We follow the simulation scenarios of Li, Zhong and Zhu (2012) [15], where we generate $\mathbf{X} = (X_1, \dots, X_d)$ from a Normal distribution with zero mean and covariance matrix $\Sigma = (\sigma_{ij})_{d\times d}$ with $\sigma_{ij} = 0.8^{|i-j|}$, and error term $\epsilon \sim N(0,1)$. Because Normally distributed covariates are used in most variable selection literature, they are used in this simulation section despite the fact that they do not meet condition C7. In consequence, $m_1(X_1) = E(Y|X_1)$ does not meet conditions C1-C5 for all models considered except Model 5. This is because the expected value of exponential functions of terms with order higher than 2 do not exist for Normal random variables ($E(e^{sX^3})$ diverges for X Normally distributed). Hence, the results of this simulation section demonstrate the robustness of the proposed method against departures from conditions C1-C5 and C7. We consider $n = 200$ and $d = 1000$ or 3000 and repeat the experiment 1000 times. The following criteria is used to evaluate the performance of the screening methods:

1. $\mathcal{S}$: the minimum model size to include all active predictors. We report the 5%, 25%, 50%, 75% and 95% quantiles of $\mathcal{S}$ out of 1000 replications.
2. $\mathcal{P}_s$: the proportion that an individual active predictor is selected for a given model size $|\hat{I}|$ in the 1000 replications.
3. $\mathcal{P}_a$: the proportion that all active predictors are selected for a given model size $|\hat{I}|$ in the 1000 replications.

We consider the following models:

$$\begin{aligned}
&1: Y = 2\beta_1 X_1 X_2 + 3\beta_2 \mathbb{1}(X_{12} < 0) + 2\beta_3 X_{22} + \epsilon,\\
&2: Y = 2\beta_1 X_1 X_2 + 2\beta_2 X_{22} + 3\beta_3 sin(X_{12}) + \epsilon\\
&3: Y = 2\beta_1 \beta_2 X_1 cos(X_2) + 2\beta_3 X_{22} + 3\beta_4 sin(X_{12}) + (X_1 + X_2)\epsilon\\
&4: Y = 2\log(|X_1|) + 2X_2 + X_{12} + \sin(X_{12}) + 2X_{22}^2 + \epsilon\\
&5: Y = X_1^2 - 10\cos(2\pi X_1) + X_2^2 - 10\cos(2\pi X_2) + X_{12}^2 - 10\cos(2\pi X_{12})\\
&\qquad\quad + X_{22}^2 - 10\cos(2\pi X_{22}) + (X_{40} + X_{50})\epsilon\\
&6: Y = -10\cos(2\pi X_1) - 10\cos(2\pi X_2) - 10\cos(2\pi X_{12}) - 10\cos(2\pi X_{22}) + \epsilon\\
&7: Y = 2\beta_1 X_1 X_2 + 5\beta_2 \mathbb{1}(0 < X_{12} < 0.2) + 5\beta_2 \mathbb{1}(1 < X_{12} < 1.2) + 2\beta_3 X_{22} + \epsilon
\end{aligned}$$

We generate $\beta_j = (-1)^U (a + |Z|)$, for j =1, 2, 3 and 4, where $a = 4log(n)/\sqrt{n}$, $U \sim Bernoulli(0.4)$, $Z \sim N(0,1)$. Table 1 presents the results of $\mathcal{S}$ and Tables 2 and 3 show the results of $\mathcal{P}_s$ and $\mathcal{P}_a$ for $d = 1000$ and $d = 3000$ respectively. For comparison purposes, the size of the superset is set to $|\hat{I}| = n/log(n) = 38$ as suggested by Fan and Lv or $|\hat{I}| = \#\{T_k \geq \lambda_n\}$, corresponding to the false positive rate of 0.05.

As expected, SIS has low performance in capturing the significance of predictors with nonlinear effects and hence its minimum model size $\mathcal{S}$ is in general

Table 1
Minimum model size $\mathcal{S}$.

		d = 1000				
Model	test	5%	25%	50%	75%	95%
1	HT-SIS	4	6	9	25	171.5
	DC-SIS	5	8	12	18	35
	NSIS	4	6	9	17	113.3
	SIS	18.9	105	378	728	954
2	HT-SIS	4	6	10	26.2	185.2
	DC-SIS	6	10	14	21	44
	NSIS	4	5	7	11	21.6
	SIS	22.9	117.5	386.5	691.2	941
3	HT-SIS	5	8	16	59.2	383.3
	DC-SIS	7	12	20.5	47.2	192.2
	NSIS	4	6	9	15	54
	SIS	32.9	184.5	426.5	696	943
4	HT-SIS	4	5	6	11	35
	DC-SIS	5	7	10	13	16
	NSIS	4	5	6	8	13
	SIS	17	79.7	302.5	618	925.1
5	HT-SIS	4	4	4	4	6
	DC-SIS	121	220	326	445	641.1
	NSIS	127.9	338.2	567.5	799	953
	SIS	403.8	692.7	838	923	989
6	HT-SIS	4	4	4	4	8
	DC-SIS	198.1	329	436	548.5	715.7
	NSIS	420.8	701	843	935	989
	SIS	482.1	699	843	928.5	989 2
7	HT-SIS	4	6	10	29	206
	DC-SIS	11	23	66	197.2	567
	NSIS	13	43	154	399.5	820
	SIS	75.9	348	642	845	980

much larger than that of other methods. For models 1 through 3, the results HT-SIS, DC-SIS and NSIS are similar up to the 50-th percentile. At the 75-th percentile, NSIS seems to obtain lower model sizes than the other methods, with DC-SIS and HT-SIS following with somewhat larger sizes. For model 4 HT-SIS, DC-SIS and NSIS perform similarly up to the 75-th percentile. For models 2 through 4, NSIS maintains a small model size at the 95-th percentile while DC-SIS and HT-SIS obtain larger sizes, however NSIS and HT-SIS have a larger 95-th percentile for model 1. For models 5 through 7 the 95-th percentile obtained by HT-SIS is by far the lowest. Although the first three models seem to cause a rapid increase in HT-SIS's model size from the 75-th to the 95-th percentiles, the proportions $\mathcal{P}_s$ of HT-SIS are maintained high, only slightly

Table 1
(continued)

		d = 3000				
Model	test	5%	25%	50%	75%	95%
1	HT-SIS	4	6	11	53.6	652.1
	DC-SIS	5	9	14	21	82.6
	NSIS	4	6	9	18	260.5
	SIS	22.7	291	1116	2200	2838
2	HT-SIS	4	6	12	39	528.1
	DC-SIS	5	10	16	23	86.2
	NSIS	4	5	7	11	32.1
	SIS	24	363	1084	2181	2873
3	HT-SIS	4	8	19	95	920
	DC-SIS	6	14	27	75	553.8
	NSIS	4	6	9	16	137.3
	SIS	92	552	1163	2008	2836.8
4	HT-SIS	4	5	7	12.5	85.4
	DC-SIS	5	7	9	13	18
	NSIS	4	5	6	8	16
	SIS	21	182.5	758	1743	2822.2
5	HT-SIS	4	4	4	4	9
	DC-SIS	363.85	670.5	989	1353.5	1870.4
	NSIS	333.1	957.5	1662	2338	2859
	SIS	1190	2011.8	2491	2765.2	2962.1
6	HT-SIS	4	4	4	4	9
	DC-SIS	575	973	1291	1625	2122
	NSIS	1188	2068	2500	2771	2949
	SIS	1189	2136	2557	2799	2962
7	HT-SIS	4	6	13	58.2	663.1
	DC-SIS	13	50	173.5	587.7	1801.2
	NSIS	16	111	435	1175	2397
	SIS	178	953	1913.5	2560	2936

lower in average than those of DC-SIS and NSIS. This suggests that the high 95-th percentile of $\mathcal{S}$ using HT-SIS for the first three models is due to the fact that one of the important predictors may have been assigned a very low rank 5% of the generated datasets. It is important to notice that for models 5 and 6 SIS, NSIS and DC-SIS fail to identify any of the important predictors in their top ranked probably due to the high frequency of the sine and cosine functions. On the other hand, it can be seen from Tables 2 and 3 that HT-SIS captures their significance at least 99% of the time, keeping an extremely low model size $\mathcal{S}$ at all percentiles. Finally, the proportion of time that the two-peak effect of X_{12} in model 7 is selected by HT-SIS is on average 83.5%, considerably higher than the 31.7% on average achieved by DC-SIS and 24% buy NSIS.

