Statistica Sinica **24** (2014), 1837-1858
doi:http://dx.doi.org/10.5705/ss.2013.112

NONPARAMETRIC LACK-OF-FIT TESTING AND CONSISTENT VARIABLE SELECTION

Adriano Zanin Zambom and Michael G. Akritas

State University of Campinas and Penn State University

Abstract: Let $\mathbf{X}$ be a d-dimensional vector of covariates and Y be the response variable. Under the nonparametric model $Y = m(\mathbf{X}) + \sigma(\mathbf{X})\epsilon$ we develop an ANOVA-type test for the null hypothesis that a particular coordinate of $\mathbf{X}$ has no influence on the regression function. The asymptotic distribution of the test statistic, using residuals based on local polynomial regression, is established under the null hypothesis and local alternatives. Simulations suggest that the test outperforms existing procedures in heteroscedastic settings. Using p-values from this test, a variable selection method based on False Discovery Rate corrections is proposed, and proved to be consistent in estimating the set of indices corresponding to the significant covariates. Simulations suggest that, under a sparse model, dimension reduction techniques can help avoid the curse of dimensionality. We also propose a backward elimination version of this procedure, called BEAMS (Backward Elimination ANOVA-type Model Selection), which performs competitively against well-established procedures in linear regression settings, and outperforms them in nonparametric settings. A data set is analyzed.

Key words and phrases: Backward elimination, dimension reduction, local polynomial regression, model checking, multiple testing, nonparametric regression.

1. Introduction

For a response variable Y and a d-dimensional vector of the available covariates $\mathbf{X}$ set $m(\mathbf{X}) = E(Y|\mathbf{X})$. The dual problems of testing for the predictive significance of a particular covariate, and identification of the set of relevant, for prediction purposes, covariates are common in applied research and in methodological investigations. Due to readily available software, these tasks are often performed under the assumption of a linear model, $m(\mathbf{X}) = \mathbf{X}\boldsymbol{\beta}$. Model checking fits naturally in the methodological context of hypothesis testing, while variable selection is typically addressed through minimization of a constrained or penalized objective function, such as Tibshirani's (1996) LASSO, Fan and Li's (2001) SCAD, Efron et al.'s (2004) least angle regression, Zou's (2006) adaptive LASSO, and Candes and Tao's (2007) Dantzig selector.

At a conceptual level, however, the two problems are intimately connected: dropping variable j from the model is equivalent to not rejecting the null hypothesis $H_0^j : \beta_j = 0$. Abramovich et al. (2006) bridged the methodological divide by

showing that application of the false discovery rate (FDR) controlling procedure of Benjamini and Hochberg (1995) on p-values resulting from testing each H_0^j can be translated into minimizing a model selection criterion of the form

$$\sum_{i=1}^{n}\Big(Y_i - \sum_{j\in S}\widehat{\beta}_j^S x_{ij}\Big)^2 + \sigma^2|S|\lambda, \tag{1.1}$$

where S is a subset of $\{1,2,\ldots,d\}$ specifying the model, $\widehat{\beta}_i^S$ denotes the least squares estimator from fitting model S, $|S|$ is the cardinality of the subset S, and the penalty parameter λ depends both on d and $|S|$. This is similar to penalty parameters used in Tibshirani and Knight (1999), Birge and Massart (2001), and Foster and Stine (2004), which also depend on both d and $|S|$, and more flexible than the proposal in Donoho and Johnstone (1994) which uses λ depending only on d, as well as AIC and Mallow's C_p which use constant λ.

Working with orthogonal designs, Abramovich et al. (2006) showed that the global minimum of the penalized least squares (1.1) with the FDR penalty parameter is asymptotically minimax for ℓ^r loss, $0 < r \leq 2$, simultaneously throughout a range of sparsity classes, provided the level q for the FDR is set to $q < 0.5$. Generalizations of this methodology to non-orthogonal designs differ mainly in the generation of the p-values for testing $H_0^j : \beta_j = 0$, and the FDR method employed. Bunea, Wegkamp, and Auguste (2006) use p-values generated from the standardized regression coefficients resulting from fitting the full model and employ Benjamini and Yekutieli's (2001) method for controlling FDR under dependency, while Benjamini and Gavrilov (2009) use p-values from a forward selection procedure where the ith stage p-to-enter is the ith stage constant in the multiple-stage FDR procedure in Benjamini, Krieger, and Yekutieli (2006).

Testing for the significance of covariates and variable selection procedures based on the assumption that the regression function is linear may fail to discern the relevance of covariates whose effect on $m(\mathbf{x})$ is nonlinear; this is demonstrated by the simulations in Section 4. Because of this, procedures for both model checking and variable selection have been developed under more general/flexible models. See, for example Kong and Xia (2007), Li and Liang (2008), Wang and Xia (2008), Huang, Horowitz, and Wei (2010), Storlie et al. (2011), and references therein. However, the methodological approaches for variable selection under these more flexible models have been distinct from those of model checking.

This paper aims at showing that a powerful model checking procedure, in the context of a heteroscedastic nonparametric regression model, can be used to construct a competitive nonparametric variable selection procedure by exploiting the aforementioned conceptual connection between model checking and variable selection. Thus, this paper has two objectives: to develop a procedure for testing

the predictive significance of each one of the d covariates, given all the other covariates are in the model, and to propose a consistent variable selection procedure based on the Benjamini and Yekutieli (2001) FDR procedure applied on the d p-values. Simulations suggest that a backward elimination version of the proposed variable selection procedure often has better performance characteristics. This version, which is recommended in practice, is called *BEAMS* for Backward Elimination ANOVA-type Model Selection.

In Section 2, we formally describe the model, introduce the hypothesis and the statistic for testing the predictive significance of a covariate, introduce a test-based variable selection procedure and a version of it based on backward elimination called BEAMS. Theoretical asymptotic properties of the test statistic under the null and local alternatives are presented in Section 3, along with the consistency of the variable selection method. In Section 4 we present a series of simulation studies where the performances of both proposed model checking and variable selection methods are compared to those of existing ones. Finally a data set is analyzed with the new and existing variable selection procedures, and the results of the different analyses are compared.

2. The Proposed Procedures

2.1. Lack of fit testing

Let Y be the response variable and $\mathbf{X} = (X_1, \ldots, X_d)$ the vector of available covariates. Set $m(\mathbf{X}) = E(Y|\mathbf{X})$ for the regression function and define

$$\zeta = Y - m(\mathbf{X}). \tag{2.1}$$

From its definition it follows that $E(\zeta|\mathbf{X}) = E(\zeta) = E(\zeta|X_j) = 0$ for all $j = 1, \ldots, d$. Setting $\sigma^2(\mathbf{X}) = \mathrm{Var}\,(\zeta|\mathbf{X})$, we have the model

$$Y = m(\mathbf{X}) + \sigma(\mathbf{X})\epsilon, \tag{2.2}$$

where ϵ is the standardized error ζ. Based on a sample $(Y_i, \mathbf{X}_i), i = 1, \ldots, n$, of iid observations from model (2.2), we consider testing the hypothesis that the regression function does not depend on the jth covariate. We set $\mathbf{X} = (\mathbf{X}_1, X_2)$, where $\mathbf{X}_1$ is of dimension $(d-1)$ and X_2 is univariate. Setting $E(Y|\mathbf{X}_1) = m_1(\mathbf{X}_1)$ the hypothesis we consider can be written as

$$H_0 : m(\mathbf{x}_1, x_2) = m_1(\mathbf{x}_1). \tag{2.3}$$

Additional insight for this hypothesis can be gained from the ANOVA decomposition

$$m(\mathbf{X}_1, X_2) = \mu + \tilde{m}_1(\mathbf{X}_1) + \tilde{m}_2(X_2) + \tilde{m}_{12}(\mathbf{X}_1, X_2), \tag{2.4}$$

