Nonparametric Independence Screening in Sparse Ultra-High Dimensional Additive Models [*]

Jianqing Fan, Yang Feng and Rui Song

December 14, 2009

Abstract

A variable screening procedure via correlation learning was proposed in 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 is called NIS, a specific member of the sure independence screening. Several closely related variable screening procedures are proposed. Under the nonparametric additive

[*] Jianqing Fan is Frederick L. Moore Professor of Finance, Department of Operations Research and Financial Engineering, Princeton University, Princeton NJ 08544 (Email: jqfan@princeton.edu). Yang Feng is Graduate Student, Department of Operations Research and Financial Engineering, Princeton University, Princeton NJ 08544 (Email: yangfeng@princeton.edu). Rui Song is Assistant Professor, Department of Statistics, Colorado State University, Fort Collins, CO 80523 (Email: song@stat.colostate.edu). The bulk of the work was conducted when Rui Song was a postdoctoral research fellow at Princeton University. The financial support from NSF grants DMS-0714554, DMS-0704337 and NIH grant R01-GM072611 are greatly acknowledged. The authors are in deep debt to Dr. Lukas Meier for sharing the codes of penGAM.

models, it is shown that under some mild technical conditions, the proposed independence screening methods enjoy a sure screening property. The extent to which the dimensionality can be reduced by independence screening is also explicitly quantified. As a methodological extension, an iterative nonparametric independence screening (INIS) is also proposed 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.

Keywords: Additive model, independent learning, nonparametric regression, sparsity, sure independence screening, nonparametric 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, microarrays, proteomics, fMRI, functional data and high frequency financial data. In all 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 it non-polynomial (NP) dimensionality or ultra-high dimensionality. What makes the under-determined statistical inference possible is the sparsity assumption: only a small set of independent variables contribute to the response. Therefore, 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, such as the Lasso

(Tibshirani, 1996), the SCAD and other folded-concave penalty (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, 2009) and related methods (Zou, 2006; Zou and Li, 2008). Nevertheless, due to the "curse of dimensionality" in terms of simultaneously challenges on the computational expediency, statistical accuracy and algorithmic stability, these methods meet their limits in 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 et al. (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 et al. (2008) also investigated the marginal bridge regression in the ordinary linear model. These methods focus on studying the marginal pseudo-likelihood and are fast but crude in terms of reducing the NP-dimensionality to a more moderate size. To enhance the performance, Fan and Lv (2008) and Fan et al. (2009) introduced some methodological extensions include iterative SIS (ISIS) and multi-stage procedures, such as SIS-SCAD and SIS-LASSO, to select variables and estimate parameters simultaneously. Nevertheless, these marginal screening methods have some methodological challenges. When the covariates are not jointly normal, even if the linear model holds in the joint regression, the marginal regression can be highly nonlinear. Therefore, sure screening based on nonparametric marginal regression becomes a natural candidate.

In practice, there is often little prior information 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). It increases substantially the flexibility of the ordinary linear model and allows a data-analytic transform of the covariates to enter into the linear model. Yet, the literature on variable selection in nonparametric additive models are limited. See, for example, Koltchinskii and Yuan (2008), Ravikumar et al. (2009), Huang et al. (2009) and Meier et al. (2009). Koltchinskii and Yuan (2008) and Ravikumar et al. (2009) are closely related with COSSO proposed in Lin and Zhang (2006) with fixed minimal signals, which does not converge to zero. Huang et al. (2009) can be viewed as an extension of adaptive lasso to additive models with fixed minimal signals. Meier et al. (2009) proposed a penalty which is a combination of sparsity and smoothness with a fixed design. Under ultra-high dimensional settings, all these methods still suffer from the aforementioned three challenges as they can be viewed as extensions of penalized pseudo-likelihood approaches to additive modeling. The commonly used algorithm in additive modeling such as backfitting makes the situation even more challenging, as it is quite computationally expensive.

In this paper, 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 non-sparsity 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 in Fan and Lv (2008) and Fan and Song (2009). Compared with these papers, the minimum distinguishable signal is related with not only the stochastic error in estimating the nonparametric components, but also approximation errors in modeling nonparametric components, which depends on the number of basis functions used for the approximation. This brings significant challenges to the theoretical development and leads to an interesting result on 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 will be demonstrated in our empirical studies.

We approximate the nonparametric additive components by using a B-spline basis. Hence, the component selection in additive models can be viewed as a functional version of the grouped variable selection. An early literature on the group variable selection using group penalized least-squares is Antoniadis and Fan (2001) (see page 966), in which blocks of wavelet coefficients are either killed or selected. The group variable selection was more intensively and thoroughly studied in Yuan and Lin (2006), Kim et al. (2006), Wei and Huang (2007) and Meier et al. (2009). Our methods and results have important implications on the group variable selections.

The rest of the paper is organized as follows. In Section 2, we introduce the nonparametric independence screening (NIS) procedure in additive models. The theoretical properties for NIS are presented in Section 3. As a methodological extension, INIS-penGAM is outlined in Section 4. 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 zero. To expeditiously identify important variables in model (1), without the "curse-of-dimensionality", we consider the following p marginal nonparametric regression problems:

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

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 via thresholding.

