Nonparametric Independence Screening in Sparse Ultra-High-Dimensional Additive Models

Jianqing FAN, Yang FENG, and Rui SONG

A variable screening procedure via correlation learning was proposed by Fan and Lv (2008) to reduce dimensionality in sparse ultra-high-dimensional models. Even when the true model is linear, the marginal regression can be highly nonlinear. To address this issue, we further extend the correlation learning to marginal nonparametric learning. Our nonparametric independence screening (NIS) is a specific type of sure independence screening. We propose several closely related variable screening procedures. We show that with general nonparametric models, under some mild technical conditions, the proposed independence screening methods have a sure screening property. The extent to which the dimensionality can be reduced by independence screening is also explicitly quantified. As a methodological extension, we also propose a data-driven thresholding and an iterative nonparametric independence screening (INIS) method to enhance the finite- sample performance for fitting sparse additive models. The simulation results and a real data analysis demonstrate that the proposed procedure works well with moderate sample size and large dimension and performs better than competing methods.

KEY WORDS: Additive model; Independent learning; Nonparametric independence screening; Nonparametric regression; Sparsity; Sure independence screening; Variable selection.

1. INTRODUCTION

With rapid advances of computing power and other modern technology, high-throughput data of unprecedented size and complexity are frequently seen in many contemporary statistical studies. Examples include data from genetic, microarray, proteomic, snd functional magnetic resonance imaging studies, functional data, and high-frequency financial data. In all of these examples, the number of variables p can grow much faster than the number of observations n. To be more specific, we assume $\log p = O(n^a)$ for some $a \in (0, 1/2)$. Following Fan and Lv (2009), we call this nonpolynomial (NP) dimensionality, or ultra-high dimensionality. What makes the underdetermined statistical inference possible is the sparsity assumption; only a small set of independent variables contribute to the response. Thus, dimension reduction and feature selection play pivotal roles in these ultra-high-dimensional problems.

The statistical literature contains numerous procedures on the variable selection for linear models and other parametric models, including the Lasso (Tibshirani 1996), the SCAD and other folded-concave penalty models (Fan 1997; Fan and Li 2001), the Dantzig selector (Candes and Tao 2007), the Elastic net (Enet) penalty (Zou and Hastie 2005), the MCP (Zhang 2010), and related methods (Zou 2006; Zou and Li 2008). Nevertheless, due to the "curse of dimensionality" in terms of simultaneous challenges to computational expediency, statistical accuracy, and algorithmic stability, these methods are limited in handling ultra-high-dimensional problems.

Motivated by these concerns, Fan and Lv (2008) introduced a new framework for variable screening via correlation learning with NP dimensionality in the context of least squares. Hall, Titterington, and Xue (2009) used a different marginal utility, derived from an empirical likelihood point of view. Hall and Miller (2009) proposed a generalized correlation ranking, which allows nonlinear regression. Huang, Horowitz, and Ma (2008) also investigated marginal bridge regression in the ordinary linear model. These methods focus on studying the marginal pseudolikelihood and are fast but crude in terms of reducing NP dimensionality to a more moderate size. To enhance performance, Fan and Lv (2008) and Fan, Samworth, and Wu (2009) introduced some methodological extensions to independence screening (SIS), including iterative SIS (ISIS) and multistage procedures, such as SIS-SCAD and SIS-LASSO, to select variables and estimate parameters simultaneously. These marginal screening methods have some methodological challenges, however. When the covariates are not jointly normal, even if the linear model holds in the joint regression, the marginal regression can be highly nonlinear. Thus, SIS based on nonparametric marginal regression becomes a natural candidate.

In practice, there is often little prior information indicating that the effects of the covariates take a linear form or belong to any other finite-dimensional parametric family. Substantial improvements are sometimes possible by using a more flexible class of nonparametric models, such as the additive model, $Y = \sum_{j=1}^{p} m_j(X_j) + \varepsilon$, introduced by Stone (1985). Thus significantly increases the flexibility of the ordinary linear model and allows a data-analytic transform of the covariates to enter into the linear model. Nonetheless, the literature on variable selection in nonparametric additive models s limited (see, e.g., Koltchinskii and Yuan 2008; Meier, Geer, and Bühlmann 2009; Ravikumar et al. 2009; Huang, Horowitz, and Wei 2010). The work of Koltchinskii and Yuan (2008) and Ravikumar et al. (2009) is closely related to he COSSO methods proposed by Lin and Zhang (2006) with fixed minimal signals, which does not converge to 0. The approach of Huang, Horowitz, and Wei (2010) can be viewed as an extension of adaptive lasso to additive models with fixed minimal signals. Meier, Geer, and

Jianqing Fan is Frederick L. Moore Professor of Finance, Department of Operations Research and Financial Engineering, Princeton University, Princeton, NJ 08544 (E-mail: *jqfan@princeton.edu*). Yang Feng is Assistant Professor, Department of Statistics, Columbia University, New York, NY 10027 (E-mail: *yangfeng@stat.columbia.edu*). Rui Song is Assistant Professor, Department of Statistics, Colorado State University, Fort Collins, CO 80523 (E-mail: *song@stat.colostate.edu*). Financial support was provided by National Science Foundation grants DMS-0714554, DMS-0704337, and DMS-1007698 and National Institutes of Health grant R01-GM072611. The authors are in deep debt to Dr. Lukas Meier for sharing the codes of penGAM. The authors thank the editor, the associate editor, and referees for their constructive comments.

Journal of the American Statistical Association
June 2011, Vol. 106, No. 494, Theory and Methods
DOI: 10.1198/jasa.2011.tm09779

Bühlmann (2009) proposed a penalty that is a combination of sparsity and smoothness with a fixed design. In ultra-high-dimensional settings, all of these methods still suffer from the aforementioned three challenges, because they can be viewed as extensions of penalized pseudolikelihood approaches to additive modeling. A commonly used algorithm in additive modeling, such as backfitting, makes the situation even more challenging, given its great computational expense.

In this article, we consider independence learning by ranking the magnitude of marginal estimators, nonparametric marginal correlations, and the marginal residual sum of squares. That is, we fit p marginal nonparametric regressions of the response Y against each covariate X_i separately and rank their importance to the joint model according to a measure of the goodness of fit of their marginal model. The magnitude of these marginal utilities can preserve the nonsparsity of the joint additive models under some reasonable conditions, even with converging minimum strength of signals. Our work can be regarded as an important and nontrivial extension of SIS procedures proposed by Fan and Lv (2008) and Fan and Song (2010). Compared with those articles, the minimum distinguishable signal is related not only to the stochastic error in estimating the nonparametric components, but also to approximation errors in modeling nonparametric components, which depend on the number of basis functions used for the approximation. This poses significant challenges to the theoretical development and leads to an interesting result regarding the extent to which the dimensionality can be reduced by nonparametric independence screening. We also propose an iterative nonparametric independence screening procedure, INIS-penGAM, to reduce the false positive rate and stabilize the computation. This two-stage procedure can deal with the aforementioned three challenges better than other methods, as we demonstrated in our empirical studies.

We approximate the nonparametric additive components using a B-spline basis. Thus the component selection in additive models can be viewed as a functional version of the grouped variable selection. An early article on group variable selection using group penalized least squares is that of Antoniadis and Fan (2001, p. 966), in which blocks of wavelet coefficients are either killed or selected. The group variable selection was studied more thoroughly by Yuan and Lin (2006), Kim, Kim, and Kim (2006), Wei and Huang (2007), and Meier, Geer, and Bühlmann (2009). Our methods and results have important implications for group variable selection, because in additive regression, each component can be expressed as a linear combination of a set of basis functions whose coefficients must be either killed or selected simultaneously.

The rest of the article is organized as follows. In Section 2 we introduce the nonparametric independence screening (NIS) procedure in additive models. We present the theoretical properties for NIS in Section 3. As a methodological extension, we outline INIS-penGAM and its greedy version g-INIS-penGAM in Section 4. Our Monte Carlo simulations and a real data analysis in Section 5 demonstrate the effectiveness of the INIS method. We conclude with a discussion in Section 6, and relegate the proofs to Section 7.

2. NONPARAMETRIC INDEPENDENCE SCREENING

Suppose that we have a random sample, $\{(\mathbf{X}_i, Y_i)\}_{i=1}^n$, from the population

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

in which $\mathbf{X} = (X_1, \ldots, X_p)^T$, ε is the random error with conditional mean 0. To expeditiously identify important variables in model (1), without the curse of dimensionality, we consider the following p marginal nonparametric regression problems:

$$\min_{f_j \in L_2(P)} E(Y - f_j(X_j))^2, \tag{2}$$

where P denotes the joint distribution of $(\mathbf{X}, Y)$ and $L_2(P)$ is the class of square integrable functions under the measure P. The minimizer of (2) is $f_j = E(Y|X_j)$, the projection of Y onto X_j. We rank the utility of covariates in model (1) according to, for example, $Ef_j^2(X_j)$, and select a small group of covariates by thresholding.

To obtain a sample version of the marginal nonparametric regression, we use a B-spline basis. Let $\mathcal{S}_n$ be the space of polynomial splines of degree $l \geq 1$ and $\{\Psi_{jk}, k = 1, \ldots, d_n\}$ denote a normalized B-spline basis with $\|\Psi_{jk}\|_\infty \leq 1$, where $\|\cdot\|_\infty$ is the sup norm. For any $f_{nj} \in \mathcal{S}_n$, we have

$$f_{nj}(x) = \sum_{k=1}^{d_n} \beta_{jk} \Psi_{jk}(x), \qquad 1 \leq j \leq p,$$

for some coefficients $\{\beta_{jk}\}_{k=1}^{d_n}$. Under some smoothness conditions, the nonparametric projections $\{f_j\}_{j=1}^p$ can be well approximated by functions in $\mathcal{S}_n$. The sample version of the marginal regression problem can be expressed as

$$\min_{f_{nj} \in \mathcal{S}_n} \mathbb{P}_n(Y - f_{nj}(X_j))^2 = \min_{\boldsymbol{\beta}_j \in \mathbb{R}^{d_n}} \mathbb{P}_n(Y - \boldsymbol{\Psi}_j^T \boldsymbol{\beta}_j)^2, \tag{3}$$

where $\boldsymbol{\Psi}_j \equiv \boldsymbol{\Psi}_j(X_j) = (\Psi_1(X_j), \ldots, \Psi_{d_n}(X_j))^T$ denotes the d_n-dimensional basis functions and $\mathbb{P}_n g(\mathbf{X}, Y)$ is the expectation with respect to the empirical measure $\mathbb{P}_n$, that is, the sample average of $\{g(\mathbf{X}_i, Y_i)\}_{i=1}^n$. This univariate nonparametric smoothing can be computed rapidly, even for NP-dimensional problems. We correspondingly define the population version of the minimizer of the componentwise least squares regression,

$$f_{nj}(X_j) = \boldsymbol{\Psi}_j^T (E\boldsymbol{\Psi}_j \boldsymbol{\Psi}_j^T)^{-1} E\boldsymbol{\Psi}_j Y, \qquad j = 1, \ldots, p,$$

where E denotes the expectation under the true model.

We now select a set of variables

$$\widehat{\mathcal{M}}_{\nu_n} = \{1 \leq j \leq p : \|\hat{f}_{nj}\|_n^2 \geq \nu_n\}, \tag{4}$$

where $\|\hat{f}_{nj}\|_n^2 = n^{-1} \sum_{i=1}^n \hat{f}_{nj}(X_{ij})^2$ and ν_n is a predefined threshold value. Such an independence screening ranks the importance according to the marginal strength of the marginal nonparametric regression. This screening also can be viewed as ranking by the magnitude of the correlation of the marginal nonparametric estimate $\{\hat{f}_{nj}(X_{ij})\}_{i=1}^n$ with the response $\{Y_i\}_{i=1}^n$, because $\|\hat{f}_{nj}\|_n^2 = \|Y\hat{f}_{nj}\|_n$. In this sense, the proposed NIS procedure is related to the correlation learning proposed by Fan and Lv (2008).

