Error Variance Estimation in Ultrahigh Dimensional Additive Models

Zhao Chen, Jianqing Fan and Runze Li

August 1, 2016

Abstract

Error variance estimation plays an important role in statistical inference for high dimensional regression models. This paper 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 which 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 (Fan, Guo and Hao, 2012). 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.

Key words: Feature screening, Refitted cross-validation, Sparse additive model, Variance estimation

*Zhao Chen is Research Associate, Department of Statistics, The Pennsylvania State University at University Park, PA 16802-2111, USA. Email: zuc4@psu.edu. Chen's research was supported by NSF grant DMS-1206464 and NIH grants R01-GM072611. Jianqing Fan is Frederick L. Moore'18 Professor of Finance, Department of Operations Research & Financial Engineering, Princeton University, Princeton, NJ 08544, USA and Honorary Professor, School of Data Science, Fudan University, and Academy of Mathematics and System Science, Chinese Academy of Science, Beijing, China (E-mail: jqfan@princeton.edu). Fan's research was supported by NSF grant DMS-1206464 and NIH grants R01-GM072611 and R01GM100474-01. Runze Li is Verne M. Willaman Professor, Department of Statistics and The Methodology Center, The Pennsylvania State University, University Park, PA 16802-2111. Email: rzli@psu.edu. His research was supported by a NSF grant DMS 1512422, National Institute on Drug Abuse (NIDA) grants P50 DA039838, P50 DA036107, and R01 DA039854. The authors thank the Editor, the AE and reviewers for their constructive comments, which have led to a dramatic improvement of the earlier version of this paper. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NSF, NIH and NIDA.

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 of 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 papers 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 paper aims to study issues of error variance estimation in ultrahigh dimensional nonparametric regression settings.

In this paper, 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 paper 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 semi-varying 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 squares 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 squares 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 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 of Fan and Lv (2008).

In this paper, 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 in 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 $\sqrt{d_n n^{-1}\log(pd_n)}$ by using one most correlated predictor with d_n basis functions. If s variables are

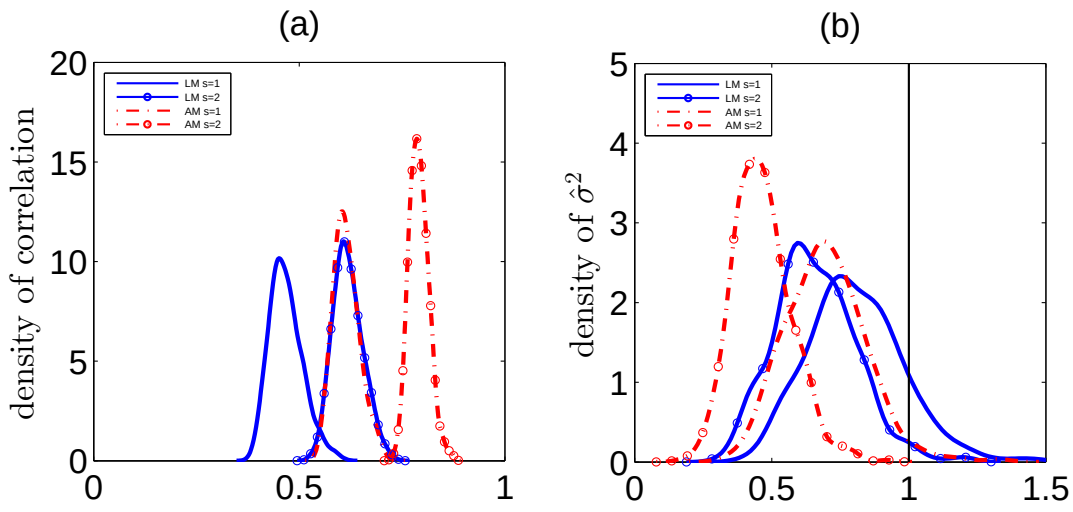

Figure 1: Distributions of the maximum "linear" and "nonparametric" spurious correlations for $s = 1$ and $s = 2$ (left panel, $n = 50$ and $p = 1000$) and their consequences on the estimating of noise variances (right panel). The legend 'LM' stands for linear model, and 'AM' stands for additive model,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

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 in Fan, Guo and Hao (2012), we compute the maximum "linear" spurious correlation $\zeta_n^L = \max_{1\le j\le p} |\widehat{\text{corr}}(X_j, Y)|$ and the maximum "nonparametric" spurious correlation $\zeta_n^N = \max_{1\le j\le p} |\widehat{\text{corr}}(\widehat{f}_j(X_j), Y)|$, where $\widehat{f}_j(X_j)$ 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 data sets, Figure 1 depicts the results, which 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 overfitting. 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 paper 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 paper 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 paper 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 $\mathbf{x} = (X_1, \cdots, X_p)^T$ be a predictor vector. The additive model assumes that