Table 2
Proportions $\mathcal{P}_s$ and $\mathcal{P}_a$. $d = 1000$

		$\lvert\hat{I}\rvert = n/log(n)$					$\lvert\hat{I}\rvert = \#\{T_k \geq \lambda_n\}$				
		$\mathcal{P}_s$				$\mathcal{P}_a$	$\mathcal{P}_s$				$\mathcal{P}_a$
Model	test	X_1	X_2	X_{12}	X_{22}	ALL	X_1	X_2	X_{12}	X_{22}	ALL
1	HT-SIS	.97	.96	.87	.97	.80	.98	.97	.89	.97	.82
	DC-SIS	.97	.98	.98	1	.95	.98	.98	.98	1	.96
	NSIS	1	1	.88	.99	.88	1	1	.89	.99	.88
	SIS	.17	.17	.92	.99	.10	.21	.21	.93	.99	.14
2	HT-SIS	.94	.94	.96	.94	.81	.96	.96	.97	.94	.84
	DC-SIS	.95	.96	.99	.99	.93	.97	.98	.99	.99	.96
	NCSIS	1	1	.98	.99	.97	1	1	.98	.99	.97
	SIS	.14	.16	.99	.99	.11	.17	.19	.99	.99	.13
3	HT-SIS	.90	.85	.92	.92	.67	.93	.87	.94	.93	.72
	DC-SIS	.97	.74	.99	.98	.72	.97	.80	.99	.98	.78
	NSIS	.99	.97	.97	.98	.93	.99	.98	.98	.98	.94
	SIS	.68	.11	.99	.98	.05	.72	.14	.99	.98	.07
4	HT-SIS	1	.99	.95	.99	.95	1	.99	.96	1	.96
	DC-SIS	1	1	1	.99	.99	1	1	1	.99	.99
	NSIS	1	1	.99	1	.99	1	1	.99	1	.99
	SIS	1	1	1	.15	.15	1	1	1	.17	.17
5	HT-SIS	1	1	.99	.99	.99	1	1	.99	.99	.99
	DC-SIS	.20	.21	.12	.10	0	.19	.22	.11	.10	.01
	NSIS	.41	.41	.15	.14	.01	.41	.40	.15	.14	.01
	SIS	.04	.05	.05	.04	0	.03	.05	.04	.04	0
6	HT-SIS	1	.99	1	1	.99	1	.99	1	1	.99
	DC-SIS	.08	.06	.07	.09	0	.08	.07	.08	.09	0
	NSIS	.02	.04	.04	.04	0	.03	.03	.03	.04	0
	SIS	.03	.02	.03	.04	0	.03	.02	.04	.05	0
7	HT-SIS	.98	.96	.86	.98	.79	.98	.97	.86	.98	.80
	DC-SIS	.99	.98	.37	1	.35	.99	.98	.41	1	.39
	NSIS	1	1	.22	.99	.22	1	1	.25	.99	.25
	SIS	.16	.15	.20	1	.01	.16	.16	.23	1	.02

6. Real data application

In this section we apply the proposed screening method to the cardiomyopathy dataset. This dataset has been studied in Segal, Dahlquist, and Conklin (2003) [16], Hall and Miller (2009) [13] and Li, Zhong and Zhu (2012) [15] and is composed of $n = 30$ observations of $d = 6319$ gene expressions in mice. The objective is to identify which genes contribute the most for the overexpression of Ro1, a G protein-coupled receptor. For comparison and visualization purposes, we only display the top 8 ranked predictors using HT-SIS and DC-SIS. Note that if one wishes to keep the size of the superset $|\hat{I}|$ smaller than $n = 30$, any

Table 3

Proportions $\mathcal{P}_s$ and $\mathcal{P}_a$. d = 3000

		$\lvert\hat{I}\rvert = n/log(n)$					$\lvert\hat{I}\rvert = \#\{T_k \geq \lambda_n\}$				
		$\mathcal{P}_s$				$\mathcal{P}_a$	$\mathcal{P}_s$				$\mathcal{P}_a$
Model	test	X_1	X_2	X_{12}	X_{22}	ALL	X_1	X_2	X_{12}	X_{22}	ALL
1	HT-SIS	.95	.93	.85	.94	.73	.97	.95	.88	.95	.78
	DC-SIS	.96	.97	.95	1	.90	.98	.98	.95	1	.92
	NSIS	1	.99	.85	.98	.83	1	.99	.89	.99	.88
	SIS	.12	.13	.88	1	.08	.16	.17	.90	1	.10
2	HT-SIS	.91	.91	.94	.92	.74	.93	.94	.95	.94	.80
	DC-SIS	.91	.92	1	.99	.88	.96	.96	1	.99	.94
	NSIS	.99	.99	.98	.98	.96	.99	.99	.99	.98	.97
	SIS	.09	.11	.99	.99	.07	.13	.15	.99	.99	.11
3	HT-SIS	.91	.80	.92	.89	.63	.92	.85	.93	.91	.70
	DC-SIS	.95	.62	.99	.99	.61	.97	.73	.99	.99	.71
	NSIS	.97	.96	.97	.97	.89	.97	.97	.98	.97	.92
	SIS	.59	.04	.96	.99	.02	.66	.09	.97	.99	.04
4	HT-SIS	1	.99	.92	.99	.90	1	.99	.93	.99	.93
	DC-SIS	1	1	1	.99	.99	1	1	1	.99	.99
	NSIS	1	1	.99	.99	.99	.99	.99	.99	.99	.99
	SIS	.99	1	.99	.09	.09	1	1	.99	.13	.13
5	HT-SIS	.99	.99	.99	.99	.98	1	.99	1	.99	.99
	DC-SIS	.07	.08	.04	.03	0	.10	.12	.06	.05	0
	NSIS	.25	.25	.08	.08	0	.31	.31	.11	.10	0
	SIS	.02	.02	.02	.01	0	.02	.03	.02	.01	0
6	HT-SIS	.99	1	.99	.99	.99	.99	1	.99	.99	.99
	DC-SIS	.02	.01	.02	.02	0	.03	.02	.03	.03	0
	NSIS	.01	.01	.01	.01	0	.01	.02	.01	.02	0
	SIS	.01	.01	.01	.01	0	.01	.01	.01	.01	0
7	HT-SIS	.95	.95	.79	.96	.68	.96	.96	.83	.96	.74
	DC-SIS	.97	.98	.21	1	.2	.99	.98	.28	1	.28
	NSIS	1	1	.12	.99	.12	1	1	.17	.99	.17
	SIS	.12	.11	.10	.99	.01	.15	.14	.15	.99	.02

choice of the number of false positives (less than 30) would correspond to keeping the false positive rate less than 0.5%. Figures 1 and 2 show the scatterplots of Ro1 and expression levels of the 8 most influential genes (left to right and top to bottom) ranked according to HT-SIS and DC-SIS respectively. In order to help visualize the relationships between Ro1 and the predictors, we added to each graph a cubic spline fit curve and the lowess (locally weighted polynomial regression) fit curve.