Another screening approach is to rank according to the descent order of the residual sum of squares of the componentwise nonparametric regressions, where we select a set of variables:

$$\widehat{\mathcal{N}}_{\gamma_n} = \{1 \le j \le p : u_j \le \gamma_n\},$$

with $u_j = \min_{\boldsymbol{\beta}_j} \mathbb{P}_n(Y - \boldsymbol{\Psi}_j^T \boldsymbol{\beta}_j)^2$ is the residual sum of squares of the marginal fit and γ_n is a predefined threshold value. It is straightforward to show that $u_j = \mathbb{P}_n(Y^2 - \hat{f}_{nj}^2)$. Thus the two methods are equivalent.

The NIS procedure reduces the dimensionality from p to a possibly much smaller space with model size $|\widehat{\mathcal{M}}_{\nu_n}|$ or $|\widehat{\mathcal{N}}_{\gamma_n}|$. The method is applicable to all models. The question is whether we have mistakenly deleted some active variables in model (1)—in other words, whether the procedure has a sure screening property, as postulated by Fan and Lv (2008). In the next section we show that the sure screening property indeed holds for nonparametric additive models with a limited false selection rate.

3. SURE SCREENING PROPERTIES

In this section we establish the sure screening properties for additive models, with results presented in three steps.

3.1 Preliminaries

We assume that the true regression function admits the additive structure

$$m(\mathbf{X}) = \sum_{j=1}^{p} m_j(X_j). \tag{5}$$

For identifiability, we assume that $\{m_j(X_j)\}_{j=1}^p$ have mean 0. Consequently, the response Y has mean 0 as well. Let $\mathcal{M}_\star = \{j : Em_j(X_j)^2 > 0\}$ be the true sparse model with nonsparsity size $s_n = |\mathcal{M}_\star|$. We allow p to grow with n and denote it by p_n whenever necessary.

The theoretical basis of the sure screening is that the marginal signal of the active components $(\|f_j\|, j \in \mathcal{M}_\star)$ does not vanish, where $\|f_j\|^2 = Ef_j^2$. The following conditions make this possible. For simplicity, let $[a, b]$ be the support of X_j.

A. The nonparametric marginal projections $\{f_j\}_{j=1}^p$ belong to a class of functions $\mathcal{F}$, whose rth derivative $f^{(r)}$ exists and is Lipschitz of order α,

$$\mathcal{F} = \left\{f(\cdot) : \left|f^{(r)}(s) - f^{(r)}(t)\right| \le K|s-t|^\alpha \text{ for } s, t \in [a, b]\right\}$$

for some positive constant K, where r is a nonnegative integer and $\alpha \in (0, 1]$ such that $d = r + \alpha > 0.5$.

B. The marginal density function g_j of X_j satisfies $0 < K_1 \le g_j(X_j) \le K_2 < \infty$ on $[a, b]$ for $1 \le j \le p$ for some constants K_1 and K_2.

C. $\min_{j \in \mathcal{M}_\star} E\{E(Y|X_j)^2\} \ge c_1 d_n n^{-2\kappa}$, for some $0 < \kappa < d/(2d+1)$ and $c_1 > 0$.

Under conditions A and B, the following three facts hold when $l \ge d$ and are used in this work. We state them here for readability.

Fact 1. There exists a positive constant C_1 such that (Stone 1985)

$$\|f_j - f_{nj}\|^2 \le C_1 d_n^{-2d}. \tag{6}$$

Fact 2. There exists a positive constant C_2 such that (Stone 1985; Huang, Horowitz, and Wei 2010)

$$E\Psi_{jk}^2(X_{ij}) \le C_2 d_n^{-1}. \tag{7}$$

Fact 3. There exist some positive constants D_1 and D_2 such that (Zhou, Shen, and Wolfe 1998)

$$D_1 d_n^{-1} \le \lambda_{\min}(E\boldsymbol{\Psi}_j\boldsymbol{\Psi}_j^T) \le \lambda_{\max}(E\boldsymbol{\Psi}_j\boldsymbol{\Psi}_j^T) \le D_2 d_n^{-1}. \tag{8}$$

The following lemma shows that the minimum signal of $\{\|f_{nj}\|\}_{j \in \mathcal{M}_\star}$ is at the same level of the marginal projection, provided that the approximation error is negligible.

Lemma 1. Under conditions A–C, we have

$$\min_{j \in \mathcal{M}_\star} \|f_{nj}\|^2 \ge c_1 \xi d_n n^{-2\kappa},$$

provided that $d_n^{-2d-1} \le c_1(1-\xi)n^{-2\kappa}/C_1$ for some $\xi \in (0, 1)$.

A model selection consistency result can be established with nonparametric independence screening under the partial orthogonality condition, that is, $\{X_j, j \notin \mathcal{M}_\star\}$ is independent of $\{X_i, i \in \mathcal{M}_\star\}$. In this case, there is a separation between the strengths of marginal signals $\|f_{nj}\|^2$ for active variables $\{X_j; j \in \mathcal{M}_\star\}$ and for inactive variables $\{X_j, j \notin \mathcal{M}_\star\}$, which are zero. When the separation is sufficiently large, these two sets of variables can be easily identified.

3.2 Sure Screening

In this section we establish the sure screening properties of NIS. We require the following additional conditions:

D. $\|m\|_\infty < B_1$ for some positive constant B_1, where $\|\cdot\|_\infty$ is the sup norm.

E. The random error $\{\varepsilon_i\}_{i=1}^n$ are iid with conditional mean 0, and for any $B_2 > 0$, there exists a positive constant B_3 such that $E[\exp(B_2|\varepsilon_i|)|\mathbf{X}_i] < B_3$.

F. There exist positive constants c_1 and $\xi \in (0, 1)$ such that $d_n^{-2d-1} \le c_1(1-\xi)n^{-2\kappa}/C_1$.

The following theorem gives the sure screening properties. It shows that it is only the size of nonsparse elements, s_n, that matters for the purpose of sure screening, not the dimensionality, p_n. The first result is on the uniform convergence of $\|\hat{f}_{nj}\|_n^2$ to $\|f_{nj}\|^2$.

Theorem 1. Suppose that conditions A, B, D, and E hold.

(i) For any $c_2 > 0$, there exist some positive constants c_3 and c_4 such that

$$\begin{aligned} P\Big(\max_{1 \le j \le p_n} |\|\hat{f}_{nj}\|_n^2 - \|f_{nj}\|^2| \ge c_2 d_n n^{-2\kappa}\Big) \\ \le p_n d_n\{(8 + 2d_n)\exp(-c_3 n^{1-4\kappa} d_n^{-3}) \\ + 6d_n \exp(-c_4 n d_n^{-3})\}. \end{aligned} \tag{9}$$

(ii) If, in addition, conditions C and F hold, then, by taking $\nu_n = c_5 d_n n^{-2\kappa}$ with $c_5 \le c_1\xi/2$, we have

$$\begin{aligned} P\big(\mathcal{M}_\star \subset \widehat{\mathcal{M}}_{\nu_n}\big) \ge 1 - s_n d_n\{(8 + 2d_n)\exp(-c_3 n^{1-4\kappa} d_n^{-3}) \\ + 6d_n \exp(-c_4 n d_n^{-3})\}. \end{aligned}$$

Note that the second part of the upper bound in Theorem 1 is related to the uniform convergence rates of the minimum eigenvalues of the design matrices. It gives an upper bound on the number of basis, $d_n = o(n^{1/3})$, to have the sure screening property, whereas condition F requires $d_n \geq B_4 n^{2\kappa/(2d+1)}$, where $B_4 = (c_1(1-\xi)/C_1)^{-1/(2d+1)}$.

It follows from Theorem 1 that we can handle the NP dimensionality,

$$\log p_n = o(n^{1-4\kappa} d_n^{-3} + n d_n^{-3}). \tag{10}$$

Under this condition,

$$P\big(\mathcal{M}_\star \subset \widehat{\mathcal{M}}_{\nu_n}\big) \to 1,$$

that is, the sure screening property. It is worthwhile to point out that the number of spline bases, d_n, affects the order of dimensionality; compare this with the results of Fan and Lv (2008) and Fan and Song (2010), in which univariate marginal regression is used. Equation (10) shows that the larger the minimum signal level or the smaller the number of basis functions, the higher the dimensionality that the NIS can handle. This is in line with our intuition. On the other hand, the number of basis functions can not be too small, because the approximation error can not be too large. As required by condition F, $d_n \geq B_4 n^{2\kappa/(2d+1)}$; the smoother the underlying function, the smaller the d_n that we can take and the higher the dimension that the NIS can handle. If the minimum signal does not converge to 0 (as in Lin and Zhang 2006; Koltchinskii and Yuan 2008; and Huang, Horowitz, and Wei 2010), then $\kappa = 0$. In this case, d_n can be taken to be finite as long as it is sufficiently large so that the minimum signal in Lemma 1 exceeds the noise level. Taking $d_n = n^{1/(2d+1)}$, the optimal rate for nonparametric regression (Stone 1985), we have $\log p_n = o(n^{2(d-1)/(2d+1)})$. In other words, the dimensionality can be as high as $\exp\{o(n^{2(d-1)/(2d+1)})\}$.

3.3 Controlling False Selection Rates

The sure screening property without controlling for false selection rates is not insightful. It basically states that the NIS has no false negatives. An ideal case for the vanishing false-positive rate is that

$$\max_{j \notin \mathcal{M}_\star} \|f_{nj}\|^2 = o(d_n n^{-2\kappa}),$$

so that there is a gap between active variables and inactive variables in model (1) when using the marginal nonparametric screener. In this case, by Theorem 1(i), if (9) tends to 0, with probability tending to 1 that

$$\max_{j \notin \mathcal{M}_\star} \|\hat{f}_{nj}\|_n^2 \leq c_2 d_n n^{-2\kappa} \quad \text{for any } c_2 > 0.$$

Thus, by the choice of ν_n as in Theorem 1(ii), we can achieve model selection consistency:

$$P\big(\widehat{\mathcal{M}}_{\nu_n} = \mathcal{M}_\star\big) = 1 - o(1).$$

We now deal with the more general case. The idea is to bound the size of the selected set making use of the fact that $\operatorname{var}(Y)$ is bounded. In this part, we show that the correlations among the basis functions (i.e., the design matrix of the basis functions) are related to the size of selected models.

Theorem 2. Suppose that conditions A–F hold and $\operatorname{var}(Y) = O(1)$. Then, for any $\nu_n = c_5 d_n n^{-2\kappa}$, there exist positive constants c_3 and c_4 such that

$$\begin{aligned} P\big[\big|\widehat{\mathcal{M}}_{\nu_n}\big| \leq O\{n^{2\kappa}\lambda_{\max}(\boldsymbol{\Sigma})\}\big] &\geq 1 - p_n d_n\{(8+2d_n)\exp(-c_3 n^{1-4\kappa} d_n^{-3}) \\ &\quad + 6 d_n \exp(-c_4 n d_n^{-3})\}, \end{aligned}$$

where $\boldsymbol{\Sigma} = E\boldsymbol{\Psi}\boldsymbol{\Psi}^T$ and $\boldsymbol{\Psi} = (\boldsymbol{\Psi}_1, \ldots, \boldsymbol{\Psi}_{p_n})^T$.

The significance of this result is that when $\lambda_{\max}(\boldsymbol{\Sigma}) = O(n^\tau)$, the selected model size with the sure screening property is only of polynomial order, whereas the original model size is of NP dimensionality. In other words, the false selection rate converges to 0 exponentially fast. The size of the selected variables is of order $O(n^{2\kappa+\tau})$, of the same order as in the approach of Fan and Lv (2008). Our result is an extension of the work of Fan and Lv (2008), even in this very specific case without the condition $2\kappa + \tau < 1$. The results are also consistent with that of Fan and Song (2010), with the number of selected variables related to the correlation structure of the covariance matrix.

In the specific case where the covariates are independent, then the matrix $\boldsymbol{\Sigma}$ is block diagonal with jth block $\boldsymbol{\Sigma}_j$. Thus it follows from (8) that $\lambda_{\max}(\boldsymbol{\Sigma}) = O(d_n^{-1})$.