where, if F_1, F_2 denote the marginal CDFs of $\mathbf{X}_1, X_2$, respectively, $\mu = \int\int m(\mathbf{x}_1, x_2) dF_1(\mathbf{x}_1)\, dF_2(x_2)$, $\tilde{m}_1(\mathbf{x}_1) = \int m(\mathbf{x}_1, x_2) dF_{X_2}(x_2) - \mu$, $\tilde{m}_2(x_2) = \int m(\mathbf{x}_1, x_2) dF_{\mathbf{X}_1}(\mathbf{x}_1) - \mu$, and $\tilde{m}_{12}(\mathbf{x}_1,\, x_2)$ is defined from (2.4) by subtraction. Thus, under (2.3),

$$m_1(\mathbf{x}_1) = \mu + \tilde{m}_1(\mathbf{x}_1), \text{ and } \tilde{m}_2(X_2) = \tilde{m}_{12}(\mathbf{X}_1, X_2) = 0, \tag{2.5}$$

while under the alternative,

$$m_1(\mathbf{x}_1) = \mu + \tilde{m}_1(\mathbf{x}_1) + E[\tilde{m}_2(X_2) + \tilde{m}_{12}(\mathbf{X}_1, X_2)|\mathbf{X}_1 = \mathbf{x}_1]. \tag{2.6}$$

To motivate the test statistic, note that, under the H_0 in (2.3), the null hypothesis residuals,

$$\xi_i = Y_i - m_1(\mathbf{X}_{1i}), \tag{2.7}$$

are the residuals defined in (2.1). Hence, under H_0,

$$E\left(\xi_i|X_{2i}\right) = 0. \tag{2.8}$$

There are several procedures for testing that a covariate has no predictive value for a response, that the conditional expectation of the response given the covariate is constant. Because of (2.8), any such procedure can be applied to test (2.3), treating the ξ_i as the response. Most procedures, however, are developed under homoscedasticity and become quite 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.

A procedure with good power properties against departures from the H_0 in (2.3), and which maintains its level under heteroscedasticity, considers (ξ_i, X_{2i}), $i = 1, \ldots, n$, as data from a one-way ANOVA design with ξ_i being the observation at "level" X_{2i}. An ANOVA-type statistic from Akritas and Papadatos (2004), the one for high-dimensional balanced one-way designs, can then be used; see Wang, Akritas, and Keilegom (2008). These statistics, however, require two or more observations per factor level, and in regression designs we typically have only one response per covariate value. This issue is dealt with through smoothness conditions that make it possible to augment each cell X_{2i} by including the ξ_ℓ's corresponding to covariate values that are nearest to X_{2i} on either side. The precise way of doing this is described below. An additional issue has to do with the fact that the ξ_i's have to be estimated. Let

$$\hat{\xi}_i = Y_i - \hat{m}_1(\mathbf{X}_{1i}) \tag{2.9}$$

denote the estimated null hypothesis residuals.

We conjecture that the asymptotic theory for the proposed ANOVA-type test, which is presented in the next section, remains the same (up to slightly different conditions) for a wide class of nonparametric estimators of $E(Y|\mathbf{x}_1)$, including kernel, local polynomial, spline or other basis approximation estimations, the backfitting estimator (under an additive model), or the estimators by Lafferty and Wasserman (2008) and Bertin and Lecué (2008) that accommodate a large number of covariates under local sparcity. Moreover, it is possible to use

$$\widehat{\tilde{m}}_1(\mathbf{x}_1) = \frac{1}{n}\sum_{i=1}^{n} \widehat{m}(\mathbf{x}_1, X_{2i}),$$

which is a version of an estimator proposed by Newey (1994) and Linton and Nielsen (1995), and further studied in Mammen, Linton, and Nielsen (1999) and Horowitz and Mammen (2004). Under the null hypothesis this also estimates $E(Y|\mathbf{x}_1)$ (see (2.5)), but under the alternative it estimates

$$E(m(\mathbf{x}_1, X_2)) = \mu + \tilde{m}_1(\mathbf{x}_1).$$

Thus, a test using the residuals $\tilde{\xi} = Y - \widehat{\tilde{m}}_1(\mathbf{x}_1)$ may have improved power against non-additive alternatives, since subtracting $\hat{m}_1(\mathbf{X}_{1i})$ inadvertently removes some of the effect of X_2; see (2.6).

The rest of the paper uses the residuals in (2.9) with $\hat{m}_1(\mathbf{X}_{1i})$ obtained by local polynomial regression estimation of m_1, using a bounded $(d-1)$-variate kernel function K of bounded variation and with bounded support, and a symmetric positive definite $(d-1)\times(d-1)$ bandwidth matrix H_n. Letting $K_{H_n}(\mathbf{x}) = |H_n|^{-1}K(H_n^{-1}\mathbf{x})$, the local polynomial regression estimator of order q is

$$\hat{m}_1(\mathbf{X}_i) = \mathbf{e}_1^T \left(\mathbb{X}_{\mathbf{X}_i}^T \mathbb{W}_{\mathbf{X}_i} \mathbb{X}_{\mathbf{X}_i}\right)^{-1} \mathbb{X}_{\mathbf{X}_i}^T \mathbb{W}_{\mathbf{X}_i} \mathbf{Y} = \sum_{j=1}^{n} \tilde{w}(\mathbf{X}_i, \mathbf{X}_j) Y_j,\ i = 1, \ldots, n, \quad (2.10)$$

where

$$\mathbb{X}_{\mathbf{x}} = \begin{pmatrix} 1 & (\mathbf{X}_1 - \mathbf{x})^T & \operatorname{vech}^T\left\{(\mathbf{X}_1 - \mathbf{x})(\mathbf{X}_1 - \mathbf{x})^T\right\} & \ldots \\ \vdots & \vdots & \vdots & \ldots \\ 1 & (\mathbf{X}_n - \mathbf{x})^T & \operatorname{vech}^T\left\{(\mathbf{X}_n - \mathbf{x})(\mathbf{X}_n - \mathbf{x})^T\right\} & \ldots \end{pmatrix}$$

is the $n \times \eta_d$ design matrix, with

$$\eta_d = \sum_{j=0}^{q} \sum_{\substack{k_1=0 \\ k_1+\ldots+k_d=j}}^{j} \cdots \sum_{k_d=0}^{j} 1,$$

"vech" is the half-vectorization operator, and $\mathbb{W}_{\mathbf{x}} = diag\{K_{H_n}(\mathbf{X}_1 - \mathbf{x}), \ldots, K_{H_n}(\mathbf{X}_n - \mathbf{x})\}$.

The requirement of more than one covariate value X_{2i} is dealt with by including additional $\hat{\xi}_j$'s corresponding to covariate values that are nearest to X_{2i} on either side. We consider the $(\hat{\xi}_i, X_{2i})$, $i = 1, \ldots, n$, arranged so that $X_{2i_1} < X_{2i_2}$, whenever $i_1 < i_2$, and define the *augmented* X_{2i}*-cell* to consist of $\hat{\xi}_i$ and the $\hat{\xi}_j$'s corresponding to the $(p-1)/2$ X_{2j}'s on either side of X_{2i}, for p odd. The set of indices j, with the property that $\hat{\xi}_j$ is in the augmented X_{2i}-cell, is given by

$$W_i = \left\{ j : |\hat{F}_{X_2}(X_{2j}) - \hat{F}_{X_2}(X_{2i})| \leq \frac{p-1}{2n} \right\}, \tag{2.11}$$

where $\hat{F}_{X_2}$ is the empirical distribution function of X_2. We treat these augmented cells as the "groups" in a high-dimensional one-way ANOVA design. The main differences from the usual one-way ANOVA design are that the response variables are the estimated residuals which are not independent, and that each response can belong to several groups (this is because the set of indices W_i are not disjoint), causing additional dependence between the groups. Nevertheless, the proposed test statistic is based on the difference of the treatment and error mean sum of squares, which is the typical test statistic in high-dimensional ANOVA:

$$MST - MSE = \frac{p}{n-1} \sum_{i=1}^{n} (\hat{\xi}_{i\cdot} - \hat{\xi}_{\cdot\cdot})^2 - \frac{1}{np-n} \sum_{i=1}^{n} \sum_{j \in W_i}^{p} (\hat{\xi}_j - \hat{\xi}_{i\cdot})^2, \tag{2.12}$$

where $\hat{\xi}_{i\cdot} = (1/p) \sum_{j \in W_i} \hat{\xi}_j$ and $\hat{\xi}_{\cdot\cdot} = (1/np) \sum_{i=1}^{n} \sum_{j \in W_i} \hat{\xi}_j$.