Note that, according to HT-SIS, the most influential gene is Msa.2400.0, which is ranked seventh with DC-SIS. On the other hand, DC-SIS ranks first gene Msa.2134.0, which is ranked second according to HT-SIS. To compare the

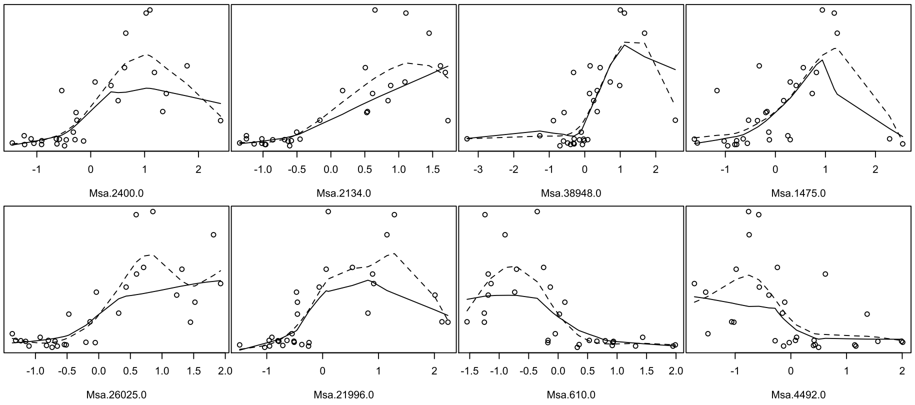

Fig 1. *Scatterplot of Ro1 and the expression of the top 8 genes ranked with HT-SIS and spline (dashed) and lowess (solid) fit curves.*

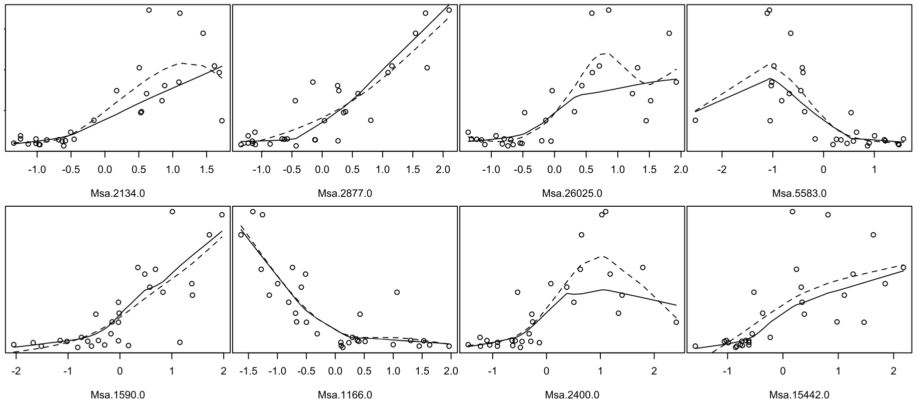

Fig 2. *Scatterplot of Ro1 and the expression of the top 8 genes ranked with DC-SIS and spline (dashed) and lowess (solid) fit curves.*

predictive significance of each method's top ranked gene, we fit a nonparametric model using penalized regression splines with a bandwidth chosen with generalized cross validation. The nonparametric regression of Ro1 on Msa.2134.0, DC-SIS's top ranked gene, yields an adjusted $R^2 = 0.657$ and deviance explained 0.698, much lower values compared to $R^2 = 0.807$ and deviance explained 0.865 for regressing Ro1 on Msa.2400.0, HT-SIS's top ranked gene. This criterium suggests that the most influential gene is in fact Msa.2400.0. Since a fully nonparametric model suffers from the curse of dimensionality, it is unfeasible to fit a nonparametric (or even an additive) model using all the top eight ranked genes with only $n = 30$ observations in this dataset. In that case, for an elementary insight into the results of the screening methods, we look at the fits resulting from a nonparametric additive model with the top 3 ranked genes for each method using package mgcv from the R software (the addition of a fourth predictor is unfeasible due to the lack of degrees of freedom). HT-SIS obtained an adjusted $R^2 = 0.944$ and deviance explained 0.975 while DC-SIS achieves 0.98 and 0.992 for the same measures respectively. Although DC-SIS achieves somewhat better results, it is clear that both methods perform comparably in ranking the most influential genes, with very high deviance explained. Note that the addition of more genes to the additive model would surely increase the R^2 and the explained deviance. Hence, the supersets obtained by HT-SIS and DC-SIS, although slightly different, consist of genes with high predictive significance with respect to Ro1.

7. Discussion

In this paper we propose a screening method based on a test statistic for the hypothesis that a covariate is influential in the prediction of the response variable. The sure independence screening property is demonstrated using a nonparametric heteroscedastic regression model. Simulations suggest that the proposed method performs well even with highly correlated predictors. However, improved versions of screening methods have been widely studied in the literature. The original idea proposed by Fan and Lv (2008) [8] is to first choose a smaller set of predictors with high predictive significance, and then iteratively, choose a subsequent small set of predictors that is significantly related to the residuals obtained from the modeling of the previous set with the response. Following such idea, an iterative HT-SIS can be easily adapted to screening nonparametric models, improving the inclusion of predictors that have little or no marginal predictive significance, but jointly with other predictors yield a significant model. Theoretical aspects of such iterative method need a more detailed appraisal.

A meaningful choice of the threshold parameter is derived and theoretically justified through the control of the false positive rate of the selection. It is interesting to note that the proposed procedure for choosing the threshold parameter is based only on the number of predictors d and the allowed false discovery rate. This fundamentally differs from the ad-hoc choices used in the literature, which are based solely on the sample size n. As observed in the microarray analysis in Section 6, for real situations with ultra-high predictor space and very small sample size, the proposed method for choosing the threshold parameter may suggest a screened superset with size larger than n. Depending on the objective of the screening, a lower false positive rate might be selected in order to keep the size of the screened superset below n. Overall, choosing the number of predictors to retain when performing variable screening is a difficult challenge that still needs further investigation.

Appendix

Throughout the appendix and the proofs herein, the notations C, c, c_1 and c_2 are generic constants, which may take different values at each appearance. Moreover, we use C_σ, C_m and $C_{m\sigma}$, which may take different values at each appearance, to denote a constant that depends on the functions $\sigma_k(\cdot), k=1,\ldots,d$. C_k may be different for instance when depending on different moments of $\sigma_k(\cdot)$.

A.1. Auxiliary lemmas

Lemma 6. *Let $X_1,\ldots,X_n$ be i.i.d. random variables with distribution F_X satisfying condition C7 and let $X_{(1)},\ldots,X_{(n)}$ be the corresponding order statistics. Then*

$$|F_X(X_{(k)}) - F_X(X_{(k-p)})| = O_p\left(\frac{1}{n}\right).$$

Proof. Note that $F_X(X_{(1)}),\ldots,F_X(X_{(n)})$ are order statistics of a Uniform distribution on (0,1), and hence $F_X(X_{(k)}) \sim Beta(k, n+1-k)$. Thus, for any $\epsilon > 0$

$$\begin{aligned} P(n|F_X(X_{(k)}) - F_X(X_{(k-p)})| \geq M) &\leq \frac{nE|F_X(X_{(k)}) - F_X(X_{(k-p)})|}{M^2} \\ &= \frac{nE[F_X(X_{(k)})] - E[F_X(X_{(k-p)})]}{M^2} \\ &= \frac{n(k/(n-1) - (k-p)/(n-1))}{M^2} < \epsilon, \end{aligned}$$

for the $M = 2p/\sqrt{\epsilon}$ and any $n > 2$. This completes the proof of Lemma 6. □

Lemma 7. *For neighboring observations x_{ki} and x_{kj} such that $|\hat{F}_{X_k}(x_{ki}) - \hat{F}_{X_k}(x_{kj})| \leq \frac{p-1}{2n}$ for a constant p, and any Lipschitz continuous function $g(x)$, under condition C7 we have that*

$$|g(x_{kj}) - g(x_{ki})| = O_p\left(\frac{M_g}{c_{f_k} n}\right)$$

uniformly in $i, j = 1,\ldots,n$.