4. INIS METHOD

4.1 Description of the Algorithm

After variable screening, the next step is naturally to select the variables using more refined techniques in the additive model. For example, the penalized method for additive model (penGAM) of Meier, Geer, and Bühlmann (2009) can be used to select a subset of active variables, resulting in NIS-penGAM. To further enhance the performance of the method in terms of false selection rates, following Fan and Lv (2008) and Fan, Samworth, and Wu (2009), we can iteratively use a large-scale screening and moderate-scale selection strategy, resulting in the INIS-penGAM.

Given the data $\{(\mathbf{X}_i, Y_i)\}, i = 1, \ldots, n$, for each component $f_j(\cdot), j = 1, \ldots, p$, we choose the same truncation term, $d_n = O(n^{1/5})$. To determine a data-driven thresholding for independence screening, we extend the random permutation idea of Zhao and Li (2010), which allows only $1 - q$ proportion (for a given $q \in [0, 1]$) of inactive variables to enter the model when $\mathbf{X}$ and Y are not related (the null model). We use random permutation to decouple $\mathbf{X}_i$ and Y_i, so that the resulting data $(\mathbf{X}_{\pi(i)}, Y_i)$ follow a null model, where $\pi(1), \ldots, \pi(n)$ are a random permutation of the index $1, \ldots, n$. The algorithm works as follows:

Step 1. For every $j \in \{1, \ldots, p\}$, compute

$$\hat{f}_{nj} = \arg\min_{f_{nj} \in \mathcal{S}_n} \mathbb{P}_n(Y - f_{nj}(X_j))^2 \quad \text{for } 1 \leq j \leq p.$$

Randomly permute the rows of $\mathbf{X}$, yielding $\tilde{\mathbf{X}}$. Let $\omega_{(q)}$ be the qth quantile of $\{\|\hat{f}_{nj}^*\|_n^2, j = 1, 2, \ldots, p\}$, where

$$\hat{f}_{nj}^* = \arg\min_{f_{nj} \in \mathcal{S}_n} \mathbb{P}_n(Y - f_{nj}(\tilde{X}_j))^2.$$

Then NIS selects the following variables:

$$\mathcal{A}_1 = \big\{j : \|\hat{f}_{nj}^*\|_n^2 \geq \omega_{(q)}\big\}.$$

In our numerical examples, we use $q = 1$ (i.e., take the maximum value of the empirical norm of the permuted estimates).

Step 2. Apply penGAM (Meier, Geer, and Bühlmann 2009) on the set $\mathcal{A}_1$ to select a subset $\mathcal{M}_1$. Inside the penGAM algorithm, the penalty parameter is selected by cross-validation.

Step 3. For every $j \in \mathcal{M}_1^c = \{1, \ldots, p\} \setminus \mathcal{M}_1$, minimize

$$\mathbb{P}_n\left(Y - \sum_{i \in \mathcal{M}_1} f_{ni}(X_i) - f_{nj}(X_j)\right)^2, \tag{11}$$

with respect to $f_{ni} \in \mathcal{S}_n$ for all $i \in \mathcal{M}_1$ and $f_{nj} \in \mathcal{S}_n$. This regression reflects the additional contribution of the jth components conditioning on the existence of the variable set $\mathcal{M}_1$. After marginally screening as in Step 1, choose a set of indices, $\mathcal{A}_2$. Here the determination of size is the same as in Step 1, except that only the variables not in $\mathcal{M}_1$ are randomly permuted. Then the penGAM algorithm is applied on the set $\mathcal{M}_1 \cup \mathcal{A}_2$ to select a subset $\mathcal{M}_2$.

Step 4. Iterate the process until $|\mathcal{M}_l| \geq s_0$ or $\mathcal{M}_l = \mathcal{M}_{l-1}$.

We have a few comments about this method. In Step 2, we use penGAM. In fact, any variable selection method for additive models, such as SpAM (Ravikumar et al. 2009) and the adaptive group LASSO for additive models of Huang, Horowitz, and Wei (2010), would work. A similar sample splitting idea as described by Fan, Samworth, and Wu (2009) can be applied here to further decrease the false selection rate.

4.2 Greedy INIS

We now propose a greedy modification to the INIS algorithm to speed up the computation and to enhance performance. Specifically, we restrict the size of the set $\mathcal{A}_j$ in the iterative screening steps to be at most p_0, a small positive integer, and the algorithm stops when none of the variables is recruited, that is, exceeding the thresholding for the null model. In the numerical studies, p_0 is taken to be 1 for simplicity. This greedy version of the INIS algorithm is called g-INIS.

When $p_0 = 1$, the g-INIS method is connected with forward selection (Efroymson 1960; Draper and Smith 1966). Recently, Wang (2009) showed that under certain technical conditions, forward selection also can achieve the sure screening property. Both g-INIS and forward selection recruit at most one new variable into the model at a time. The major difference is that unlike forward selection, which keeps a variable once selected, g-INIS includes a deletion step via penalized least squares that can remove multiple variables. This makes the g-INIS algorithm more attractive, because it is more flexible in terms of recruiting and deleting variables.

The g-INIS method is particularly effective when the covariates are highly correlated or conditionally correlated. In this case, the original INIS method tends to select many unimportant variables that have high correlation with important variables because they too have large marginal effects on the response. Although greedy, the g-INIS method is better at choosing true positives due to more stringent screening and improves the likelihood of selecting the remaining important variables in subsequent stages because of fewer false positives at each stage. This leads to conditioning on a smaller set of more relevant variables and improves the overall performance. Based on our numerical experience, the g-INIS method outperforms the original INIS method in all examples in terms of a higher true-positive rate, lower false-positive rate, and smaller prediction error.

5. NUMERICAL RESULTS

In this section we illustrate our method by studying its performance on the simulated data and in a real data analysis. Some of the simulation settings are adapted from the work of Fan and Lv (2008), Meier, Geer, and Bühlmann (2009), Huang, Horowitz, and Wei (2010), and Fan and Song (2010).

5.1 Comparison of Minimum Model Size

We first illustrate the behavior of the NIS procedure under different correlation structures. Following Fan and Song (2010), we use the minimum model size (MMS) required for the NIS procedure and the penGAM procedure to have the sure screening property (i.e., to contain the true model $\mathcal{M}^*$) as a measure of the effectiveness of a screening method. We also include the correlation screening method of Fan and Lv (2008) for comparison. The advantage of the MMS method is that we do not need to choose the thresholding parameter or penalized parameters. For NIS, we take $d_n = \lfloor n^{1/5} \rfloor + 2 = 5$. We set $n = 400$ and $p = 1000$ for all examples.

Example 1. Following Fan and Song (2010), let $\{X_k\}_{k=1}^{950}$ be iid standard normal random variables and

$$X_k = \sum_{j=1}^{s} X_j (-1)^{j+1} \Big/ 5 + \sqrt{1 - \frac{s}{25}}\varepsilon_k, \qquad k = 951, \ldots, 1000,$$

where $\{\varepsilon_k\}_{k=951}^{1000}$ are standard normally distributed. We consider the following linear model as a specific case of the additive model: $Y = \boldsymbol{\beta}^{*T}\mathbf{X} + \varepsilon$, in which $\varepsilon \sim \mathrm{N}(0, 3)$ and $\boldsymbol{\beta}^* = (1, -1, \ldots)^T$ has s nonvanishing components, taking values ± 1 alternately.

Example 2. In this example, the data are generated from the simple linear regression $Y = X_1 + X_2 + X_3 + \varepsilon$, where $\varepsilon \sim \mathrm{N}(0, 3)$. The covariates are not normally distributed, however; $\{X_k\}_{k \neq 2}$ are iid standard normal random variables, whereas $X_2 = -\frac{1}{3}X_1^3 + \tilde{\varepsilon}$, where $\tilde{\varepsilon} \sim \mathrm{N}(0, 1)$. In this case, $E(Y|X_1)$ and $E(Y|X_2)$ are nonlinear.

The MMS for each method and its associated robust estimate of the standard deviation (RSD = IQR/1.34) are given in Table 1. The columns "NIS," "penGAM," and "SIS" summarize the results for the MMS based on 100 simulations for our proposed NIS method, the penalized method for the additive model of Meier, Geer, and Bühlmann (2009), and the linear correlation

Table 1. Minimum model size and robust estimate of standard deviations (in parentheses).

Model	NIS	penGAM	SIS
Example 1 ($s = 3$, SNR ≈ 1.01)	3 (0)	3 (0)	3 (0)
Example 1 ($s = 6$, SNR ≈ 1.99)	56 (0)	1000 (0)	56 (0)
Example 1 ($s = 12$, SNR ≈ 4.07)	66 (7)	1000 (0)	62 (1)
Example 1 ($s = 24$, SNR ≈ 8.20)	269 (134)	1000 (0)	109 (43)
Example 2 (SNR ≈ 0.83)	3 (0)	3 (0)	360 (361)

ranking method of Fan and Lv (2008). For Example 1, when the nonsparsity size $s > 5$, the irrepresentable condition required for the model selection consistency of LASSO fails. For these cases, penGAM fails to even include the true model until the last step. In contrast, our proposed NIS method performs reasonably well. It is also worth noting that SIS performs better than NIS in the first example, particularly for $s = 24$. This is due to the fact that the true model is linear and the covariates are jointly normally distributed, which implies that the marginal projection is linear as well. In this case, NIS selects variables from pd_n parameters, whereas SIS selects only from p parameters. However, for the nonlinear problem as in Example 2, both the nonlinear NIS method and penGAM behave nicely, whereas SIS fails badly even though the underlying true model is indeed linear.

5.2 Comparison of Model Selection and Estimation

As in the previous section, we set $n = 400$ and $p = 1000$ for all of the examples to demonstrate the power of our proposed INIS and g-INIS methods. Here, in the NIS step, we fix $d_n = 5$ as in the previous section. The number of simulations is 100. Here we use five-fold cross-validation in Step 2 of the INIS algorithm. For simplicity of notation, we let

$$g_1(x) = x, \qquad g_2(x) = (2x-1)^2, \qquad g_3(x) = \frac{\sin(2\pi x)}{2-\sin(2\pi x)},$$

and

$$g_4(x) = 0.1\sin(2\pi x) + 0.2\cos(2\pi x) + 0.3\sin(2\pi x)^2 + 0.4\cos(2\pi x)^3 + 0.5\sin(2\pi x)^3.$$

Example 3. Following Meier, Geer, and Bühlmann (2009), we generate the data from the following additive model:

$$Y = 5g_1(X_1) + 3g_2(X_2) + 4g_3(X_3) + 6g_4(X_4) + \sqrt{1.74}\varepsilon.$$

The covariates $X = (X_1, \ldots, X_p)^T$ are simulated according to the random-effects model

$$X_j = \frac{W_j + tU}{1+t}, \qquad j = 1, \ldots, p,$$

where $W_1, \ldots, W_p$ and U are iid Unif(0, 1), and $\varepsilon \sim \mathrm{N}(0, 1)$. When $t = 0$, the covariates are all independent, and when $t = 1$, the pairwise correlation of covariates is 0.5.

Example 4. Again, we adapt the simulation model of Meier, Geer, and Bühlmann (2009). This example is a more difficult case than Example 3, because it has 12 important variables with different coefficients

$$\begin{aligned} Y = {} & g_1(X_1) + g_2(X_2) + g_3(X_3) + g_4(X_4) \\ & + 1.5g_1(X_5) + 1.5g_2(X_6) + 1.5g_3(X_7) + 1.5g_4(X_8) \\ & + 2g_1(X_9) + 2g_2(X_{10}) + 2g_3(X_{11}) + 2g_4(X_{12}) \\ & + \sqrt{0.5184}\varepsilon, \end{aligned}$$

where $\varepsilon \sim \mathrm{N}(0, 1)$. The covariates are simulated as in Example 3.