Remark 1. Simulations suggest that the choice of the "group" size p of the augmented high-dimensional ANOVA design does not much influence the performance of the test procedure, 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 the simulations we used $p = 7$. 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.

According to Theorem 2, the asymptotic mean of the test statistic under both additive and general local alternatives is positive. This suggests that null hypothesis should be rejected for "large" values of the test statistic. In particular, if

$$Z = \frac{n^{1/2}(MST - MSE)}{\widehat{\tau}\sqrt{2p(2p-1)/(3(p-1))}} \tag{2.13}$$

denotes the standardized test statistic, where $\hat{\tau}$ is given in (3.1), its p-value is computed from

$$\pi = 1 - \Phi(Z). \tag{2.14}$$

2.2 Test based variable selection

In this section we propose a test-based variable selection method that is shown to be consistent in Section 3. A similar procedure was proposed by Bunea, Wegkamp, and Auguste (2006) in the context of a homoscedastic linear model.

Let $I_d = \{1, \ldots, d\}$ denote the set of indices of the d available predictors and, for any subset $I \subseteq I_d$, let $\mathbf{X}_I$ denote the subset of the vector of covariates with indices in I. Suppose that the true regression function, m, is a function of $d_0 \leq d$ covariates,

$$m(\mathbf{X}) = m(\mathbf{X}_{I_0}),$$

where $I_0 = \{j_1, \ldots, j_{d_0}\}$ is the (unknown) subset of indices corresponding to the d_0 significant covariates and, with an abuse of notation, the number of arguments of the function m is determined from the dimension of the vector it is applied to. Thus, the true underlying model can be written as

$$Y = m(\mathbf{X}_{I_0}) + \sigma(\mathbf{X})\epsilon. \tag{2.15}$$

The objective of the proposed variable selection method is to identify the subset I_0. Thus, we are interested in identifying the set of covariates with predictive significance in a model where heteroscedasticity, as well as other aspects of the conditional distribution of the response, are allowed to depend on all available covariates.

Let $I_0^j = I_0$, if $j \notin I_0$, and $I_0^j = I_0 - \{j\}$, if $j \in I_0$, and set

$$H_0^j : m(\mathbf{x}_{I_0}) = m(\mathbf{x}_{I_0^j}) \tag{2.16}$$

for the null hypothesis that the regression function does not depend on the jth covariate. Let Z_j and $\pi_j = 1 - \Phi(Z_j)$, $j = 1, \ldots, d$, denote the test statistic and p-value for testing H_0^j; see (2.13) and (2.14). Let $H_0^{(j)}$ denote the null hypothesis corresponding to the p-value $\pi_{(j)}$, $j = 1, \ldots, d$, where $\pi_{(1)} \leq \cdots \leq \pi_{(d)}$ denote the ordered p-values. The FDR procedure of Benjamini and Hochberg (1995) and Benjamini and Yekutieli (2001) computes

$$k = \max\left\{ j : \pi_{(j)} \leq \frac{j}{d} \frac{q}{\sum_{l=1}^{d} l^{-1}} \right\}, \tag{2.17}$$

for a choice of the level q, and rejects the hypotheses $H_0^{(j)}$, $j = 1, \ldots, k$. If no such k exists, no hypotheses are rejected. The proposed variable selection method selects the variables with indices corresponding to the k rejected null hypotheses. Thus, I_0 is estimated by the set $\hat{I}$ of indices corresponding to the first k ordered p-values.

2.3. BEAMS

The test-based variable selection method described in Section 2.2 works well when the sample size is large, in fact it is asymptotically consistent for selecting the true predictors (see Section 3.2). However, when the sample size is small or the dimension of the predictor space is large, this method is often outperformed by a backward elimination version of the procedure. (In fact, a backward elimination version of the Bunea, Wegkamp, and Auguste (2006) procedure also improved it. See also Li, Cook, and Nachtsheim (2005) who also used backward elimination, though not based on multiple testing ideas.) Thus *BEAMS*, which stands for Backward Elimination ANOVA-type Model Selection, is the procedure we recommend in practice. Application of the BEAMS procedure consists of the following steps:

1. Obtain p-values, $\pi_1, \ldots, \pi_d$, from testing each of the hypotheses $H_0^j : m(\mathbf{x}) = m(\mathbf{x}_{-j})$, $j = 1, \ldots, d$, where x_{-j} is obtained from $\mathbf{x}$ by omitting the jth coordinate.
2. Compute k as in (2.17). If $k = d$, stop and retain all variables. If $k < d$, update $\mathbf{x}$ by eliminating the covariate corresponding to $\pi_{(d)}$, set $d = d - 1$, and return to Step 1.

Using dimension reduction techniques

Simulations suggest that, under a sparse model, the variable selection procedure and its modification, BEAMS, with local linear regression for generating the p-values, can be applied with a large number of covariates, provided the dimensionality is suitably reduced. Following the seminal paper of Li (1991), a number of dimension reduction (DR) methods have been proposed. Because we are interested in identifying covariates with predictive significance, methods such as Hristache et al. (2001), Xia et al. (2002), and Xia (2008) that target the regression function, are particularly relevant. Essentially, these methods express the conditional mean $m(\mathbf{x}_{I_0})$ as a function of $K \leq d_0$ linear combinations of the coordinates of $\mathbf{x}$,

$$m(\mathbf{x}) = g(\mathbf{B}\mathbf{x}),$$

where $\mathbf{B}$ is a $K \times d$ matrix. Thus, if $K < d$ the effective dimension of the problem is K, not d. Finally, in order to reduce the computational time, all simulations reported in Section 4.2 employed a variable screening method prior to applying a DR method. The variable screening consists of performing the marginal test of Wang, Akritas, and Keilegom (2008) for the significance of each variable, and keeping those variables for which the p-value is less than 0.5. The description of BEAMS, with screening and dimension reduction as it is applied in the simulations of Section 4.2, is as follows.

1. Apply the variable screening procedure described above. Update d to the number of remaining covariates, and $\mathbf{x}$ to the vector of remaining covariates.
2. Use a DR method to estimate $\mathbf{B}$. Let $\widehat{\mathbf{B}}$ denote the estimator.
3. Obtain p-values, $\pi_1, \ldots, \pi_d$, from testing each of the hypotheses $H_0^j : m(\mathbf{x}) = m(\mathbf{x}_{-j})$, $j = 1, \ldots, d$, using residuals formed by a local linear regression estimator on the variables $\widehat{\mathbf{B}}_{-j}\mathbf{x}_{-j}$, where $\widehat{\mathbf{B}}_{-j}$ is the $K \times (d-1)$ matrix obtained by omitting the jth column of $\widehat{\mathbf{B}}$.
4. Compute k as in (2.17). If $k = d$ stop and retain all variables. If $k < d$, set $d = d - 1$, update $\mathbf{x}$ by eliminating the covariate corresponding to $\pi_{(d)}$, update $\widehat{\mathbf{B}}$ by eliminating the column corresponding to the deleted variable, and return to Step 3.

In the simulations of Section 4.2 data were generated so the covariates influenced only the regression function, and not any other aspects of the conditional distribution of the response given the covariates. In such settings, Li's (1991) SIR, which is available in the R package *dr*, may also be used (and we did). Another option is to use the method of Hristache et al. (2001) which is available in the R package *EDR*.