Proof. First note that by the Lipschitz continuity and the Mean Value Theorem, for $|x_{ki}| \leq c_n, |x_{kj}| \leq c_n$ we have

$$\begin{aligned} |g(x_{kj}) - g(x_{ki})| &\leq M_g|x_{kj} - x_{ki}| \leq M_g|F_{X_k}(x_{kj}) - F_{X_k}(x_{ki})|/f_{X_k}(\tilde{x}_{ij}) \\ &\leq M_g|F_{X_k}(x_{kj}) - F_{X_k}(x_{ki})|/f_{X_k}(c_n), \\ &\leq M_g|F_{X_k}(x_{kj}) - F_{X_k}(x_{ki})|/c_{f_k}, \end{aligned}$$

for some Lipschitz constant M_g, where $c_{f_k} = \inf_{x\in\mathcal{X}_k} f_{X_k}(x)$, and $\tilde{x}_{ij}$ is between x_{kj} and x_{ki}. Thus, for x_{ki} and x_{kj} such that $|\hat{F}_{X_k}(x_{ki}) - \hat{F}_{X_k}(x_{kj})| \leq \frac{p-1}{2n}$, we have

$$|g(x_{kj}) - g(x_{ki})| \leq M_g \frac{|F_{X_k}(x_{kj}) - F_{X_k}(x_{ki})|}{c_{f_k}} = O_p\left(\frac{M_g}{c_{f_k} n}\right)$$

where the last equality follows from Lemma 6 and assumption C7. □

A.2. Proofs of lemmas and theorems

Proof of Lemma 1

The subscript k is dropped from the test statistic T_k and the univariate functions $m_k(x)$ and $\sigma_k(x)$ throughout the proof for ease of notation (not to be confused with the multivariate functions $m(\mathbf{x})$ and $\sigma(\mathbf{x})$ in equation (1)). Letting $\xi_i = Y_i - m(X_i)$, we can write

$$T = MST - MSE = \mathbf{Y}_W^T A \mathbf{Y}_W = \boldsymbol{\xi}_W^T A \boldsymbol{\xi}_W + 2\mathbf{m}_W^T A \boldsymbol{\xi}_W + \mathbf{m}_W^T A \mathbf{m}_W, \quad (10)$$

where $\mathbf{Y}_W$ is the vector of $(n-p+1)p$ augmented observations

$$\mathbf{Y}_W = (Y_i, i \in W_1, \dots, Y_i, i \in W_n)^T \quad (11)$$

in the one-way ANOVA, $\boldsymbol{\xi}_W$ and $\mathbf{m}_W$ are defined as in (11) but using ξ_i and $m(X_i)$ instead of Y_i, and the matrix A is defined as in (6). After some algebra, we can write the first term on the right hand side of (10) as

$$\begin{aligned}\boldsymbol{\xi}_W^T A \boldsymbol{\xi}_W \quad &= \quad \frac{(np-1)}{n(n-1)p(p-1)} \sum_{i=1}^n \left[\sum_{j=1}^n \xi_j I(j \in W_i) \right]^2 \\ &\quad - \frac{p}{n(n-1)} \left[\sum_{i=1}^n \xi_i \right]^2 - \frac{p}{n(p-1)} \sum_{i=1}^n \xi_i^2 := T_1 - T_2 - T_3, \quad (12)\end{aligned}$$

and note that we can write

$$\begin{aligned} T_1 &= c_{n1} \left[\frac{1}{n} \sum_{i=1}^n \sum_{j=1}^n \xi_j^2 I(j \in W_i) + \frac{1}{n} \sum_{i=1}^n \sum_{j_1 \neq j_2}^n \xi_{j_1} \xi_{j_2} I(j_1, j_2 \in W_i) \right] \\ &= c_{n1} T_{11}^* + c_{n1} T_{12}^*, \\ T_2 + T_3 &= \frac{p}{n(n-1)} \sum_i \sum_{j \neq i} \xi_i \xi_j + \frac{p(p-1) + p(n-1)}{n(n-1)(p-1)} \sum_{i=1}^n \xi_i^2 \\ &= \frac{p}{n(n-1)} \sum_i \sum_{j \neq i} \xi_i \xi_j + c_{n2} \frac{1}{n} \sum_{i=1}^n \xi_i^2 := pT_2^* + c_{n2} T_3^*, \end{aligned}$$

where $c_{n1} = (np-1)/[(n-1)p(p-1)] \to 1/(p-1)$ and $c_{n2} = [p(p-1) + p(n-1)]/[(n-1)(p-1)] \to p/(p-1)$, which do not influence the asymptotic convergence rates of the test statistic.

Here we follow steps similar to those in Li, Zhong, and Zhu (2012). We first deal with T_3^*. Decompose T_3^* into two parts

$$T_3^* = T_{3a}^* + T_{3b}^* = \frac{1}{n} \sum_{i=1}^n \xi_i^2 I(\xi_i^2 \le M) + \frac{1}{n} \sum_{i=1}^n \xi_i^2 I(\xi_i^2 > M),$$

where M will be specified later. By the Markov Inequality and the fact that ξ_i are i.i.d., for any $\varepsilon > 0$ and $t > 0$

$$\begin{aligned} P(T^*_{3a} - E(T^*_{3a}) \geq \varepsilon) &\leq \exp\{-t\varepsilon\}\exp\{-tE(T^*_{3a})\}E(\exp\{tT^*_{3a}\}) \\ &= \exp\{-t\varepsilon\}\exp\{-tE(T^*_{3a})\}E^n\left(\exp\left\{\frac{t}{n}\xi_1^2 I(\xi_1^2 \leq M)\right\}\right) \\ &= \exp\{-t\varepsilon\}E^n\left(\exp\left\{\frac{t}{n}\left(\xi_1^2 I(\xi_1^2 \leq M) - E(T^*_{13a})\right)\right\}\right) \\ &\leq \exp\{-t\varepsilon + t^2M^2/(8n)\}, \end{aligned}$$

where the last inequality follows from Lemma 5.6.1A in Serfling (1980).

Choosing $t = 4\varepsilon n/M^2$ we have

$$P(T^*_{3a} - E(T^*_{3a}) \geq \varepsilon) \leq \exp\{-4\varepsilon^2 n/M^2 + 16\varepsilon^2 nM^2/M^4 8\} = \exp\{-2\varepsilon^2 n/M^2\},$$

and by the symmetry of T^*_{3a} $P(|T^*_{3a} - E(T^*_{3a})| \geq \varepsilon) \leq 2\exp\{(-2\varepsilon^2 n/M^2\}$.