Example 5. We follow the simulation model of Fan, Samworth, and Wu (2009), in which $Y = \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \varepsilon$ is simulated, where $\varepsilon \sim \mathrm{N}(0, 1)$. The covariates $X_1, \ldots, X_p$ are jointly Gaussian, marginally N(0, 1), and with $\mathrm{corr}(X_i, X_4) = 1/\sqrt{2}$ for all $i \neq 4$ and $\mathrm{corr}(X_i, X_j) = 1/2$ if i and j are distinct elements of $\{1, \ldots, p\} \setminus \{4\}$. The coefficients $\beta_1 = 2, \beta_2 = 2, \beta_3 = 2, \beta_4 = -3\sqrt{2}$, and $\beta_j = 0$ for $j > 4$ are taken so that X_4 is independent of Y, even though it is the most important variable in the joint model in terms of the regression coefficient.

For each example, we compare the performance of INIS-penGAM, g-INIS-penGAM proposed in this article, penGAM (Meier, Geer, and Bühlmann 2009), and ISIS-SCAD (Fan, Samworth, and Wu 2009), which aims for a sparse linear model. The results are given in Table 2, in which the true positives (TP), false positives (FP), prediction error (PE), and computation time (Time) are reported for each method. Here the prediction error is calculated on an independent test dataset of size $n/2$.

For the greedy modification, g-INIS-penGAM, the number of false-positive variables is approximately 1 for all examples. The number of false-positive variables for both INIS-penGAM and ISIS-SCAD is much smaller than that for penGAM. In terms of false positives, we can see that in Examples 3 and 4, INIS-penGAM and penGAM have similar performance, whereas penGAM misses one variable most of the time in Example 5. The linear method ISIS-SCAD misses important variables in the nonlinear models in Examples 3 and 4.

Note that in Example 4 ($t = 1$), even INIS and g-INIS miss more than one variable on average. To explore the reason for this, we closely examined the iterative process for this example and found that the variables X_1 and X_2 were missed quite often. The explanation for this is that although the overall signal-to-noise ratio (SNR) for this example is approximately 10.89, the individual contributions to the total signal vary significantly. We now introduce the notion of individual SNR. For example, $\mathrm{var}(m_1(X_1))/\mathrm{var}(\varepsilon)$ in the additive model

$$Y = m_1(X_1) + \cdots + m_p(X_p) + \varepsilon$$

is the individual SNR for the first component under the oracle model where $m_2, \ldots, m_p$ are known. In Example 4 ($t = 1$), the variances of all 12 components are as follows:

1	2	3	4	5	6	7	8	9	10	11	12
0.08	0.09	0.21	0.26	0.19	0.20	0.47	0.58	0.33	0.36	0.84	1.03

We can see that the variance varies significantly among the 12 components, which leads to very different marginal SNRs. For example, the individual SNR for the first component is merely $0.08/0.518 = 0.154$, th detecton of which is very challenging. With the overall SNR fixed, the individual SNRs play an important role in determining the difficulty of selecting individual variables.

From the perspective of the prediction error, INIS-penGAM, g-INIS-penGAM, and penGAM outperform ISIS-SCAD in the nonlinear models but perform worse than ISIS-SCAD in the

Table 2. Average values of the numbers of true positives (TP), false positives (FP), prediction error (PE), and time (in seconds). Robust standard deviations are given in parentheses

Model	Method	TP	FP	PE	Time
Example 3 ($t = 0$)	INIS	4.00 (0.00)	2.58 (2.24)	3.02 (0.34)	18.50 (7.22)
(SNR $\approx$ 9.02)	g-INIS	4.00 (0.00)	0.67 (0.75)	2.92 (0.30)	25.03 (4.87)
	penGAM	4.00 (0.00)	31.86 (23.51)	3.30 (0.40)	180.63 (6.92)
	ISIS	3.03 (0.00)	29.97 (0.00)	15.95 (1.74)	12.95 (4.18)
Example 3 ($t = 1$)	INIS	3.98 (0.00)	15.76 (6.72)	2.97 (0.39)	78.80 (26.91)
(SNR $\approx$ 7.58)	g-INIS	4.00 (0.00)	0.98 (1.49)	2.61 (0.26)	33.89 (9.99)
	penGAM	4.00 (0.00)	39.21 (24.63)	2.97 (0.28)	254.06 (13.06)
	ISIS	3.01 (0.00)	29.99 (0.00)	12.91 (1.39)	18.59 (4.37)
Example 4 ($t = 0$)	INIS	11.97 (0.00)	3.22 (1.49)	0.97 (0.11)	73.60 (25.77)
(SNR $\approx$ 8.67)	g-INIS	12.00 (0.00)	0.73 (0.75)	0.91 (0.10)	160.75 (19.94)
	penGAM	11.99 (0.00)	80.10 (18.28)	1.27 (0.14)	233.72 (10.25)
	ISIS	7.96 (0.75)	25.04 (0.75)	4.70 (0.40)	12.89 (5.00)
Example 4 ($t = 1$)	INIS	10.01 (1.49)	15.56 (0.93)	1.03 (0.13)	125.11 (39.99)
(SNR $\approx$ 10.89)	g-INIS	10.78 (0.75)	1.08 (1.49)	0.87 (0.11)	156.37 (28.58)
	penGAM	10.51 (0.75)	62.11 (26.31)	1.13 (0.12)	278.61 (16.93)
	ISIS	6.53 (0.75)	26.47 (0.75)	4.30 (0.44)	17.02 (4.01)
Example 5	INIS	3.99 (0.00)	21.96 (0.00)	1.62 (0.18)	94.50 (7.12)
(SNR $\approx$ 6.11)	g-INIS	4.00 (0.00)	1.04 (1.49)	1.16 (0.12)	39.78 (12.45)
	penGAM	3.00 (0.00)	195.03 (21.08)	1.93 (0.28)	1481.12 (181.93)
	ISIS	4.00 (0.00)	29.00 (0.00)	1.40 (0.17)	17.78 (3.85)

linear model of Example 5. Overall, the greedy modification g-INIS clearly is a competitive variable selection method in ultra-high-dimensional additive models with a very low false selection rate, small prediction error, and fast computation.

5.3 d_n and SNR

Here we report a simulation study conducted to investigate the performance of the INIS-penGAM estimator under different SNR settings using different numbers (d_n) of basis functions.

Example 6. We generate the data from the following additive model:

$$Y = 3g_1(X_1) + 3g_2(X_2) + 2g_3(X_3) + 2g_4(X_4) + C\sqrt{3.3843}\varepsilon,$$

where the covariates $X = (X_1, \ldots, X_p)^T$ are simulated according to Example 3. Here C takes a series of different values ($C^2 = 2, 1, 0.5, 0.25$) to make the corresponding SNR $= 0.5, 1, 2, 4$. The results of using different numbers of basis functions, $d_n = 2, 4, 6, 8$, are reported in Tables A.1 and A.2 in the Appendix.

From Table A.1 in the Appendix, in which all of the variables are independent, both methods have very good true positives under various SNRs when d_n is not too large. However, for the case of SNR $= 0.5$ and $d_n = 16$, INIS and penGAM perform poorly in terms of a low true-positive rate. This is due to the fact that when d_n is large, the estimation variance is large, which makes it difficult to differentiate the active variables from inactive variables when the signals are weak.

We now consider the more difficult case shown in Table A.2 (in the Appendix) where pairwise correlation between variables is 0.5. INIS has a competitive performance under various SNR values except when $d_n = 16$. When SNR $= 0.5$, we cannot achieve sure screening with the current sample size and configuration, for the aforementioned reasons.

5.4 An Analysis on Affymetric GeneChip Rat Genome 230 2.0 Array

We use the dataset reported by Scheetz et al. (2006) and analyzed by Huang, Horowitz, and Wei (2010) to illustrate an application of our proposed method. For this dataset, 120 12-week-old male rats were selected for harvesting of tissue from the eyes and subsequent microarray analysis. The microarrays used to analyze the RNA from the eyes of these animals contain more than 31,042 different probe sets (Affymetric GeneChip Rat Genome 230 2.0 Array). The intensity values were normalized using the robust multichip averaging method (Irizarry et al. 2003) method to obtain summary expression values for each probe set. Gene expression levels were analyzed on a logarithmic scale.

Following Huang, Horowitz, and Wei (2010), we were interested in finding the genes that are related to the TRIM32 gene, which was recently found to cause Bardet–Biedl syndrome (Chiang et al. 2006) and is a genetically heterogeneous disease of multiple organ systems, including the retina. Although more than 30,000 probe sets are represented on the Rat Genome 230 2.0 Array, many of these are not expressed in the eye tissue. We focused only on the 18,975 probes that are expressed in the eye tissue. We used our INIS-penGAM method directly on this dataset, where $n = 120$ and $p = 18{,}975$, and designated this method INIS-penGAM ($p = 18{,}975$). The fitted regression functions are shown in Figure 1. Direct application of penGAM approach on the whole dataset is too slow. Following Huang, Horowitz, and Wei (2010), we used 2000 probe sets that are expressed in the eye and have the greatest marginal correlation with TRIM32 in the analysis. On the subset of the data ($n = 120, p = 2000$), we applied INIS-penGAM

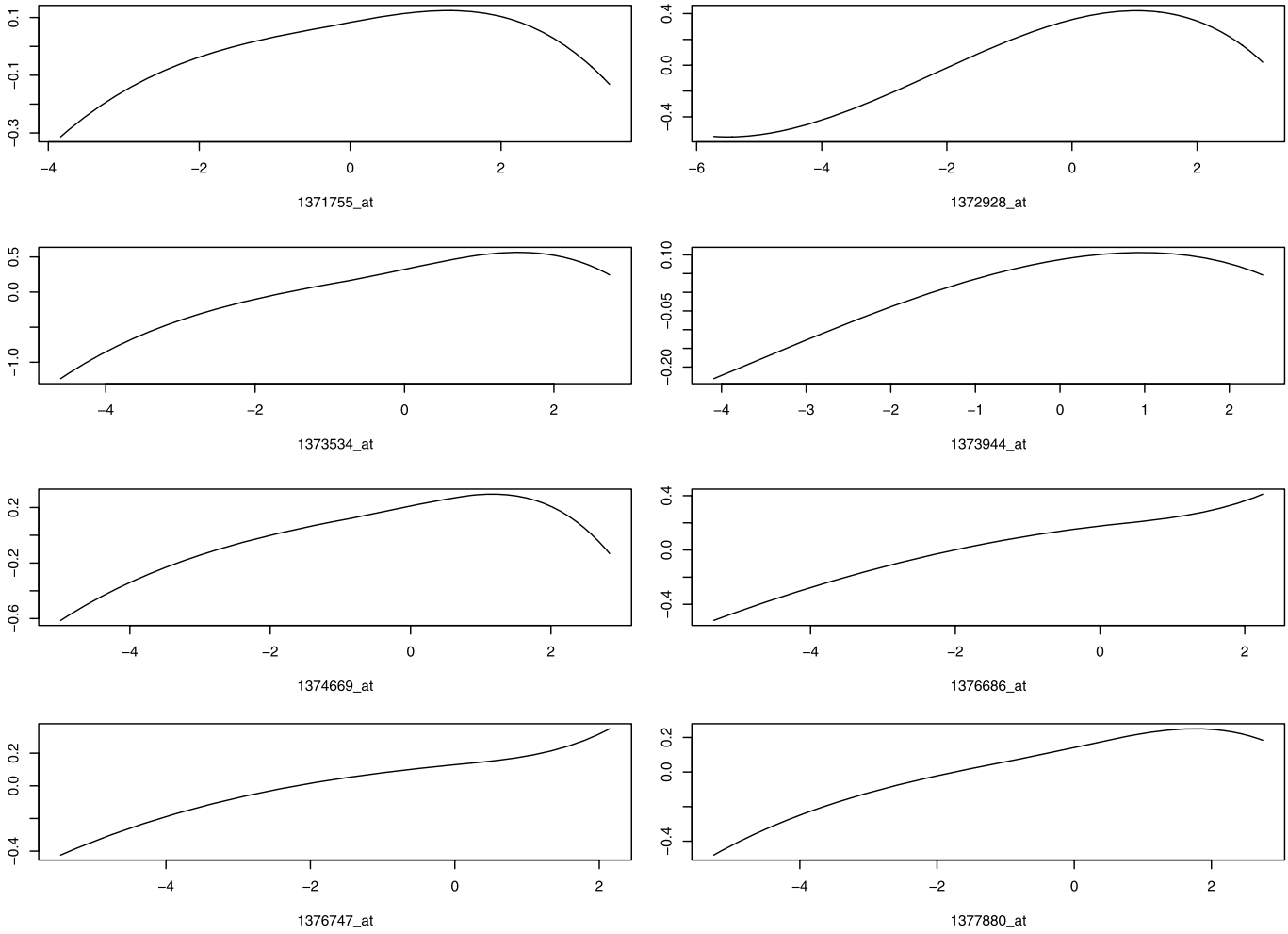

Figure 1. Fitted regression functions for the eight probes selected by INIS-penGAM ($p = 18{,}975$).

