



## Error Variance Estimation in Ultrahigh-Dimensional Additive Models

Zhao Chen, Jianqing Fan & Runze Li

To cite this article: Zhao Chen, Jianqing Fan & Runze Li (2018) Error Variance Estimation in Ultrahigh-Dimensional Additive Models, Journal of the American Statistical Association, 113:521, 315-327, DOI: [10.1080/01621459.2016.1251440](https://doi.org/10.1080/01621459.2016.1251440)

To link to this article: <https://doi.org/10.1080/01621459.2016.1251440>



View supplementary material [↗](#)



Accepted author version posted online: 16 Dec 2016.  
Published online: 26 Sep 2017.



Submit your article to this journal [↗](#)



Article views: 601



View Crossmark data [↗](#)



## Error Variance Estimation in Ultrahigh-Dimensional Additive Models

Zhao Chen<sup>a,c</sup>, Jianqing Fan<sup>b,c</sup>, and Runze Li<sup>d</sup>

<sup>a</sup>Department of Statistics, The Pennsylvania State University at University Park, PA; <sup>b</sup>School of Data Science, Fudan University; <sup>c</sup>Department of Operations Research & Financial Engineering, Princeton University, Princeton, NJ; <sup>d</sup>Department of Statistics and The Methodology Center, The Pennsylvania State University, University Park, PA

### ABSTRACT

Error variance estimation plays an important role in statistical inference for high-dimensional regression models. This article concerns with error variance estimation in high-dimensional sparse additive model. We study the asymptotic behavior of the traditional mean squared errors, the naive estimate of error variance, and show that it may significantly underestimate the error variance due to spurious correlations that are even higher in nonparametric models than linear models. We further propose an accurate estimate for error variance in ultrahigh-dimensional sparse additive model by effectively integrating sure independence screening and refitted cross-validation techniques. The root  $n$  consistency and the asymptotic normality of the resulting estimate are established. We conduct Monte Carlo simulation study to examine the finite sample performance of the newly proposed estimate. A real data example is used to illustrate the proposed methodology. Supplementary materials for this article are available online.

### ARTICLE HISTORY

Received December 2013  
Revised August 2016

### KEYWORDS

Feature screening; Refitted cross-validation; Sparse additive model; Variance estimation

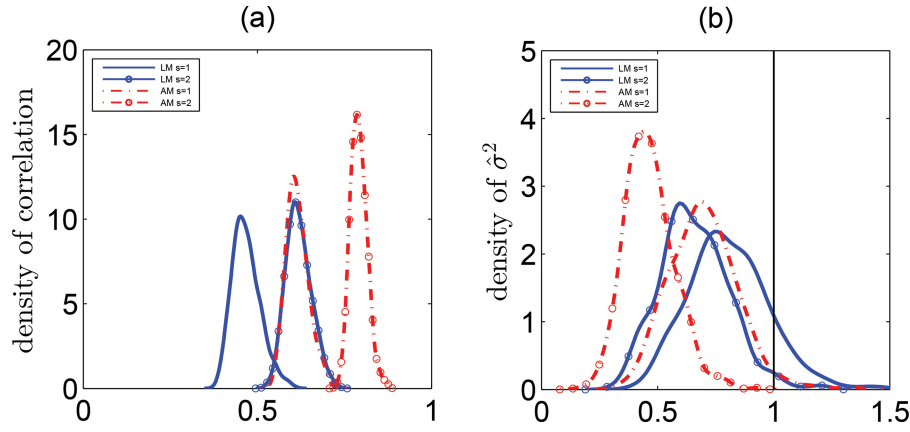
### 1. Introduction

Statistical inference on regression models typically involves the estimation of the variance of its random error. Hypothesis testing on regression functions, confidence/prediction interval construction, and variable selection all require an accurate estimate of the error variance. In the classical linear regression analysis, the adjusted mean squared error is an unbiased estimate of the error variance, and it performs well when the sample size is much larger than the number of predictors, or more accurately when the degree of freedom is large. It has been empirically observed that the mean squared error estimator leads to an underestimation of the error variance when model is significantly over-fitted. This has been further confirmed by the theoretical analysis by Fan, Guo, and Hao (2012), in which the authors demonstrated the challenges of error variance estimation in the high-dimensional linear regression analysis, and further developed an accurate error variance estimator by introducing refitted cross-validation techniques.

Fueled by the demand in the analysis of genomic, financial, health, and image data, the analysis of high-dimensional data has become one of the most important research topics during last two decades (Donoho 2000; Fan and Li 2006). There have been a huge number of research articles on high-dimensional data analysis in the literature. It is impossible for us to give a comprehensive review here. Readers are referred to Fan and Lv (2010), Bühlmann and Van de Geer (2011), and references therein. Due to the complex structure of high-dimensional data, the high-dimensional linear regression analysis may be a good start, but it may not be powerful to explore nonlinear features inherent into data. Nonparametric regression modeling

provides valuable analysis for high-dimensional data (Ravikumar et al. 2009; Hall and Miller 2009; Fan, Feng, and Song 2011). This is particularly the case for error variance estimation, as nonparametric modeling reduces modeling biases in the estimate, but creates stronger spurious correlations. This article aims to study issues of error variance estimation in ultrahigh-dimensional nonparametric regression settings.

In this article, we focus on sparse additive model. Our primary interest is to develop an accurate estimator for error variance in ultrahigh-dimensional additive model. The techniques developed in this article are applicable to other nonparametric regression models such as sparse varying coefficient models and some commonly used semiparametric regression models such as sparse partial linear additive models and sparse semivarying coefficient partial linear models. Since its introduction by Friedman and Stuetzle (1981), additive model has been popular, and many statistical procedures have been developed for sparse additive models in the recent literature. Lin and Zhang (2006) proposed COSSO method to identify significant variables in multivariate nonparametric models. Bach (2008) studied penalized least-square regression with group Lasso-type penalty for linear predictors and regularization on reproducing kernel Hilbert space norms, which is referred to as multiple kernel learning. Xue (2009) studied variable selection problem in additive models by integrating a group-SCAD penalized least-square method (Fan and Li 2001) and the regression spline technique. Ravikumar et al. (2009) modified the backfitting algorithm for sparse additive models, and further established the model selection consistency of their procedure. Meier, Van de Geer and Bühlmann (2009) studied the model selection



**Figure 1.** Distributions of the maximum “linear” and “nonparametric” spurious correlations for  $s = 1$  and  $s = 2$  (left panel,  $n = 50$  and  $p = 1000$ ) and their consequences on the estimating of noise variances (right panel). The legend “LM” stands for linear model, and “AM” stands for additive model, that is, nonparametric model.

and estimation of additive models with a diverging number of significant predictors. They proposed a new sparsity and smoothness penalty and proved that their method can select all nonzero components with probability approaching to 1 as the sample size tends to infinity. With the ordinary group Lasso estimator as the initial estimator, Huang, Horowitz and Wei (2010) applied adaptive group Lasso to additive model under the setting in which there are only finite fixed number of significant predictors. Fan, Feng and Song (2011) proposed a nonparametric independent screening procedure for sparse ultrahigh-dimensional data, and established its sure screening property in the terminology by Fan and Lv (2008).

In this article, we propose an error variance estimate in ultrahigh-dimensional additive models. It is typical to assume sparsity in ultrahigh-dimensional data analysis. By sparsity, it means that the regression function depends only on a few significant predictors, and the number of significant predictors is assumed to be much smaller than the sample size. Because of the basis expansion in nonparametric fitting, the actual number of terms significantly increases in additive models. Therefore, the spurious correlation documented by Fan, Guo, and Hao (2012) increases significantly. This is indeed demonstrated in Lemma 1, which shows that the spurious correlation with the response increases from  $\sqrt{n^{-1} \log(p)}$  using one most correlated predictor among  $p$  variables to  $\sqrt{d_n n^{-1} \log(p d_n)}$  by using one most correlated predictor with  $d_n$  basis functions. If  $s$  variables are used, the spurious correlation may increase to its upper bound at an exponential rate of  $s$ .

To quantify this increase and explain more clearly the concept and the problem, we simulate  $n = 50$  data points from the independent normal covariates  $\{X_j\}_{j=1}^p$  (with  $p = 1000$ ) and also independently normal response  $Y$ . In this null model, all covariates  $\{X_j\}_{j=1}^p$  and the response  $Y$  are independent and follow the standard normal distribution. As by Fan, Guo, and Hao (2012), we compute the maximum “linear” spurious correlation  $\zeta_n^L = \max_{1 \leq j \leq p} |\widehat{\text{corr}}(X_j, Y)|$  and the maximum “nonparametric” spurious correlation  $\zeta_n^N = \max_{1 \leq j \leq p} |\widehat{\text{corr}}(\hat{f}_j(X_j), Y)|$ , where  $\hat{f}_j(X_j)$  is the best cubic spline fit of variable  $X_j$  to the response  $Y$ , using 3 equally spaced knots in the range of the variable  $X_j$  which create  $d_n = 6$  B-spline bases for  $X_j$ . The concept of the maximum spurious “linear” and spurious “nonparametric” (additive) correlations can easily be extended to  $s$  variables,

which are the correlation between the response and fitted values using the best subset of  $s$ -variables. Based on 500 simulated datasets, Figure 1 depicts the results that show the big increase of spurious correlations from linear to nonparametric fit. As the result, the noise variance is significantly underestimated.

The above reasoning and evidence show that the naive estimation of error variance is seriously biased. This is indeed shown in Theorem 1. This prompts us to propose a two-stage refitted cross-validation procedure to reduce spurious correlation. In the first stage, we apply a sure independence screening procedure to reduce the ultrahigh dimensionality to relative large dimensional regression problem. In the second stage, we apply refitted cross-validation technique, which was proposed for linear regression model by Fan, Guo, and Hao (2012), for the dimension-reduced additive models obtained from the first stage. The implementation of the newly proposed procedure is not difficult. However, it is challenging in establishing its sampling properties. This is because the dimensionality of ultrahigh-dimensional sparse additive models becomes even higher.

We propose using B-splines to approximate the nonparametric functions, and first study the asymptotic properties of the traditional mean squared error, a naive estimator of the error variance. Under some mild conditions, we show that the mean squared error leads to a significant underestimate of the error variance. We then study the sampling properties of the proposed refitted cross-validation estimate, and establish its asymptotic normality. From our theoretical analysis, it can be found that the refitted cross-validation techniques can eliminate the side effects due to over-fitting. We also conduct Monte Carlo simulation studies to examine the finite sample performance of the proposed procedure. Our simulation results show that the newly proposed error variance estimate may perform significantly better than the mean squared error.

This article makes the following major contributions. (a) We show the traditional mean squared errors as a naive estimation of error variance is seriously biased. Although this is expected, the rigorous theoretical development indeed is challenging rather than straightforward. (b) We propose a refitted cross-validation error variance estimation for ultrahigh-dimensional nonparametric additive models, and further establish the asymptotic normality of the proposed estimator. The asymptotic normality implies that the proposed estimator is

asymptotic unbiased and root  $n$  consistent. The extensions of refitted cross-validation error variance estimation from linear models to nonparametric models are interesting, and not straightforward in terms of theoretical development because the bias due to approximation error calls for new techniques to establish the theory. Furthermore, the related techniques developed in this article may be further applied for refitted cross-validation error variance estimation in other ultrahigh-dimensional nonparametric regression models such as varying coefficient models and ultrahigh-dimensional semiparametric regression models such as partially linear additive models and semiparametric partially linear varying coefficient models.

This article is organized as follows. In [Section 2](#), we propose a new error variance estimation procedure, and further study its sampling properties. In [Section 3](#), we conduct Monte Carlo simulation studies to examine the finite sample performance of the proposed estimator, and demonstrate the new estimation procedure by a real data example. Some concluding remarks are given in [Section 4](#). Technical conditions and proofs are given in the [Appendix](#).

## 2. New Procedures for Error Variance Estimation

Let  $Y$  be a response variable, and  $\mathbf{x} = (X_1, \dots, X_p)^T$  be a predictor vector. The additive model assumes that

$$Y = \mu + \sum_{j=1}^p f_j(X_j) + \varepsilon, \quad (2.1)$$

where  $\mu$  is intercept term,  $\{f_j(\cdot), j = 1, \dots, p\}$  are the unknown functions and  $\varepsilon$  is the random error with  $E(\varepsilon) = 0$  and  $\text{var}(\varepsilon) = \sigma^2$ . Following the convention in the literature, it is assumed throughout this article that  $E f_j(X_j) = 0$  for  $j = 1, \dots, p$  so that model (2.1) is identifiable. This assumption implies that  $\mu = E(Y)$ . Thus, a natural estimator for  $\mu$  is the sample average of  $Y$ 's. This estimator is root  $n$  consistent, and its rate of convergence is faster than that for the estimator of nonparametric function  $f_j$ 's. Without loss of generality, we further assume  $\mu = 0$  for ease of notation. The goal of this section is to develop an estimation procedure for  $\sigma^2$  for additive models.