To obtain a sample version of the marginal nonparametric regression, we employ a B-Spline basis. Let $\mathcal{S}_n$ be the space of polynomial splines of degree $l \geq 1$ and $\{\Psi_{jk},\ k = 1, \cdots, d_n\}$ 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

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

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

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

where $\boldsymbol{\Psi}_j \equiv \boldsymbol{\Psi}_j(X_j) = (\Psi_1(X_j), \cdots, \Psi_{d_n}(X_j))^T$ denotes the d_n dimensional basis functions and $\mathbb{P}_n g(\mathbf{X}, Y)$ is the expectation with respect to the empirical measure $\mathbb{P}_n$, i.e., the sample average of $\{g(\mathbf{X}_i, Y_i)\}_{i=1}^n$. The least square estimator $\hat{f}_{nj}$ of (3) can thus be viewed as a projection by smoothing the response. This can be rapidly computed, even for NP-dimensional problems. We correspondingly define the population version of the minimizer of the componentwise least square regression,

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

where E denotes the expectation under the true model.

We now select a set of variables

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

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 can also be viewed as ranking by the magnitude of the correlation of the marginal nonparametric estimate $\{\hat{f}_{nj}(X_{ij})\}_{i=1}^n$ (note that it is different from the joint nonparametric component) with the response $\{Y_i\}_{i=1}^n$, since $\|\hat{f}_{nj}\|_n^2 = \|Y\hat{f}_{nj}\|_n$. In this sense, the proposed NIS procedure is related to the correlation learning proposed in 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:

$$\widehat{\mathcal{M}}_{\gamma_n} = \{1 \leq j \leq p : u_j \leq \gamma_n\},$$

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 $u_j = \mathbb{P}_n(Y^2 - \hat{f}_{nj}^2)$. Hence, the two methods are equivalent.

The nonparametric independence screening reduces the dimensionality from p to a possibly much smaller space with model size $|\widehat{\mathcal{M}}_{\nu_n}|$ or $|\widehat{\mathcal{M}}_{\gamma_n}|$. It 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 will show that the sure screening property indeed holds for nonparametric additive models with 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 now 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 $\{m_j(X_j)\}_{j=1}^{p}$ have mean zero. Consequently, the response Y has zero mean, too. Let $\mathcal{M}_\star = \{j : Em_j(X_j)^2 > 0\}$ be the true sparse model with non-sparsity size $s_n = |\mathcal{M}_\star|$. We allow p to grow with n and denote it as p_n whenever needed.

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:

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

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

B. The marginal density function g_j of X_j satisfies $0 < K_1 \leq g_j(X_j) \leq$

$K_2 < \infty$ on $[a, b]$ for $1 \leq j \leq p$ for some constants K_1 and K_2.

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

Under conditions A and B, the following three facts hold when $l \geq d$ and will be used in the paper. We state them here for readability.

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

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

Fact 2. There exists a positive constant C_2 such that (Stone, 1985; Huang et al., 2009)

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

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

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

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

Lemma 1. *Under conditions A–C, we have*

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

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

A model selection consistency result can be established with nonparametric independence screening under the partial orthogonality condition, i.e., $\{X_j,\ j \notin \mathcal{M}_\star\}$ is independent of $\{X_i,\ i \in \mathcal{M}_\star\}$. In this case, there is a separation between the strength of marginal signals $\|f_{nj}\|^2$ for active variables $\{X_j; j \in \mathcal{M}_\star\}$ and 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 the nonparametric independence screening (NIS). We need 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 i.i.d. with conditional mean zero 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 a positive constant c_1 and $\xi \in (0,1)$ such that $d_n^{-2d-1} \leq c_1(1-\xi)n^{-2\kappa}/C_1$.

The following theorem gives the sure screening properties. It reveals that it is only the size of non-sparse 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

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

(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

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

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 $d_n = o(n^{1/3})$ in order to have the sure screening property, whereas Condition F requires $d_n \ge B_4 n^{2\kappa/(2d+1)}$, where $B_4 = (c_1(1-\xi)/C_1)^{-1/(2d+1)}$.

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

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

Under this condition,

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

i.e., the sure screening property. It is worthwhile to point out that the number of spline basis d_n affects the order of dimensionality, comparing with the results of Fan and Lv (2008) and Fan and Song (2009) 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 dimensionality the nonparametric independence screening (NIS) can handle. This is in line with our intuition. On the other hand, the number of basis functions can not be too small, since 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 d_n we can take and the higher the dimension that the NIS can handle. If the minimum signal does not converge to zero, as in Lin and Zhang (2006), Koltchinskii and Yuan (2008) and Huang et al. (2009), then $\kappa = 0$. In this case, d_n can be taken to be finite as long as it is sufficiently large so that minimum signal in Lemma 1 exceeds the noise level and threshold by a large enough margin. By 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 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 zero, with probability tending to one that

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

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

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

We now deal with the more general case. The idea is to bound the size of the selected set by using the fact that 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 directly related to the dimension reduction with additive models.

Theorem 2. *Suppose Conditions A–F hold and 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*

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

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

The significance of the result is that when $\lambda_{\max}(\mathbf{\Sigma}) = O(n^\tau)$, 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 zero exponentially fast. The size of the selected variables is of order $O(n^{2\kappa+\tau})$. This is of the same order as in Fan and Lv (2008). Our result is an extension 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 in Fan and Song (2009): the number of selected variables is related to the correlation structure of the covariance matrix.

In the specific case where the covariates are independent, then the matrix $\mathbf{\Sigma}$

is block diagonal with j-th block $\mathbf{\Sigma}_j$. Hence, it follows from (8) that $\lambda_{\max}(\mathbf{\Sigma}) = O(d_n^{-1})$. In general, since the B-spline basis is local, the covariance between $\Psi_{jk}(X_j)$ and $\Psi_{il}(X_i)$ is small and hence $\lambda_{\max}(\mathbf{\Sigma})$ can not grow too quickly.

4 INIS Method

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) in Meier et al. (2009) can be employed to select a subset of active variables. This results in NIS-penGAM. To further enhance the performance of the method, in terms of false selection rates, following Fan and Lv (2008) and Fan et al. (2009), we can iteratively employ the large-scale screening and moderate-scale selection strategy, resulting in the INIS-penGAM.