and penGAM to model the relationships between the expression of TRIM32 and expression of the 2000 genes. For simplicity, we did not implement g-INIS-penGAM. Before the analysis, we standardized each probe to be of mean 0 and variance 1. We now have three different estimators: INIS-penGAM ($p = 18{,}975$), INIS-penGAM ($p = 2000$), and penGAM ($p = 2000$). INIS-penGAM ($p = 18{,}975$) selects the following eight probes: 1371755_at, 1372928_at, 1373534_at, 1373944_at, 1374669_at, 1376686_at, 1376747_at, 1377880_at. INIS-penGAM ($p = 2000$) selects the following eight probes: 1376686_at, 1376747_at, 1378590_at, 1373534_at, 1377880_at, 1372928_at, 1374669_at, 1373944_at. In contrast, penGAM ($p = 2000$) selects 32 probes: The residual sum of squares (RSS) for these fittings are 0.24 for INIS-penGAM ($p = 18{,}975$), 0.26 for INIS-penGAM ($p = 2000$), and 0.1 for penGAM ($p = 2000$).

To further evaluate the performance of the two methods, we used cross-validation and compared the prediction mean squared error (PE). We randomly partitioned the data into a training set of 100 observations and a test set of 20 observations. We computed the number of probes selected using the 100 observations and the PEs on these 20 test sets. This process was repeated 100 times. Table 3 presents the average values and their associated robust standard deviations over 100 replications. As clearly shown in the table, the INIS-penGAM approach selects far fewer genes and has a smaller PE. Thus, in this example, INIS-penGAM provides the biological investigator with a more targeted list of probe sets, which could be very useful in subsequent studies.

6. DISCUSSION

In this article we have studied the NIS method for variable selection in additive models. We used B-spline basis functions for fitting the marginal nonparametric components. Our proposed marginal projection criteria represent an important extension of the marginal correlation. We also have proposed iterative NIS procedures in which variable selection and coefficient estimation can be achieved simultaneously. By applying the INIS-penGAM method, we can preserve the sure screening property

Table 3. Mean model size (MS) and prediction error (PE) over 100 repetitions and their robust standard deviations (in parentheses) for INIS ($p = 18{,}975$), INIS ($p = 2000$), and penGAM ($p = 2000$)

Method	MS	PE
INIS ($p = 18975$)	7.73 (0.00)	0.47 (0.13)
INIS ($p = 2000$)	7.68 (0.75)	0.44 (0.15)
penGAM ($p = 2000$)	26.71 (14.93)	0.48 (0.16)

and substantially reduce the false selection rate. We have proposed a greedy modification of the g-INIS-penGAM method to further reduce the false selection rate. Moreover, this method can deal with the case where some variable is marginally uncorrelated but jointly correlated with the response. The proposed method can be easily generalized to generalized additive models with appropriate conditions.

Given that in this work the additive components are specifically approximated by truncated series expansions with B-spline bases, the theoretical results should hold in general, and the proposed framework can be readily adapted to other smoothing methods with additive models (Silverman 1984; Horowitz, Klemelä, and Mammen 2006), such as local polynomial regression (Fan and Jiang 2005), wavelet approximations (Antoniadis and Fan 2001; Sardy and Tseng 2004), and smoothing splines (Speckman 1985). This is an interesting topic for future research.

7. PROOFS

Proof of Lemma 1. By the least squares property, $E(Y - f_{nj})f_{nj} = 0$ and $E(Y - f_j)f_{nj} = 0$. Therefore,

$$Ef_{nj}(f_j - f_{nj}) = E(Y - f_{nj})f_{nj} - E(Y - f_j)f_{nj} = 0.$$

It follows from this and the orthogonal decomposition $f_j = f_{nj} + (f_j - f_{nj})$ that

$$\|f_{nj}\|^2 = \|f_j\|^2 - \|f_j - f_{nj}\|^2.$$

The desired result follows from condition C together with Fact 1.

The following types of Bernstein's inequality in van der Vaart and Wellner (1996) are needed:

Lemma 2 (Bernstein's inequality, lemma 2.2.9, van der Vaart and Wellner 1996). For independent random variables $Y_1, \ldots, Y_n$ with bounded ranges $[-M, M]$ and 0 means,

$$P(|Y_1 + \cdots + Y_n| > x) \le 2\exp\{-x^2/(2(v + Mx/3))\}$$

for $v \ge \operatorname{var}(Y_1 + \cdots + Y_n)$.

Lemma 3 (Bernstein's inequality, lemma 2.2.11, van der Vaart and Wellner 1996). Let $Y_1, \ldots, Y_n$ be independent random variables with 0 mean such that $E|Y_i|^m \le m!M^{m-2}v_i/2$, for every $m \ge 2$ (and all i) and some constants M and v_i. Then

$$P(|Y_1 + \cdots + Y_n| > x) \le 2\exp\{-x^2/(2(v + Mx))\}$$

for $v \ge v_1 + \cdots + v_n$.

The following lemmas are needed to prove Theorem 1.

Lemma 4. Under conditions A, B, and D, for any $\delta > 0$, there exist some positive constants c_6 and c_7 such that

$$P\big(|(\mathbb{P}_n - E)\Psi_{jk}Y| \ge \delta n^{-1}\big) \le 4\exp(-\delta^2/2(c_6 n d_n^{-1} + c_7\delta))$$

for $k = 1, \ldots, d_n, j = 1, \ldots, p$.

Proof. Write $T_{jki} = \Psi_{jk}(X_{ij})Y_i - E\Psi_{jk}(X_{ij})Y_i$. Because $Y_i = m(\mathbf{X}_i) + \varepsilon_i$, we can write $T_{jki} = T_{jki1} + T_{jki2}$, where

$$T_{jki1} = \Psi_{jk}(X_{ij})m(\mathbf{X}_i) - E\Psi_{jk}(X_{ij})m(\mathbf{X}_i),$$

and $T_{jki2} = \Psi_{jk}(X_{ij})\varepsilon_i$.

By conditions A, B, D, and Fact 2, recalling that $\|\Psi_{jk}\|_\infty \le 1$, we have

$$\begin{aligned} |T_{jki1}| &\le 2B_1, \\ \operatorname{var}(T_{jki1}) &\le E\Psi_{jk}^2(X_{ij})m_i(X_{ij})^2 \le B_1^2 C_2 d_n^{-1}. \end{aligned} \tag{12}$$

By Bernstein's inequality (Lemma 2), for any $\delta_1 > 0$,

$$P\left(\left|\sum_{i=1}^n T_{jki1}\right| > \delta_1\right) \le 2\exp\left(-\frac{1}{2}\frac{\delta_1^2}{nB_1^2C_2d_n^{-1} + 2B_1\delta_1/3}\right). \tag{13}$$

We next bound the tails of T_{jki2}. For every $r \ge 2$,

$$\begin{aligned} E|T_{jki2}|^r &\le E|\Psi_{jk}(X_{ij})|^2 E(|\varepsilon_i|^r|\mathbf{X}_i) \\ &\le r!B_2^{-r}E|\Psi_{jk}(X_{ij})|^2 E\exp(B_2|\varepsilon_i||\mathbf{X}_i) \\ &\le B_3C_2d_n^{-1}r!B_2^{-r}, \end{aligned}$$

where the last inequality uses condition E and Fact 2. By Bernstein's inequality (Lemma 3), for any $\delta_2 > 0$,

$$P\left(\left|\sum_{i=1}^n T_{jki2}\right| > \delta_2\right) \le 2\exp\left(-\frac{1}{2}\frac{\delta_2^2}{2nB_2^{-2}B_3C_2d_n^{-1} + B_2^{-1}\delta_2}\right). \tag{14}$$

Combining (13) and (14), the desired result follows by taking $c_6 = \max(B_1^2C_2, 2B_2^{-2}B_3C_2)$ and $c_7 = \max(2/3B_1, B_2^{-1})$.

Throughout the rest of the proof, for any matrix $\mathbf{A}$, let $\|\mathbf{A}\| = \sqrt{\lambda_{\max}(\mathbf{A}^T\mathbf{A})}$ be the operator norm and $\|\mathbf{A}\|_\infty = \max_{i,j}|A_{ij}|$ be the infinity norm. The next lemma is about the tail probability of the eigenvalues of the design matrix.

Lemma 5. Under conditions A and B, for any $\delta > 0$,

$$\begin{aligned} &P\big(|\lambda_{\min}(\mathbb{P}_n\Psi_j\Psi_j^T) - \lambda_{\min}(E\Psi_j\Psi_j^T)| \ge d_n\delta/n\big) \\ &\qquad \le 2d_n^2\exp\left\{-\frac{1}{2}\frac{\delta^2}{C_2nd_n^{-1} + \delta/3}\right\}. \end{aligned}$$

In addition, for any given constant c_4, there exists some positive constant c_8 such that

$$\begin{aligned} &P\big\{\big|\|(\mathbb{P}_n\Psi_j\Psi_j^T)^{-1}\| - \|(E\Psi_j\Psi_j^T)^{-1}\|\big| \ge c_8\|(E\Psi_j\Psi_j^T)^{-1}\|\big\} \\ &\qquad \le 2d_n^2\exp(-c_4nd_n^{-3}). \end{aligned} \tag{15}$$

Proof. For any symmetric matrices $\mathbf{A}$ and $\mathbf{B}$ and any $\|\mathbf{x}\| = 1$, where $\|\cdot\|$ is the Euclidean norm,

$$\mathbf{x}^T(\mathbf{A} + \mathbf{B})x = \mathbf{x}^T\mathbf{A}\mathbf{x} + \mathbf{x}^T\mathbf{B}\mathbf{x} \ge \min_{\|\mathbf{x}\|=1}\mathbf{x}^T\mathbf{A}\mathbf{x} + \min_{\|\mathbf{x}\|=1}\mathbf{x}^T\mathbf{B}\mathbf{x}.$$

Taking the minimum among $\|\mathbf{x}\| = 1$ on the left side, we have

$$\min_{\|\mathbf{x}\|=1}\mathbf{x}^T(\mathbf{A} + \mathbf{B})\mathbf{x} \ge \min_{\|\mathbf{x}\|=1}\mathbf{x}^T\mathbf{A}\mathbf{x} + \min_{\|\mathbf{x}\|=1}\mathbf{x}^T\mathbf{B}\mathbf{x},$$

which is equivalent to $\lambda_{\min}(\mathbf{A} + \mathbf{B}) \ge \lambda_{\min}(\mathbf{A}) + \lambda_{\min}(\mathbf{B})$.