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

where μ is intercept term, $\{f_j(\cdot),\ j = 1, \cdots, p\}$ are the unknown functions and ε 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 paper that $Ef_j(X_j) = 0$ for $j = 1, \cdots, p$ so that model (2.1) is identifiable. This assumption implies that $\mu = E(Y)$. Thus, a natural estimator for μ 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 σ^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 paper, we employ B-spline method throughout this paper. Readers are referred to De Boor (1978) for detailed procedure of B-spline construction. Let $\{B_{jk}(x),\ k = 1, \cdots, d_j,\ a \le x \le b\}$ be B-spline basis of space $\mathcal{S}_j^l([a, b])$ 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

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

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

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

Suppose that $\{(\mathbf{x}_i, Y_i)\}$, $i = 1, \cdots, n$ 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 $\|f_j\|^2 = Ef_j^2(X_j) \neq 0$ 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 σ^2. This is due to the fact that extra variables are actually selected to predict the

realized noises (Fan, Guo and Hao, 2012). After feature screening, a direct estimate of σ^2 is the mean squared errors of the least squares 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 $\widehat{\mathcal{D}}^*$ the indices of the selected predictors respectively, satisfying the sure screening property $\mathcal{D}^* \subset \widehat{\mathcal{D}}^*$. Then, we minimize the following least squares function with respect to $\boldsymbol{\gamma}$:

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

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

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

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

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

See Lemma 1 and Theorem 1 in Section 2.2 for details. Our first aim is to propose a new estimation procedure of σ^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 data sets denoted by $\mathcal{I}_1$ and $\mathcal{I}_2$ with approximately equal size. Without loss of generality, assume through this paper 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 $\widehat{\mathcal{D}}_1$ and $\widehat{\mathcal{D}}_2$. 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 $\widehat{\mathcal{D}}_{3-l}$ for $l = 1$ and 2 by the least squares method. These results in two least squares estimate $\widehat{\boldsymbol{\gamma}}^{(3-l)}$ based on $\mathcal{I}_l$, respectively. In the second step, we calculate the mean squared errors for each fit:

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

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

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

This estimator is adapted from the one proposed in Fan, Guo and Hao (2012) for linear regression models, however, it is much more challenge in establishing the asymptotic property of $\widehat{\sigma}_{\mathrm{RCV}}^2$ 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 $\widehat{\sigma}_{\widehat{\mathcal{D}}^*}^2$ and $\widehat{\sigma}_{\mathrm{RCV}}^2$. 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 $\mathrm{E}\left\{\exp(A_1 |\varepsilon|) \middle| \mathbf{x}\right\} \leq A_2$.

(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

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

for a given constant $L > 0$, where $r \leq l$ is the "integer part" of d and $\alpha \in (0,1]$ such that