### 2.1. Refitted Cross-Validation

In this section, we propose a strategy to estimate the error variance when the predictor vector is ultrahigh-dimensional. Since  $f_j$ 's are nonparametric smoothing functions, it is natural to use smoothing techniques to estimate  $f_j$ . In this article, we employ B-spline method throughout this article. Readers are referred to De Boor (1978) for detailed procedure of B-spline construction. Let  $\{B_{jk}(x), k = 1, \dots, d_j, a \leq x \leq b\}$  be B-spline basis of space  $S_j^l([a, b])$  with knots depending on  $j$ , the polynomial spline space defined on finite interval  $[a, b]$  with degree  $l \geq 1$ . Approximate  $f_j$  by its spline expansion

$$f_j(x) \approx \sum_{k=1}^{d_j} \gamma_{jk} B_{jk}(x) \quad (2.2)$$

for some  $d_j \geq 1$ . In practice,  $d_j$  is allowed to grow with the sample size  $n$ , and therefore denoted by  $d_{jn}$  to emphasize the dependence of  $n$ . With slightly abuse of notation, we use  $d_n$  stands for  $d_{jn}$  for ease of notation. Thus, model (2.1) can be written as

$$Y \approx \sum_{j=1}^p \sum_{k=1}^{d_n} \gamma_{jk} B_{jk}(X_j) + \varepsilon. \quad (2.3)$$

Suppose that  $\{(\mathbf{x}_i, Y_i)\}, i = 1, \dots, n$  is a random sample from the additive model (2.1). Model (2.3) is not estimable when  $pd_n > n$ . It is common to assume sparsity in ultrahigh-dimensional data analysis. By sparsity in additive model, it means that only a few  $\|f_j\|^2 = E f_j^2(X_j) \neq 0$  and other  $\|f_j\| = 0$ . A general strategy to reduce ultrahigh dimensionality is sure independent feature screening, which enables one to reduce ultrahigh dimension to large or high dimension. Some existing feature screening procedures can be directly applied for ultrahigh-dimensional sparse additive models. Fan, Feng, and Song (2011) proposed nonparametric sure independent (NIS) screening method and further showed that the NIS screening method possesses sure screening property for ultrahigh-dimensional additive models. That is, under some regularity conditions, with an overwhelming probability, the NIS is able to retain all active predictors after feature screening. Li, Zhong, and Zhu (2012) proposed a model-free feature screening procedure based on distance correlation sure independent screening (DC-SIS). The DC-SIS is also shown to have sure screening property. Both NIS and DC-SIS can be used for feature screening with ultrahigh-dimensional sparse additive models, although we will use DC-SIS in our numerical implementation due to its intuitive and simple implementation.

Hereafter we always assume that all important variables have been selected by screening procedure. Under such assumption, we will overfit the response variable  $Y$  and underestimate the error variance  $\sigma^2$ . This is because extra variables are actually selected to predict the realized noises (Fan, Guo, and Hao 2012). After feature screening, a direct estimate of  $\sigma^2$  is the mean squared errors of the least-square approach. That is, we apply a feature screening procedure such as DC-SIS and NIS to screen  $x$ -variables and fit the data to the corresponding selected spline regression model. Denoted by  $\mathcal{D}^*$  the indices of all true predictors and  $\hat{\mathcal{D}}^*$  the indices of the selected predictors, respectively, satisfying the sure screening property  $\mathcal{D}^* \subset \hat{\mathcal{D}}^*$ . Then, we minimize the following least-square function with respect to  $\gamma$ :

$$\sum_{i=1}^n \left\{ Y_i - \sum_{j \in \hat{\mathcal{D}}^*} \sum_{k=1}^{d_n} \gamma_{jk} B_{jk}(X_{ij}) \right\}^2. \quad (2.4)$$

Denote by  $\hat{\gamma}_{jk}$  the resulting least-square estimate. Then, the nonparametric residual variance estimator is

$$\hat{\sigma}_{\hat{\mathcal{D}}^*}^2 = \frac{1}{n - |\hat{\mathcal{D}}^*| \cdot d_n} \sum_{i=1}^n \left\{ Y_i - \sum_{j \in \hat{\mathcal{D}}^*} \sum_{k=1}^{d_n} \hat{\gamma}_{jk} B_{jk}(X_{ij}) \right\}^2.$$

Hereafter  $|\mathcal{D}|$  stands for the cardinality of a set  $\mathcal{D}$  and we have implicitly assumed that the choice of  $\hat{\mathcal{D}}^*$  and  $d_n$  is such that

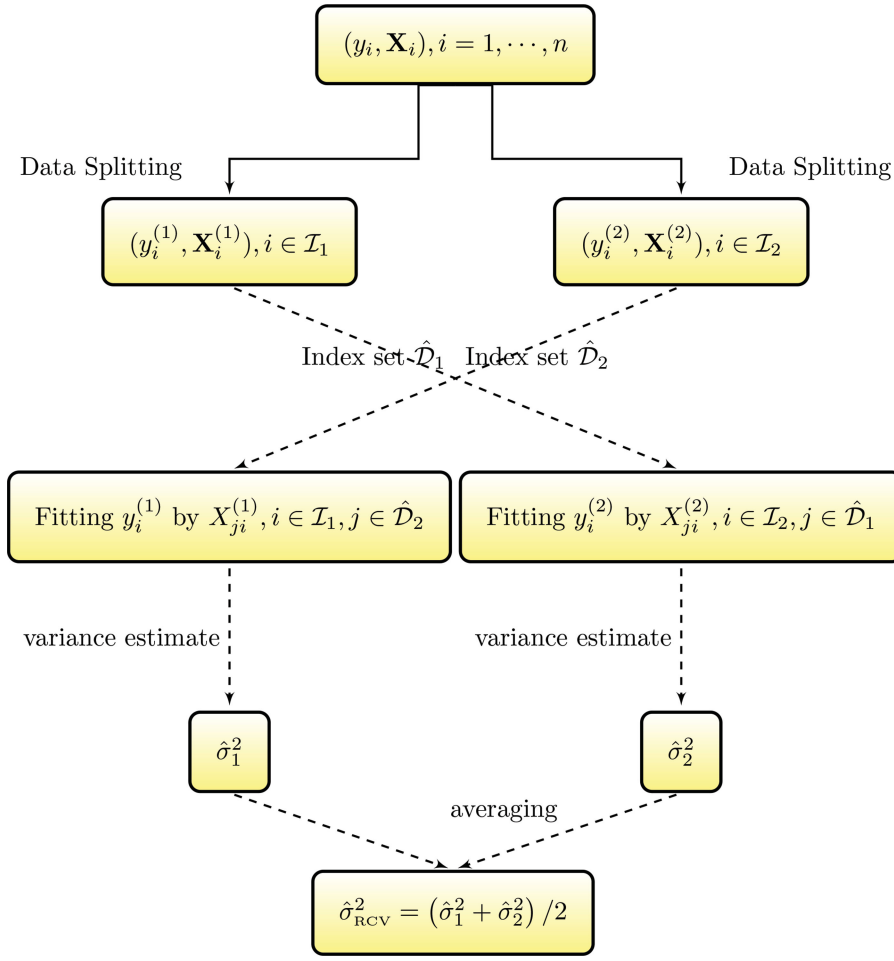


Figure 2. Refitted cross-validation procedure.

$n \gg |\hat{\mathcal{D}}^*| \cdot d_n$ . It will be shown in Theorem 1 that  $\hat{\sigma}_{\hat{\mathcal{D}}^*}^2$  significantly underestimates  $\sigma^2$ , due to spurious correlation between the realized but unobserved noises and the spline bases. Indeed we will show that  $\hat{\sigma}_{\hat{\mathcal{D}}^*}^2$  is inconsistent estimate when  $|\hat{\mathcal{D}}^*| \cdot d_n$  is large. Specifically, let  $\mathbf{P}_{\hat{\mathcal{D}}^*}$  be the corresponding projection matrix of model (2.4) with the entire samples. Denoted by  $\hat{\gamma}_n^2 = \boldsymbol{\varepsilon}^T \mathbf{P}_{\hat{\mathcal{D}}^*} \boldsymbol{\varepsilon} / \boldsymbol{\varepsilon}^T \boldsymbol{\varepsilon}$ , where  $\boldsymbol{\varepsilon} = (\varepsilon_1, \dots, \varepsilon_n)^T$ . We will show that  $\hat{\sigma}_{\hat{\mathcal{D}}^*}^2 / (1 - \hat{\gamma}_n^2)$  converges to  $\sigma^2$  with root  $n$  convergence rate, yet the spurious correlation  $\hat{\gamma}_n^2$  is of order

$$\hat{\gamma}_n^2 = O\left(\left(\frac{2}{1-\delta}\right)^{|\hat{\mathcal{D}}^*|} \frac{d_n \log(pd_n)}{n}\right), \text{ for some } \delta \in (0, 1). \quad (2.5)$$

See Lemma 1 and Theorem 1 in Section 2.2 for details. Our first aim is to propose a new estimation procedure of  $\sigma^2$  by using refitted cross-validation technique (Fan, Guo, and Hao 2012).

The refitted cross-validation procedure is to randomly split the random samples into two datasets denoted by  $\mathcal{I}_1$  and  $\mathcal{I}_2$  with approximately equal size. Without loss of generality, assume through this article that  $\mathcal{I}_1$  and  $\mathcal{I}_2$  have the same sample size  $n/2$ . We apply a feature screening procedure (e.g., DC-SIS or NIS) for each set, and obtain two index sets of selected  $x$ -variables, denoted by  $\hat{\mathcal{D}}_1$  and  $\hat{\mathcal{D}}_2$ . Both of them retain all important predictors. The refitted cross-validation procedure consists of three steps. In the first step, we fit data in  $\mathcal{I}_l$  to the selected additive

model  $\hat{\mathcal{D}}_{3-l}$  for  $l = 1$  and  $2$  by the least-square method. These results in two least-square estimate  $\hat{\gamma}^{(3-l)}$  based on  $\mathcal{I}_l$ , respectively. In the second step, we calculate the mean squared errors for each fit:

$$\hat{\sigma}_l^2 = \frac{1}{n/2 - |\hat{\mathcal{D}}_{3-l}| \cdot d_n} \sum_{i \in \mathcal{I}_l} \left\{ Y_i - \sum_{j \in \hat{\mathcal{D}}_{3-l}} \sum_{k=1}^{d_n} \hat{\gamma}_{jk}^{(3-l)} B_{jk}(X_{ij}) \right\}^2$$

for  $l = 1$  and  $2$ . Then the refitted cross-validation estimate of  $\sigma^2$  is defined by

$$\hat{\sigma}_{\text{RCV}}^2 = (\hat{\sigma}_1^2 + \hat{\sigma}_2^2) / 2.$$

This estimator is adapted from the one proposed by Fan, Guo, and Hao (2012) for linear regression models, however, it is much more challenge in establishing the asymptotic property of  $\hat{\sigma}_{\text{RCV}}^2$  for the large dimensional additive models than that for linear regression models. The major hurdle is to deal with the approximation error in nonparametric modeling as well as the correlation structure induced by the B-spline bases. The procedure of refitted cross-validation is illustrated schematically in Figure 2.



## 2.2. Sampling Properties

We next study the asymptotic properties of  $\hat{\sigma}_{\hat{\mathcal{D}}^*}^2$  and  $\hat{\sigma}_{\text{RCV}}^2$ . The following technical conditions are needed to facilitate the proofs, although they may not be the weakest.

(C1) There exist two positive constants  $A_1$  and  $A_2$  such that  $E\{\exp(A_1|\varepsilon|)|\mathbf{x}\} \leq A_2$ .

(C2) For all  $j$ ,  $f_j(\cdot) \in \mathcal{C}^d([a, b])$ , which consists of functions whose  $r$ th derivative  $f_j^{(r)}$  exists and satisfies

$$|f_j^{(r)}(s) - f_j^{(r)}(t)| \leq L|s - t|^\alpha, \text{ for } s, t \in [a, b], j = 1, \dots, p, \quad (2.6)$$

for a given constant  $L > 0$ , where  $r \leq l$  is the “integer part” of  $d$  and  $\alpha \in (0, 1]$  such that  $d = r + \alpha \geq 2$ . Furthermore, it is assumed that  $d_n = O(n^{1/(2d+1)})$ , the optimal nonparametric rate (Stone 1985).

(C3) The joint distribution of predictors  $\mathbf{X}$  is absolutely continuous and its density  $g$  is bounded by two positive numbers  $b$  and  $B$  satisfying that  $b \leq g \leq B$ . The predictor  $X_j$ ,  $j = 1, \dots, p$  has a continuous density function  $g_j$ , which satisfies that for any  $x \in [a, b]$ ,  $0 < A_3 \leq g_j(x) \leq A_4 < \infty$  for two positive constants  $A_3$  and  $A_4$ .