We then have

$$\lambda_{\min}(\mathbf{A}) \ge \lambda_{\min}(\mathbf{B}) + \lambda_{\min}(\mathbf{A} - \mathbf{B}),$$

which is the same as

$$\lambda_{\min}(\mathbf{A}-\mathbf{B}) \le \lambda_{\min}(\mathbf{A}) - \lambda_{\min}(\mathbf{B}).$$

By switching the roles of $\mathbf{A}$ and $\mathbf{B}$, we also have

$$\lambda_{\min}(\mathbf{B}-\mathbf{A}) \le \lambda_{\min}(\mathbf{B}) - \lambda_{\min}(\mathbf{A}).$$

In other words,

$$|\lambda_{\min}(\mathbf{A}) - \lambda_{\min}(\mathbf{B})| \le \max\bigl\{|\lambda_{\min}(\mathbf{A}-\mathbf{B})|, |\lambda_{\min}(\mathbf{B}-\mathbf{A})|\bigr\}. \quad (16)$$

Let $\mathbf{D}_j = \mathbb{P}_n\boldsymbol{\Psi}_j\boldsymbol{\Psi}_j^T - E\boldsymbol{\Psi}_j\boldsymbol{\Psi}_j^T$. It then follows from (16) that

$$|\lambda_{\min}(\mathbb{P}_n\boldsymbol{\Psi}_j\boldsymbol{\Psi}_j^T) - \lambda_{\min}(E\boldsymbol{\Psi}_j\boldsymbol{\Psi}_j^T)| \le \max\bigl\{|\lambda_{\min}(\mathbf{D}_j)|, |\lambda_{\min}(-\mathbf{D}_j)|\bigr\}. \quad (17)$$

We now bound the right side of (17). Let $\mathbf{D}_j^{(i,l)}$ be the (i,l) entry of $\mathbf{D}_j$. Then it is easy to see that for any $\|\mathbf{x}\| = 1$,

$$|\mathbf{x}^T\mathbf{D}_j\mathbf{x}| \le \|\mathbf{D}_j\|_\infty\left(\sum_{i=1}^{d_n}|x_i|\right)^2 \le d_n\|\mathbf{D}_j\|_\infty. \quad (18)$$

Thus,

$$\lambda_{\min}(\mathbf{D}_j) = \min_{\|\mathbf{x}\|=1}\mathbf{x}^T\mathbf{D}_j\mathbf{x} \le d_n\|\mathbf{D}_j\|_\infty.$$

On the other hand, using (18) again, we have

$$\lambda_{\min}(\mathbf{D}_j) = -\max_{\|\mathbf{x}\|=1}(-\mathbf{x}^T\mathbf{D}_j\mathbf{x}) \ge -d_n\|\mathbf{D}_j\|_\infty.$$

We conclude that

$$|\lambda_{\min}(\mathbf{D}_j)| \le d_n\|\mathbf{D}_j\|_\infty.$$

The same bound on $|\lambda_{\min}(-\mathbf{D}_j)|$ can be obtained using the same argument. Thus, by (17), we have

$$|\lambda_{\min}(\mathbb{P}_n\boldsymbol{\Psi}_j\boldsymbol{\Psi}_j^T) - \lambda_{\min}(E\boldsymbol{\Psi}_j\boldsymbol{\Psi}_j^T)| \le d_n\|\mathbf{D}_j\|_\infty. \quad (19)$$

We now use Bernstein's inequality to bound the right side of (19). Because $\|\Psi_{jk}\|_\infty \le 1$, and using Fact 2, we have that

$$\operatorname{var}(\Psi_{jk}(X_j)\Psi_{jl}(X_j)) \le E\Psi_{jk}^2(X_j)\Psi_{jl}^2(X_j) \le E\Psi_{jk}^2(X_j) \le C_2d_n^{-1}.$$

By Bernstein's inequality (Lemma 2), for any $\delta > 0$,

$$P\bigl(|(\mathbb{P}_n - E)\Psi_{jk}(X_j)\Psi_{jl}(X_j)| > \delta/n\bigr) \le 2\exp\left\{-\frac{\delta^2}{2(C_2nd_n^{-1} + \delta/3)}\right\}. \quad (20)$$

It follows from (19), (20), and the union bound of probability that

$$P\bigl(|\lambda_{\min}(\mathbb{P}_n\boldsymbol{\Psi}_j\boldsymbol{\Psi}_j^T) - \lambda_{\min}(E\boldsymbol{\Psi}_j\boldsymbol{\Psi}_j^T)| \ge d_n\delta/n\bigr) \le 2d_n^2\exp\left\{-\frac{\delta^2}{2(C_2nd_n^{-1} + \delta/3)}\right\}.$$

This completes the proof of the first inequality.

To prove the second inequality, take $\delta = c_9D_1nd_n^{-2}$ in (20), where $c_9 \in (0,1)$. Recalling Fact 3, it follows that

$$P\bigl(|\lambda_{\min}(\mathbb{P}_n\boldsymbol{\Psi}_j\boldsymbol{\Psi}_j^T) - \lambda_{\min}(E\boldsymbol{\Psi}_j\boldsymbol{\Psi}_j^T)| \ge c_9\lambda_{\min}(E\boldsymbol{\Psi}_j\boldsymbol{\Psi}_j^T)\bigr) \le 2d_n^2\exp(-c_4nd_n^{-3}) \quad (21)$$

for some positive constant c_4. The second part of the lemma thus follows from the fact that $\lambda_{\min}(\mathbf{H})^{-1} = \lambda_{\max}(\mathbf{H}^{-1})$, if we establish

$$P\bigl(\bigl|\{\lambda_{\min}(\mathbb{P}_n\boldsymbol{\Psi}_j\boldsymbol{\Psi}_j^T)\}^{-1} - \{\lambda_{\min}(E\boldsymbol{\Psi}_j\boldsymbol{\Psi}_j^T)\}^{-1}\bigr| \ge c_8\{\lambda_{\min}(E\boldsymbol{\Psi}_j\boldsymbol{\Psi}_j^T)\}^{-1}\bigr) \le 2d_n^2\exp(-c_4nd_n^{-3}), \quad (22)$$

by using (21), where $c_8 = 1/(1-c_9) - 1$.

We now deduce (22) from (21). Let $A = \lambda_{\min}(\mathbb{P}_n\boldsymbol{\Psi}_j\boldsymbol{\Psi}_j^T)$ and $B = \lambda_{\min}(E\boldsymbol{\Psi}_j\boldsymbol{\Psi}_j^T)$. Then $A > 0$ and $B > 0$. We aim to show that for $a \in (0,1)$,

$$|A^{-1} - B^{-1}| \ge cB^{-1} \quad \text{implies} \quad |A - B| \ge aB,$$

where $c = 1/(1-a) - 1$.

Because

$$|A^{-1} - B^{-1}| \ge (1/(1-a) - 1)B^{-1},$$

we have

$$A^{-1} - B^{-1} \le -(1/(1-a) - 1)B^{-1} \quad \text{or} \quad \ge (1/(1-a) - 1)B^{-1}.$$

Note that for $a \in (0,1)$, we have $1 - 1/(1+a) < 1/(1-a) - 1$. It then follows that

$$A^{-1} - B^{-1} \le -(1 - 1/(1+a))B^{-1} \quad \text{or} \quad \ge (1/(1-a) - 1)B^{-1},$$

which is equivalent to $|A - B| \ge aB$.

This concludes the proof of the lemma.

Proof of Theorem 1. We first prove part (a). Recall that

$$\|\hat{f}_{nj}\|_n^2 = (\mathbb{P}_n\boldsymbol{\Psi}_jY)^T(\mathbb{P}_n\boldsymbol{\Psi}_j\boldsymbol{\Psi}_j^T)^{-1}\mathbb{P}_n\boldsymbol{\Psi}_jY$$

and

$$\|f_{nj}\|^2 = (E\boldsymbol{\Psi}_jY)^T(E\boldsymbol{\Psi}_j\boldsymbol{\Psi}_j^T)^{-1}E\boldsymbol{\Psi}_jY.$$

Let $\mathbf{a}_n = \mathbb{P}_n\boldsymbol{\Psi}_jY$, $\mathbf{B}_n = (\mathbb{P}_n\boldsymbol{\Psi}_j\boldsymbol{\Psi}_j^T)^{-1}$, $\mathbf{a} = E\boldsymbol{\Psi}_jY$, and $\mathbf{B} = (E\boldsymbol{\Psi}_j\boldsymbol{\Psi}_j^T)^{-1}$. By some algebra,

$$\mathbf{a}_n^T\mathbf{B}_n\mathbf{a}_n - \mathbf{a}^T\mathbf{B}\mathbf{a} = (\mathbf{a}_n - \mathbf{a})^T\mathbf{B}_n(\mathbf{a}_n - \mathbf{a}) + 2(\mathbf{a}_n - \mathbf{a})^T\mathbf{B}_n\mathbf{a} + \mathbf{a}_n^T(\mathbf{B}_n - \mathbf{B})\mathbf{a},$$

we have

$$\|\hat{f}_{nj}\|_n^2 - \|f_{nj}\|^2 = S_1 + S_2 + S_3, \quad (23)$$

where

$$S_1 = (\mathbb{P}_n\boldsymbol{\Psi}_jY - E\boldsymbol{\Psi}_jY)^T(\mathbb{P}_n\boldsymbol{\Psi}_j\boldsymbol{\Psi}_j^T)^{-1}(\mathbb{P}_n\boldsymbol{\Psi}_jY - E\boldsymbol{\Psi}_jY),$$

$$S_2 = 2(\mathbb{P}_n\boldsymbol{\Psi}_jY - E\boldsymbol{\Psi}_jY)^T(\mathbb{P}_n\boldsymbol{\Psi}_j\boldsymbol{\Psi}_j^T)^{-1}E\boldsymbol{\Psi}_jY,$$

$$S_3 = (E\boldsymbol{\Psi}_jY)^T((\mathbb{P}_n\boldsymbol{\Psi}_j\boldsymbol{\Psi}_j^T)^{-1} - (E\boldsymbol{\Psi}_j\boldsymbol{\Psi}_j^T)^{-1})E\boldsymbol{\Psi}_jY.$$

Note that

$$S_1 \le \|(\mathbb{P}_n\boldsymbol{\Psi}_j\boldsymbol{\Psi}_j^T)^{-1}\| \cdot \|\mathbb{P}_n\boldsymbol{\Psi}_jY - E\boldsymbol{\Psi}_jY\|^2. \quad (24)$$

By Lemma 4 and the union bound of probability,

$$P(\|\mathbb{P}_n\boldsymbol{\Psi}_jY - E\boldsymbol{\Psi}_jY\|^2 \ge d_n\delta^2n^{-2}) \le 4d_n\exp(-\delta^2/2(c_6nd_n^{-1} + c_7\delta)). \quad (25)$$

Recall the result in Lemma 5 that, for any given constant c_4, there exists a positive constant c_8 such that

$$P\{|\|(\mathbb{P}_n\Psi_j\Psi_j^T)^{-1}\| - \|(E\Psi_j\Psi_j^T)^{-1}\|| \ge c_8\|(E\Psi_j\Psi_j^T)^{-1}\|\} \le 2d_n^2\exp(-c_4nd_n^{-3}).$$

Because by Fact 3,

$$\|(E\Psi_j\Psi_j^T)^{-1}\| \le D_1^{-1}d_n,$$

it follows that

$$P\{\|(\mathbb{P}_n\Psi_j\Psi_j^T)^{-1}\| \ge (c_8+1)D_1^{-1}d_n\} \le 2d_n^2\exp(-c_4nd_n^{-3}). \quad (26)$$

Combining (24)–(26) and the union bound of probability, we have

$$P(S_1 \ge (c_8+1)D_1^{-1}d_n^2\delta^2/n^2) \le 4d_n\exp(-\delta^2/2(c_6nd_n^{-1}+c_7\delta)) + 2d_n^2\exp(-c_4nd_n^{-3}). \quad (27)$$

To bound S_2, we note that

$$|S_2| \le 2\|\mathbb{P}_n\Psi_jY - E\Psi_jY\|\cdot\|(\mathbb{P}_n\Psi_j\Psi_j^T)^{-1}E\Psi_jY\| \le 2\|\mathbb{P}_n\Psi_jY - E\Psi_jY\|\cdot\|(\mathbb{P}_n\Psi_j\Psi_j^T)^{-1}\|\cdot\|E\Psi_jY\|. \quad (28)$$