Given the data $\{(\mathbf{X}_i, Y_i)\}, i = 1, \cdots, n$, for each component $f_j(\cdot), j = 1, \cdots, p$, we choose the same truncation term $d_n = O(n^{1/5})$. In the algorithm, a predetermined sparsity size parameter s_0 for the NIS procedure is needed. It is recommended to take $s_0 = O(n/\log(n))$. However, it can be adjusted accordingly depending whether sure screening or reducing false selection rate is more important. The algorithm works as follows:

Step 1: For every $j \in \{1, \cdots, p\}$, we compute

$$\min_{f_{nj} \in \mathcal{S}_n} \mathbb{P}_n \Big(Y - f_{nj}(X_j) \Big)^2, \text{ for } 1 \le j \le p.$$

Using the NIS, we can pick a set $\mathcal{A}_1$ of indices of size k_1. In our implementation, we choose $k_1 = \lfloor 2s_0/3 \rfloor$ to guarantee it will take at least two iterations.

Step 2: We apply further the penalized method for additive model (penGAM) in Meier et al. (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: Instead of computing residuals, for every $j \in \mathcal{M}_1^c = \{1, \cdots, p\} \backslash \mathcal{M}_1$, we minimize

$$\mathbb{P}_n\Big(Y - \sum_{i \in \mathcal{M}_1} f_{ni}(X_i) - f_{nj}(X_j)\Big)^2, \text{ for } 1 \leq j \leq p, \tag{11}$$

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 j-th components conditioning on the existence of the variable set $\mathcal{M}_1$. After marginally screening as in the first step, we can pick a set $\mathcal{A}_2$ of indices of size $k_2 = s_0 - |\mathcal{M}_1|$. Then we apply further the penGAM algorithm on the set $\mathcal{M}_1 \bigcup \mathcal{A}_2$ to select a subset $\mathcal{M}_2$.

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

Here are a few comments about the method. In Step 2, we use the penGAM method. In fact, any variable selection method for additive models will work such as the SpAM in Ravikumar et al. (2009) and also the adaptive group LASSO for additive models in Huang et al. (2009). A similar sample splitting idea as described in Fan et al. (2009) can be applied here to further reduce false selection rate.

5 Numerical Results

In this section, we will illustrate our method by studying the performance on the simulated data and a real data analysis. Part of the simulation settings are adapted from Fan and Lv (2008), Meier et al. (2009), Huang et al. (2009), and Fan and Song (2009).

5.1 Comparison of Minimum Model Size

We first illustrate the behavior of the NIS procedure under different correlation structures. Following Fan and Song (2009), 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 $\mathcal{M}^*$, is used as a measure of the effectiveness of a screening method. We also include the correlation screening 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 (2009), let $\{X_k\}_{k=1}^{950}$ be i.i.d standard normal random variables and

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

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 $\varepsilon \sim N(0, 1)$ and $\boldsymbol{\beta}^* = (3, -3, \cdots)^T$ has s non-vanishing components, taking values ±3 alternately.

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

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

Model	NIS	PenGAM	SIS
Ex 1 ($s = 3$)	3(0)	3(0)	3(0)
Ex 1 ($s = 6$)	56(0)	103(703)	56(0)
Ex 1 ($s = 12$)	63(3)	—	62(0)
Ex 1 ($s = 24$)	228(130)	—	102(34)
Ex 2	3(0)	3(0)	297(357)

The minimum model size(MMS) for each method and its associated robust estimate of the standard deviation($RSD = IQR/1.34$) are shown in Table 1. The column "NIS", "penGAM", and "SIS" summarizes the results on the MMS based on 100 simulations, respectively for the nonparametric independence screening in the paper, penalized method for additive model of Meier et al. (2009), and the linear correlation 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 performs poorly. When $s = 12$ and $s = 24$, penGAM fails even to include the true model until the last step. In contrast, the proposed nonparametric independence screening 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 also linear. In this case, NIS selects variables from pd_n parameters whereas SIS selects only from p parameters. However, for the nonlinear problem like Example 2, we can see that both nonlinear method NIS and penGAM behave nicely. However, SIS fails badly even though the underlying true model is indeed linear.

5.2 Comparison of Model Selection and Estimation

As in the last section, we set $n = 400$ and $p = 1000$ for all the examples to demonstrate the power of our newly proposed method. The number of simulations is 100. Here, we use ten-fold cross validation in Step 2 of the INIS algorithm. For simplicity of notations, we let

$$f_1(x) = x, \quad f_2(x) = (2x-1)^2, \quad f_3(x) = \frac{\sin(2\pi x)}{2 - \sin(2\pi x)}$$

and

$$f_4(x) = 0.1\sin(2\pi x) + 0.2\cos(2\pi x) + 0.3\sin(2\pi x)^2 + 0.4\cos(2\pi x)^3 + 0.5\sin(2\pi x)^3.$$

Example 3. Following Meier et al. (2009), we generate the data from the following additive model:

$$Y = 5f_1(X_1) + 3f_2(X_2) + 4f_3(X_3) + 6f_4(X_4) + \sqrt{1.74}\varepsilon$$

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

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

where $W_1, \cdots, W_p$ and U are i.i.d. Unif$(0,1)$ and $\varepsilon \sim N(0,1)$. When $t = 0$, the covariates are all independent, and when $t = 1$ the pairwise correlation of covariates is 0.5.