Condition (C1) is a tail condition on the random error. Condition (C2) is a typical smoothness condition in the literature of regression splines. Condition (C3) is a mild condition on the densities of the predictors, and this condition was imposed by Stone (1985) for low-dimensional additive models, and implies that there is no collinearity between the candidate predictors with probability one. The asymptotic properties of  $\hat{\sigma}_{\hat{\mathcal{D}}^*}^2$  are given in the following theorem, in which we use  $p_n$  to stand for  $p$  to emphasize that the dimension  $p$  of the predictor vector may depend on  $n$ . Since the DC-SIS and the NIS possess sure screening property, the resulting subset of predictors selected by the used screening procedure contains all active predictors, with probability tending to one. Thus, we assume that all active predictors are retained in the stage of feature screening in the following two theorems. This can be achieved by imposing the conditions by Li, Zhong, and Zhu (2012) for the DC-SIS and the conditions by Fan, Feng, and Song (2011) for the NIS. We first derive the orders of  $\boldsymbol{\varepsilon}^T \mathbf{P}_{\hat{\mathcal{D}}^*} \boldsymbol{\varepsilon}$  and  $\hat{\gamma}_n^2$  in next lemma, which plays a critical role in the proofs of Theorems 1 and 2. The proofs of Lemma 1 and Theorems 1 and 2 will be given in the Appendix.

**Lemma 1.** Under Conditions (C1)|(C3), it follows that

$$\boldsymbol{\varepsilon}^T \mathbf{P}_{\hat{\mathcal{D}}^*} \boldsymbol{\varepsilon} = O_p \left\{ \left( \frac{2}{1-\delta} \right)^{\hat{s}} d_n \log(p d_n) \right\},$$

$$\text{and } \hat{\gamma}_n^2 = \frac{\boldsymbol{\varepsilon}^T \mathbf{P}_{\hat{\mathcal{D}}^*} \boldsymbol{\varepsilon}}{\boldsymbol{\varepsilon}^T \boldsymbol{\varepsilon}} = O_p \left\{ \left( \frac{2}{1-\delta} \right)^{\hat{s}} \frac{d_n \log(p d_n)}{n} \right\},$$

where  $\delta \in (\sqrt{1 - b^2 \zeta_0 / B^2}, 1)$  for some constant  $\zeta_0 \in (0, 1)$  with  $b$  and  $B$  being given in Condition (C3).

**Lemma 1** clearly shows that the spurious correlation  $\hat{\gamma}_n^2$  increases to its upper bound at an exponential rate of  $\hat{s}$  since  $\delta \in (0, 1)$  and  $2/(1-\delta) > 2$ .

**Theorem 1.** Assume that  $\limsup_{n \rightarrow \infty} \hat{\gamma}_n^2 < 1$ . Let  $\hat{s} = |\hat{\mathcal{D}}^*|$  be the number of elements in the estimated active index set  $\hat{\mathcal{D}}^*$ .

Assume that all active predictors are retained in the stage of feature screening. That is,  $\hat{\mathcal{D}}^*$  contains all active predictors. Under Conditions (C1)–(C3), the following statements hold:

- (i) If  $\log(p_n) = O(n^\zeta)$ ,  $0 \leq \zeta < 1$  and  $\hat{s} = O_p(\log(n))$ , then  $\hat{\sigma}_{\hat{\mathcal{D}}^*}^2 / (1 - \hat{\gamma}_n^2)$  converges to  $\sigma^2$  in probability as  $n \rightarrow \infty$ ;
- (ii) If  $\log(p_n) = O(n^\zeta)$ ,  $0 \leq \zeta < 3/(2d+1)$  and  $\hat{s} = O_p(\log(n))$ , then it follows that

$$\sqrt{n} \left( \hat{\sigma}_{\hat{\mathcal{D}}^*}^2 / (1 - \hat{\gamma}_n^2) - \sigma^2 \right) \xrightarrow{\mathcal{L}} \mathcal{N}(0, E\varepsilon_1^4 - \sigma^4), \quad (2.7)$$

where  $\xrightarrow{\mathcal{L}}$  stands for convergence in law.

**Theorem 1** (i) clearly indicates that the naive error variance estimator  $\hat{\sigma}_{\hat{\mathcal{D}}^*}^2$  underestimates  $\sigma^2$  by a factor of  $(1 - \hat{\gamma}_n^2)$ , yet by **Lemma 1**,  $\hat{\gamma}_n^2$  is of order given in (2.5) and is not small. Since  $\hat{\gamma}_n^2$  cannot be estimated directly from the data, it is challenging to derive an adjusted error variance by modifying the commonly used mean squared errors. On the other hand, the refitted cross-validation method provides an automatic bias correction via refitting and hence a consistent estimator, as we now show.

**Theorem 2.** Assume that  $\hat{\mathcal{D}}_j^*$  contains all active predictors, for  $j = 1$  and  $2$ . Let  $\hat{s}_j = |\hat{\mathcal{D}}_j^*|$  be the number of elements in  $\hat{\mathcal{D}}_j^*$ . Under Conditions (C1)–(C3), if  $\hat{s}_1 = o(n^{(2d-1)/4(2d+1)})$ , and  $\hat{s}_2 = o(n^{(2d-1)/4(2d+1)})$ , then

$$\sqrt{n} (\hat{\sigma}_{\text{RCV}}^2 - \sigma^2) \xrightarrow{\mathcal{L}} \mathcal{N}(0, E\varepsilon_1^4 - \sigma^4). \quad (2.8)$$

Comparing with the result in **Theorem 1**, the refitted cross-validation method can eliminate the side-effect of the selected redundant variables to correct the bias of the naive variance estimator through the contributions of refitting. This bias factor can be nontrivial.

**Remark 1.** This remark provides some implications and limitations of **Theorems 1** and **2** and some clarification of conditions implicitly required by **Theorem 2**.

- (a) From the proof of **Theorems 1** and **2**, it has been shown that  $\hat{\sigma}_{\hat{\mathcal{D}}^*}^2 / (1 - \hat{\gamma}_n^2) = \sigma^2 + O_p(1/\sqrt{n})$  and  $\hat{\sigma}_{\text{RCV}}^2 = \sigma^2 + O_p(1/\sqrt{n})$ . As a result, the ratio of RCV estimate to the naive estimator may be used to provide one an estimate of the shrinkage factor  $1 - \hat{\gamma}_n^2$ .
- (b) **Theorem 2** is applicable provided that the active index sets  $\hat{\mathcal{D}}_j^*$ ,  $j = 1$  and  $2$  include all active predictor variables. Here, we emphasize that the RCV method can be integrated with any dimension reduction procedure to effectively correct the bias of naive error variance estimate, and do not directly impose condition on the dimension  $p_n$ . In practical implementation, the assumption that both two active index sets include all important variables implies further condition on  $p_n$ . In particular, the condition  $\log(p_n) = o(n)$  is necessary for DC-SIS (Li, Zhong, and Zhu 2012) to achieve sure screening property. This condition is also necessary for other sure screening procedures such as the NIS (Fan, Feng, and Song 2011) to achieve sure screening property. In **Theorems 1** and **2**,

**Table 1.** Simulation results for different  $\hat{s}$  ( $\sigma^2 = 1$ )

Method	$a = 0$			
	$\hat{s} = 20$	$\hat{s} = 30$	$\hat{s} = 40$	$\hat{s} = 50$
Oracle	1.0042 (0.0618)*	1.0042 (0.0618)	1.0042 (0.0618)	1.0042 (0.0618)
Naive	0.8048 (0.0558)	0.7549 (0.0589)	0.7138 (0.0584)	0.6771 (0.0584)
RCV	1.0022 (0.0656)	0.9994 (0.0666)	0.9990 (0.0698)	0.9967 (0.0705)
$a = 1/\sqrt{3}$				
Oracle	1.0049 (0.0617)	1.0049 (0.0617)	1.0049 (0.0617)	1.0049 (0.0617)
Naive	0.9054 (0.0572)	0.8683 (0.0592)	0.8387 (0.0615)	0.8143 (0.0644)
RCV	1.0704 (0.1300)	1.0493 (0.1187)	1.0374 (0.1095)	1.0273 (0.1106)
$a = 2/\sqrt{3}$				
Oracle	1.0072 (0.0618)	1.0072 (0.0618)	1.0072 (0.0618)	1.0072 (0.0618)
Naive	0.9618 (0.0647)	0.9618 (0.0647)	0.9306 (0.0687)	0.9194 (0.0780)
RCV	1.0026 (0.0657)	1.0026 (0.0657)	1.0020 (0.0735)	1.0013 (0.0779)

NOTE: \*Values in parentheses are standard errors.

we have imposed conditions on  $\hat{s}$ ,  $\hat{s}_1$ , and  $\hat{s}_2$ . These conditions may implicitly require extra conditions on the DC-SIS to ensure that the size of the subset selected by DC-SIS is of order required by the conditions. For NIS, by Theorem 2 by Fan, Feng, and Song (2011), we need to impose some explicit conditions on the signal strength as well as the growth of the operator norm of the covariance matrix of covariates.

- (c) The RCV method can be combined with any feature screening methods such as DC-SIS and NIS and variable selection methods such as grouped LASSO and grouped SCAD (Xue 2009) for ultrahigh-dimensional additive models. The NIS method needs to choose a smoothing parameter for each predictor. The grouped LASSO and the grouped SCAD methods are expensive in terms of computational cost. We focus only on DC-SIS in the numerical studies to save space.
- (d) For sure independent screening procedures such as the SIS and DC-SIS, the authors recommended to set  $\hat{s} = n/\log(n)$ . The diverging rate of  $\hat{s}$ ,  $\hat{s}_1$ , and  $\hat{s}_2$  required in Theorems 1 and 2 are slower than this due to the nonparametric nature. It seems that it is difficult to further relax the conditions in Theorems 1 and 2. This can be viewed as a limitation of our theoretical results. From our simulation studies and real data examples, the performance of the naive method certainly relies on the choice of  $\hat{s}$ , while the RCV method performs well for a wide range of  $\hat{s}_1$  and  $\hat{s}_2$ . As shown in Tables 1 and 2, the resulting estimate of the RCV method is very close to the oracle estimate across all scenarios in the tables. Theoretical studies on how to determine  $\hat{s}_1$  and  $\hat{s}_2$  are more related to the topic of feature screening than the variance estimation and we do not intend to pursue further in this article. In practical implementation, the choices of these parameters should take into account of the degree of freedoms in the refitting stage so that the residual variance can be estimated with a reasonable accuracy. We would recommend considering several possible choices of  $\hat{s}_1$  and  $\hat{s}_2$  to examine whether the resulting variance estimate is relatively stable to the choices of  $\hat{s}_1$  and  $\hat{s}_2$ . This is implemented in the real data example in Section 3.2.

**Table 2.** Simulation results with different  $n$  ( $\sigma^2 = 1$ )

Method	$a = 0$	
	$n = 400$	$n = 600$
Oracle	1.0044 (0.0646)*	0.9924 (0.0575)
Naive	0.6969 (0.0610)	0.7340 (0.0542)
RCV	0.9905 (0.0837)	0.9845 (0.0729)
$a = 1/\sqrt{3}$		
Oracle	1.0047 (0.0737)	0.9970 (0.0552)
Naive	0.8390 (0.0815)	0.8533 (0.0555)
RCV	1.1273 (0.1528)	1.0144 (0.0954)
$a = 2/\sqrt{3}$		
Oracle	0.9903 (0.0687)	1.0075 (0.0643)
Naive	0.9013 (0.0785)	0.9340 (0.0691)
RCV	1.0241 (0.1886)	1.0031 (0.0780)

NOTE: \*Values in parentheses are standard errors.

### 3. Numerical Studies

In this section, we investigate the finite sample performances of the newly proposed procedures. We further illustrate the proposed procedure by an empirical analysis of a real data example. In our numerical studies, we report only results of the proposed RCV method with DC-SIS to save space, although the NIS method, the grouped LASSO, and the grouped SCAD (Xue 2009) can be used to screen or select variables. All numerical studies are conducted using Matlab code.