Because by condition D,

$$\|E\Psi_jY\|^2 = \sum_{k=1}^{d_n}(E\Psi_{jk}Y)^2 = \sum_{k=1}^{d_n}(E\Psi_{jk}m)^2 \le \sum_{k=1}^{d_n}B_1^2E\Psi_{jk}^2 \le B_1^2C_2, \quad (29)$$

it follows from (25), (26), (28), (29), and the union bound of probability that

$$P(|S_2| \ge 2(c_8+1)D_1^{-1}C_2^{1/2}B_1d_n^{3/2}\delta/n) \le 4d_n\exp(-\delta^2/2(c_6nd_n^{-1}+c_7\delta)) + 2d_n^2\exp(-c_4nd_n^{-3}). \quad (30)$$

We now bound S_3. Note that

$$S_3 = (E\Psi_jY)^T(\mathbb{P}_n\Psi_j\Psi_j^T)^{-1}\times(E-\mathbb{P}_n)\Psi_j\Psi_j^T(E\Psi_j\Psi_j^T)^{-1}E\Psi_jY. \quad (31)$$

By the fact that $\|\mathbf{AB}\| \le \|\mathbf{A}\|\cdot\|\mathbf{B}\|$, we have

$$|S_3| \le \|(\mathbb{P}_n-E)\Psi_j\Psi_j^T\|\cdot\|(\mathbb{P}_n\Psi_j\Psi_j^T)^{-1}\|\cdot\|(E\Psi_j\Psi_j^T)^{-1}\|\cdot\|E\Psi_jY\|^2. \quad (32)$$

For any $\|\mathbf{x}\| = 1$ and d_n-dimensional square matrix $\mathbf{D}$,

$$\mathbf{x}^T\mathbf{D}^T\mathbf{D}\mathbf{x} = \sum_i\Big(\sum_j d_{ij}x_j\Big)^2 \le \|\mathbf{D}\|_\infty^2 d_n\Big(\sum_{j=1}^{d_n}|x_i|\Big)^2 \le d_n^2\|\mathbf{D}\|_\infty^2.$$

Thus $\|\mathbf{D}\| \le d_n\|\mathbf{D}\|_\infty$. We conclude that

$$\|(\mathbb{P}_n-E)\Psi_j\Psi_j^T\| \le d_n\|(\mathbb{P}_n-E)\Psi_j\Psi_j^T\|_\infty. \quad (33)$$

By (20), (26), (29), (32), (33), and the union bound of probability, it follows that

$$P(|S_3| \ge (c_8+1)D_1^{-2}B_1^2C_2d_n^3\delta/n) \le 2d_n^2\exp(-\delta^2/2(c_6nd_n^{-1}+c_7\delta)) + 2d_n^2\exp(-c_4nd_n^{-3}). \quad (34)$$

It follows from (23), (27), (30), (34), and the union bound of probability that for some positive constants c_{10}, c_{11}, and c_{12},

$$P(|\|\hat{f}_{nj}\|_n^2 - \|f_{nj}\|^2| \ge c_{10}d_n^2\delta^2/n^2 + c_{11}d_n^{3/2}\delta/n + c_{12}d_n^3\delta/n) \le (8d_n+2d_n^2)\exp(-\delta^2/2(c_6nd_n^{-1}+c_7\delta)) + 6d_n^2\exp(-c_4nd_n^{-3}). \quad (35)$$

In (35), let $c_{10}d_n^2\delta^2/n^2 + c_{11}d_n^{3/2}\delta/n + c_{12}d_n^3\delta/n = c_2d_nn^{-2\kappa}$ for any given $c_2 > 0$ (i.e., taking $\delta = n^{1-2\kappa}d_n^{-2}c_2/c_{12}$), there exist some positive constants c_3 and c_4 such that

$$P(|\|\hat{f}_{nj}\|_n^2 - \|f_{nj}\|^2| \ge c_2d_nn^{-2\kappa}) \le (8d_n+2d_n^2)\exp(-c_3n^{1-4\kappa}d_n^{-3}) + 6d_n^2\exp(-c_4nd_n^{-3}).$$

The first part thus follows the union bound of probability.

To prove the second part, note that on the event

$$A_n \equiv \Big\{\max_{j\in\mathcal{M}_\star}|\|\hat{f}_{nj}\|_n^2 - \|f_{nj}\|^2| \le c_1\xi d_nn^{-2\kappa}/2\Big\},$$

by Lemma 1, we have

$$\|\hat{f}_{nj}\|_n^2 \ge c_1\xi d_nn^{-2\kappa}/2 \quad \text{for all } j\in\mathcal{M}_\star. \quad (36)$$

Thus, by the choice of ν_n, we have $\mathcal{M}_\star \subset \widehat{\mathcal{M}}_{\nu_n}$. The result now follows from a simple union bound:

$$P(A_n^c) \le s_n\{(8d_n+2d_n^2)\exp(-c_3n^{1-4\kappa}d_n^{-3}) + 6d_n^2\exp(-c_4nd_n^{-3})\}.$$

This completes the proof.

Proof of Theorem 2. The key idea of the proof is to show that

$$\|E\Psi Y\|^2 = O(\lambda_{\max}(\mathbf{\Sigma})). \quad (37)$$

If so, by definition and $\|\Psi_{jk}\|_\infty \le 1$, we have

$$\sum_{j=1}^{p_n}\|f_{nj}\|^2 \le \max_{1\le j\le p_n}\lambda_{\max}\{(E\Psi_j\Psi_j^T)^{-1}\}\|E\Psi Y\|^2 = O(d_n\lambda_{\max}(\mathbf{\Sigma})).$$

This implies that the number of $\{j:\|f_{nj}\|^2 > \varepsilon d_nn^{-2\kappa}\}$ cannot exceed $O(n^{2\kappa}\lambda_{\max}(\mathbf{\Sigma}))$ for any $\varepsilon > 0$. Thus, on the set

$$B_n = \Big\{\max_{1\le j\le p_n}|\|\hat{f}_{nj}\|_n^2 - \|f_{nj}\|^2| \le \varepsilon d_nn^{-2\kappa}\Big\},$$

the number of $\{j:\|\hat{f}_{nj}\|_n^2 > 2\varepsilon d_nn^{-2\kappa}\}$ cannot exceed the number of $\{j:\|f_{nj}\|^2 > \varepsilon d_nn^{-2\kappa}\}$, which is bounded by $O\{n^{2\kappa}\lambda_{\max}(\mathbf{\Sigma})\}$.

By taking $\varepsilon = c_5/2$, we have

$$P\big[\big|\widehat{\mathcal{M}}_{\nu_n}\big| \le O\{n^{2\kappa}\lambda_{\max}(\boldsymbol{\Sigma})\}\big] \ge P(B_n).$$

The conclusion follows from Theorem 1(i).

It remains to prove (37). Note that (37) is more related to the joint regression rather than the marginal regression. Let

$$\boldsymbol{\alpha}_n = \arg\min_{\boldsymbol{\alpha}} E(Y - \boldsymbol{\Psi}^T\boldsymbol{\alpha})^2,$$

which is the joint regression coefficients in the population. By the score equation of $\boldsymbol{\alpha}_n$, we get

$$E\boldsymbol{\Psi}(Y - \boldsymbol{\Psi}^T\boldsymbol{\alpha}_n) = 0.$$

Thus

$$\|E\boldsymbol{\Psi}Y\|^2 = \boldsymbol{\alpha}_n^T E\boldsymbol{\Psi}\boldsymbol{\Psi}^T E\boldsymbol{\Psi}\boldsymbol{\Psi}^T\boldsymbol{\alpha}_n \le \lambda_{\max}(\boldsymbol{\Sigma})\boldsymbol{\alpha}_n^T E\boldsymbol{\Psi}\boldsymbol{\Psi}^T\boldsymbol{\alpha}_n.$$

It now follows from the orthogonal decomposition that

$$\operatorname{var}(Y) = \operatorname{var}(\boldsymbol{\Psi}^T\boldsymbol{\alpha}_n) + \operatorname{var}(Y - \boldsymbol{\Psi}^T\boldsymbol{\alpha}_n).$$

Because $\operatorname{var}(Y) = O(1)$, we conclude that $\operatorname{var}(\boldsymbol{\Psi}^T\boldsymbol{\alpha}_n) = O(1)$, that is,

$$\boldsymbol{\alpha}_n^T E\boldsymbol{\Psi}\boldsymbol{\Psi}^T\boldsymbol{\alpha}_n = O(1).$$

This completes the proof.

APPENDIX: TABLES FOR SIMULATION RESULTS OF SECTION 5.3

Table A.1. Average values of the numbers of TP, FP, PE, and computation time for Example 6 ($t = 0$). Robust standard deviations are given in parentheses

SNR	d_n	Method	TP	FP	PE	Time
0.5	2	INIS	3.96 (0.00)	2.28 (1.49)	7.74 (0.79)	16.09 (5.32)
		penGAM	4.00 (0.00)	27.85 (16.98)	8.07 (0.92)	354.46 (31.48)
	4	INIS	3.93 (0.00)	2.29 (1.68)	7.90 (0.81)	21.68 (8.95)
		penGAM	3.99 (0.00)	25.61 (13.62)	8.21 (0.84)	421.17 (35.71)
	8	INIS	3.81 (0.00)	2.59 (2.24)	8.16 (1.08)	33.10 (15.79)
		penGAM	3.95 (0.00)	34.59 (20.34)	8.49 (0.82)	484.17 (179.70)
	16	INIS	3.38 (0.75)	2.02 (1.49)	8.60 (1.13)	42.69 (20.13)
		penGAM	3.74 (0.00)	33.48 (23.88)	9.04 (0.93)	685.97 (267.43)
1.0	2	INIS	4.00 (0.00)	2.16 (2.24)	3.98 (0.34)	16.03 (5.74)
		penGAM	4.00 (0.00)	26.51 (14.18)	4.20 (0.46)	284.85 (20.30)
	4	INIS	4.00 (0.00)	2.08 (1.49)	3.97 (0.45)	20.80 (8.57)
		penGAM	4.00 (0.00)	28.33 (15.49)	4.24 (0.47)	362.02 (81.43)
	8	INIS	4.00 (0.00)	2.72 (2.24)	4.04 (0.43)	35.79 (18.38)
		penGAM	4.00 (0.00)	36.50 (21.83)	4.37 (0.47)	427.60 (152.53)
	16	INIS	4.00 (0.00)	1.80 (1.49)	4.26 (0.45)	46.81 (21.47)
		penGAM	4.00 (0.00)	38.60 (19.78)	4.80 (0.57)	595.87 (197.06)
2.0	2	INIS	4.00 (0.00)	2.03 (2.24)	2.12 (0.17)	15.92 (5.42)
		penGAM	4.00 (0.00)	25.89 (13.06)	2.25 (0.24)	235.69 (13.32)
	4	INIS	4.00 (0.00)	2.38 (2.24)	2.06 (0.22)	23.54 (9.08)
		penGAM	4.00 (0.00)	30.37 (17.16)	2.21 (0.26)	341.13 (19.44)
	8	INIS	4.00 (0.00)	2.79 (2.24)	2.03 (0.21)	38.56 (19.58)
		penGAM	4.00 (0.00)	38.51 (16.42)	2.24 (0.26)	396.84 (20.51)
	16	INIS	4.00 (0.00)	1.77 (1.49)	2.17 (0.25)	48.40 (24.65)
		penGAM	4.00 (0.00)	42.58 (16.60)	2.54 (0.30)	540.89 (165.39)
4.0	2	INIS	4.00 (0.00)	2.06 (2.24)	1.19 (0.13)	17.74 (6.42)
		penGAM	4.00 (0.00)	28.57 (14.37)	1.27 (0.15)	213.43 (12.09)
	4	INIS	4.00 (0.00)	2.33 (1.49)	1.09 (0.10)	23.28 (9.37)
		penGAM	4.00 (0.00)	30.75 (17.35)	1.18 (0.14)	300.69 (12.21)
	8	INIS	4.00 (0.00)	2.88 (2.24)	1.02 (0.12)	39.21 (19.17)
		penGAM	4.00 (0.00)	40.51 (17.54)	1.14 (0.14)	340.06 (11.49)
	16	INIS	4.00 (0.00)	1.72 (1.49)	1.10 (0.12)	49.79 (25.78)
		penGAM	4.00 (0.00)	45.77 (19.03)	1.33 (0.16)	481.19 (141.51)