Example 4. We adapt the simulation model from Meier et al. (2009) but reduce the variance of the error from 0.5184 to 0.5184/4=0.1296, since we decrease sample size and increase the scale of the model from $n = 100, p = 60$ to $n = 400, p = 1000$:

$$\begin{aligned} Y &= f_1(X_1) + f_2(X_2) + f_3(X_3) + f_4(X_4) \\ &+ 1.5f_1(X_5) + 1.5f_2(X_6) + 1.5f_3(X_7) + 1.5f_4(X_8) \\ &+ 2f_1(X_9) + 2f_2(X_{10}) + 2f_3(X_{11}) + 2f_4(X_{12}) + \sqrt{0.1296}\varepsilon, \end{aligned}$$

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

Example 5. We follow the simulation model of Fan et al. (2009), in which $Y = \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \varepsilon$ is simulated, where $\varepsilon \sim N(0,1)$. The covariates $X_1, \cdots, X_p$ are jointly Gaussian, marginally $N(0,1)$, and with $\text{corr}(X_i, X_4) = 1/\sqrt{2}$ for all $i \neq 4$ and $\text{corr}(X_i, X_j) = 1/2$ if i and j are distinct elements of $\{1, \cdots, p\}\backslash\{4\}$. The coefficients $\beta_1 = 4, \beta_2 = 4, \beta_3 = 4, \beta_4 = -6\sqrt{2}$, 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 performances of INIS-penGAM proposed in the paper, penGAM(Meier et al., 2009), and ISIS-SCAD (Fan et al., 2009) which aims for sparse linear model. Their results are shown respectively in the rows "INIS", "penGAM", and "ISIS" of Table 2, in which the True Positives(TP), False Positives(FP), and the Prediction Error(PE) are reported for each method. Here the prediction error is calculated on an independent test data set of size $n/2$. First of all, it is easy to notice that the number of false positive for both INIS-penGAM and ISIS-SCAD are 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 almost always misses one variable in Example 5. The linear method ISIS-SCAD missed important variables in the nonlinear models in Examples 3 and 4. In the perspective of the prediction error, INIS-penGAM and penGAM outperforms ISIS-SCAD in the nonlinear models whereas their performances are worse than ISIS-SCAD in the linear model, Example 5. Overall, in the designed simulation settings, the ISIS-SCAD and INIS-penGAM outperform the penGAM in terms of smaller false selection rates.

5.3 Boston Housing Data Analysis

The Boston housing data were collected to study house values in the suburbs of Boston. There are 506 observations with 10 covariates. To save space, we omit the description of those variables. Interested readers can find more details from `http://lib.stat.cmu.edu/datasets/boston`. The dataset has been studied by many other authors (Härdle et al., 2004; Lin and Zhang, 2006; Ravikumar et al., 2009), with various transformations proposed for different covariates. To demonstrate the effectiveness of our method, following a similar idea of Ravikumar et al. (2009), we add 90 irrelevant variables, which are randomly drawn from Uniform$(0, 1)$. Therefore, there are 100 covariates, among which the last 90 covariates are known to be irrelevant variables and the first 10 variables might be relevant to the housing value. Therefore, we fit the 100-dimensional sparse additive model, using INIS-penGAM and penGAM.

Table 2: Average values of the numbers of true (TP) and false (FP) positives. Robust standard deviations are given in parentheses.

Model	Method	TP	FP	PE
Ex 3 ($t = 0$)	INIS	4.00(0.00)	28.96(0.00)	2.57(0.33)
	penGAM	4.00(0.00)	45.81(26.31)	2.46(0.28)
	ISIS	3.03(0.00)	29.97(0.00)	15.92(1.66)
Ex 3 ($t = 1$)	INIS	3.99(0.00)	29.00(0.00)	2.59(0.31)
	penGAM	4.00(0.00)	55.41(17.35)	2.62(0.29)
	ISIS	3.01(0.00)	29.99(0.00)	12.89(1.46)
Ex 4 ($t = 0$)	INIS	11.97(0.00)	21.03(0.00)	0.32(0.04)
	penGAM	12.00(0.00)	109.94(21.64)	0.52(0.09)
	ISIS	8.24(0.75)	24.76(0.75)	4.08(0.37)
Ex 4 ($t = 1$)	INIS	11.58(0.75)	21.42(0.75)	0.33(0.05)
	penGAM	11.31(0.75)	96.81(26.87)	0.54(0.09)
	ISIS	6.89(1.49)	26.11(1.49)	3.73(0.43)
Ex 5	INIS	4.00(0.00)	29.00(0.00)	3.26(0.62)
	penGAM	3.05(0.00)	200.37(9.89)	5.61(0.87)
	ISIS	4.00(0.00)	29.00(0.00)	1.35(0.15)

We randomly divide the data into two parts, the first 2/3 as the training data and the remaining 1/3 as the test data. The above experiment is repeated 100 times to test the stability of the method.

Among the 100 experiments (fitting sparse 100-variable additive model 100 times), the false selection of the 90 known irrelevant variables is recorded and so are the relevant covariates selected. Among 10 potential relevant covariates, the variables nox(nitric oxides concentration), rm(average number of rooms per dwelling), dis(weighted distances to five Boston employment centres), tax(full-value property-tax rate per \$10,000), ptratio(pupil-teacher ratio by town), b ($1000(\mathrm{Bk} - 0.63)^2$ where Bk is the proportion of blacks by town) and lstat(% lower status of the population) are always selected by both INIS-penGAM and penGAM in the 100 numerical experiments. The selected frequencies for other three variables are given in Table 3. The average number of false positive and prediction error on the test data along with their robust estimate of standard deviations are also depicted in Table 3. This shows our method has a smaller false selection rate, and at the same time, it gives us better prediction accuracy. Using ten-fold cross validation as tuning criterion, we discovered that indus and probably crim for the penGAM method are estimated to be irrelevant, which is consistent with Ravikumar et al. (2009). The non-vanishing estimated additive components in a typical experiment are shown in Figure 1.