Now we investigate T^*_{3b}. Note that for any $c_2 > 0$,

$$\begin{aligned} E^2(T^*_{3b}) \leq E(\xi_i^4)P(\xi_i^2 > M) &\leq E(\xi_i^4)E(\exp\{c_2\xi_i^2\})/\exp\{c_2 M\} \\ &= E(\sigma^4(X_i)\epsilon_i^4)E(\exp\{s\sigma^2(X_i)\epsilon_i^2\})\exp\{-c_2 M\}. \end{aligned}$$

In view of assumptions C1-C5 and C8, if we choose $M = cn^\gamma$ for $0 < \gamma < 1/2 - k$, then $E(T^*_{3b}) \leq \varepsilon/2$ when n is sufficiently large. Consequently

$$\begin{aligned} P(|T^*_{3b} - E(T^*_{3b})| > \varepsilon) &\leq P(|T^*_{3b}| > \varepsilon/2) \leq P(\cup\{\xi_i^2 > M\}) \\ &\leq nP(\xi^2 > M) = nP(\exp(c_2\xi^2) > \exp(c_2 M)) \\ &\leq n\exp\{-c_2 M\}E(\exp\{c_2\sigma^2(X_i)\epsilon_i^2\}) \\ &\leq nC_\sigma\exp\{-c_2 M\}, \end{aligned}$$

where C_σ is a constant that depends on the moments of $\sigma_k(\cdot)$, and hence

$$P(|T^*_3 - E(T^*_3)| \geq 2\varepsilon) \leq 2\exp(-2\varepsilon^2 n^{1-2\gamma}) + nC_\sigma\exp(-c_2 n^\gamma).$$

For T^*_2 we write

$$\begin{aligned} T^*_2 = T^*_{2a} + T^*_{2b} &= \frac{1}{n(n-1)}\sum_{i=1}\sum_{j\neq i}\xi_i\xi_j I(|\xi_i\xi_j| \leq M) \\ &\quad + \frac{1}{n(n-1)}\sum_{i=1}\sum_{j\neq i}\xi_i\xi_j I(|\xi_i\xi_j| > M). \end{aligned}$$

Note that $E(T^*_2) = 0$. Since T^*_2 is a (symmetric) U-statistic of second order, using the fact that $P(T^*_{2a} \geq \epsilon) \leq e^{-t\epsilon}E(e^{\sum_{i=1}\sum_{j\neq i}\xi_i\xi_j I(0\leq\xi_i\xi_j\leq M)/(n(n-1))})$, with steps similar to those for T^*_3, for constants $c_1 > 0$ and $c_2 > 0$

$$P(|T^*_2 - E(T^*_2)| \geq 2\varepsilon) \leq 2\exp(-c_1\varepsilon^2 n^{1-2\gamma}) + nC_\sigma\exp(-c_2 n^\gamma),$$

and because all windows W_i are of finite size (p), we have

$$P(|T_{11}^* - E(T_{11}^*)| \geq 2\varepsilon) \leq 2\exp(-c_1\varepsilon^2 n^{1-2\gamma}) + nC_\sigma \exp(-c_2 n^\gamma).$$

Consider now T_{12}^*. Write

$$T_{12}^* \quad = \quad \frac{1}{n}\sum_{i=1}^n A_i + \frac{1}{n}\sum_{i=1}^n B_i$$

where $A_i = \sum_{j_1 \neq j_2}^n \xi_{j_1}\xi_{j_2} I(j_1, j_2 \in W_i) I(|\xi_{j_1}\xi_{j_2}| \leq M)$ and $B_i = \sum_{j_1 \neq j_2}^n \xi_{j_1}\xi_{j_2} I(j_1, j_2 \in W_i) I(|\xi_{j_1}\xi_{j_2}| > M)$. Define

$$\begin{aligned} U_{ni} &= A_{(i-1)(6p)+1} + \ldots + A_{(i-1)(6p)+3p} \\ V_{ni} &= A_{(i-1)(6p)+3p+1} + \ldots + A_{i(6p)}. \end{aligned}$$

Then

$$\frac{1}{n}\sum_{i=1}^n A_i = \frac{1}{n}\sum_{i=1}^{n/(6p)} U_{ni} + \frac{1}{n}\sum_{i=1}^{n/(6p)} V_{ni},$$

where $U_{ni}, i = 1, \ldots, n/(6p)$, are independent and also $V_{ni}, i = 1, \ldots, n/(6p)$, are independent. Thus, by the Markov and Cauchy Schwarz inequalities and the choice of $t = 4\varepsilon n/M^2$ and a constant $c_3 = 1/(12p^3(p-1)^2)$,

$$\begin{aligned} P\left(\frac{1}{n}\sum_{i=1}^n (A_i - E(A_i)) \geq \frac{\varepsilon}{c_3}\right) \quad &\leq \quad e^{-\frac{t\varepsilon}{c_3}} E\big(e^{\frac{t}{n}\sum_{i=1}^n (A_i - E(A_i))}\big) \\ &= e^{-\frac{t\varepsilon}{c_3}} E\big(e^{\frac{t}{n}\sum_{i=1}^{n/(6p)} (U_i - E(U_i))} e^{\frac{t}{n}\sum_{i=1}^{n/(6p)} (V_i - E(V_i))}\big) \\ &\leq e^{-\frac{t\varepsilon}{c_3}} \sqrt{E\big(e^{\frac{2t}{n}\sum_{i=1}^{n/(6p)} (U_i - E(U_i))}\big) E\big(e^{\frac{2t}{n}\sum_{i=1}^{n/(6p)} (V_i - E(V_i))}\big)} \\ &= e^{-\frac{t\varepsilon}{c_3}} \sqrt{E^{n/(6p)}\big(e^{\frac{2t}{n}(U_1 - E(U_1))}\big) E^{n/(6p)}\big(e^{\frac{2t}{n}(V_1 - E(V_1))}\big)} \\ &= e^{-\frac{t\varepsilon}{c_3}} \exp\left\{\frac{n}{6p}\frac{4t^2}{n^2}(3p)^2 p^2 (p-1)^2 M^2/8\right\} \\ &= \exp\left\{-\frac{t\varepsilon}{c_3} + \frac{t^2}{8c_3 n}M^2\right\} = \exp\left\{-\frac{4\varepsilon^2 n}{M^2 c_3} + \frac{2\varepsilon^2 n}{M^2 c_3}\right\}. \end{aligned}$$

Using steps similar to those for T_{3b}^*, under assumptions C1-C5 and C8, with the choice of $M = cn^\gamma$, for a constant $c_2 > 0$

$$P\left(\left|\frac{1}{n}\sum_{i=1}^n B_i - E(B_i)\right| > \varepsilon\right) \leq nC_\sigma \exp(-c_2 M),$$

and hence

$$P(|T_{12}^* - E(T_{12}^*)| \geq \varepsilon) \leq 2\exp\left\{-\frac{2\varepsilon^2 n}{M^2 c_3}\right\} + nC_\sigma \exp(-c_2 M),$$

so that

$$\begin{aligned}&P(|T_1 - T_2 - T_3 - E(T_1 - T_2 - T_3)| \geq \varepsilon)\\ &\leq P(|T_1 - E(T_1)| \geq \varepsilon) + P(|T_2 - E(T_2)| \geq \varepsilon) + P(|T_3 - E(T_3)| \geq \varepsilon)\\ &= O(\exp(-c_1\varepsilon^2 n^{1-2\gamma}) + nC_\sigma \exp(-c_2 n^\gamma)).\end{aligned}$$

Using similar steps, it is easy to show that the second and last terms on the right hand side of (10) have the same convergence rates, that is

$$P(|\mathbf{m}_W^T A\boldsymbol{\xi}_W - E(\mathbf{m}_W^T A\boldsymbol{\xi}_W)| \geq \varepsilon) \leq O(\exp(-c_1\varepsilon^2 n^{1-2\gamma}) + nC_{m\sigma}\exp(-c_2 n^\gamma)),$$

and

$$P(|\mathbf{m}_W^T A\mathbf{m}_W - E(\mathbf{m}_W^T A\mathbf{m}_W)| \geq \varepsilon) \leq O(\exp(-c_1\varepsilon^2 n^{1-2\gamma}) + nC_m\exp(-c_2 n^\gamma)).$$