#### 3.1. Monte Carlo Simulation

Since there is little work to study the variance estimate for ultrahigh-dimensional nonparametric additive model, this simulation study is designed to compare the finite sample performances of two-stage naive variance estimate and refitted cross-validation variance estimate. In our simulation study, data were generated from the following sparse additive model:

$$y = a (X_1 + 0.75X_2^2 + 2.25 \cos(X_5)) + \varepsilon, \quad (3.1)$$

where  $\varepsilon \sim N(0, 1)$ , and  $\{X_1, \dots, X_p\} \sim N_p(0, \Sigma)$  with  $\Sigma = \{\rho_{ij}\}_{i,j=1}^p$  where  $\rho_{ii} = 1$  and  $\rho_{ij} = 0.2$  for  $i \neq j$ . We set  $p = 2000$  and  $n = 600$ . We take  $a = 0$ ,  $1/\sqrt{3}$ , and  $2/\sqrt{3}$  to examine the impact of signal-to-noise ratio to error variance estimation. When  $a = 0$ , the DC-SIS always can pick up the active sets and the challenge is to reduce spurious correlation, while when  $a = 2/\sqrt{3}$ , the signal is strong enough to pick up active sets so that DC-SIS performs very well. The case  $a = 1/\sqrt{3}$  corresponds to the signal-to-noise equaling to 1. This is a difficult case to distinguish signals and noises and is the most challenge one for DC-SIS among these three cases considered: the first and the third case are easy to achieve sure screening with relative fewer number of selected variables and this reduces the biases of the RCV method and leaves more degrees of freedoms for estimating the residual variance. We intended to design such a case to challenge our proposed procedure, as sure screening is harder to achieve.

As a benchmark, we include the oracle estimator in our simulation. Here the oracle estimator corresponds to the mean squared errors for the fitting of the oracle model that includes

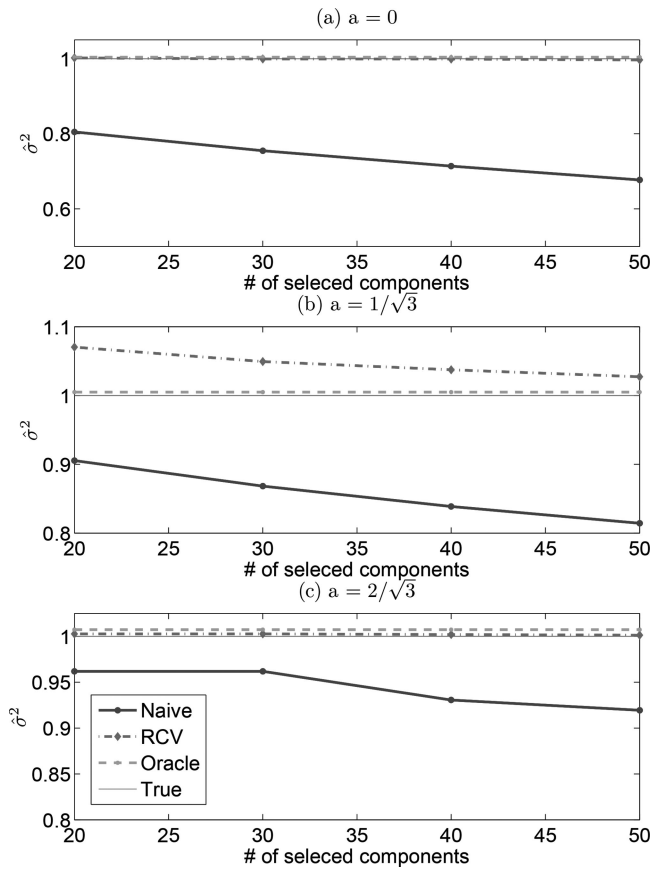


Figure 3. Variance estimators for different signal-to-noise ratios.

only  $X_1$ ,  $X_2$ , and  $X_5$  for  $a \neq 0$ , and include none of predictors when  $a = 0$ . In our simulation, we employ the distance correlation to rank importance of predictors, and screen out  $p - \hat{s}$  predictors with low distance correlation. Thus, the resulting model includes  $\hat{s}$  predictors. We consider  $\hat{s} = 20, 30, 40$ , and  $50$  to illustrate the impact of choices of  $\hat{s}$  on the performance of the naive estimator and the refitted cross-validation estimator.

In our simulation, each function  $f_j(\cdot)$  is approximated by a linear combination of an intercept and five cubic B-splines bases with three knots equally spaced between the minimum and maximum of the  $j$ th variable. Thus, when  $\hat{s} = 50$ , the reduced model actually has 251 terms, which is near half of the sample size. Table 1 depicts the average and the standard error of 150 estimates over the 150 simulations. To get an overall picture how the error variance estimates change over  $\hat{s}$ , Figure 3 depicts the overall average of the 150 estimates. In Table 1 and Figure 3, “Oracle” stands for the oracle estimate based on nonparametric additive models using only active variables, “Naive” for the naive estimate, and “RCV” for the refitted cross-validation estimate.

Table 1 and Figure 3 clearly show that the naive two-stage estimator significantly underestimates the error variance in the presence of many redundant variables. The larger the value  $\hat{s}$ , the bigger the spurious correlation  $\gamma_n^2$ , and hence the larger the bias of the naive estimate. The performance of the naive estimate also depends on the signal-to-noise ratio. In general, it performs better when the signal-to-noise ratio is large. The RCV estimator performs much better than the naive estimator. Its performance is very close to that of the oracle estimator for all cases listed in Table 1.

In practice, we have to choose one  $\hat{s}$  in data analysis. Fan and Lv (2008) suggested  $\hat{s} = \lceil n / \log(n) \rceil$  for their sure independence screening procedure based on Pearson correlation ranking. We modify their proposal and set  $\hat{s} = \lceil n^{4/5} / \log(n^{4/5}) \rceil$  to take into account effective sample size in nonparametric regression. Table 2 depicts the average and the standard error of 150 estimates over the 150 simulations when the sample size  $n = 400$  and  $600$ . The caption of Table 2 is the same as that in Table 1. Results in Table 2 clearly show that the RCV performs as well as the oracle procedure, and outperforms the naive estimate.

We further studied the impact of randomly splitting data strategy on the resulting estimate. As an alternative, one may repeat the proposed procedure several times, each randomly splitting data into two parts, and then take the average as the estimate of  $\sigma^2$ . Our findings from our simulations study are consistent with the discussion by Fan, Guo, and Hao (2012): (a) the estimates of  $\sigma^2$  for different numbers of repetitions are almost the same; and (b) as the number of repetitions increases, the variation slightly reduces at the price of computational cost. This implies that it is unnecessary to repeat the proposed procedure several times. As another alternative, one may randomly split the sample data into  $k$  groups. Specifically, the case  $k = 2$  is the proposed RCV methods in the article. Similarly, we can use data in one group to select useful predictors, data in other groups to fit the additive model. We refer this splitting strategy as multi-folder splitting. Our simulation results implies that the multi-folder splitting leads to (a) less accurate estimate for the coefficients and (b) increased variation of  $\hat{\sigma}_l^2$  used to construct the RCV estimate. This is because this strategy splits the data into many subsets with even smaller sample size. If the sample size  $n$  is large, as nowadays Big Data, it may be worth to try multiple random splits, otherwise we do not recommend it.

### 3.2. A Real Data Example

In this section, we illustrate the proposed procedure by an empirical analysis of a supermarket dataset (Wang 2009). The dataset contain a total of  $n = 464$  daily records of the number of customers ( $Y_i$ ) and the sale amounts of  $p = 6398$  products, denoted as  $X_{i1}, \dots, X_{ip}$ , which will be used as predictors. Both the response and predictors are standardized so that they have zero sample mean and unit sample variance. We fit the following additive model in our illustration.

$$Y_i = \mu + f_1(X_1) + \dots + f_p(X_p) + \varepsilon_i,$$

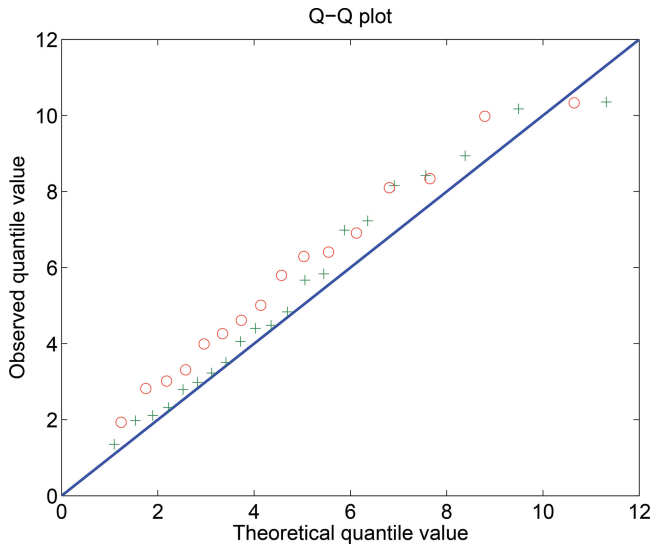
where  $\varepsilon_i$  is a random error with  $E(\varepsilon_i) = 0$  and  $\text{var}(\varepsilon_i | \mathbf{x}_i) = \sigma^2$ .

Since the sample size  $n = 464$ , we set  $\hat{s} = \lceil n^{4/5} / \log(n^{4/5}) \rceil = 28$ . The naive error variance estimate equals 0.0938, while the RCV error variance estimate equals 0.1340, a 43% increase of the estimated value when the spurious correlation is reduced. Table 3 depicts the resulting estimates of the error variance with different values of  $\hat{s}$ , and clearly shows that the RCV estimate

Table 3. Error variance estimate for market data.

$\hat{s}$	40	35	30	28	25
Naive	0.0866	0.0872	0.0910	0.0938	0.0990
RCV	0.1245	0.1104	0.1277	0.1340	0.1271





**Figure 4.** Quantile–quantile plot of  $\chi^2$ -test values. “o” stands for  $\chi^2$ -test using naive error variance estimate. “+” stands for  $\chi^2$ -test using RCV error variance estimate.

of error variance is stable with different choices of  $\hat{s}$ , while the estimate of error variance by the naive method reduces as  $\hat{s}$  increases. This is consistent with our theoretical and simulation results.

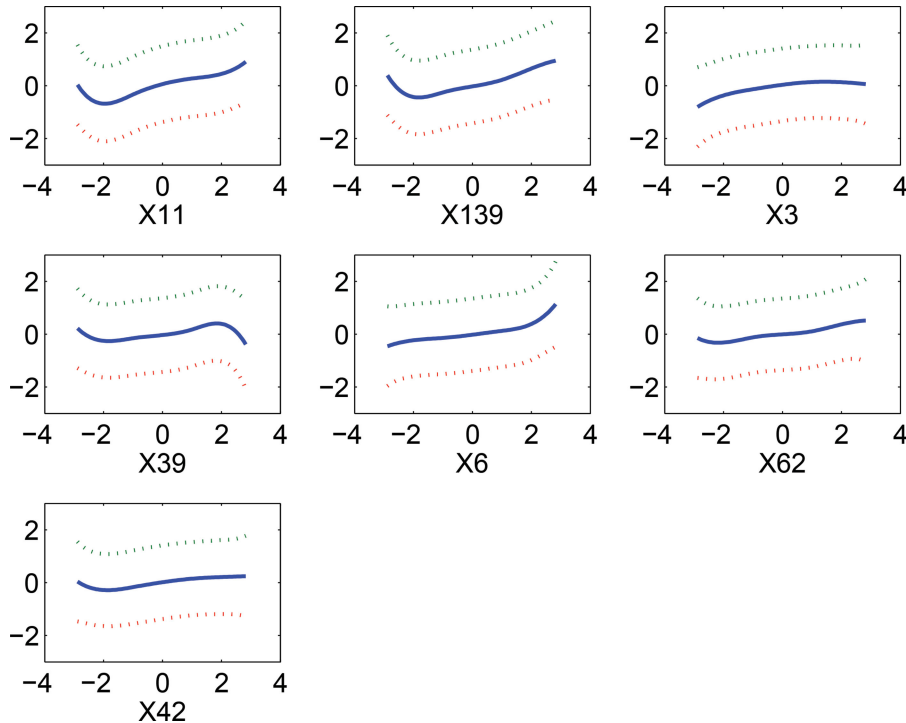
Regarding the selected models with  $\hat{s}$  predictors as a correct model and ignoring the approximation errors (if any) due to B-spline, we further employ the Wald’s  $\chi^2$ -test for hypothesis whether  $(\gamma_{j1}, \dots, \gamma_{jd_j})^T$  equals zero, namely, whether the  $j$ th variable is active in presence of the rest variables. Such Wald’s  $\chi^2$  statistics offer us a rough picture whether  $X_j$  is significant or not. The Wald’s  $\chi^2$ -test with the naive error variance estimate concludes 12 significant predictors at significant level 0.05, while the Wald’s  $\chi^2$ -test with the RCV error variance estimate

concludes seven significant predictors at the same significant level. Figure 4 depicts the Q-Q plot of values of the  $\chi^2$ -test statistic of those insignificant predictors identified by the Wald’s test. Figure 4 clearly shows that the  $\chi^2$ -test values using naive error variance estimate systematically deviate from the 45-degree line. This implies that the naive method results in an underestimate of error variance, while the RCV method results in a good estimate of error variance.