Table 3: Selected frequency for variables "crim", "indus" and "age" and average selected model size and average number of false positives and average prediction error. Robust estimates of standard deviations are given in parentheses.

Method	crim	indus	age	model size	false positive	prediction error
INIS	0.77	0.15	0.62	17.02(2.99)	8.48 (2.24)	17.48(0.98)
penGAM	0.17	0.02	0.35	17.63(9.89)	10.09(8.58)	18.45 (0.41)

6 Remarks

In this paper, we studied the nonparametric independence screening (NIS) method for variable selection in additive models. B-spline basis functions are used for fitting the marginal nonparametric components. The proposed marginal projection criteria is an important extension of the marginal correlation. Iterative NIS procedures are also proposed such that variable selection and coefficient estimation can be achieved simultaneously. By applying the

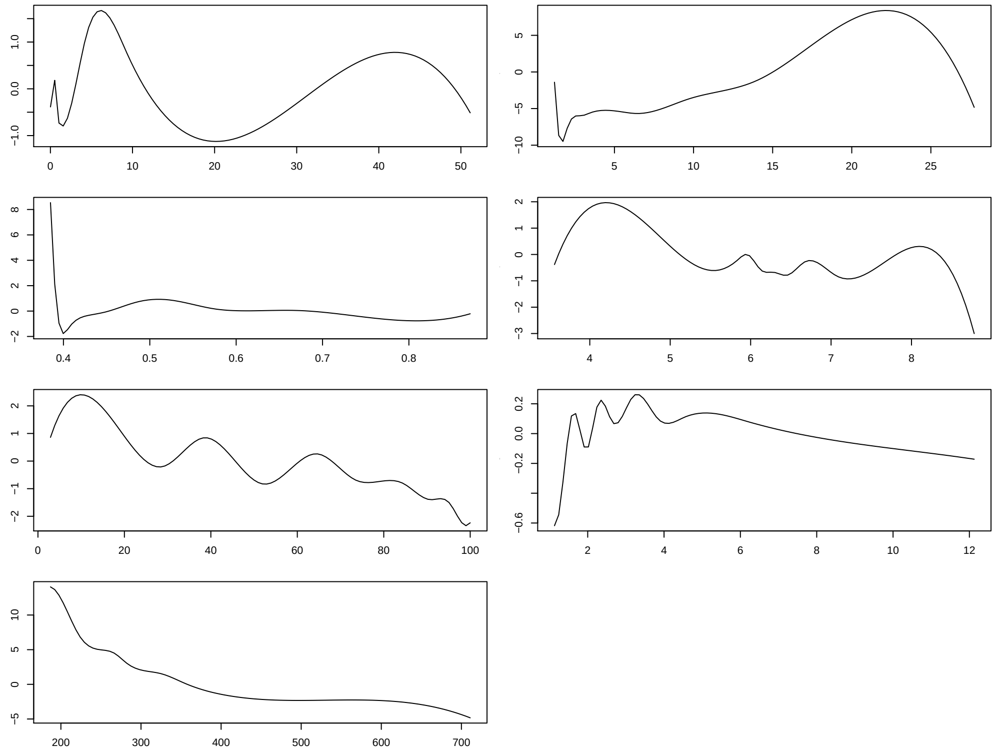

Figure 1: Non-vanishing estimated additive components in one repetition.

INIS-penGAM method, we can preserve the sure screening property and substantially reduce the false selection rate. Moreover, we 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 model with appropriate conditions.

As the additive components are specifically approximated by truncated series expansions with B-spline bases in this paper, the theoretical results should hold in general and the proposed framework can be readily adaptive to other smoothing methods with additive models (Horowitz et al., 2006;

Silverman, 1984), such as local polynomial regression (Fan and Jiang, 2005), wavelets approximations(Antoniadis and Fan, 2001; Sardy and Tseng, 2004) and smoothing spline (Speckman, 1985). This is an interesting topic for future research.

7 Proofs

Proof of Lemma 1.

By the property of the least-squares, $E(Y - f_{nj})f_{nj} = 0$ 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 two types of Bernstein's inequality in van der Vaart and Wellner (1996) will be needed. We reproduce them here for the sake of readability.

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

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

for $v \geq var(Y_1 + \cdots + Y_n)$.

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

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

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

The following two lemmas will be 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*

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

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

Proof of Lemma 4.

Denote by $T_{jki} = \Psi_{jk}(X_{ij})Y_i - E\Psi_{jk}(X_{ij})Y_i$. Since $Y_i = m(\mathbf{X}_i) + \varepsilon_i$, 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 $\|\Psi_{jk}\|_\infty \leq 1$, we have

$$|T_{jki1}| \leq 2B_1, \quad \mathrm{var}(T_{jki1}) \leq E\Psi_{jk}^2(X_{ij})m_i(X_{ij})^2 \leq B_1^2 C_2 d_n^{-1}. \tag{12}$$

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

$$P(\Big|\sum_{i=1}^n T_{jki1}\Big| > \delta_1) \leq 2\exp\Big(-\frac{1}{2}\frac{\delta_1^2}{nB_1^2 C_2 d_n^{-1} + 2B_1\delta_1/3}\Big). \tag{13}$$