Let $\varepsilon = cn^{-\alpha}$, where $0 < \alpha < 1/2 - \gamma$. Thus

$$\begin{aligned}P(\max_{k=1,\ldots,d}|T_k - E(T_k)| \geq cn^{-\alpha}) \quad &\leq \quad d\max_{k=1,\ldots,d} P(|T_k - E(T_k)| \geq cn^{-\alpha})\\ &\leq O(d[\exp(-c_1 n^{1-2(\gamma+\alpha)}) + nC_{m\sigma}\exp(-c_2 n^\gamma)]).\end{aligned}$$

Proof of Lemma 2

Using Lemma 7, we have

$$\begin{aligned}E(\hat{v}_k|X_k = x_k) \quad &= \quad E\left[\frac{1}{4(n-3)}\sum_{j=2}^{n-2}(Y_j - Y_{j-1})^2(Y_{j+2} - Y_{j+1})^2|\mathbf{X}_k = \mathbf{x}_k\right]\\ &= \quad \frac{1}{4(n-3)}\sum_{j=2}^{n-2} E\Big[(\sigma_k^2(x_{kj}) + \sigma_k^2(x_{k(j-1)})) \times\\ &\quad \times(\sigma_k^2(x_{k(j+2)}) + \sigma_k^2(x_{k(j+1)}))|\mathbf{X}_k = \mathbf{x}_k\Big]\\ &= \frac{1}{n-3}\sum_{j=2}^{n-2} E[\sigma_k^4(x_{kj})|\mathbf{X}_k = \mathbf{x}_k] + O_p\left(\frac{C_{\sigma_k}}{c_{f_k} n}\right),\end{aligned}$$

where C_{σ_k} is the Lipschitz constant for $\sigma_k(\cdot)$, and $c_{f_k} = \inf_{x\in\mathcal{X}_k} f_k(x)$. Taking the expected value with respect to X_k completes the proof of Lemma 2, since the expected value of the $O_p(\cdot)$ term is $O\left(\frac{C_{\sigma_k}}{c_{f_k} n}\right)$ by steps similar to those in Lemma 6.

Proof of Lemma 3

First note that for any $\epsilon > 0$

$$P(|\sqrt{\hat{v}_k} - \sqrt{v_k}| \geq \epsilon) = P\left(\frac{|\hat{v}_k - v_k|}{|\sqrt{\hat{v}_k} + \sqrt{v_k}|} \geq \epsilon\right) \leq P\left(|\hat{v}_k - v_k| \geq \epsilon L\right),$$

where L is the lower bound for v_k, that is $v_k \geq L$.

Let $A = \{(Y_j - Y_{j-1})^2(Y_{j+2} - Y_{j+1})^2 \leq M\}$ and write

$$\begin{aligned} \hat{v}_k &= \frac{2p(2p-1)}{3(p-1)}\frac{1}{4(n-3)}\sum_{j=2}^{n-2}(Y_j - Y_{j-1})^2(Y_{j+2} - Y_{j+1})^2 \\ &= \frac{2p(2p-1)}{3(p-1)}\frac{1}{4(n-3)}\sum_{j=2}^{n-2}(Y_j - Y_{j-1})^2(Y_{j+2} - Y_{j+1})^2 I(A) \\ &+ \frac{2p(2p-1)}{3(p-1)}\frac{1}{4(n-3)}\sum_{j=2}^{n-2}(Y_j - Y_{j-1})^2(Y_{j+2} - Y_{j+1})^2 I(A^c) := \hat{v}_{k1} + \hat{v}_{k2}. \end{aligned}$$

Let $v_k = v_{k1} + v_{k2}$, where v_{k1} and v_{k2} are the decomposition of v_k corresponding to the decomposition $\hat{v}_k = \hat{v}_{k1} + \hat{v}_{k2}$. Using steps similar to those for term T_{12} in the proof of Lemma 1, one can show that, for constants $c_1 > 0$ and $c_2 > 0$, there exists a constant C_σ such that

$$\begin{aligned} &P(|\hat{v}_{k1} - v_{k1}| \geq \epsilon) \leq 2\exp\left\{-c_1\epsilon^2 n/M^2\right\} \text{ and} \\ &P(|\hat{v}_{k2} - v_{k2}| > \varepsilon) \leq nC_\sigma \exp\{-c_2 M\}, \end{aligned}$$

so that

$$P(|\hat{v}_k - v_k)| \geq 2\varepsilon) \leq 2\exp(-c_1\varepsilon^2 n^{1-2\gamma}) + nC_\sigma \exp(-c_2 n^\gamma).$$

Let $\varepsilon = cn^{-\alpha}$, where $0 < \alpha < 1/2 - \gamma$. Thus

$$\begin{aligned} P(\max_{k=1,\ldots,d}|\hat{v}_k - v_k| \geq cn^{-\alpha}) &\leq d\max_{k=1,\ldots,d}P(|\hat{v}_k - v_k| \geq cn^{-\alpha}) \\ &\leq O(d[\exp(-c_1 n^{1-2(\gamma+\alpha)}) + nC_\sigma\exp(-c_2 n^\gamma)]). \end{aligned}$$

This completes the proof of Lemma 3.

Proof of Theorem 2

By Lemma 1 we have $P(|T_k - E(T_k)| \geq cn^{-\alpha}) \leq O(\exp(-c_1 n^{1-2(\gamma+\alpha)}) + nC_\sigma\exp(-c_2 n^\gamma))$, and by Lemma 3, we have $P(|\sqrt{\hat{v}}_k - \sqrt{v}_k| \geq cn^{-\alpha}) \leq O(\exp(-c_1 n^{1-2(\gamma+\alpha)}) + nC_\sigma\exp(-c_2 n^\gamma))$. Hence the convergence rate of $T_k/\sqrt{\hat{v}}_k - E(T_k)/\sqrt{v}_k$ has the same form. Using condition C8 and Lemma 3.0.9 in Zambom and Akritas (2014) we have, for any $\varepsilon' > 0$, taking $\varepsilon = \varepsilon'/d_0$,

$$\begin{aligned} P(I_0 \subseteq \hat{I}) \geq\ & P(\max_{k\in I_0}|T_k/\sqrt{\hat{v}}_k - \mathrm{Var}(m_k(x))/\sqrt{v}_k| \leq cn^{-\alpha}) \\ &= 1 - P(\max_{k\in I_0}|T_k/\sqrt{\hat{v}}_k - \mathrm{Var}(m_k(x))/\sqrt{v}_k| \geq cn^{-\alpha}) \\ &= 1 - d_0 P(|T_k/\sqrt{\hat{v}}_k - \mathrm{Var}(m_k(x))/\sqrt{v}_k| \geq cn^{-\alpha}) \\ &= 1 - d_0 P(|T_k/\sqrt{\hat{v}}_k - E(T_k)/\sqrt{v}_k + O_p(n^{-1/2})| \geq cn^{-\alpha}) \end{aligned}$$

$$
\begin{aligned}
&\geq 1 - d_0 P(|T_k/\sqrt{\hat{v}_k} - E(T_k)/\sqrt{v_k}| \geq cn^{-\alpha} - O_p(n^{-1/2})) \\
&= 1 - d_0 P(\{|T_k/\sqrt{\hat{v}_k} - E(T_k)/\sqrt{v_k}| \geq cn^{-\alpha} - O_p(n^{-1/2})\} \cap \\
&\qquad\qquad \cap \{O_p(n^{-1/2}) \leq cn^{-1/2}\}) \\
&\qquad -d_0 P(O_p(n^{-1/2}) > cn^{-1/2}) \\
&\geq 1 - O\left(d_0\left[\exp\left(-c_1 n^{1-2(\gamma+\alpha)}\right) + nC_{m\sigma}\exp(-c_2 n^{\gamma})\right]\right) - \varepsilon,
\end{aligned}
$$

where d_0 is the cardinality of I_0, and the last inequality follows from Lemma 1 and the definition of $O_p(n^{-1/2})$ for a constant c.