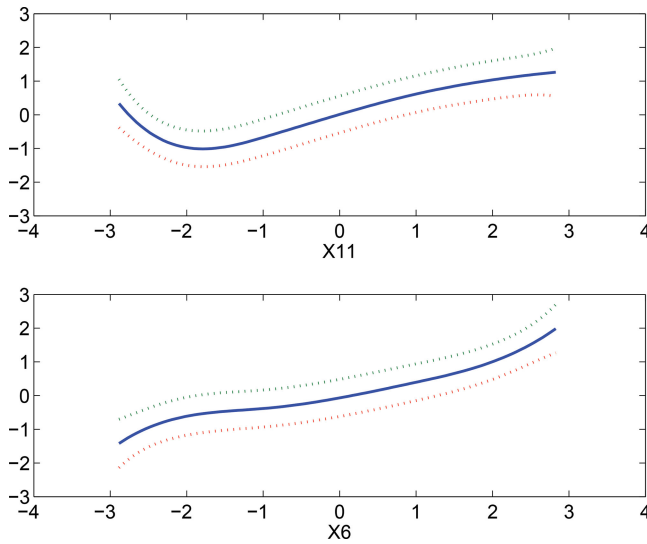
The Wald’s test at level 0.05 is in favor that seven predictors,  $X_{11}$ ,  $X_{139}$ ,  $X_3$ ,  $X_{39}$ ,  $X_6$ ,  $X_{62}$ , and  $X_{42}$ , are significant. We refit the data with the additive model with these seven predictors. The corresponding mean squared errors is 0.1207, which is close to the  $\hat{\sigma}_{\text{RCV}}^2 = 0.1340$ . Note that  $\sigma^2$  is the minimum possible prediction error. It provides a benchmark for other methods to compare with and is achievable when modeling bias and estimation errors are negligible.

To see how the above selected variables perform in terms of prediction, we further use the leave-one-out cross-validation (CV) and five-fold CV to estimate the mean squared prediction errors (MSPE). The leave-one-out CV yields  $\text{MSPE} = 0.1414$ , and the average of the MSPE obtained from five-fold CV based on 400 randomly splitting data yields is 0.1488 with the 2.5th percentile and 97.5 percentile being 0.1411 and 0.1626, respectively. The MSPE is slightly greater than  $\hat{\sigma}_{\text{RCV}}^2$ . This is expected as the uncertainty of parameter estimation has not been accounted. This bias can be corrected from the theory of linear regression analysis.

Suppose that  $\{\mathbf{x}_i, Y_i\}$ ,  $i = 1, \dots, n$  is an independent and identically distributed random sample from a linear regression model  $Y = \mathbf{x}^T \boldsymbol{\beta} + \varepsilon$ , the linear predictor  $\hat{Y} = \mathbf{x}^T \hat{\boldsymbol{\beta}}$ , where  $\hat{\boldsymbol{\beta}}$  is the least-square estimate of  $\boldsymbol{\beta}$ , has prediction error at a new observation  $\{\mathbf{x}_*, y_*\}$ :  $E\{(y_* - \mathbf{x}_*^T \hat{\boldsymbol{\beta}})^2 | \mathbf{X}\} = \sigma^2(1 + \mathbf{x}_*^T (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{x}_*)$ , where  $\sigma^2$  is the error variance and  $\mathbf{X}$  is the



**Figure 5.** Estimated functions based on 7 variables selected from 28 variables that survive DC-SIS screening by the  $\chi^2$ -test with the RCV error variance estimator.



**Figure 6.** Estimated functions based on 2 variables selected from 28 variables that survive DC-SIS screening by the  $\chi^2$ -test with the RCV error variance estimator and the Bonferroni adjustment.

corresponding design matrix. This explains why the MSPE is slightly greater than  $\hat{\sigma}_{RCV}^2$ . To further gauge the accuracy of the RCV estimate of  $\sigma^2$ , define weighted prediction error  $|y_* - \mathbf{x}_*^T \hat{\boldsymbol{\beta}}| / \sqrt{1 + \mathbf{x}_*^T (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{x}_*}$ . Then the leave-one-out method leads to the mean squared weighted predictor error (MSWPE) 0.1289 and the average of five-fold CV based on 400 randomly splitting data yields MSWPE 0.1305 with the 2.5th percentile and 97.5 percentile being 0.1254 and 0.1366, respectively. These results imply (a) the seven selected variables achieve the benchmark prediction; (b) the modeling biases using the additive models of these seven variables are negligible; (c)  $\hat{\sigma}_{RCV}^2$  provides a very good estimate for  $\sigma^2$ .

Their estimated functions  $\hat{f}_j(x_j)$  are depicted in Figure 5, from which it seems that all predictors shown in Figure 5 are not significant since zero crosses the entire confidence interval. This can be because we have used too many variables, which increases the variance of the estimate.

We further employ the Wald's test with Bonferroni correction for 28 null hypotheses. This leads only two significant predictors,  $X_{11}$  and  $X_6$ , at level 0.05. We refit the data with the two selected predictors. Figure 6 depicts the plot of  $\hat{f}_{11}(x_{11})$  and  $\hat{f}_6(x_6)$ .

## 4. Discussions

In this article, we proposed an error variance estimator in ultrahigh-dimensional additive model by using refitted cross-validation technique. This is particularly important given the high level of spurious correlation induced by the nonparametric models (see Figure 1 and Lemma 1). We established the root  $n$  consistency and asymptotic normality of the resulting estimator, and examined the empirical performance of the proposed estimator by Monte Carlo simulation. We further demonstrated the proposed methodology via an empirical analysis of supermarket data. The proposed estimator performs well with moderate sample size. However, when the sample size is very small, the refitted cross-validation procedure may be unstable. How to construct an accurate error variance estimate with very small sample

size is challenging and will be an interesting topic for future research.

## Appendix: Proofs

### A.1 Proofs of Lemma 1 and Theorem 1

Let  $\Psi$  be the corresponding design matrix of model (2.3). Specifically,  $\Psi$  is a  $n \times (pd_n)$  matrix with  $i$ th row being  $(B_{11}(X_{i1}), \dots, B_{1d_n}(X_{i1}), B_{21}(X_{i2}), \dots, B_{pd_n}(X_{ip}))$ . Denote by  $\Psi^{(\hat{D}^*)}$  the corresponding design matrix of model  $\hat{D}^*$ , and  $\mathbf{P}_{\hat{D}^*}$  the corresponding projection matrix. That is,  $\mathbf{P}_{\hat{D}^*} = \Psi^{(\hat{D}^*)} (\Psi^{(\hat{D}^*)T} \Psi^{(\hat{D}^*)})^{-1} \Psi^{(\hat{D}^*)T}$ . Denote  $\mathbf{P}_{\hat{D}^*}^c = \mathbf{I}_n - \mathbf{P}_{\hat{D}^*}$ . Without loss of generality, assume that the first  $s$  non-parametric components are nonzero and others are all zero. By the assumption that all active predictors are retained by DC-SIS screening procedure. For ease of notation and without loss of generality, assume that  $\hat{D}^* = \{1, 2, \dots, \hat{s}\}$ , where  $\hat{s} = |\hat{D}^*|$ .

*Proof of Lemma 1.* Note that

$$\begin{aligned} \boldsymbol{\varepsilon}^T \mathbf{P}_{\hat{D}^*} \boldsymbol{\varepsilon} &= \boldsymbol{\varepsilon}^T \Psi^{(\hat{D}^*)} \left( \Psi^{(\hat{D}^*)T} \Psi^{(\hat{D}^*)} \right)^{-1} \Psi^{(\hat{D}^*)T} \boldsymbol{\varepsilon} \\ &\leq \lambda_{\min}^{-1} \left( \Psi^{(\hat{D}^*)T} \Psi^{(\hat{D}^*)} \right) \left\| \Psi^{(\hat{D}^*)T} \boldsymbol{\varepsilon} \right\|_2^2, \end{aligned} \quad (\text{A.1})$$

where  $\lambda_{\min}(\mathbf{A})$  stands for the minimal eigenvalue of matrix  $\mathbf{A}$ . To show Lemma 1, we need to derive the bound of eigenvalue of matrix  $\Psi^{(\hat{D}^*)T} \Psi^{(\hat{D}^*)}$ . Note that  $\Psi^{(\hat{D}^*)} = (\Psi_1, \dots, \Psi_{\hat{s}})$  with

$$\Psi_j = \begin{pmatrix} B_{j1}(X_{j1}) & \cdots & B_{jd_n}(X_{j1}) \\ \vdots & \ddots & \vdots \\ B_{j1}(X_{jn}) & \cdots & B_{jd_n}(X_{jn}) \end{pmatrix}, \quad j = 1, \dots, \hat{s}. \quad (\text{A.2})$$

Let  $\mathbf{b} = (\mathbf{b}_1^T, \dots, \mathbf{b}_{\hat{s}}^T)^T$  and  $\|\mathbf{b}\|_2^2 = \mathbf{b}^T \mathbf{b} = 1$ . Then we have  $\Psi^{(\hat{D}^*)} \mathbf{b} = \Psi_1 \mathbf{b}_1 + \dots + \Psi_{\hat{s}} \mathbf{b}_{\hat{s}}$ . As shown in Lemma S.5 in the supplemental material of this paper, it follows that

$$\begin{aligned} &\left( \frac{1-\delta}{2} \right)^{\hat{s}-1} (\|\Psi_1 \mathbf{b}_1\|_2 + \dots + \|\Psi_{\hat{s}} \mathbf{b}_{\hat{s}}\|_2)^2 \\ &\leq \|\Psi_1 \mathbf{b}_1 + \dots + \Psi_{\hat{s}} \mathbf{b}_{\hat{s}}\|_2^2 = \mathbf{b}^T \Psi^{(\hat{D}^*)T} \Psi^{(\hat{D}^*)} \mathbf{b}. \end{aligned} \quad (\text{A.3})$$

This yields that

$$\left( \frac{1-\delta}{2} \right)^{\hat{s}-1} \left( \sum_{i=1}^{\hat{s}} \mathbf{b}_i \Psi_i^T \Psi_i \mathbf{b}_i \right) \leq \mathbf{b}^T \Psi^{(\hat{D}^*)T} \Psi^{(\hat{D}^*)} \mathbf{b}, \quad (\text{A.4})$$

since  $\|\Psi_i \mathbf{b}_i\|_2 \geq 0$ . Furthermore,

$$\begin{aligned} &\left( \frac{1-\delta}{2} \right)^{\hat{s}-1} \left( \sum_{i=1}^{\hat{s}} \mathbf{b}_i \Psi_i^T \Psi_i \mathbf{b}_i \right) \\ &\geq \left( \frac{1-\delta}{2} \right)^{\hat{s}-1} \left( \sum_{\mathbf{b}_i^T \mathbf{b}_i \neq 0} \lambda_{\min}(\Psi_i^T \Psi_i) \mathbf{b}_i^T \mathbf{b}_i \right). \end{aligned}$$

Recalling Lemma 6.2 of Zhou, Shen and Wolfe (1998), there exists two positive constants  $C_1$  and  $C_2$  such that, for any  $1 \leq i \leq \hat{s}$ ,

$$C_1 d_n^{-1} n \leq \lambda_{\min}(\Psi_i^T \Psi_i) \leq \lambda_{\max}(\Psi_i^T \Psi_i) \leq C_2 d_n^{-1} n. \quad (\text{A.5})$$

Thus,

$$\begin{aligned} & \left( \frac{1-\delta}{2} \right)^{\hat{s}-1} \left( \sum_{\mathbf{b}_i^T \mathbf{b}_i \neq 0} \lambda_{\min}(\Psi_i^T \Psi_i) \mathbf{b}_i^T \mathbf{b}_i \right) \\ & \geq C_1 \left( \frac{1-\delta}{2} \right)^{\hat{s}-1} d_n^{-1} n \sum_{\mathbf{b}_i^T \mathbf{b}_i \neq 0} \mathbf{b}_i^T \mathbf{b}_i = C_1 \left( \frac{1-\delta}{2} \right)^{\hat{s}-1} d_n^{-1} n. \end{aligned} \quad (\text{A.6})$$

The last equation is valid due to  $\|\mathbf{b}\|_2^2 = \mathbf{b}^T \mathbf{b} = 1$ . Combining the equation (A.4) and (A.6), we have

$$\lambda_{\min}(\Psi^{(\hat{D}^*)^T} \Psi^{(\hat{D}^*)}) \geq C_1 \left( \frac{1-\delta}{2} \right)^{\hat{s}-1} d_n^{-1} n. \quad (\text{A.7})$$

Thus, it follows by using (A.1) that

$$\boldsymbol{\varepsilon}^T \mathbf{P}_{\hat{D}^*} \boldsymbol{\varepsilon} \leq C_1^{-1} \left( \frac{2}{1-\delta} \right)^{\hat{s}-1} d_n n^{-1} \left\| \Psi^{(\hat{D}^*)^T} \boldsymbol{\varepsilon} \right\|_2^2. \quad (\text{A.8})$$