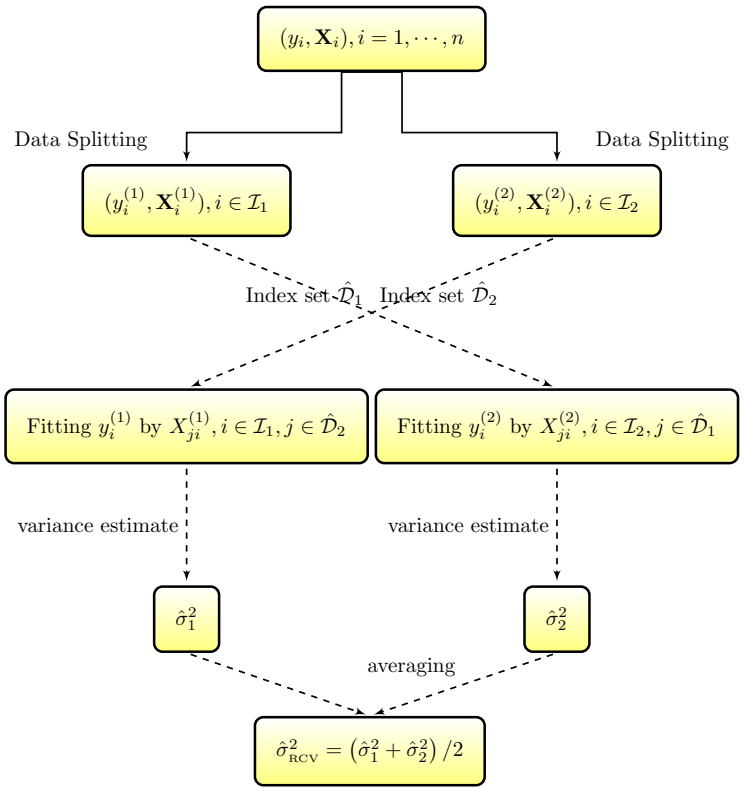

Figure 2: Refitted Cross Validation Procedure

$d = r + \alpha \geq 2$. 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 $X_j,\ j = 1, \cdots, p$ 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 in 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 $\widehat{\sigma}^2_{\widehat{\mathcal{D}}^*}$ 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 utilized 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 in Li, Zhong and Zhu (2012) for the DC-SIS and the conditions in Fan, Feng and Song (2011) for the NIS. We first derive the orders of $\boldsymbol{\varepsilon}^T\mathbf{P}_{\widehat{\mathcal{D}}^*}\boldsymbol{\varepsilon}$ and $\widehat{\gamma}_n^2$ in next lemma, which plays a critical role in the proofs of Theorems 1 and 2 below. 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

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

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 $\widehat{\gamma}_n^2$ increases to its upper bound at an exponential rate of $\widehat{s}$ since $\delta \in (0,1)$ and $2/(1-\delta) > 2$.

Theorem 1. Assume that $\limsup_{n\to\infty} \widehat{\gamma}_n^2 < 1$. Let $\widehat{s} = \left|\widehat{\mathcal{D}}^*\right|$ be the number of elements in the estimated active index set $\widehat{\mathcal{D}}^*$. Assume that all active predictors are retained in the stage of feature screening. That is, $\widehat{\mathcal{D}}^*$ contains all active predictors. Under Conditions (C1)-(C3), the following statements hold:

(i) If $\log(p_n) = O(n^\zeta)$, $0 \le \zeta < 1$ and $\widehat{s} = O_p(\log(n))$, then $\widehat{\sigma}^2_{\widehat{\mathcal{D}}^*}/(1-\widehat{\gamma}_n^2)$ converges to σ^2 in probability as $n \to \infty$;

(ii) If $\log(p_n) = O(n^\zeta)$, $0 \le \zeta < 3/(2d+1)$ and $\widehat{s} = O_p(\log(n))$, then it follows that

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

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

Theorem 1 (i) clearly indicates that the naive error variance estimator $\widehat{\sigma}^2_{\widehat{\mathcal{D}}^*}$ underestimates σ^2 by a factor of $(1-\widehat{\gamma}_n^2)$, yet by Lemma 1, $\widehat{\gamma}_n^2$ is of order given in (2.5) and is not small. Since $\widehat{\gamma}_n^2$ can not 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 $\widehat{\mathcal{D}}_j^*$ contains all active predictors, for $j = 1$ and 2. Let $\widehat{s}_j = \left|\widehat{\mathcal{D}}_j^*\right|$ be the number of elements in $\widehat{\mathcal{D}}_j^*$. Under Conditions (C1)-(C3), if $\widehat{s}_1 = o(n^{(2d-1)/4(2d+1)})$, and $\widehat{s}_2 = o(n^{(2d-1)/4(2d+1)})$, then

$$\sqrt{n}\left(\widehat{\sigma}^2_{\mathrm{RCV}} - \sigma^2\right) \xrightarrow{\mathcal{L}} \mathcal{N}(0, \mathrm{E}\varepsilon_1^4 - \sigma^4) \tag{2.8}$$

Comparing with the result in the 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 non-trivial.

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 $\widehat{\sigma}^2_{\widehat{\mathcal{D}}^*}/(1-\widehat{\gamma}_n^2) = \sigma^2 + O_p(1/\sqrt{n})$ and $\widehat{\sigma}^2_{\mathrm{RCV}} = \sigma^2 + O_p(1/\sqrt{n})$. As a result, the ratio of RCV estimate to the naive estimator may be used to provide one an estimate of the shrinkage factor $1-\widehat{\gamma}_n^2$.