Proof of Lemma 4

We omit the proof of this Lemma, as it follows using arguments similar to those in Wang, Akritas and Van Keilegom (2008).

Proof of Lemma 5

Assume without loss of generality that the constant C_k in (2) is equal to 0. Note that

$$
\sqrt{n}\mathbf{Y}_{W_k}^T A_d \mathbf{Y}_{W_k} \quad = \quad \frac{\sqrt{n}}{n(p-1)} \sum_{i=1}^{n} \sum_{j_1 \neq j_2}^{n} Y_{j_1} Y_{j_2} I(j_1, j_2 \in W_i) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} D_i,
$$

where $D_i = \sum_{j_1 \neq j_2} Y_{j_1} Y_{j_2} I(j_1, j_2 \in W_i)/(p-1)$. Since D_i are dependent (on only a few other D_i), we will make use of the block Markov techinique to show normality of the test statistic. Write

$$
\begin{aligned}
E_{ni} &= D_{(i-1)(n^\beta+3p)+1} + \ldots + D_{(i-1)(n^\beta+3p)+n^\beta} \\
F_{ni} &= D_{(i-1)(n^\beta+3p)+n^\beta+1} + \ldots + D_{i(n^\beta+3p)},
\end{aligned}
$$

where $0 < \beta < 1$ is a constant. The choice of beta determines the rate of convergence of the test statistic to the normal distribution and the rate at which the small blocks composed by F_{ni} go to 0. Now we have

$$
\sqrt{n}\mathbf{Y}_{W_k}^T A_d \mathbf{Y}_{W_k} = \frac{1}{\sqrt{n}} \sum_{i=1}^{r_n} E_{ni} + \frac{1}{\sqrt{n}} \sum_{i=1}^{r_n} F_{ni},
$$

where $r_n \sim n/(n^\beta + 3p)$. Note that

$$
\begin{aligned}
P\left(\left|\frac{1}{\sqrt{n}} \sum_{i=1}^{r_n} F_{ni}\right| \geq \varepsilon\right) \leq \sum_{i=1}^{r_n} P\left(|F_{ni}| \geq \varepsilon\sqrt{n} r_n^{-1}\right) &\leq KC_\sigma \varepsilon^{-4} n^{-2} r_n^5 \\
&= O(C_\sigma \varepsilon^{-4} n^{5(1-\beta)-2}),
\end{aligned}
$$

where the last inequality follows from the Markov's inequality and assumptions C1 - C5. Hence

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{r_n} F_i = O_p(C_\sigma \varepsilon^{-4} n^{5(1-\beta)-2}).$$

It is easy to establish the Lyapunov condition for $\sum_{i=1}^{r_n} E_i/\sqrt{n}$ (see Zambom and Akritas 2014). Note that $E(E_i)=0$. Write

$$\begin{aligned}
&\sup_x \left| P\left(\sqrt{n}\mathbf{Y}_{W_k}^T A_d \mathbf{Y}_{W_k}/\sqrt{v_k} \le x\right) - \Phi(x)\right| \\
= &\sup_x \left| P\left(\sum_{i=1}^{r_n} E_{ni}/\sqrt{nv_k} + \sum_{i=1}^{r_n} F_{ni}/\sqrt{nv_k} \le x\right) - \Phi(x)\right| \\
\le &\sup_x \left| P\left(\sum_{i=1}^{r_n} E_{ni}/\sqrt{nv_k} \le x\right) - \Phi(x)\right| \\
&+\sup_x \left| P\left(\sum_{i=1}^{r_n} \frac{E_{ni}}{\sqrt{nv_k}} + \sum_{i=1}^{r_n} \frac{F_{ni}}{\sqrt{nv_k}} \le x\right) - P\left(\sum_{i=1}^{r_n} \frac{E_{ni}}{\sqrt{nv_k}} \le x\right)\right| \quad (13)
\end{aligned}$$

Using the Berry Essen theorem (Berry, 1941), the first term in (13) is bounded by a term of order $O(r_n^{-1/2}E(|E_{ni}|^3)Var(E_{ni})^{-3/2})$. We have

$$\begin{aligned}
Var(E_{ni}) \quad = \quad E(E_{ni}^2) = E\Big[\sum_{i=1}^{n^\beta}\sum_{j=1}^{n^\beta}\sum_{k_1\ne k_2,\ell_1\ne\ell_2} Y_{k_1}Y_{k_2}Y_{\ell_1}Y_{\ell_2} \times \\
\times I(k_1,k_2\in W_i)I(\ell_1,\ell_2\in W_j)\Big],
\end{aligned}$$

which is only different from 0 if $Y_{k_1}Y_{k_2}Y_{\ell_1}Y_{\ell_2}$ consists of two pairs of equal observations. Hence, the order of $Var(E_{ni})$ is $O(n^\beta C_\sigma)$. Using similar steps, and the fact that (Cauchy Schwarz)

$$E\left(|E_{ni}|^3\right) = E\left(|\sum_{i=1}^{n^\beta} D_i|^3\right) \le \sqrt{E\left[\left(\sum_{i=1}^{n^\beta} D_i\right)^4\right] E\left[\left(\sum_{i=1}^{n^\beta} D_i\right)^2\right]}$$

it can be shown that $E(|E_{ni}|^3)$ is of order $O(n^{3\beta/2}C_\sigma)$. Hence, the Berry Essen bound for the first term in (13) is $O(C_\sigma n^{-(1/2)(1-\beta)+3\beta/2-(3/2)\beta}) = O(C_\sigma n^{\beta/2-1/2})$.

For any $\epsilon>0$, the second term in (13) is equal to

$$\begin{aligned}
\sup_x \Bigg| \quad & P\left(\sum_{i=1}^{r_n} E_{ni}/\sqrt{nv_k} + \sum_{i=1}^{r_n} F_{ni}/\sqrt{nv_k} \le x, |\sum_{i=1}^{r_n} F_{ni}/\sqrt{n}| \ge \varepsilon\right) \\
&+P\left(\sum_{i=1}^{r_n} E_{ni}/\sqrt{nv_k} + \sum_{i=1}^{r_n} F_{ni}/\sqrt{nv_k} \le x, |\sum_{i=1}^{r_n} F_{ni}/\sqrt{n}| \le \varepsilon\right)
\end{aligned}$$

$$
\begin{aligned}
& -P\left(\sum_{i=1}^{r_n} E_{ni}/\sqrt{nv_k} \leq x\right)\Bigg| \\
\leq \quad & \sup_x P\left(|\sum_{i=1}^{r_n} F_{ni}/\sqrt{n}| \geq \varepsilon\right) \\
& + \sup_x \left| P\left(\sum_{i=1}^{r_n} E_{ni}/\sqrt{nv_k} \leq x - \varepsilon/\sqrt{v_k}\right) - P\left(\sum_{i=1}^{r_n} E_{ni}/\sqrt{nv_k} \leq x\right)\right| \\
& + \sup_x \left| P\left(\sum_{i=1}^{r_n} E_{ni}/\sqrt{nv_k} \leq x + \varepsilon/\sqrt{v_k}\right) - P\left(\sum_{i=1}^{r_n} E_{ni}/\sqrt{nv_k} \leq x\right)\right|
\end{aligned}
$$

For a choice of $0 < \beta$ large enough say $\beta = 9/10$ and $\varepsilon = n^{-(\beta-3/5)} = n^{-3/10}$, we have convergence of $P(|(1/n)\sum_{i=1}^{r_n} F_{ni}| \geq \epsilon)$ of order $O(C_\sigma n^{4(3/10)+5(1-\beta)-2}$ $= O(C_\sigma n^{-3/10})$.