By the notation (A.2), we have

$$\Psi_i^T \boldsymbol{\varepsilon} = \begin{pmatrix} \sum_{k=1}^n B_{i1}(X_{ik}) \varepsilon_k \\ \sum_{k=1}^n B_{i2}(X_{ik}) \varepsilon_k \\ \vdots \\ \sum_{k=1}^n B_{id_n}(X_{ik}) \varepsilon_k \end{pmatrix}. \quad (\text{A.9})$$

Recalling that  $0 \leq B_{ij}(\cdot) \leq 1$ , for any  $i, j$  and  $\mathbb{E} |B_{ij}(X_{ik})|^2 \leq C_4 d_n^{-1}$  (Stone, 1985), we note the fact that for  $m \geq 2$ ,  $\mathbb{E} |B_{ij}(X_{ik})|^m \leq \mathbb{E} |B_{ij}(X_{ik})|^2 \leq C_4 d_n^{-1}$ . Observe that, using Condition (C1), for any integers  $i$  and  $j$

$$\begin{aligned} \mathbb{E} |B_{ij}(X_{ik}) \varepsilon_k|^m &= \mathbb{E} |B_{ij}(X_{ik})|^m \cdot \mathbb{E} |\varepsilon_k|^m \\ &\leq \mathbb{E} |B_{ij}(X_{ik})|^m \mathbb{E} (m! a^m \exp\{|\varepsilon_1|/a\}). \end{aligned} \quad (\text{A.10})$$

Taking  $A_1 = 1/a$  and  $A_2 = b$  in Condition (C1), it follows that the right hand side of above inequality will not exceed

$$C_4 m! a^m d_n^{-1} \mathbb{E} (\exp\{|\varepsilon_1|/a\}) \leq \frac{C_4}{2} m! (2 d_n^{-1} b a^2) a^{m-2}. \quad (\text{A.11})$$

Using Bernstein's Inequality (see Lemma 2.2.11 of Van der Vaart and Wellner 1996), we have

$$\begin{aligned} & \mathbb{P} \left( \max_{\substack{1 \leq i \leq p \\ 1 \leq j \leq d_n}} \left| \sum_{k=1}^n B_{ij}(X_{ik}) \varepsilon_k \right| \geq M \right) \\ & \leq \sum_{i=1}^p \sum_{j=1}^{d_n} \mathbb{P} \left( \left| \sum_{k=1}^n B_{ij}(X_{ik}) \varepsilon_k \right| \geq M \right) \\ & \leq 2 p d_n \exp \left\{ -\frac{M^2}{2(2 d_n^{-1} b a^2 n + a M)} \right\} \\ & = 2 \exp \left\{ \log(p d_n) \left( 1 - \frac{1}{4 \log(p d_n) n d_n^{-1} b a^2 M^{-2} + 2 \log(p d_n) a M^{-1}} \right) \right\}. \end{aligned}$$

When we take  $M = C_5 \sqrt{n \log(p d_n)/d_n}$ , with  $\frac{d_n \log(p d_n)}{n} \rightarrow 0$  and sufficiently large  $C_5$ , the power in the last equation goes to negative infinity. Thus, with probability approaching to one, we have

$$\max_{\substack{1 \leq i \leq p \\ 1 \leq j \leq d_n}} \left| \sum_{k=1}^n B_{ij}(X_{ik}) \varepsilon_k \right| \leq C_5 \sqrt{n \log(p d_n)/d_n} \text{ and}$$

$$\begin{aligned} \boldsymbol{\varepsilon}^T \mathbf{P}_{\hat{D}^*} \boldsymbol{\varepsilon} &\leq C_1^{-1} \left( \frac{2}{1-\delta} \right)^{\hat{s}-1} d_n n^{-1} \left\| \Psi^{(\hat{D}^*)^T} \boldsymbol{\varepsilon} \right\|_2^2 \\ &\leq C_5^2 C_1^{-1} \left( \frac{2}{1-\delta} \right)^{\hat{s}-1} d_n \log(p d_n). \end{aligned} \quad (\text{A.12})$$

Due to the independent and identically distributed random errors with mean 0 and variance  $\sigma^2$ , by the law of large numbers, we have

$$\frac{1}{n} \sum_{i=1}^n \varepsilon_i \xrightarrow{\text{a.s.}} 0, \quad \frac{1}{n} \sum_{i=1}^n \varepsilon_i^2 \xrightarrow{\text{a.s.}} \sigma^2. \quad (\text{A.13})$$

Thus, we obtain that

$$\hat{\gamma}_n^2 = \frac{\boldsymbol{\varepsilon}^T \mathbf{P}_{\hat{D}^*} \boldsymbol{\varepsilon}}{\boldsymbol{\varepsilon}^T \boldsymbol{\varepsilon}} = O_p \left\{ \left( \frac{2}{1-\delta} \right)^{\hat{s}} \frac{d_n \log(p d_n)}{n} \right\}. \quad (\text{A.14})$$

□

*Proof of Theorem 1.* Note that

$$\hat{\sigma}_{\hat{D}^*}^2 = \frac{1}{n - \hat{s} d_n} \left[ \sum_{j=1}^{\hat{s}} \mathbf{f}_j^T(\mathbf{X}_j) \mathbf{P}_{\hat{D}^*}^c \sum_{j=1}^{\hat{s}} \mathbf{f}_j(\mathbf{X}_j) + 2 \boldsymbol{\varepsilon}^T \mathbf{P}_{\hat{D}^*}^c \sum_{j=1}^{\hat{s}} \mathbf{f}_j(\mathbf{X}_j) + \boldsymbol{\varepsilon}^T \mathbf{P}_{\hat{D}^*}^c \boldsymbol{\varepsilon} \right],$$

where  $\mathbf{f}_j(\mathbf{X}_j) = (f_j(X_{j1}), \dots, f_j(X_{jn}))^T$ ,  $j = 1, \dots, p$ . To simplify the first term in  $\hat{\sigma}_{\hat{D}^*}^2$ , let  $\Delta_1 = \sum_{j=1}^{\hat{s}} \mathbf{f}_j^T(\mathbf{X}_j) \mathbf{P}_{\hat{D}^*}^c \sum_{j=1}^{\hat{s}} \mathbf{f}_j(\mathbf{X}_j)$ . Then

$$\begin{aligned} \Delta_1 &= \left\{ \sum_{j=1}^{\hat{s}} \mathbf{f}_j(\mathbf{X}_j) - \sum_{j=1}^{\hat{s}} \mathbf{f}_{nj}(\mathbf{X}_j) + \sum_{j=1}^{\hat{s}} \mathbf{f}_{nj}(\mathbf{X}_j) \right\}^T \mathbf{P}_{\hat{D}^*}^c \\ &\quad \times \left\{ \sum_{j=1}^{\hat{s}} \mathbf{f}_j(\mathbf{X}_j) - \sum_{j=1}^{\hat{s}} \mathbf{f}_{nj}(\mathbf{X}_j) + \sum_{j=1}^{\hat{s}} \mathbf{f}_{nj}(\mathbf{X}_j) \right\}, \end{aligned}$$

where  $\mathbf{f}_{nj}(\mathbf{X}_j) = (f_{nj}(X_{j1}), \dots, f_{nj}(X_{jn}))^T = (\mathbf{B}_j(X_{j1})^T \boldsymbol{\Gamma}_j, \dots, \mathbf{B}_j(X_{jn})^T \boldsymbol{\Gamma}_j)^T$ ,  $j = 1, \dots, p$ . Define

$$\begin{aligned} \Delta_{11} &= \left\{ \sum_{j=1}^{\hat{s}} \mathbf{f}_j(\mathbf{X}_j) - \sum_{j=1}^{\hat{s}} \mathbf{f}_{nj}(\mathbf{X}_j) \right\}^T \mathbf{P}_{\hat{D}^*}^c \left\{ \sum_{j=1}^{\hat{s}} \mathbf{f}_j(\mathbf{X}_j) - \sum_{j=1}^{\hat{s}} \mathbf{f}_{nj}(\mathbf{X}_j) \right\}, \\ \Delta_{12} &= \left\{ \sum_{j=1}^{\hat{s}} \mathbf{f}_{nj}(\mathbf{X}_j) \right\}^T \mathbf{P}_{\hat{D}^*}^c \left\{ \sum_{j=1}^{\hat{s}} \mathbf{f}_{nj}(\mathbf{X}_j) \right\}, \\ \Delta_{13} &= 2 \left\{ \sum_{j=1}^{\hat{s}} \mathbf{f}_j(\mathbf{X}_j) - \sum_{j=1}^{\hat{s}} \mathbf{f}_{nj}(\mathbf{X}_j) \right\}^T \mathbf{P}_{\hat{D}^*}^c \left\{ \sum_{j=1}^{\hat{s}} \mathbf{f}_{nj}(\mathbf{X}_j) \right\}. \end{aligned}$$

Then  $\Delta_1 = \Delta_{11} + \Delta_{12} + \Delta_{13}$ . Note that  $\mathbf{P}_{\hat{D}^*}^c$  is a projection matrix on the complement of the linear space of  $\Psi^{(\hat{D}^*)}$ , and therefore  $\mathbf{P}_{\hat{D}^*}^c \{\sum_{j=1}^{\hat{s}} \mathbf{f}_{nj}(\mathbf{X}_j)\} = 0$ . Thus, both  $\Delta_{12}$  and  $\Delta_{13}$  equal 0. We next calculate the order of  $\Delta_{11}$ . By the property of B-spline (Stone, 1985), there exists a constant  $c_1 > 0$  such that  $\|f_j - f_{nj}\|^2 \leq c_1 d_n^{-2d}$ . Since  $\mathbf{P}_{\hat{D}^*}^c$  is a projection matrix, its eigenvalues equal either 0 or 1. By the Cauchy-Schwarz inequality and some straightforward calculation, it follows that  $\Delta_{11} = O_p(\hat{s}^2 n d_n^{-2d})$ . Therefore  $\Delta_1 = O_p(\hat{s}^2 n d_n^{-2d})$ . Under conditions in Theorem 1(i),  $O_p(\hat{s}^2 n d_n^{-2d}) = o_p(1)$ . As a result,  $\Delta_1 = o_p(n)$ . Under conditions in Theorem 1(ii),  $\hat{s} = o(n^{(2d-1)/(4(2d+1))})$  and therefore  $\Delta_1 = o_p(\sqrt{n})$ .

Now we deal with the second term in  $\hat{\sigma}_{\hat{\mathcal{D}}^*}^2$ . Denote  $\Delta_2 = 2\mathbf{e}^T \mathbf{P}_{\hat{\mathcal{D}}^*}^c \sum_{j=1}^{\hat{s}} \mathbf{f}_j(\mathbf{X}_j)$ . Since  $\mathbf{P}_{\hat{\mathcal{D}}^*}^c \{\sum_{j=1}^{\hat{s}} \mathbf{f}_{nj}(\mathbf{X}_j)\} = 0$ , it follows that

$$\Delta_2 = 2\mathbf{e}^T \mathbf{P}_{\hat{\mathcal{D}}^*}^c \left( \sum_{j=1}^{\hat{s}} \mathbf{f}_j(\mathbf{X}_j) - \sum_{j=1}^{\hat{s}} \mathbf{f}_{nj}(\mathbf{X}_j) \right).$$

Denote  $\Delta_{21} = \sum_{j=1}^{\hat{s}} \sum_{i=1}^n (f_j(X_{ji}) - f_{nj}(X_{ji})) \varepsilon_i$  and  $\Delta_{22} = (\mathbf{e}^T \mathbf{P}_{\hat{\mathcal{D}}^*}) (\sum_{j=1}^{\hat{s}} \mathbf{f}_j(\mathbf{X}_j) - \sum_{j=1}^{\hat{s}} \mathbf{f}_{nj}(\mathbf{X}_j))$ . Thus,  $\Delta_2 = 2(\Delta_{21} - \Delta_{22})$ . To deal with  $\Delta_{21}$ , we bound the tails of  $(f_j(X_{ji}) - f_{nj}(X_{ji})) \varepsilon_i$ ,  $i = 1, \dots, n$ ,  $j = 1, \dots, \hat{s}$ . For any  $m \geq 2$ , because  $f_j \in \mathcal{C}^d([a, b])$  and  $f_{nj}$  belongs to the spline space  $\mathcal{S}^l([a, b])$ , we have

$$\mathbb{E} |(f_j(X_{ji}) - f_{nj}(X_{ji})) \varepsilon_i|^m = \mathbb{E} \left( |f_j(X_{ji}) - f_{nj}(X_{ji})|^m \mathbb{E} (|\varepsilon_i|^m | \mathbf{x}_i) \right),$$

which is bounded by  $C_6^{m-2} \mathbb{E} (|f_j(X_{ji}) - f_{nj}(X_{ji})|^2 \mathbb{E} (|\varepsilon_i|^m | \mathbf{x}_i))$  for some constant  $C_6$  by the property of B-spline approximation. There exists a constant  $c_1 > 0$  such that  $\|f_j - f_{nj}\|^2 \leq c_1 d_n^{-2d}$  by the property of B-spline (Stone, 1985). Applying Condition (C1) for  $\mathbb{E} \{\exp(A_1 |\varepsilon_i|) | \mathbf{x}_i\}$ , it follows that

$$\mathbb{E} |(f_j(X_{ji}) - f_{nj}(X_{ji})) \varepsilon_i|^m \leq m! \left( \frac{C_6}{A_1} \right)^{m-2} \frac{A_2}{A_1^2} c_1 d_n^{-2d}.$$