3 Asymptotic Results

3.1. Asymptotic distribution of the test statistic

Consider the following conditions

(a) $E|Y|^\rho < \infty$ for some $\rho > 2$.

(b) The marginal densities $f_{\mathbf{X}_1}$, f_{X_2} of $\mathbf{X}_1$, X_2, respectively, are bounded away from zero.

(c) $f_{\mathbf{X}_1}$ is uniformly continuous and bounded.

(d) The $q+1$ derivatives of $m_1(\mathbf{x}_1)$ exist and are Lipschitz uniformly continuous and bounded.

(e) $\sigma^2(., x_2) := E(\xi^2 | X_2 = x_2)$ is Lipschitz continuous, $\sup_{\mathbf{u}} \sigma^2(\mathbf{u}) < \infty$, and $E(\epsilon_i^4) < \infty$.

We assume that the eigenvalues, λ_i, $i = 1, \ldots, d-1$, of the bandwidth matrix $H_n^{1/2}$ defined in (2.10), converge to zero at the same rate and satisfy

$$(1)\ n\lambda_i^{4(q+1)} \to 0\ i = 1, \ldots, d-1,$$

$$(2)\ \frac{n\lambda_i^{2(d-1)}}{(\log n)^2} \to \infty,\ i = 1, \ldots, d-1,$$

$$(3)\ \frac{n^{1-2/\rho}\lambda_i^{d-1}}{\ln n[\ln n(\ln \ln n)^{1+\delta}]^{2/\rho}} \to \infty,\ i = 1, \ldots, d-1.$$

The proofs of the asymptotic normality results stated in Theorems 1 and 2 are given in the supplementary material available online.

Theorem 1. *If* (a)−(e) *and* (1)−(3) *hold, then, under* H_0 *in* (2.3)*, the asymptotic distribution of the test statistic in* (2.12) *is*

$$n^{1/2}(MST - MSE) \xrightarrow{d} N\left(0, \frac{2p(2p-1)}{3(p-1)}\tau^2\right),$$

where $\tau = \int \left[\sigma^2(., x_2)\right]^2 f_{X_2}(x_2)dx_2$.

An estimate of τ^2 can be obtained by modifying Rice's (1984) estimator as

$$\hat{\tau}^2 = \frac{1}{4(n-3)} \sum_{j=2}^{n-2} (\hat{\xi}_j - \hat{\xi}_{j-1})^2 (\hat{\xi}_{j+2} - \hat{\xi}_{j+1})^2. \tag{3.1}$$

Asymptotics under local alternatives

The local additive alternatives and the general local alternatives are of the form

$$H_1^A : m(\mathbf{x}_1, x_2) = m_1(\mathbf{x}_1) + \delta_n \tilde{m}_2(x_2), \tag{3.2}$$

$$H_1^G : m(\mathbf{x}_1, x_2) = m_1(\mathbf{x}_1) + \delta_n(\tilde{m}_2(x_2) + \tilde{m}_{12}(\mathbf{x}_1, x_2)), \tag{3.3}$$

where the functions $\tilde{m}_2$, $\tilde{m}_{12}$ satisfy $E\left(\tilde{m}_2(X_2)\right) = 0 = E\left(\tilde{m}_{12}(\mathbf{x}_1, X_2)\right)$.

Theorem 2. *Suppose that* $\tilde{m}_2(x)$ *is Lipschitz continuous,* $\tilde{m}_{12}(\mathbf{x}_1, x_2)$ *is Lipschitz continuous on* x_2 *uniformly on* $\mathbf{x}_1$*, the assumptions of Theorem* 1 *hold, and* $\delta_n = n^{-1/4}$.

1. *Under* H_1^A *in* (3.2)*, as* $n \to \infty$,

$$n^{1/2}(MST - MSE) \xrightarrow{d} N\left(p\mathrm{Var}\left(\tilde{m}_2(X_2)\right), \frac{2p(2p-1)}{3(p-1)}\tau^2\right).$$

2. *Under* H_1^G *in* (3.3)*, as* $n \to \infty$,

$$n^{1/2}(MST - MSE) \xrightarrow{d} N\left(pVar(\tilde{m}_2(X_2) + \tilde{m}_{12}(\mathbf{X}_1, X_2)), \frac{2p(2p-1)}{3(p-1)}\tau^2\right).$$

Remark 2. If δ_n goes to 0 any faster than $n^{-1/4}$, then the asymptotic mean of the test statistic under local alternatives shrinks to 0. On the other hand, if δ_n goes to 0 any slower than $n^{-1/4}$, then the asymptotic mean tends to infinity.

3.2. Consistency of the test based variable selection

Let R denote the total number of rejected hypotheses: $R = k$ if k in (2.17) exists, and $R = 0$ otherwise. Let V be the number of falsely rejected hypotheses, and set

$$Q = \begin{cases} \frac{V}{R} & \text{if } R > 0, \\ 0 & \text{otherwise,} \end{cases}$$

for the proportion of falsely rejected hypotheses. By definition, the false discovery rate is $E(Q)$, and Benjamini and Yekutieli (2001) showed that $E(Q) \le q(d - d_0)/d \le q$.

The variable selection procedure, and $\hat{I}$, are called consistent if $P(\hat{I} = I_0) \to 1$. The consistency result presented here allows the significance of the predictors to be diminishing with n, where we quantify the significance of a predictor X_j by

$$C_j = \frac{\operatorname{Var}\left(m(\mathbf{X}_{I_0}) - m(\mathbf{X}_{I_0^j})\right)}{\tau_j \sqrt{2p(2p-1)/(3(p-1))}}, \tag{3.4}$$

where τ_j is defined for the predictor X_j as the τ of Theorem 1. Note that $C_j = 0$ for all $j \notin I_0$. This quantification of significance is justified by Theorem 2. In what follows, we allow each C_j to tend to zero with n, but its dependence on n is suppressed for convenience.

Lemma 1. *Let Z_j and $\pi_j = 1 - \Phi(Z_j)$ be the test statistic and p-value for testing H_0^j, as in* (2.13) *and* (2.14).

(a) *For $j \notin I_0$ and any $\gamma > 0$, we have $P(\pi_j \le \gamma) = \gamma + o(1)$.*

(b) *For $j \in I_0$, let $\gamma_n > 0$, $n \ge 1$. Then if*

$$n^{1/2} C_j \to \infty, \quad \text{and} \quad \gamma_n > \frac{\exp(-nC_j^2/4)}{n^{1/2} C_j}$$

we have $P(\pi_j > \gamma_n) = o(1)$.

Proof. (a) The result follows from Theorem 1 by noting that, for $j \notin I_0$, the null hypothesis H_0^j is true.

(b) From the proof of Theorem 2 we have that the standardized test statistic for testing the significance of X_j has a representation of the form

$$Z_j = Z_j^{H_0^j} + n^{1/2} C_j + o_p(1),$$

where $Z_j^{H_0^j} \xrightarrow{d} N(0,1)$. Thus, for any sequence $a_n \to \infty$, $a_n = o(n^{1/2} C_j)$,

$$\begin{aligned} P(\pi_j > \gamma_n) &= P(1 - \Phi(Z_j^{H_0^j} + n^{1/2} C_j + o_p(1)) > \gamma_n) \\ &\le P(1 - \Phi(Z_j^{H_0^j} + n^{1/2} C_j + o_p(1)) > \gamma_n, Z_j^{H_0^j} + o_p(1) \ge -a_n) \\ &\quad + P(Z_j^{H_0^j} + o_p(1) < -a_n) \\ &\le P(1 - \Phi(n^{1/2} C_j - a_n) > \gamma_n) + o(1) = o(1), \end{aligned}$$

by the choice of γ_n since, using also $a_n < n^{1/2}C_j/4$, it can be shown that $1 - \Phi(n^{1/2}C_j - a_n) \leq \exp(-nC_{jn}^2/4)/n^{1/2}C_{jn}$.