Next, we bound the tails of T_{jki2}. For every $r \geq 2$,

$$\begin{aligned} E|T_{jki2}|^r &\leq E|\Psi_{jk}(X_{ij})|^2 E(|\varepsilon_i|^r|\mathbf{X}_i) \\ &\leq r!B_2^{-r}E|\Psi_{jk}(X_{ij})|^2 E\exp(B_2|\varepsilon_i||\mathbf{X}_i) \\ &\leq B_3 C_2 d_n^{-1} r! B_2^{-r}, \end{aligned}$$

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

$$P(\Big|\sum_{i=1}^n T_{jki2}\Big| > \delta_2) \leq 2\exp\Big(-\frac{1}{2}\frac{\delta_2^2}{2nB_2^{-2}B_3 C_2 d_n^{-1} + B_2^{-1}\delta_2}\Big). \tag{14}$$

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

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 $\delta > 0$,*

$$\begin{aligned} & P(|\lambda_{\min}(\mathbb{P}_n \mathbf{\Psi}_j \mathbf{\Psi}_j^T) - \lambda_{\min}(E\mathbf{\Psi}_j \mathbf{\Psi}_j^T)| \geq d_n \delta / n) \\ \leq & \ 2d_n^2 \exp\Big\{-\frac{1}{2}\frac{\delta^2}{C_2 n d_n^{-1} + \delta/3}\Big\}. \end{aligned}$$

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

$$\begin{aligned} & P\left\{\Big|\big\|(\mathbb{P}_n \mathbf{\Psi}_j \mathbf{\Psi}_j^T)^{-1}\big\| - \big\|(E\mathbf{\Psi}_j \mathbf{\Psi}_j^T)^{-1}\big\|\Big| \geq c_8 \big\|(E\mathbf{\Psi}_j \mathbf{\Psi}_j^T)^{-1}\big\|\right\} \\ \leq & \ 2d_n^2 \exp\Big(-c_4 n d_n^{-3}\Big). \end{aligned} \quad (15)$$

Proof of Lemma 5.

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

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

Taking minimum among $\|x\| = 1$ on both sides, we have

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

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

Then we have

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

which is the same as

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

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

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

In other words,

$$|\lambda_{\min}(\mathbf{A}) - \lambda_{\min}(\mathbf{B})| \leq \max\{|\lambda_{\min}(\mathbf{A}-\mathbf{B})|, |\lambda_{\min}(\mathbf{B}-\mathbf{A})|\} \tag{16}$$

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

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

We now bound the right-hand 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 $\|\mathbf{x}\| = 1$,

$$|\mathbf{x}^T\mathbf{D}_j\mathbf{x}| \leq \|\mathbf{D}_j\|_\infty \Big(\sum_{i=1}^{d_n} |x_i|\Big)^2 \leq d_n \|\mathbf{D}_j\|_\infty. \tag{18}$$

Thus,

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

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

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

We conclude that

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

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

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

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

$$\text{var}(\Psi_{jk}(X_j)\Psi_{jl}(X_j)) \leq E\Psi_{jk}^2(X_j)\Psi_{jl}^2(X_j) \leq E\Psi_{jk}^2(X_j) \leq C_2 d_n^{-1}.$$

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

$$P(|(\mathbb{P}_n - E)\Psi_{jk}(X_j)\Psi_{jl}(X_j)| > \delta/n) \leq 2\exp\Big\{-\frac{\delta^2}{2(C_2 n d_n^{-1} + \delta/3)}\Big\}. \tag{20}$$

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

$$\begin{aligned} & P(|\lambda_{\min}(\mathbb{P}_n \boldsymbol{\Psi}_j \boldsymbol{\Psi}_j^T) - \lambda_{\min}(E\boldsymbol{\Psi}_j \boldsymbol{\Psi}_j^T)| \geq d_n \delta/n) \\ \leq \;\; & 2d_n^2 \exp\Big\{-\frac{\delta^2}{2(C_2 n d_n^{-1} + \delta/3)}\Big\}. \end{aligned}$$

This completes the proof of the first inequality.

To prove the second inequality, let us take $\delta = c_9 D_1 n d_n^{-2}$ in (20), where $c_9 \in (0,1)$. By recalling Fact 3, it follows that

$$\begin{aligned} & P(|\lambda_{\min}(\mathbb{P}_n \boldsymbol{\Psi}_j \boldsymbol{\Psi}_j^T) - \lambda_{\min}(E\boldsymbol{\Psi}_j \boldsymbol{\Psi}_j^T)| \geq c_9 \lambda_{\min}(E\boldsymbol{\Psi}_j \boldsymbol{\Psi}_j^T)) \\ \leq \;\; & 2d_n^2 \exp\Big(-c_4 n d_n^{-3}\Big), \end{aligned} \tag{21}$$

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

$$\begin{aligned} & P\left(\left|\left\{\lambda_{\min}(\mathbb{P}_n \boldsymbol{\Psi}_j \boldsymbol{\Psi}_j^T)\right\}^{-1} - \left\{\lambda_{\min}(E\boldsymbol{\Psi}_j \boldsymbol{\Psi}_j^T)\right\}^{-1}\right| \geq c_8 \left\{\lambda_{\min}(E\boldsymbol{\Psi}_j \boldsymbol{\Psi}_j^T)\right\}^{-1}\right) \\ \leq \; & 2d_n^2 \exp\left(-c_4 n d_n^{-3}\right), \end{aligned} \tag{22}$$

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 for $a \in (0,1)$,

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

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

Since

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

we have

$$A^{-1} - B^{-1} \leq -(1/(1-a) - 1)B^{-1}, \quad \text{or} \quad \geq (1/(1-a) - 1)B^{-1}.$$

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

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

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

This concludes the proof of the lemma. □

Proof of Theorem 1.