Denote  $C_7 = c_1 A_2 / A_1^2$ , and  $C_8 = C_6 / A_1$ . Using the Bernstein's inequality, for some  $M$ , we have

$$\begin{aligned} & \mathbb{P} \left( \max_{1 \leq j \leq \hat{s}} \left| \sum_{i=1}^n (f_j(X_{ji}) - f_{nj}(X_{ji})) \varepsilon_i \right| > M \right) \\ & \leq 2p \exp \left\{ -\frac{1}{2} \frac{M^2}{2 C_7 n d_n^{-2d} - C_8 M} \right\}. \end{aligned} \quad (\text{A.15})$$

If we take  $M = C_9 \sqrt{\log(p) n d_n^{-2d}}$ , and for sufficiently large  $C_9$ , then the tail probability (A.15) goes to zero. Thus,

$$\Delta_{21} = O_p \left( \hat{s} \sqrt{\log(p) n d_n^{-2d}} \right). \quad (\text{A.16})$$

Under condition of Theorem 1(i),  $\hat{s} = o(n^{(4d+1)/(2(2d+1))})$  with  $\zeta < 1$ . Thus,  $O_p(\hat{s} d_n^{-d} \sqrt{\log(p d_n)}) = o_p(\sqrt{n})$ . Following the similar arguments dealing with  $\Delta_{11}$ , it follows that  $\Delta_{21} = o_p(n)$ . Under condition of Theorem 1(ii),  $\hat{s} = o(n^{d/(2d+1)-\zeta/2})$  with  $\zeta < 3/(2d+1)$ . Thus,  $\Delta_{21} = o_p(\sqrt{n})$ . By the Cauchy-Schwarz inequality, it follows by Lemma 1 that

$$\begin{aligned} \Delta_{22} & \leq \|\mathbf{e}^T \mathbf{P}_{\hat{\mathcal{D}}^*}^c\|_2 \cdot \left\| \sum_{j=1}^{\hat{s}} \mathbf{f}_j(\mathbf{X}_j) - \sum_{j=1}^{\hat{s}} \mathbf{f}_{nj}(\mathbf{X}_j) \right\|_2 \\ & = O_p \left( \left( \frac{2}{1-\delta} \right)^{\hat{s}} \sqrt{d_n \log(p d_n)} \right) \cdot O_p(\hat{s} n^{1/2} d_n^{-d}) \\ & = O_p \left( \left( \frac{2}{1-\delta} \right)^{\hat{s}} \sqrt{\log(p d_n) d_n^{-d+1/2}} \right). \end{aligned}$$

When  $\zeta < 4d/(2d+1)$ , and  $\hat{s} = O_p(\log(n^\alpha))$ ,  $\alpha \leq 4d/(2d+1) - \zeta$ , it follows that  $\Delta_{22} = o_p(n)$  under condition of Theorem 1(i). When  $\zeta < (2d-1)/(2(2d+1))$  and  $\hat{s} = \log(n^\alpha)$ ,  $\alpha \leq (2d-1)/(2(2d+1)) - \zeta$ ,  $(2/(1-\delta))^{\hat{s}} n^{1/2} \sqrt{\log(p d_n) d_n^{-d+1/2}} = o_p(\sqrt{n})$ . Thus,  $\Delta_{22} = o_p(\sqrt{n})$  under condition of Theorem 1(ii). Comparing the order of  $\Delta_{11}$ ,  $\Delta_{21}$  and  $\Delta_{22}$ , we

obtain the order of  $\hat{s}$  in Theorem 1. Therefore, we have

$$\begin{aligned} & \mathbf{Y}^T \left( I_n - \Psi^{(\hat{\mathcal{D}}^*)} \left( \Psi^{(\hat{\mathcal{D}}^*)T} \Psi^{(\hat{\mathcal{D}}^*)} \right)^{-1} \Psi^{(\hat{\mathcal{D}}^*)T} \right) \mathbf{Y} \\ & = \mathbf{e}^T (I_n - \mathbf{P}_{\hat{\mathcal{D}}^*}) \mathbf{e} + \Delta_1 + \Delta_2 \\ & = \mathbf{e}^T (I_n - \mathbf{P}_{\hat{\mathcal{D}}^*}) \mathbf{e} + O_p(\hat{s}^2 n d_n^{-2d}) + O_p \left( \hat{s} \sqrt{\log(p) n d_n^{-2d}} \right) + \Delta_{22}. \end{aligned}$$

and it follows by the definition of  $\hat{\gamma}_n^2$  that

$$\begin{aligned} \hat{\sigma}_{\hat{\mathcal{D}}^*}^2 & = \frac{1}{n - \hat{s} d_n} \mathbf{Y}^T \left( I_n - \Psi^{(\hat{\mathcal{D}}^*)} \left( \Psi^{(\hat{\mathcal{D}}^*)T} \Psi^{(\hat{\mathcal{D}}^*)} \right)^{-1} \Psi^{(\hat{\mathcal{D}}^*)T} \right) \mathbf{Y} \\ & = \frac{1}{n - \hat{s} d_n} \mathbf{e}^T \mathbf{e} (1 - \hat{\gamma}_n^2) + O_p \left( \frac{\hat{s}^2 n d_n^{-2d}}{n - \hat{s} d_n} \right) \\ & \quad + O_p \left( \frac{\sqrt{\log(p) \hat{s}^2 n d_n^{-2d}}}{n - \hat{s} d_n} \right) + \frac{\Delta_{22}}{n - \hat{s} d_n}. \end{aligned}$$

Since  $\hat{s} d_n = o_p(n)$  and  $\limsup \hat{\gamma}_n^2 < 1$ , we have

$$\begin{aligned} \frac{\hat{\sigma}_{\hat{\mathcal{D}}^*}^2}{(1 - \hat{\gamma}_n^2)} & = \frac{1}{n - \hat{s} d_n} \mathbf{e}^T \mathbf{e} + O_p(\hat{s}^2 d_n^{-2d}) \\ & \quad + O_p(\sqrt{\log(p) \hat{s} n^{-1/2} d_n^{-d}}) + O_p \left( \frac{\Delta_{22}}{n} \right). \end{aligned} \quad (\text{A.17})$$

Under conditions of Theorem 1(i), the small order term in (A.17) is bounded by  $o_p(1)$ . We have

$$\frac{\hat{\sigma}_{\hat{\mathcal{D}}^*}^2}{1 - \hat{\gamma}_n^2} \rightarrow \mathbb{P} \sigma^2. \quad (\text{A.18})$$

To establish the asymptotic normality, we should study the asymptotic bias of the estimator. By the Central Limit Theorem, it follows that

$$\frac{1}{\sqrt{n}} \sum_{i=1}^n (\varepsilon_i^2 - \sigma^2) \xrightarrow{\mathcal{L}} \mathcal{N}(0, \mathbb{E} \varepsilon_1^4 - \sigma^4). \quad (\text{A.19})$$

Note that under conditions of Theorem 1(ii), the small order term in (A.17) is bounded by  $o_p(n^{-1/2})$ . Therefore, the asymptotic normality holds.  $\square$

## A.2 Proof of Theorem 2

Define events  $\mathcal{A}_{n1} = \{\mathcal{D}^* \subset \hat{\mathcal{D}}_1^*\}$ ,  $\mathcal{A}_{n2} = \{\mathcal{D}^* \subset \hat{\mathcal{D}}_2^*\}$  and  $\mathcal{A}_n = \mathcal{A}_{n1} \cap \mathcal{A}_{n2}$ . Unless specifically mentioned, our analysis and calculation are based on the event  $\mathcal{A}_n$ .

Let  $\Psi^{(\hat{\mathcal{D}}_1^*)}$  be the design matrix corresponding to  $\hat{\mathcal{D}}_1^*$ ,  $\mathbf{P}_{\hat{\mathcal{D}}_1^*} = \Psi^{(\hat{\mathcal{D}}_1^*)} (\Psi^{(\hat{\mathcal{D}}_1^*)T} \Psi^{(\hat{\mathcal{D}}_1^*)})^{-1} \Psi^{(\hat{\mathcal{D}}_1^*)T}$ , and  $\mathbf{P}_{\hat{\mathcal{D}}_1^*}^c = I - \mathbf{P}_{\hat{\mathcal{D}}_1^*}$ . Note that  $\mathbf{P}_{\hat{\mathcal{D}}_1^*}^c (\sum_{j=1}^{\hat{s}_1} \mathbf{f}_{nj}(\mathbf{X}_j^{(2)})) = 0$ . Thus,

$$\begin{aligned} (n/2 - \hat{s}_1 d_n) \hat{\sigma}_{\hat{\mathcal{D}}_1^*}^2 & = \mathbf{e}^{(2)T} \mathbf{P}_{\hat{\mathcal{D}}_1^*}^c \mathbf{e}^{(2)} \\ & \quad + \left( \sum_{j=1}^{\hat{s}_1} \mathbf{f}_j(\mathbf{X}_j^{(2)}) - \sum_{j=1}^{\hat{s}_1} \mathbf{f}_{nj}(\mathbf{X}_j^{(2)}) \right)^T \mathbf{P}_{\hat{\mathcal{D}}_1^*}^c \left( \sum_{j=1}^{\hat{s}_1} \mathbf{f}_j(\mathbf{X}_j^{(2)}) - \sum_{j=1}^{\hat{s}_1} \mathbf{f}_{nj}(\mathbf{X}_j^{(2)}) \right). \end{aligned}$$

By the same argument as that in the proof of Theorem 1, the second term in the above equation is of the order  $O_p(\hat{s}_1^2 n d_n^{-2d})$ . Thus,

$$\begin{aligned} (n/2 - \hat{s}_1 d_n) (\hat{\sigma}_{\hat{\mathcal{D}}_1^*}^2 - \sigma^2) & = \left( \mathbf{e}^{(2)T} \mathbf{e}^{(2)} - \frac{n}{2} \sigma^2 \right) - \left( \mathbf{e}^{(2)T} \mathbf{P}_{\hat{\mathcal{D}}_1^*} \mathbf{e}^{(2)} - \hat{s}_1 d_n \sigma^2 \right) + O_p(\hat{s}_1^2 n d_n^{-2d}). \end{aligned}$$

We next calculate the order of  $(\boldsymbol{\varepsilon}^{(2)T} \mathbf{P}_{\hat{\mathcal{D}}_1^*} \boldsymbol{\varepsilon}^{(2)} - \hat{s}_1 d_n \sigma^2)$ . Note that

$$\mathbb{E} \left( \boldsymbol{\varepsilon}^{(2)T} \mathbf{P}_{\hat{\mathcal{D}}_1^*} \boldsymbol{\varepsilon}^{(2)} - \sigma^2 \text{tr}(\mathbf{P}_{\hat{\mathcal{D}}_1^*}) \mid \mathbf{X}_{\hat{\mathcal{D}}_1^*}^{(2)} \right) = 0.$$

We now calculate its variance

$$\begin{aligned} \text{Var} \left( \boldsymbol{\varepsilon}^{(2)T} \mathbf{P}_{\hat{\mathcal{D}}_1^*} \boldsymbol{\varepsilon}^{(2)} - \sigma^2 \text{tr}(\mathbf{P}_{\hat{\mathcal{D}}_1^*}) \mid \mathbf{X}_{\hat{\mathcal{D}}_1^*}^{(2)} \right) \\ = \mathbb{E} \left( \left( \boldsymbol{\varepsilon}^{(2)T} \mathbf{P}_{\hat{\mathcal{D}}_1^*} \boldsymbol{\varepsilon}^{(2)} \right)^2 \mid \mathbf{X}_{\hat{\mathcal{D}}_1^*}^{(2)} \right) - \sigma^4 \text{tr}^2 \mathbf{P}_{\hat{\mathcal{D}}_1^*}. \end{aligned} \quad (\text{A.20})$$