Lemma 2. *Let $\mathcal{E}_n$ be the event where the smallest d_0 p-values are the p-values corresponding to the d_0 significant covariates, with $I_0 = \{j_1, \ldots, j_{d_0}\}$,*

$$\mathcal{E}_n = \big[\{\pi_{(1)}, \ldots, \pi_{(d_0)}\} = \{\pi_{j_1}, \ldots, \pi_{j_{d_0}}\}\big].$$

Then $\lim_{n\to\infty} P(\mathcal{E}_n) = 1$.

Proof. Let γ be any number between 0 and 1, and write

$$\begin{aligned}
P(\mathcal{E}_n^c) &\leq \sum_{j\in I_0}\sum_{i\notin I_0} P(\pi_i < \pi_j) \\
&= \sum_{j\in I_0}\sum_{i\notin I_0} \big[P([\pi_i < \pi_j] \cap [\pi_i \leq \gamma]) + P([\pi_i < \pi_j] \cap [\pi_i > \gamma])\big] \\
&\leq \sum_{j\in I_0}\sum_{i\notin I_0} [P(\pi_i \leq \gamma) + P(\pi_j > \gamma)] \\
&\leq \sum_{j\in I_0}\sum_{i\notin I_0} [\gamma + o(1)] \quad \text{(by Lemma 1)} \\
&= d_0(d - d_0)\gamma + o(1).
\end{aligned}$$

Since γ is arbitrary, this shows that $\lim_{n\to\infty} P(\mathcal{E}_n^c) = 0$, completing the proof.

Theorem 3. *With C_j at* (3.4)*, and q the chosen bound of FDR (see* (2.17)*), assume that $n^{1/4}C_j \to \infty$ and $q \to 0$, as $n \to \infty$, in such a way that*

$$q > \frac{d}{d_0}\left(\sum_{l=1}^{d} l^{-1}\right)\frac{\exp(-nC_j^2/4)}{n^{1/2}C_j}.$$

Then, $\lim_{n\to\infty} P(\hat{I} = I_0) = 1$.

Proof. If the estimator $\hat{I}$ is equal to the set I_0, we have exactly d_0 rejections ($R = d_0$) with none of them being erroneous ($V = 0$). Therefore, consistency of $\hat{I}$ is verified by proving

$$P(\hat{I} = I_0) = P(R = d_0, V = 0) \to 1, \text{ as } n \to \infty. \tag{3.5}$$

This follows by showing that both $P(R \neq d_0)$ and $P(V \geq 1)$ are asymptotically negligible. By Lemma 2.1 in Bunea, Wegkamp, and Auguste (2006), we have that

$$P(V \geq 1) \leq P(R \neq d_0) + \frac{d_0(d - d_0)}{d}q.$$

Thus, in order to show consistency of $\hat{I}$ we need only show that $P(R \neq d_0) \to 0$. Following Bunea, Wegkamp, and Auguste (2006), let $q_d = q/\sum_{l=1}^{d} l^{-1}$ and write

$$P(\{R \neq d_0\}) \leq P\left(\pi_{(d_0)} > q_d \frac{d_0}{d}\right) + \sum_{j=d_0+1}^{d} P\left(\pi_{(j)} \leq q_d \frac{j}{d}\right), \tag{3.6}$$

where the first term on the right hand side bounds the probability of $\{R < d_0\}$, and the second term bounds the probability of $\{R > d_0\}$. With the $\mathcal{E}_n$ of Lemma 2, the first term in the right hand side of (3.6) is

$$\begin{aligned} &P\left(\pi_{(d_0)} > q_d \frac{d_0}{d} \cap \mathcal{E}_n\right) + P\left(\pi_{(d_0)} > q_d \frac{d_0}{d} \cap \mathcal{E}_n^c\right) \\ &\quad \leq d_0 \max_{j \in I_0} P\left(\pi_j \geq q_d \frac{d_0}{d}\right) + P\left(\mathcal{E}_n^c\right) = o(1), \end{aligned}$$

by Lemmas 1 and 2. For the second term we have

$$\begin{aligned} &\sum_{j=d_0+1}^{d} P\left(\pi_{(j)} \leq q_d \frac{j}{d} \cap \mathcal{E}_n\right) + \sum_{j=d_0+1}^{d} P\left(\pi_{(j)} \leq q_d \frac{j}{d} \cap \mathcal{E}_n^c\right) \\ &\quad \leq \sum_{j \notin I_0} P(\pi_j \leq q_d) + (d - d_0) P\left(\mathcal{E}_n^c\right) \leq (d - d_0) q_d + (d - d_0) P\left(\mathcal{E}_n^c\right) = o(1), \end{aligned}$$

by Lemmas 1 and 2.

4. Simulation Studies

4.1. Model checking procedures

Literature review

Let $\mathbf{X} = (\mathbf{X}_1, \mathbf{X}_2)$ be the vector of d available predictors, with $\mathbf{X}_1$ being d_1-dimensional. The problem of assessing the usefulness of $\mathbf{X}_2$ has been approached from different angles by many authors. The literature is extensive, so only a brief summary is given. For additional references see Hart (1997) and Racine, Hart, and Li (2006).

One class of procedures is based on the idea that the null hypothesis residuals, $\xi = Y - m_1(\mathbf{X}_1)$, satisfy $E(\xi|\mathbf{X}) = 0$ under H_0 and $E(\xi|\mathbf{X}) = m(\mathbf{X}) - m_1(\mathbf{X}_1)$ under the alternative. Thus, $E(\xi E(\xi|\mathbf{X})|\mathbf{X}) = (m(\mathbf{X}) - m_1(\mathbf{X}_1))^2$ under the alternative and zero under the null. Using this idea, Fan and Li (1996) propose a test statistic based on estimating $E[\xi f_1(\mathbf{X}_1) E(\xi f_1(\mathbf{X}_1)|\mathbf{X}) f(\mathbf{X})]$, which equals $E[(m(\mathbf{X}) - m_1(\mathbf{X}_1))^2 f_1(\mathbf{X})^2 f(\mathbf{X})]$ under the alternative and zero under the null. Their test statistic is

$$\frac{1}{n} \sum_i [\tilde{\xi}_i \tilde{f}_1(\mathbf{X}_{1i})] \Big[\frac{1}{(n-1) h_n^d} \sum_{j \neq i} [\tilde{\xi}_j \tilde{f}_1(\mathbf{X}_{1j})] K\Big(\frac{\mathbf{X}_i - \mathbf{X}_j}{h_n} \Big) \Big],$$

where $\tilde{f}_1$ is the estimated density of $\mathbf{X}_1$, $\tilde{\xi}_i$ is the estimated residuals under the null hypothesis, and K is a kernel function. Fan and Li (1996) show that their test statistic is asymptotically normal under H_0. Lavergne and Vuong (2000) propose a test statistic based on different estimator of the same quantity as Fan and Li (1996), as

$$\frac{(n-4)!}{n!}\sum_a (Y_i - Y_k)(Y_j - Y_l) L_n\left(\frac{\mathbf{X}_{1i}-\mathbf{X}_{1k}}{g_n}\right) L_n\left(\frac{\mathbf{X}_{1j}-\mathbf{X}_{1l}}{g_n}\right) K_n\left(\frac{\mathbf{X}_i-\mathbf{X}_j}{h_n}\right),$$

where $\sum_a$ is the sum over all permutations of four distinct elements chosen from n, $L_n = g_n^{-d_1} L$ for a kernel L on $\mathbf{R}^{d_1}$, and $K_n = h_n^{-d} K$ for a kernel K on $\mathbf{R}^d$. Lavergne and Vuong (2000) show that their test statistic is also asymptotically normal under H_0.