We first show part (i). Recall that

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

and

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

Let $\mathbf{a}_n = \mathbb{P}_n \boldsymbol{\Psi}_j Y$, $\mathbf{B}_n = (\mathbb{P}_n \boldsymbol{\Psi}_j \boldsymbol{\Psi}_j^T)^{-1}$, $\mathbf{a} = E\boldsymbol{\Psi}_j Y$ 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

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

where

$$\begin{aligned} S_1 &= \Big(\mathbb{P}_n \boldsymbol{\Psi}_j Y - E\boldsymbol{\Psi}_j Y\Big)^T (\mathbb{P}_n \boldsymbol{\Psi}_j \boldsymbol{\Psi}_j^T)^{-1} \Big(\mathbb{P}_n \boldsymbol{\Psi}_j Y - E\boldsymbol{\Psi}_j Y\Big), \\ S_2 &= 2\Big(\mathbb{P}_n \boldsymbol{\Psi}_j Y - E\boldsymbol{\Psi}_j Y\Big)^T (\mathbb{P}_n \boldsymbol{\Psi}_j \boldsymbol{\Psi}_j^T)^{-1} E\boldsymbol{\Psi}_j Y, \\ S_3 &= (E\boldsymbol{\Psi}_j Y)^T \Big((\mathbb{P}_n \boldsymbol{\Psi}_j \boldsymbol{\Psi}_j^T)^{-1} - (E\boldsymbol{\Psi}_j \boldsymbol{\Psi}_j^T)^{-1}\Big) E\boldsymbol{\Psi}_j Y. \end{aligned}$$

Note that

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

By Lemma 4 and the union bound of probability,

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

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

$$\begin{aligned} & P\left\{\left|\|(\mathbb{P}_n\boldsymbol{\Psi}_j\boldsymbol{\Psi}_j^T)^{-1}\| - \|(E\boldsymbol{\Psi}_j\boldsymbol{\Psi}_j^T)^{-1}\|\right| \geq c_8\|(E\boldsymbol{\Psi}_j\boldsymbol{\Psi}_j^T)^{-1}\|\right\} \\ \leq \ & 2d_n^2\exp\Big(-c_4nd_n^{-3}\Big). \end{aligned}$$

Since by Fact 3,

$$\left\|(E\boldsymbol{\Psi}_j\boldsymbol{\Psi}_j^T)^{-1}\right\| \leq D_1^{-1}d_n,$$

it follows that

$$P\left\{\left\|(\mathbb{P}_n\boldsymbol{\Psi}_j\boldsymbol{\Psi}_j^T)^{-1}\right\| \geq (c_8+1)D_1^{-1}d_n\right\} \leq 2d_n^2\exp\Big(-c_4nd_n^{-3}\Big). \quad (26)$$

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

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

To bound S_2, we note that

$$\begin{aligned} |S_2| & \leq 2\|\mathbb{P}_n\boldsymbol{\Psi}_jY - E\boldsymbol{\Psi}_jY\| \cdot \|(\mathbb{P}_n\boldsymbol{\Psi}_j\boldsymbol{\Psi}_j^T)^{-1}E\boldsymbol{\Psi}_jY\| \\ & \leq 2\|\mathbb{P}_n\boldsymbol{\Psi}_jY - E\boldsymbol{\Psi}_jY\| \cdot \|(\mathbb{P}_n\boldsymbol{\Psi}_j\boldsymbol{\Psi}_j^T)^{-1}\| \cdot \|E\boldsymbol{\Psi}_jY\|. \quad (28) \end{aligned}$$

Since by Condition D,

$$\|E\mathbf{\Psi}_jY\|^2 = \sum_{k=1}^{d_n}(E\Psi_{jk}Y)^2 = \sum_{k=1}^{d_n}(E\Psi_{jk}m)^2 \leq \sum_{k=1}^{d_n} B_1^2 E\Psi_{jk}^2 \leq B_1^2C_2, \tag{29}$$

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

$$\begin{aligned}&P(|S_2| \geq 2(c_8+1)D_1^{-1}C_2^{1/2}B_1d_n^{3/2}\delta/n)\\ &\leq 4d_n\exp(-\delta^2/2(c_6nd_n^{-1}+c_7\delta)) + 2d_n^2\exp\Big(-c_4nd_n^{-3}\Big).\end{aligned} \tag{30}$$

Now we bound S_3. Note that

$$S_3 = (E\mathbf{\Psi}_jY)^T(\mathbb{P}_n\mathbf{\Psi}_j\mathbf{\Psi}_j^T)^{-1}\Big(E-\mathbb{P}_n\Big)\mathbf{\Psi}_j\mathbf{\Psi}_j^T(E\mathbf{\Psi}_j\mathbf{\Psi}_j^T)^{-1}E\mathbf{\Psi}_jY. \tag{31}$$

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

$$|S_3| \leq \|(\mathbb{P}_n-E)\mathbf{\Psi}_j\mathbf{\Psi}_j^T\|\cdot\|(\mathbb{P}_n\mathbf{\Psi}_j\mathbf{\Psi}_j^T)^{-1}\|\cdot\|(E\mathbf{\Psi}_j\mathbf{\Psi}_j^T)^{-1}\|\cdot\|E\mathbf{\Psi}_jY\|^2. \tag{32}$$

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

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

Therefore, $\|\mathbf{D}\| \leq d_n\|\mathbf{D}\|_\infty$. We conclude that

$$\Big\|\Big(\mathbb{P}_n-E\Big)\mathbf{\Psi}_j\mathbf{\Psi}_j^T)\Big\| \leq d_n\|(\mathbb{P}_n-E)\mathbf{\Psi}_j\mathbf{\Psi}_j^T\|_\infty. \tag{33}$$

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

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

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},

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

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

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

The first part thus follows the union bound of probability.