Denote by  $P_{ij}$  the  $(i, j)$ th entry of matrix  $\mathbf{P}_{\hat{\mathcal{D}}_1^*}$ . The first term in the right-hand side of the last equation can be written as

$$\mathbb{E} \left( \sum_{i,j,k,l} \varepsilon_i \varepsilon_j \varepsilon_k \varepsilon_l P_{ij} P_{kl} \mid \mathbf{X}_{\hat{\mathcal{D}}_1^*}^{(2)} \right).$$

It follows by the independence between  $\mathbf{X}$  and  $\boldsymbol{\varepsilon}$  that

$$\begin{aligned} \mathbb{E} \left( \boldsymbol{\varepsilon}^{(2)T} \mathbf{P}_{\hat{\mathcal{D}}_1^*} \boldsymbol{\varepsilon}^{(2)} \mid \mathbf{X}_{\hat{\mathcal{D}}_1^*}^{(2)} \right) \\ = \mathbb{E} \varepsilon_1^4 \sum_{i=1}^{n/2} P_{ii}^2 + \sigma^4 \sum_{i \neq j} P_{ii} P_{jj} + 2\sigma^4 \sum_{i \neq j} P_{ij}^2. \end{aligned}$$

Therefore, it follows that the equation (A.20) equals to

$$\begin{aligned} \mathbb{E} \varepsilon_1^4 \sum_{i=1}^{n/2} P_{ii}^2 + \sigma^4 \sum_{i \neq j} P_{ii} P_{jj} + 2\sigma^4 \sum_{i \neq j} P_{ij}^2 - \sigma^4 \left( \sum_{i=1}^{n/2} P_{ii} \right)^2 \\ = \mathbb{E} \varepsilon_1^4 \sum_{i=1}^{n/2} P_{ii}^2 + 2\sigma^4 \sum_{i \neq j} P_{ij}^2 - \sigma^4 \sum_{i=1}^{n/2} P_{ii}^2. \end{aligned}$$

Noting the fact that  $\sigma^4 = (\mathbb{E} \varepsilon^2)^2 \leq \mathbb{E} \varepsilon^4$ , the last equation is bounded by

$$\begin{aligned} \mathbb{E} \varepsilon_1^4 \sum_{i=1}^{n/2} P_{ii}^2 - \sigma^4 \sum_{i=1}^{n/2} P_{ii}^2 + \sigma^4 \sum_{i \neq j} P_{ij}^2 + \mathbb{E} \varepsilon_1^4 \sum_{i \neq j} P_{ij}^2 \\ = (\mathbb{E} \varepsilon_1^4 + \sigma^4) \sum_{i=1}^{n/2} \sum_{j=1}^{n/2} P_{ij}^2 - 2\sigma^4 \sum_{i=1}^{n/2} P_{ii}^2. \end{aligned} \quad (\text{A.21})$$

Note that

$$\begin{aligned} \text{tr}(\mathbf{P}_{\hat{\mathcal{D}}_1^*}^2) &= \text{tr}(\mathbf{P}_{\hat{\mathcal{D}}_1^*}^T \mathbf{P}_{\hat{\mathcal{D}}_1^*}) = \sum_{i=1}^{n/2} \sum_{j=1}^{n/2} P_{ij}^2, \\ \text{tr}(\mathbf{P}_{\hat{\mathcal{D}}_1^*}) &= \text{tr}(\mathbf{P}_{\hat{\mathcal{D}}_1^*}^2) = \sum_{i=1}^{n/2} P_{ii}, \\ \text{tr}^2(\mathbf{P}_{\hat{\mathcal{D}}_1^*}) &= \left( \text{tr}(\mathbf{P}_{\hat{\mathcal{D}}_1^*}) \right)^2 = \sum_{i=1}^{n/2} P_{ii}^2 + \sum_{i \neq j} P_{ii} P_{jj}. \end{aligned}$$

and that  $\text{tr}^2(\mathbf{P}_{\hat{\mathcal{D}}_1^*}) = \left( \sum_{i=1}^{n/2} P_{ii} \right)^2 \leq n \sum_{i=1}^{n/2} P_{ii}^2$ . It follows that

$$\begin{aligned} \text{Var} \left( \boldsymbol{\varepsilon}^{(2)T} \mathbf{P}_{\hat{\mathcal{D}}_1^*} \boldsymbol{\varepsilon}^{(2)} - \sigma^2 \text{tr}(\mathbf{P}_{\hat{\mathcal{D}}_1^*}) \mid \mathbf{X}_{\hat{\mathcal{D}}_1^*}^{(2)} \right) \\ \leq (\mathbb{E} \varepsilon_1^4 + \sigma^4) \text{tr}(\mathbf{P}_{\hat{\mathcal{D}}_1^*}) - \frac{2\sigma^4}{n} \text{tr}^2(\mathbf{P}_{\hat{\mathcal{D}}_1^*}) \\ \leq (\mathbb{E} \varepsilon_1^4 + \sigma^4) \hat{s}_1 d_n. \end{aligned}$$

since for the projection matrix  $\mathbf{P}_{\hat{\mathcal{D}}_1^*}$ ,  $\text{tr}(\mathbf{P}_{\hat{\mathcal{D}}_1^*}) = \hat{s}_1 d_n$ . Consequently, by Markov's inequality, we obtain

$$\boldsymbol{\varepsilon}^{(2)T} \mathbf{P}_{\hat{\mathcal{D}}_1^*} \boldsymbol{\varepsilon}^{(2)} - \sigma^2 \hat{s}_1 d_n = O_p \left( \sqrt{\hat{s}_1 d_n} \right) \quad (\text{A.22})$$

Therefore, we have that

$$\left( \frac{n}{2} - \hat{s}_1 d_n \right) \left( \hat{\sigma}_{\hat{\mathcal{D}}_1^*}^2 - \sigma^2 \right) = \boldsymbol{\varepsilon}^{(2)T} \boldsymbol{\varepsilon}^{(2)} - \frac{n}{2} \sigma^2 + O_p \left( \sqrt{\hat{s}_1 d_n} \right) + O_p \left( \hat{s}_1^2 n d_n^{-2d} \right).$$

Similarly, it follows that

$$\left( \frac{n}{2} - \hat{s}_2 d_n \right) \left( \hat{\sigma}_{\hat{\mathcal{D}}_2^*}^2 - \sigma^2 \right) = \boldsymbol{\varepsilon}^{(1)T} \boldsymbol{\varepsilon}^{(1)} - \frac{n}{2} \sigma^2 + O_p \left( \sqrt{\hat{s}_2 d_n} \right) + O_p \left( \hat{s}_2^2 n d_n^{-2d} \right).$$

Finally, we deal with  $\sqrt{n}(\hat{\sigma}_{\text{RCV}}^2 - \sigma^2)$ . Take  $\hat{s}_1 = o(n^{(2d-1)/4(2d+1)})$ , and  $\hat{s}_2 = o(n^{(2d-1)/4(2d+1)})$  so that  $n/(n - 2\hat{s}_1 d_n) = 1 + o_p(1)$  and  $n/(n - 2\hat{s}_2 d_n) = 1 + o_p(1)$ . Then

$$\begin{aligned} \sqrt{n}(\hat{\sigma}_{\text{RCV}}^2 - \sigma^2) \\ = \frac{\sqrt{n}}{n - 2\hat{s}_1 d_n} \left( \boldsymbol{\varepsilon}^{(2)T} \boldsymbol{\varepsilon}^{(2)} - \frac{n}{2} \sigma^2 + O_p \left( \sqrt{\hat{s}_1 d_n} \right) + O_p \left( \hat{s}_1^2 n d_n^{-2d} \right) \right) \\ + \frac{\sqrt{n}}{n - 2\hat{s}_2 d_n} \left( \boldsymbol{\varepsilon}^{(1)T} \boldsymbol{\varepsilon}^{(1)} - \frac{n}{2} \sigma^2 + O_p \left( \sqrt{\hat{s}_2 d_n} \right) + O_p \left( \hat{s}_2^2 n d_n^{-2d} \right) \right) \\ = \left\{ \frac{1}{\sqrt{n}} \sum_{i=1}^n (\varepsilon_i^2 - \sigma^2) \right\} \{1 + o_p(1)\} + o_p(1) \\ \xrightarrow{\mathcal{L}} \mathcal{N}(0, \mathbb{E} \varepsilon_1^4 - \sigma^4), \quad \text{as } n \rightarrow \infty. \end{aligned}$$

This completes the proof of Theorem 2.

## Supplementary Materials

The supplementary material consists of a rigorous proof of (A.3).

## Acknowledgments

The authors thank the editor, the AE, and reviewers for their constructive comments that have led to a dramatic improvement of the earlier version of this article. Jianqing Fan is the corresponding author. All authors equally contribute to this paper, and are listed in alphabetic order.

## Funding

Chen's research was supported by NSF grant DMS-1206464 and NIH grants R01-GM072611. Fan's research was supported by NSF grant DMS-1206464 and NIH grants R01-GM072611 and R01GM100474-01. Li research was supported by a NSF grant DMS 1512422, National Institute on Drug Abuse (NIDA) grants P50 DA039838, P50 DA036107, and R01 DA039854, and National Nature Science Foundation of China, 11690015. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NSF, NIH, and NIDA.

## References

- Bach, F. R. (2008), "Consistency of the Group Lasso and Multiple Kernel Learning," *The Journal of Machine Learning Research*, 9, 1179–1225. [315]
- Bühlmann, P., and Van de Geer, S. (2011), *Statistics for High-Dimensional Data*, Berlin: Springer. [315]
- De Boor, C. (1978), *A Practical Guide to Splines* (Vol. 27), New York: Springer-Verlag. [317]



- Donoho, D. (2000), "High-Dimensional Data Analysis: The Curses and Blessings of Dimensionality," in *AMS Math Challenges Lecture*, pp. 1–32. [315]
- Fan, J., Feng, Y., and Song, R. (2011), "Nonparametric Independence Screening in Sparse Ultra-High-Dimensional Additive Models," *Journal of the American Statistical Association*, 106, 544–557. [315,317,319]
- Fan, J., Guo, S., and Hao, N. (2012), "Variance Estimation Using Refitted Cross-Validation in Ultrahigh Dimensional Regression," *Journal of the Royal Statistical Society, Series B*, 74, 37–65. [315,316,317,318,321]
- 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. [315]
- (2006), "Statistical Challenges with High Dimensionality: Feature Selection in Knowledge Discovery," in *Proceedings of the International Congress of Mathematicians*, eds. M. Sanz-Sole, J. Soria, J. L. Varona, and J. Verdera, pp. 595–622. [315]
- Fan, J., and Lv, J. (2008), "Sure Independence Screening for Ultrahigh Dimensional Feature Space," *Journal of the Royal Statistical Society, Series B*, 70, 849–911. [316,321]
- (2010), "A Selective Overview of Variable Selection in High Dimensional Feature Space," *Statistica Sinica*, 20, 101–148. [315]
- Friedman, J., and Stuetzle, W. (1981), "Projection Pursuit Regression," *Journal of the American Statistical Association*, 76, 817–823. [315]
- Hall, P., and Miller, H. (2009), "Using Generalized Correlation to Effect Variable Selection in Very High Dimensional Problems," *Journal of Computational and Graphical Statistics*, 18, 533–550. [315]
- Huang, J., Horowitz, J. L., and Wei, F. (2010), "Variable Selection in Nonparametric Additive Models," *Annals of Statistics*, 38, 2282–2313. [316]
- Li, R., Zhong, W., and Zhu, L. (2012), "Feature Screening via Distance Correlation Learning," *Journal of the American Statistical Association*, 107, 1129–1139. [317,319]
- Lin, Y., and Zhang, H. H. (2006), "Component Selection and Smoothing in Multivariate Nonparametric Regression," *The Annals of Statistics*, 34, 2272–2297. [315]
- Meier, L., Van de Geer, S., and Bühlmann, P. (2009), "High-Dimensional Additive Modeling," *The Annals of Statistics*, 37, 3779–3821. [315]
- Ravikumar, P., Lafferty, J., Liu, H., and Wasserman, L. (2009), "Sparse Additive Models," *Journal of the Royal Statistical Society, Series B*, 71, 1009–1030. [315]
- Stone, C. J. (1985), "Additive Regression and Other Nonparametric Models," *The Annals of Statistics*, 13, 689–705. [319]
- Van Der Vaart, A. W., and Wellner, J. A. (1996), *Weak Convergence and Empirical Processes*, New York: Springer. [324]
- Wang, H. (2009), "Forward Regression for Ultra-High Dimensional Variable Screening," *Journal of the American Statistical Association*, 104, 1512–1524. [321]
- Xue, L. (2009), "Consistent Variable Selection in Additive Models," *Statistica Sinica*, 19, 1281–1296. [315]
- Zhou, S., Shen, X., and Wolfe, D. A. (1998), "Local Asymptotics for Regression Splines and Confidence Regions," *The Annals of Statistics*, 26, 1760–1782. [323]