A related class of procedures is based on direct estimation of $E[(m(\mathbf{X}) - m_1(\mathbf{X}_1))^2 W(\mathbf{X})]$, for some weight function W. See, for example, Aït-Sahalia, Bickel, and Stoker (2001). The use of such a test statistics is complicated by the need to correct for bias. See also the bootstrap-based procedure of Delgado and Manteiga (2001). Because of the computer intensive nature of bootstrap-based procedures, these are not included in our comparisons.

An additional class of test procedures uses alternatives based on Stone's (1985) additive model. We consider the procedure proposed by Fan and Jiang (2005). This is based on Fan, Zhang, and Zhang's (2001) Generalized Likelihood Ratio Test (GLR), using a local polynomial approximation and the backfitting algorithm for estimating the additive components.

Numerical comparison

In this section we compare the proposed ANOVA-type statistic to the statistics proposed by Lavergne and Vuong (2000) (LV), Fan and Li (1996) (FL), Fan and Jiang (2005), (GLR), and the classical F-test for linear regression.

The data was generated according to (also used in Lavergne and Vuong (2000))

$$\text{Model } j: Y = -X_1 + X_1^3 + f_j(X_2) + \epsilon,\ j = 0, 1, 2, 3, 4, 5, 6, \tag{4.1}$$

where X_1, X_2 are iid N$(0, 1)$ and $\epsilon \sim N(0, 4)$. Here, $f_0(x) = 0$, which corresponds to the null hypothesis $H_0 : m(x_1, x_2) = m(x_1)$; $f_1(X_2) = 0.5X_2$, $f_2(X_2) = X_2$ and $f_3(X_2) = 2X_2$ give three linear alternatives, and $f_4(X_2) = \sin(2\pi X_2)$, $f_5(X_2) = \sin(\pi X_2)$, and $f_6(X_2) = \sin(2/3\pi X_2)$ give three non-linear alternatives. We used $p = 7$ throughout, since after a random permutation of the response in all cases, the test showed accurate levels for p between 5 and 11. For the estimation of m_1 in the proposed procedure, the Nadaraya-Watson kernel regression estimator was used, with a uniform kernel on $(-0.5, 0.5)$, and the bandwidth was selected

Table 1. Rejection rates under H_0, linear and non-linear alternatives

			linear			sine		
n	test	H_0	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6
100	LV	0.041	0.098	0.482	0.991	0.182	0.266	0.319
	FL	0.021	0.051	0.271	0.970	0.126	0.168	0.187
	GLRT	0.044	0.365	0.951	1	0.123	0.497	0.645
	F-test	0.051	0.695	0.997	1	0.046	0.055	0.222
	ANOVA	0.056	0.244	0.780	0.999	0.432	0.527	0.551
200	LV	0.054	0.208	0.875	1	0.386	0.540	0.678
	FL	0.025	0.083	0.695	1	0.289	0.395	0.471
	GLRT	0.036	0.656	1	1	0.188	0.877	0.936
	F-test	0.052	0.931	1	1	0.051	0.053	0.340
	ANOVA	0.055	0.374	0.95	0.999	0.73	0.778	0.788

through leave-one-out cross validation. The rejection rates shown in Table 1 for LV, FL, and F-tests are taken from the simulation results reported in the LV paper (based on 2,000 runs). In each simulation setting, the LV paper reports several rejection rates for the LV and FL tests, each corresponding to different values of smoothing parameters. The rejection rates reported in Table 1 are the most accurate alpha level achieved over all constants, and the best power achieved overall constants for each alternative. For comparison purposes, the rejection rates for the ANOVA-type tests and the GLR test are also based on 2,000 simulation runs.

As expected, the F test achieved the best results for the three linear alternatives and the worse results for the three non-linear alternatives. The GLR test had higher power than the ANOVA-type tests against linear alternatives (which is partly explained by the fact it is based on normal likelihood), but was much less powerful against the first of the non-linear alternatives. As the non-linearity decreases (Model 5 and Model 6) the power of the GLR test improved.

The GLR test is designed for additive models, the simulation setting of Table 1. Under non-additive alternatives it can perform poorly, as indicated by the simulations reported in Table 2. These simulations used sample size $n = 200$ with data generated from the model $Y = X_1^{X_2}(1+\theta X_3)+X_2^{(1+\theta X_3)}/X_2+\epsilon$, where $\epsilon \sim N(0, 0.1)$, and X_1, X_2, X_3 are i.i.d. $U(0.5, 2.5)$. The hypothesis tested was $m(X_1, X_2, X_3) = m_1(X_1, X_2)$. The residuals for the ANOVA-type test in Table 2 were based on a Nadaraya-Watson fit with kernel the uniform on $(-0.5, 0.5) \times (-0.5, 0.5)$ and the common bandwidth selected through leave-one-out cross validation.

The GLR test does not maintain its level under heteroscedasticity. In simulations, reported in Table 3, under the additive but heteroscedastic model $Y = X_1^2 + \theta\cos(\pi X_2) + X_2\epsilon$, X_1, X_2 i.i.d. $\mathrm{N}(0, 1)$, $\epsilon \sim \mathrm{N}(0, 0.5)$, using sample size $n = 200$,

Table 2. Rejection rates for non-additive models

	θ				
test	0	0.02	0.04	0.06	0.08
ANOVA	0.052	0.176	0.609	0.940	0.994
GLRT	0.048	0.082	0.110	0.189	0.304

Table 3. Rejection rates for heteroscedastic models

	θ				
test	0	0.025	0.05	0.1	0.2
ANOVA	0.053	0.067	0.124	0.485	0.998
GLRT	0.465	0.511	0.624	0.908	1

the GLR test was very liberal while the ANOVA-type test maintained an accurate level.

4.2. Variable selection procedures

In this section we compare the proposed variable selection procedure (ANOVA based V.S.) and BEAMS with LASSO, SCAD, adaptive LASSO, Lin and Zhang (2006)'s COSSO, Chen, Zou, and Cook (2010)'s CISE, the FDR-based variable selection method proposed by Bunea, Wegkamp, and Auguste (2006) (BWA), and a version of the BWA procedure which uses backward elimination (BWA+BE). The simulations used sample sizes of $n = 40$ and $n = 110$. The parameter q for BEAMS, BWA and BWA+BE was set to 0.07, so FDR is below 0.056 in Table 4, and below 0.045 in Table 5. The comparison criterion is the mean number of correctly and incorrectly excluded variables. All comparisons are based on 2,000 simulated data sets.

For LASSO we found that the R code in `http://cran.r-project.org/web/packages/glmnet/index.html`, with the lambda.lse option for selecting lambda, gave the best results; for adaptive LASSO we used the R code from `http://www4.stat.ncsu.edu/~boos/var.select/lasso.adaptive.html`; for SCAD we used the function scadglm of the package SIS in R; for COSSO we used the R package "cosso" (`http://cran.r-project.org/web/packages/cosso/index.html`); and for CISE we used the matlab function available in `http://users.stat.umn.edu/~chen0982/`.