To prove the second part, note that on the event

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

by Lemma 1, we have

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

Hence, 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:

$$P(A_n^c) \leq s_n\Big\{(8d_n + 2d_n^2)\exp\Big(-c_3 n^{1-4\kappa} d_n^{-3}\Big) + 6d_n^2 \exp\Big(-c_4 n d_n^{-3}\Big)\Big\}.$$

This completes the proof. □

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

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

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

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

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

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

the number of $\{j : \|\hat{f}_{nj}\|_n^2 > 2\varepsilon d_n n^{-2\kappa}\}$ can not exceed the number of $\{j : \|f_{nj}\|^2 > \varepsilon d_n n^{-2\kappa}\}$, which is bounded by $O\{n^{2\kappa}\lambda_{\max}(\mathbf{\Sigma})\}$. By taking $\varepsilon = c_5/2$, we have

$$P[|\widehat{\mathcal{M}}_{\nu_n}| \leq O\{n^{2\kappa}\lambda_{\max}(\mathbf{\Sigma})\}] \geq P(B_n).$$

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

$$\boldsymbol{\alpha}_n = \mathrm{argmin}_{\boldsymbol{\alpha}} E\Big(Y - \boldsymbol{\Psi}^T\boldsymbol{\alpha}\Big)^2,$$

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

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

Hence

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

Now, it follows from the orthogonal decomposition that

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

Since $\mathrm{var}(Y) = O(1)$, we conclude that $\mathrm{var}(\boldsymbol{\Psi}^T\boldsymbol{\alpha}_n) = O(1)$, i.e.

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

This completes the proof. □.

References

Antoniadis, A. and Fan, J. (2001). Regularization of wavelet approximations. *Journal of the American Statistical Association*, **96** 939–967.

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.

Fan, J. (1997). Comments on "wavelets in statistics: A review" by a. antoniadis. j. *Journal of the American Statistical Association*, **6** 131–138.

Fan, J. and Jiang, J. (2005). Nonparametric inferences for additive models. *Journal of the American Statistical Association*, **100** 890–907.

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.

Fan, J. and Lv, J. (2008). Sure independence screening for ultrahigh dimensional feature space. *Journal of the Royal Statistical Society, Series B: Statistical Methodology*, **70** 849–911.

Fan, J. and Lv, J. (2009). Non-concave penalized likelihood with np-dimensionality. *Manuscript*.

Fan, J., Samworth, R. and Wu, Y. (2009). Ultra-dimensional variable selection via independent learning: beyond the linear model. *Journal of Machine Learning Research*. To appear.

Fan, J. and Song, R. (2009). Sure independence screening in generalized linear models with np-dimensionality. *Submitted*.

Hall, P. and Miller, H. (2009). Using generalised correlation to effect variable selection in very high dimensional problems. *The Journal of Computational and Graphical Statistics*. To appear.

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, Series B: Statistical Methodology*. To appear.

HÄRDLE, W., MÜLLER, M., SPERLICH, S. and WERWATZ, A. (2004). *Nonparametric and Semiparametric Models*. Springer-Verlag Inc.

HOROWITZ, J., KLEMELÄ, J. and MAMMEN, E. (2006). Optimal estimation in additive regression models. *Bernoulli*, **12** 271–298.

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.

HUANG, J., HOROWITZ, J. and WEI, F. (2009). Variable selection in nonparametric additive models. *manuscript*.

KIM, Y., KIM, J. and KIM, Y. (2006). Blockwise sparse regression. *Statistica Sinica*, **16** 375–390.

KOLTCHINSKII, V. and YUAN, M. (2008). Sparse recovery in large ensembles of kernel machines. *In CLOT (eds. R.A. Servedio and T. Zhang), Omnipress.* 229–238.

LIN, Y. and ZHANG, H. H. (2006). Component selection and smoothing in multivariate nonparametric regression. *The Annals of Statistics*, **34** 2272–2297.

MEIER, L., GEER, V. and BÜHLMANN, P. (2009). High-dimensional additive modeling. *The Annals of Statistics*. To appear.

RAVIKUMAR, P., LIU, H., LAFFERTY, J. and WASSERMAN, L. (2009). Spam: Sparse additive models. *Journal of the Royal Statistical Society: Series B*. To appear.

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.

SILVERMAN, B. (1984). Spline smoothing: The equivalent variable kernel method. *The Annals of Statistics*, **12** 898–916.

SPECKMAN, P. (1985). Spline smoothing and optimal rates of convergence in nonparametric regression models. *The Annals of Statistics*, **13** 970–983.

STONE, C. (1985). Additive regression and other nonparametric models. *The Annals of Statistics*, **13** 689–705.

TIBSHIRANI, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B*, **58** 267–288.

VAN DER VAART, A. W. and WELLNER, J. A. (1996). *Weak Convergence and Empirical Processes*. Springer, New York.

WEI, F. and HUANG, J. (2007). Consistent group selection in high-dimensional linear regression. *Technical Report No.387*. University of Iowa.

YUAN, M. and LIN, Y. (2006). Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B*, **68** 49–67.

ZHANG, C.-H. (2009). Nearly unbiased variable selection under minimax concave penalty. *The Annals of Statisitics*. To appear.

Zhou, S., Shen, X. and Wolfe, D. A. (1998). Local asymptotics for regression splines and confidence regions. *The Annals of Statistics*, **26** 1760–1782.

Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American Statistical Association*, **101** 1418–1429.

Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society, Series B: Statistical Methodology*, **67** 768–768.

Zou, H. and Li, R. (2008). One-step sparse estimates in nonconcave penalized likelihood models. *The Annals of Statistics*, **36** 1509–1533.