Proof of Theorem 3

For the proof of this Theorem, we follow Zhao and Li (2012). We have

$$
\begin{aligned}
P(I_0 \subseteq \hat{I}) &= P(\min_{k\in I_0} |T_k/\sqrt{\hat{v}_k}| \geq \lambda_n) = 1 - P(\min_{k\in I_0} |T_k/\sqrt{\hat{v}_k}| < \lambda_n) \\
&\geq 1 - P(\max_{k\in I_0} |T_k/\sqrt{\hat{v}_k} - \mathrm{Var}(m_k(x))/\sqrt{v_k}| \geq 2cn^{-\alpha} - \lambda_n),
\end{aligned}
$$

where the last inequality follows from the fact that $2cn^{-\alpha} - |T_k/\sqrt{\hat{v}_k}| \leq |T_k/\sqrt{\hat{v}_k} - \mathrm{Var}(m_k(x))/\sqrt{v_k}|$, which follows from assumption C8. For any $\lambda_n \leq cn^{-\alpha}$, Theorem 2 holds. For the choice of $\lambda_n = n^{-1/2}\Phi^{-1}(1 - r/d)$, this entails

$$
n^{-1/2}\Phi^{-1}(1 - r/d) \leq cn^{-\alpha} \iff d \leq r(1 - \Phi(cn^{1/2-\alpha}))^{-1}
$$

Using the fact that $1 - \Phi(x) \leq x^{-1}\exp(-x^2/2)$, this inequality is satisfied if $d \leq r\exp\{c^2 n^{1-2\alpha}/2\}$.

Without loss of generality, consider the constant C_k in (2) to be equal to 0. Note that for $k \in I_0^c$, $n^{1/2}T_k/\hat{v}_k = n^{1/2}\mathbf{Y}_{W_k}^T A\mathbf{Y}_{W_k}/\hat{v}_k$ we can use Lemma 4 and Lemma 5, to find

$$
\sup_x |P(n^{1/2}\mathbf{Y}_{W_k}^T A\mathbf{Y}_{W_k}/\sqrt{v_k} \leq x) - \Phi(x)| \leq C_\sigma n^{-3/10},
$$

for a constant C_σ. Then (9) implies that

$$
E\left(\frac{|\hat{I} \cap I_0^c|}{|I_0^c|}\right) \leq \frac{1}{d - d_0}\sum_{k\in I_0^c}(1 - \Phi(\gamma_n) + C_\sigma n^{-3/10}).
$$

The theorem follows if we choose $\gamma_n = n^{-1/2}\Phi^{-1}(1 - r/d)$.

Acknowledgments

This paper was partially supported with FAPESP grant 2012/10808-2, FAEPEX grant 519.292-573/13 and NSF Grant No. DMS-1209059.

References

[1] Benjamini, Y. and Hochberg, Y. (1995). Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society - B*, vol. 57, pp. 289–300. MR1325392

[2] Berry, A. C. (1941). The Accuracy of the Gaussian Approximation to the Sum of Independent Variates. *Transactions of the American Mathematical Society*, 49, 122–136 MR0003498

[3] Bunea, F., Wegkamp M. H. and Auguste, A. (2006). Consistent variable selection in high dimensional regression via multiple testing. *Journal of Statistical Planning and Inference*, vol. 136, pp. 4349–4364.

[4] Candes, E., and Tao, T. (2007). The Dantzig Selector: Statistical Estimation When p is Much Larger Than n (with discussion), *The Annals of Statistics*, vol. 35, pp. 2313–2404. MR2382647

[5] Dvoretzky, A., Kiefer, J. and Wolfowitz, J. (1956), Asymptotic minimax character of the sample distribution function and of the classical multinomial estimator. *Annals of Mathematical Statistics*, vol. 27, pp. 642–669.

[6] Efron, B., Hastie, T., Johnstone, I., and Tibshirani, R. (2004). Least Angle Regression (with discussion). *The Annals of Statistics*, vol. 32, pp. 409–499. MR2060166

[7] Fan, J., and Li, R. (2001). Variable Selection via Nonconcave Penalized Likelihood and Its Oracle Properties, *Journal of the American Statistical Association*, vol. 96, pp. 1348–1360.

[8] Fan, J. and Lv, J. (2008). Sure independence screening for ultrahigh dimensional feature space (with discussion). *Journal of the Royal Statistical Society - B*, vol. 70, pp. 849–911.

[9] Fan, J., Samworth, R., and Wu, Y. (2009). Ultrahigh Dimensional Feature Selection: Beyond the Linear Model. *Journal of Machine Learning Research*, vol. 10, pp. 1829–1853.

[10] Fan, J., and Song, R. (2010). Sure Independence Screening in Generalized Linear Models With NP-Dimensionality. *The Annals of Statistics*, vol. 38, pp. 3567–3604.

[11] Fan, J., Feng, Y., and Song, R. (2011). Nonparametric Independence Screening in Sparse Ultra-High Dimensional Additive Models. *Journal of the American Statistical Association*, vol. 106, pp. 544–557.

[12] Gorst-Rasmussen, A. and Scheike, T. (2012). Independent screening for single-index hazard rate models with ultrahigh dimensional features. *Journal of the Royal Statistical Society: Series B*, vol. 75, pp. 217–245.

[13] Hall, P., and Miller,H. (2009). Using Generalized Correlation to Effect Variable Selection in Very High Dimensional Problems. *Journal of Computational and Graphical Statistics*, vol. 18, pp. 533–550. MR2751640

[14] He, X., Wang, L. and Hong, H. G. (2013). Quantile-adaptive model free variable screening for high-dimensional heterogeneous data. *The Annals of Statistics*, vol. 41, pp. 342–369.

[15] Li, R., Zhong, W. and Zhu, L. (2012). Feature Screening via Distance Correlation Learning. *Journal of the American Statistical Association*, vol. 107, pp. 1129–1139.

[16] Segal, M. R., Dahlquist, K. D., and Conklin, B. R. (2003). Regression Approach for Microarray Data Analysis. *Journal of Computational Biology*, vol. 10, pp. 961–980.

[17] Tibshirani, R. (1996). Regression Shrinkage and Selection via LASSO, *Journal of the Royal Statistical Society, Series B*, vol. 58, 267–288.

[18] Wang, H. (2012). Factor profiled sure independence screening. *Biometrika*, 99, 15–28.

[19] Wang, L., Akritas, M. G. and Keilegom, I. V. (2008). An ANOVA-type Nonparametric Diagnostic Test for Heterocedastic Regression Models. *Journal of Nonparametric Statistics* vol. 00, pp. 1–19.

[20] Zambom, A. Z. and Akritas, M. G. (2014) Nonparametric lack-of-fit testing and consistent variable selection. *Statistica Sinica*, 24, 1837–1858. MR3308665

[21] Zhao, S. D. and Li, Y. (2012). Principled sure independence screening for Cox models with ultra-high-dimensional covariates. *Journal of Multivariate Analysss*, vol. 105, pp. 397–411.

[22] Zhong, W. (2014). Robust sure independence screening for ultrahigh dimensional non-normal data. *Acta Mathematica Sinica, English Series*, vol. 30, pp. 1885–1896.

[23] Zhu, L. P., Li, L., Li, R., and Zhu, L. X. (2011). Model-Free Feature Screening for Ultrahigh Dimensional Data. *Journal of the American Statistical Association* vol. 106, pp. 1464–1475. MR2896849

[24] Zou, H., and Hastie, T. (2005). Regularization and Variable Selection via the Elastic Net, *Journal of the Royal Statistical Society, Series B*, vol. 67, pp. 301–320.