Table A.2. Average values of the numbers of TP, FP, PE, and computation time (Time) for Example 6 ($t = 1$). Robust standard deviations are given in parentheses

SNR	d_n	Method	TP	FP	PE	Time
0.5	2	INIS	3.35 (0.75)	33.67 (8.96)	9.49 (1.28)	196.87 (91.48)
		penGAM	3.10 (0.00)	17.74 (15.11)	7.92 (0.89)	1107.78 (385.95)
	4	INIS	3.02 (0.00)	20.22 (2.43)	8.70 (1.14)	109.51 (56.11)
		penGAM	2.78 (0.00)	15.91 (10.07)	7.99 (0.91)	734.08 (227.55)
	8	INIS	2.51 (0.75)	10.48 (0.75)	8.37 (0.89)	65.12 (16.64)
		penGAM	2.59 (0.75)	16.47 (9.70)	8.13 (0.90)	624.31 (56.23)
	16	INIS	2.10 (0.00)	4.47 (0.75)	8.44 (1.00)	46.84 (15.61)
		penGAM	2.41 (0.75)	15.56 (10.63)	8.42 (0.97)	786.45 (244.02)
1.0	2	INIS	3.83 (0.00)	32.46 (9.70)	4.86 (0.60)	164.97 (64.14)
		penGAM	3.64 (0.75)	24.61 (21.08)	4.19 (0.49)	849.23 (294.03)
	4	INIS	3.56 (0.75)	20.53 (1.68)	4.42 (0.52)	118.14 (43.97)
		penGAM	3.46 (0.75)	22.07 (16.04)	4.18 (0.49)	614.93 (97.36)
	8	INIS	3.09 (0.00)	10.67 (0.75)	4.28 (0.49)	71.16 (32.10)
		penGAM	3.12 (0.00)	19.92 (10.63)	4.30 (0.50)	548.60 (33.88)
	16	INIS	2.68 (0.75)	4.18 (0.75)	4.45 (0.52)	46.08 (15.35)
		penGAM	2.95 (0.00)	16.39 (11.19)	4.57 (0.55)	710.56 (199.86)
2.0	2	INIS	3.99 (0.00)	29.45 (11.57)	2.55 (0.38)	139.67 (70.45)
		penGAM	3.97 (0.00)	36.57 (22.57)	2.25 (0.28)	626.84 (210.44)
	4	INIS	3.93 (0.00)	19.12 (3.73)	2.26 (0.24)	111.01 (21.82)
		penGAM	3.91 (0.00)	31.31 (20.52)	2.19 (0.23)	481.87 (52.11)
	8	INIS	3.50 (0.75)	10.29 (0.75)	2.21 (0.23)	78.06 (32.23)
		penGAM	3.71 (0.75)	27.06 (19.03)	2.28 (0.29)	448.38 (26.63)
	16	INIS	2.93 (0.00)	4.07 (0.00)	2.42 (0.32)	51.69 (1.10)
		penGAM	3.22 (0.00)	19.51 (12.13)	2.53 (0.30)	661.93 (46.27)
4.0	2	INIS	4.00 (0.00)	29.47 (11.38)	1.45 (0.21)	144.22 (72.54)
		penGAM	4.00 (0.00)	37.27 (20.71)	1.27 (0.17)	533.98 (69.29)
	4	INIS	3.99 (0.00)	17.36 (5.22)	1.17 (0.12)	102.97 (32.71)
		penGAM	4.00 (0.00)	38.71 (20.34)	1.16 (0.11)	403.32 (28.29)
	8	INIS	3.78 (0.00)	10.00 (0.00)	1.13 (0.16)	88.79 (12.02)
		penGAM	3.99 (0.00)	41.42 (15.86)	1.19 (0.13)	402.92 (16.94)
	16	INIS	3.02 (0.00)	3.98 (0.00)	1.36 (0.15)	49.13 (1.85)
		penGAM	3.72 (0.75)	29.58 (19.40)	1.43 (0.18)	556.31 (35.48)

[Received December 2009. Revised November 2010.]

REFERENCES

Antoniadis, A., and Fan, J. (2001), "Regularization of Wavelet Approximations," *Journal of the American Statistical Association*, 96, 939–967. [545,552]

Candes, E., and Tao, T. (2007), "The Dantzig Selector: Statistical Estimation When p Is Much Larger Than n" (with discussion), *The Annals of Statistics*, 35, 2313–2404. [544]

Chiang, A. P., Beck, J. S., Yen, H.-J., Tayeh, M. K., Scheetz, T. E., Swiderski, R., Nishimura, D., Braun, T. A., Kim, K.-Y., Huang, J., Elbedour, K., Carmi, R., Slusarski, D. C., Casavant, T. L., Stone, E. M., and Sheffield, V. C. (2006), "Homozygosity Mapping With SNP Arrays Identifies TRIM32, an E3 Ubiquitin Ligase, as a Bardet–Biedl Syndrome Gene (BBS11)," *Proceedings of the National Academy of Sciences*, 103, 6287–6292. [550]

Draper, N. R., and Smith, H. (1966), *Applied Regression Analysis*, New York: Wiley. [548]

Efroymson, M. A. (1960), "Multiple Regression Analysis," in *Mathematical Methods for Digital Computers*, New York: Wiley, pp. 191–203. [548]

Fan, J. (1997), Comment on "Wavelets in Statistics: A Review," by A. Antoniadis, *Journal of the Italian Statistical Society*, 6, 131–138. [544]

Fan, J., and Jiang, J. (2005), "Nonparametric Inferences for Additive Models," *Journal of the American Statistical Association*, 100, 890–907. [552]

Fan, J., and Li, R. (2001), "Variable Selection via Nonconcave Penalized Likelihood and Its Oracle Properties," *Journal of the American Statistical Association*, 96, 1348–1360. [544]

Fan, J., and Lv, J. (2008), "Sure Independence Screening for Ultrahigh Dimensional Feature Space," *Journal of the Royal Statistical Society, Ser. B*, 70, 849–911. [544-549]

——— (2009), "Non-Concave Penalized Likelihood With NP-Dimensionality," manuscript, Princeton University. [544]

Fan, J., and Song, R. (2010), "Sure Independence Screening in Generalized Linear Models With NP-Dimensionality," *The Annals of Statistics*, 38, 3567–3604. [545,547,548]

Fan, J., Samworth, R., and Wu, Y. (2009), "Ultra-Dimensional Variable Selection via Independent Learning: Beyond the Linear Model," *Journal of Machine Learning Research*, 10, 1829–1853. [544,547-549]

Hall, P., and Miller, H. (2009), "Using Generalised Correlation to Effect Variable Selection in Very High Dimensional Problems," *Journal of Computational and Graphical Statistics*, 18, 533–550. [544]

Hall, P., Titterington, D., and Xue, J. (2009), "Tilting Methods for Assessing the Influence of Components in a Classifier," *Journal of the Royal Statistical Society, Ser. B*, 71, 783–803. [544]

Horowitz, J., Klemelä, J., and Mammen, E. (2006), "Optimal Estimation in Additive Regression Models," *Bernoulli*, 12, 271–298. [552]

Huang, J., Horowitz, J., and Ma, S. (2008), "Asymptotic Properties of Bridge Estimators in Sparse High-Dimensional Regression Models," *The Annals of Statistics*, 36, 587–613. [544]

Huang, J., Horowitz, J., and Wei, F. (2010), "Variable Selection in Nonparametric Additive Models," *The Annals of Statistics*, 38, 2282–2313. [544,546-548,550]

Irizarry, R. A., Hobbs, B., Collin, F., Beazer-Barclay, Y. D., Antonellis, K. J., Scherf, U., and Speed, T. P. (2003), "Exploration, Normalization, and Summaries of High Density Oligonucleotide Array Probe Level Data," *Biostatistics*, 4, 249–264. [550]

Kim, Y., Kim, J., and Kim, Y. (2006), "Blockwise Sparse Regression," *Statistica Sinica*, 16, 375–390. [545]

Koltchinskii, V., and Yuan, M. (2008), "Sparse Recovery in Large Ensembles of Kernel Machines," in *21st Annual Conference on Learning Theory—COLT 2008, Helsinki, Finland, July 9–12, 2008*, eds. R. A. Servedio and T. Zhang, Omnipress, pp. 229–238. [544,547]

Lin, Y., and Zhang, H. H. (2006), "Component Selection and Smoothing in Multivariate Nonparametric Regression," *The Annals of Statistics*, 34, 2272–2297. [544,547]

Meier, L., Geer, V., and Bühlmann, P. (2009), "High-Dimensional Additive Modeling," *The Annals of Statistics*, 37, 3779–3821. [544,545,547-549]

Ravikumar, P., Liu, H., Lafferty, J., and Wasserman, L. (2009), "Spam: Sparse Additive Models," *Journal of the Royal Statistical Society, Ser. B*, 71, 1009–1030. [544,548]

Sardy, S., and Tseng, P. (2004), "AMlet, RAMlet, and GAMlet: Automatic Nonlinear Fitting of Additive Models, Robust and Generalized, With Wavelets," *Journal of Computational and Graphical Statistics*, 13, 283–309. [552]

Scheetz, T. E., Kim, K.-Y. A., Swiderski, R. E., Philp, A. R., Braun, T. A., Knudtson, K. L., Dorrance, A. M., DiBona, G. F., Huang, J., Casavant, T. L., Sheffield, V. C., and Stone, E. M. (2006), "Regulation of Gene Expression in the Mammalian Eye and Its Relevance to Eye Disease," *Proceedings of the National Academy of Sciences*, 103, 14429–14434. [550]

Silverman, B. (1984), "Spline Smoothing: The Equivalent Variable Kernel Method," *The Annals of Statistics*, 12, 898–916. [552]

Speckman, P. (1985), "Spline Smoothing and Optimal Rates of Convergence in Nonparametric Regression Models," *The Annals of Statistics*, 13, 970–983. [552]

Stone, C. (1985), "Additive Regression and Other Nonparametric Models," *The Annals of Statistics*, 13, 689–705. [544,546,547]

Tibshirani, R. (1996), "Regression Shrinkage and Selection via the Lasso," *Journal of the Royal Statistical Society, Ser. B*, 58, 267–288. [544]

van der Vaart, A. W., and Wellner, J. A. (1996), *Weak Convergence and Empirical Processes*, New York: Springer. [552]

Wang, H. (2009), "Forward Regression for Ultra-High Dimensional Variable Screening," *Journal of the American Statistical Association*, 104, 1512–1524. [548]

Wei, F., and Huang, J. (2007), "Consistent Group Selection in High-Dimensional Linear Regression," Technical Report 387, University of Iowa. [545]

Yuan, M., and Lin, Y. (2006), "Model Selection and Estimation in Regression With Grouped Variables," *Journal of the Royal Statistical Society, Ser. B*, 68, 49–67. [545]

Zhang, C.-H. (2010), "Nearly Unbiased Variable Selection Under Minimax Concave Penalty," *The Annals of Statistics*, 38, 894–942. [544]

Zhao, D. S., and Li, Y. (2010), "Principled Sure Independence Screening for Cox Models With Ultra-High-Dimensional Covariates," manuscript, Harvard University. [547]

Zhou, S., Shen, X., and Wolfe, D. A. (1998), "Local Asymptotics for Regression Splines and Confidence Regions," *The Annals of Statistics*, 26, 1760–1782. [546]

Zou, H. (2006), "The Adaptive Lasso and Its Oracle Properties," *Journal of the American Statistical Association*, 101, 1418–1429. [544]

Zou, H., and Hastie, T. (2005), "Addendum: Regularization and Variable Selection via the Elastic Net," *Journal of the Royal Statistical Society, Ser. B*, 67, 768. [544]

Zou, H., and Li, R. (2008), "One-Step Sparse Estimates in Nonconcave Penalized Likelihood Models," *The Annals of Statistics*, 36, 1509–1533. [544]