In Table 4, data sets of size $n = 110$ were generated from the linear model $Y = \boldsymbol{\beta}^T\mathbf{X} + \epsilon$, where $\epsilon \sim N(0, 3^2)$, the dimension of $\mathbf{X}$ is $d = 25$, and

$$\boldsymbol{\beta}^T = (3, 1.5, 0, 0, 2, 0, 2, 0, 0, 0, 0, 0, 0, 0, 0, 0, 3, 0, 0, 0, 0, 0, 0, 0, 0).$$

The covariates were generated from a multivariate normal with marginal means zero and covariances as shown in the table. It is seen the ANOVA based variable

Table 4. Comparisons using a linear model: $d = 25$, $n = 110$

	$\Sigma = I$		$\Sigma = (0.5^{\lvert i-j\rvert})$	
test	correct	incorrect	correct	incorrect
SCAD	19.48	0.026	19.37	0.023
LASSO	18.29	0.005	18.28	0.004
Adaptive LASSO	19.28	0.005	19.26	0.025
BWA	19.99	1.02	19.97	1.41
BWA+BE	19.55	0.001	19.49	0.041
COSSO	18.28	3.85	18.53	3.91
CISE	19.93	0.144	19.95	0.202
ANOVA based V.S.	19.66	0.948	19.78	2.52
BEAMS	19.46	0.630	19.30	0.440

Table 5. Comparisons using nonlinear models: $d = 8$, $n = 40$

	g_1		g_2	
test	correct	incorrect	correct	incorrect
SCAD	6.74	0.96	5.71	1.79
LASSO	6.59	0.92	5.72	1.80
Adaptive LASSO	6.65	0.95	5.62	1.73
BWA	6.99	1	5.99	1.99
BWA+BE	6.65	0.94	5.70	1.75
COSSO	5.36	0.48	4.70	0.95
CISE	5.82	0.81	4.97	1.46
ANOVA based V.S.	6.87	0.001	5.82	0.23
BEAMS	6.39	0.001	5.71	0.08

selection had poor performance for the linear model with dependent covariates, but achieved results slightly worse than BEAMS for the independent case. The proposed nonparametric variable selection procedure BEAMS correctly excluded, on average, about 19.5 out of the 20 nonsignificant predictors. This is about as good as the procedures designed for linear models. Moreover, BEAMS incorrectly excluded, on average, about 0.5 of the 5 significant predictors, which is more than the linear procedures (with the exception of BWA), and also more than CISE.

In Table 5, data sets of size $n = 40$ were generated from the models $Y = g_\ell(\mathbf{X}) + \epsilon$, $\ell = 1, 2$, where $\epsilon \sim N(0, 0.3^2)$, the dimension of $\mathbf{X}$ is $d = 8$, and

$$g_1(\mathbf{x}) = \sin(\pi x_1), \quad g_2(\mathbf{x}) = \sin(\frac{3}{4}\pi x_1) - 3\Phi(-|x_5|^3).$$

The covariates were generated as normal with marginal means zero and covariance matrix $\Sigma = (0.5^{|i-j|})$. It is seen that the linear model-based procedures failed to select the significant predictor(s) almost always. On the other hand, BEAMS always selected the one relevant predictor under model g_1, and

Table 6. Results for Body Fat example

Predictor	LASSO	Adpt. LASSO	SCAD	BWA	CISE
Age	0.06499	0	0.001061	0	0
Weight	0	-0.09511	-0.11688	-0.1356	0.015
Height	-0.1591	0	-0.05818	0	0
Neck	-0.2579	0	0	0	0
Chest	0	0	0	0	0
Abdomen	0.7079	0.9113	0.9052	0.9958	-0.127
Hip	0	0	0	0	0
Thigh	0	0	0	0	0
Knee	0	0	0	0	0
Ankle	0	0	0	0	0
Biceps	0	0	0	0	0
Forearm	0.21756	0	0	0.4729	0
Wrist	-1.5353	-0.9871	0	-1.5056	0

excluded incorrectly 0.08 out of the two important predictors under model g_2. The ANOVA based variable selection outperformed all the existing procedures and had performance slightly better than that of BEAMS for g_1, however, for g_2 it incorrectly excluded 3 times more than BEAMS. The other nonparametric methods, COSSO and CISE, failed to set to 0 on average 1.5 of the 7 irrelevant predictors for g_1 and 1.2 out of the 6 in g_2 and, moreover, had a poor performance in selecting the significant predictors compared to the proposed procedures.

5 Data Example: Body Fat

The Body Fat data was supplied by Dr. A. Garth Fisher for non-commercial purposes, and it can be found at "http://lib.stat.cmu.edu/datasets/bodyfat". The data set contains measurements of percent body fat (using Siri's (1956) method), Age (years), Weight (lbs), Height (inches), circunferences of Neck (cm), Chest (cm), Abdomen (cm), Hip (cm), Thigh (cm), Knee (cm), Ankle (cm), Biceps (cm), Forearm (cm), and Wrist (cm), from 252 men. The response variable was the percentage of body fat.

We compared the results of SCAD, LASSO, Adaptive LASSO, BWA with backward elimination, CISE, and BEAMS. Table 6 shows the estimated coefficients for LASSO, SCAD, Adaptive LASSO, BWA, and CISE. COSSO is not included in the comparison as the function *cosso* in the R package threw an error of computationally singular system.

Abdomen and Weight seem to be the most important predictors: Abdomen is selected by all, and Weight is selected by all except LASSO. The results of Adaptive LASSO and BWA differ only in the selection of Forearm by BWA, and both results differ considerably from those of SCAD. CISE selects only the two most important predictors.

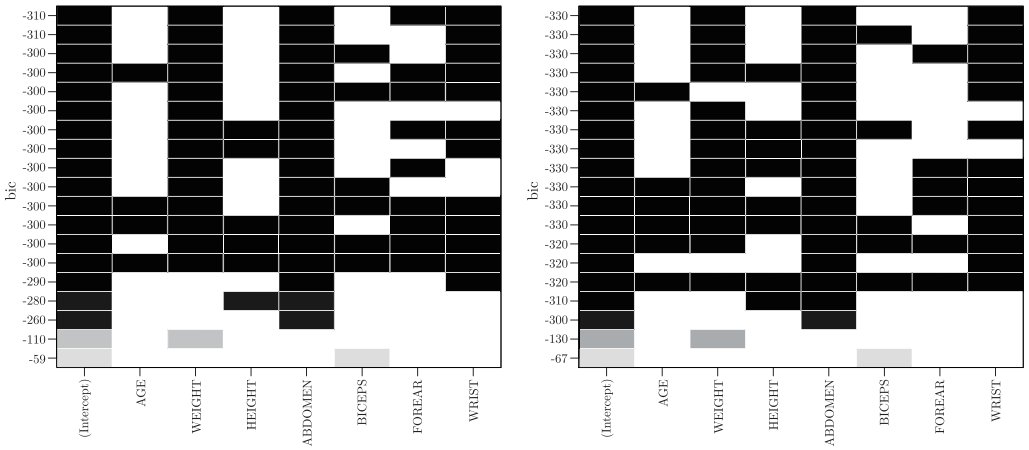

Figure 1. Best subset regression with OLS (left panel) and WLS (right panel)

BEAMS was applied using SIR, with the number of slices ranging from 2 to 100, and cell size 9 for the ANOVA-type test. Abdomen, Weight, Biceps, and Knee were selected on 99, 87, 88, and 23, respectively, of the 99 runs of the procedure (one run for each number of slices used). All other variables were selected less than 15 times. On the basis of these results we recommend a model based on Abdomen, Weight, and Biceps.

To explain the fact that Biceps was not selected by any of the other methods, we investigated possible violations of the assumptions of the multiple linear regression model, which four of the five methods in Table 6 use. The data are heteroscedastic, and, more importantly, the predictors affect the response in a nonlinear fashion and there is interaction among them. Ignoring any of them affects the results of variable selection. For example, using only the seven variables identified by more than one of the methods in Table 6, the BIC criterion gave different best models for ordinary and weighted least squares; see Figure 1. Under WLS, Biceps was included in the second best model, whose BIC value is a virtual tie with that of the best model. Additional insight is gained by running backward elimination, with p-to-remove 0.15, using WLS in models that include progressively more structure, based on the same seven variables included in Figure 1. Using a multiple linear regression model, Biceps is the first variable to be removed (p-value 0.188). Using a model that includes polynomials of degree 5 for each of the seven variables, and using p-values for the significance of the entire polynomial for each variable, the final model includes Wrist, Biceps, Abdomen, and Age. The p-value for Biceps in the final model is 0.0098. Finally, adding first order interaction terms, in addition to the polynomial terms, and using p-values for the entire polynomial as well as all interaction terms for each variable, the

final model includes Wrist, Biceps, Abdomen and Weight. The p-value for Biceps in the final model is 0.043. In conclusion, Biceps becomes a significant predictor when the complexity of the model is accounted for. The proposed method does that automatically.

Acknowledgement

This research was partially supported by CAPES/Fulbright grant 15087657, FAPESP 2012/22603-6 and 2012/10808-2, and NSF grants DMS-0805598 and DMS-1209059.

References

Aït-Sahalia, Y., Bickel, P. J. and Stoker, T.M. (2001). Goodness-of-fit tests for kernel regression with an application to option implied volatilities. *J. Econometrics* **105**, 363-412.

Abramovich, F., Benjamini, Y., Donoho, D. L. and Johnstone, I. M. (2006). Adapting to unknown sparsity by controlling the false discovery rate. *Ann. Statist.* **34**, 584-653.

Akritas, M. G. and Papadatos, N. (2004). Heterocedastic One-Way ANOVA and Lack-of-Fit Tests. *J. Amer. Statist. Assoc.* **99**, Theory and Methods.

Benjamini, Y. and Gavrilov, Y. (2009). A simple forward selection procedure based on false discovery rate control. *Ann. Appl. Statist.* **3**, 179-198.

Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. Roy. Statist. Soc. Ser. B* **57**, 289-300.

Benjamini, Y., Krieger, A. M. and Yekutieli, D. (2006). Adaptive linear step-up false discovery rate controlling procedures. *Biometrika* **93**, 491-507.

Benjamini, Y. and Yekutieli, D. (2001). The control of the false discovery rate in multiple testing under dependency. *Ann. Statist.* **29**, 1165-1188.

Bertin, K. and Lecué, G. (2008). Selection of variables and dimension reduction in high-dimensional non-parametric regression. *Electronic J. Statist.* **2**, 1224-1241.

Birge, L. and Massart, P. (2001). A generalized Cp criterion for Gaussian model. Technical report, Lab. De Probabilities, Univ. Paris VI. (`http://www.proba.jussieu.fr/mathdoc/preprints/index.html#2001`)

Bunea, F., Wegkamp, M. and Auguste, A. (2006). Consistent variable selection in high dimensional regression via multiple testing. *J. Statist. Plann. Inference* **136**, 4349-4364.

Candes, E., and Tao, T. (2007). The Dantzig selector: Statistical estimation when p is much larger than n. *Ann. Statist.* **35**, 2313-2351.

Chen, B. X., Zou, C. and Cook, R. D. (2010). Coordinate-independent sparse sufficient dimension reduction and variable selection. *Ann. Statist.* **38**, 3696-3723.

Delgado, M. A. and Manteiga, W. G. (2001). Significance testing in nonparametric regression based on the bootstrap. *Ann. Statist.* **29**, 1469-1507.

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

Efron, B., Hastie, T., Johnstone, I., and Tibshirani, R. (2004) Least angle regression. *Ann. Statist.* **32**, 407-499.

Fan, J., and Jiang, J.(2005). Nonparametric inferences for additive models. *J. Amer. Statist. Assoc.* **100**, 890-907.

Fan, J., and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *J. Amer. Statist. Assoc.* **96**, 1348-1360.

Fan, J., Zhang, C. M. and Zhang, J. (2001). Generalized likelihood ratio statistics and Wilks phenomenon. *Ann. Statist.* **29**, 153193.

Fan, Y. and Li, Q. (1996). Consistent model specification tests: Omitted variables and semiparametric functional forms. *Econometrica* **64**, 865-890.

Foster, D. P. and Stine, R. A. (2004). Variable selection in data mining: building a predictive model for bankruptcy. *J. Amer. Statist. Assoc.* **99**, 303-313.

Hart, J. D. (1997). *Nonparametric Smoothing and Lack of fit tests*. Springer.

Horowitz, J. L. and Mammen, E. (2004). Nonparametric estimation of an additive model with a link function. *Ann. Statist.* **32**, 2412-2443.

Hristache, M., Juditsky, A., Polzehl, J. and Spokoiny, V. (2001). Structure adaptive approach for dimension reduction. *Ann. Statist.* **29**, 1537-1566.

Huang, J., Horowitz, J. L. and Wei, F. (2010). Variable selection in nonparametric additive models. Available at
`http://faculty.wcas.northwestern.edu/~jlh951/papers/HHW-npam.pdf`.

Kong, E. and Xia, Y. (2007). Variable selection for the single-index model. *Biometrika* **94**, 217-229.

Lafferty, J. and Wasserman, L. (2008). Rodeo: sparse, greedy nonparametric regression. *Ann. Statist.* **36**, 28-63.

Lavergne, P. and Vuong, Q. (2000). Nonparametric significance testing. *Econometric Theory* **16**, 576-601.

Li, K. C. (1991). Sliced inverse regression for dimension reduction. *J. Amer. Statist. Assoc.* **86**, 316-327.

Li, L., Cook, R. D. and Nachtsheim, C. (2005). Model-free variable selection. *J. Roy. Statist. Soc. Ser. B* **67**, 285-299.

Li, R. and Liang, H. (2008). Variable selection in semiparametric regression modeling. *Ann. Statist.* **36**, 261-286.

Lin, Y. and Zhang, H. (2006). Component selection and smoothing in multivariate nonparametric regression. *Ann. Statist.* **34**, 2272-2297.

Linton, O. and Nielsen, J. P. (1995). A kernel method of estimating structured nonparametric regression based on marginal integration. *Biometrika* **82**, 93-100.

Mammen, E., Linton, O. and Nielsen, J. (1999). The existence and asymptotic properties of a backfitting projection algorithm under weak conditions. *Ann. Statist.* **27**, 1443-1490.

Newey, W. K. (1994). Kernel estimation of partial means. *Econom. Theory* **10**, 233-253.

Racine, J., Hart, J. D. and Li, Q. (2006). Testing the significance of categorical predictor variables in nonparametric regression models. *Econometric Rev.* **25**, 523-544.

Rice, J. (1984). Bandwidth choice for nonparametric regression. *Ann. Statist.* **12**, 1215-1230.

Stone, C. (1985). Additive regression and other nonparametric models. *Ann. Statist.* **13**, 689-705.

Storlie, C. B, Bondell, H. D, Reich, B. J. and Zhang, H. H. (2011). Surface estimation, variable selection, and the nonparametric oracle property. *Statist. Sinica* **21**, 679-705.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *J. Roy. Statist. Soc. Ser. B* **58**, 267-288.

Tibshirani, R. and Knight, K. (1999). The covariance inflation criterion for adaptive model selection. *J. Roy. Statist. Soc. Ser. B* **61**, 529-546.

Wang, H. and Xia, Y. (2008). Shrinkage estimation of the varying coefficient model. *J. Amer. Statist. Assoc.* **104**,747-757.

Wang, L., Akritas, M. G. and Keilegom, I. V. (2008). An ANOVA-type nonparametric diagnostic test for heterocedastic regression models. *J. Nonparametr. Stat.* **20**, 365-382.

Xia, Y. (2008). A multiple-index model and dimension reduction. *J. Amer. Statist. Assoc.* **103**, 1631-1640.

Xia, Y., Tong, H., Li, W. K. and Zhu, L.-X. (2002). An adaptive estimation of dimension reduction space. *J. Roy. Statist. Soc. Ser. B* **64**, 363-410.

Zou, H. (2006). The adaptive lasso and its oracle properties. *J. Amer. Statist. Assoc.* **101**, 1418-1429.

Department of Statistics, State University of Campinas, Rua Sérgio Buarque de Holanda, 651. Cidade Universitária "Zeferino Vaz" - Distr. Barão Geraldo. Campinas, São Paulo, Brasil-13083-859.

E-mail: zambom@ime.unicamp.br

Department of Statistics, Penn State University, 326 Thomas Bldg., University Park, PA 16802, USA.

E-mail: mga@stat.psu.edu

(Received May 2013; accepted January 2014)