(b) Theorem 2 is applicable provided that the active index sets $\widehat{\mathcal{D}}_j^*$, $j = 1$ 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, we have imposed conditions on $\widehat{s}$, $\widehat{s}_1$ and $\widehat{s}_2$. 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 of 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 $\widehat{s} = n/\log(n)$. The diverging rate of $\widehat{s}$, $\widehat{s}_1$ and $\widehat{s}_2$ 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 $\widehat{s}$, while the RCV method performs well for a wide range of $\widehat{s}_1$ and $\widehat{s}_2$. 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 $\widehat{s}_1$ and $\widehat{s}_2$ are more related to the topic of feature screening than the variance estimation and we do not intend to pursue further in this paper. 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 $\widehat{s}_1$ and $\widehat{s}_2$ to examine whether the resulting variance estimate is relatively stable to the choices of $\widehat{s}_1$ and $\widehat{s}_2$. This is implemented in the real data example in Section 3.2.

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 ultra-high 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

$$y = a\,\left(X_1 + 0.75X_2^2 + 2.25\cos(X_5)\right) + \varepsilon, \tag{3.1}$$

where $\varepsilon \sim N(0,1)$, and $\{X_1, \cdots, X_p\} \sim N_p(0, \Sigma)$ 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 $2/\sqrt{3}$ in order 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 equalling 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 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 $p - \widehat{s}$ predictors with low distance correlation. Thus, the resulting model includes $\widehat{s}$ predictors. We consider $\widehat{s}$=20, 30, 40 and 50 in order to illustrate the impact of choices of $\widehat{s}$ 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 5 cubic B-splines bases with 3 knots equally spaced between the minimum and maximum of the j^{th} variable. Thus, when $\widehat{s} = 50$, 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 $\widehat{s}$, 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: Simulation Results for different $\widehat{s}$ ($\sigma^2 = 1$)

	$\widehat{s} = 20$	$\widehat{s} = 30$	$\widehat{s} = 40$	$\widehat{s} = 50$
Method	$a = 0$			
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)

*Values in parentheses are standard errors

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 $\widehat{s}$, the bigger the spurious correlation γ_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

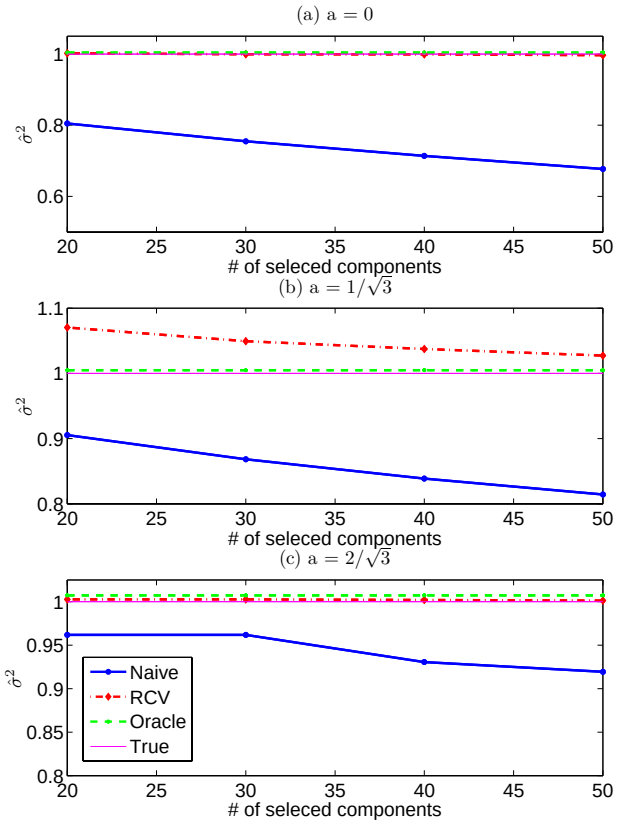

Figure 3: Variance estimators for different signal-to-noise ratios

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

	$n = 400$	$n = 600$
Method	$a = 0$	
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)

*Values in parentheses are standard errors

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 $\widehat{s}$ in data analysis. Fan and Lv (2008) suggested $\widehat{s} = [n/\log(n)]$ for their sure independence screening procedure based on Pearson correlation ranking. We modify their proposal and set $\widehat{s} = [n^{4/5}/\log(n^{4/5})]$ 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 alterative, one may repeat the proposed procedure several times, each randomly splitting data into two parts, and then take the average as the estimate of σ^2. Our findings from our simulations study are consistent with the discussion in Fan, Guo and Hao (2012): (a) the estimates of σ^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 paper. 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 to 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 $\widehat{\sigma}_l^2$ 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 data set (Wang, 2009). The data set contains a total of $n = 464$ daily records of the number of

customers (Y_i) and the sale amounts of $p = 6,398$ products, denoted as $X_{i1}, \cdots, X_{ip}$, 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.

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

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

Table 3: Error Variance Estimate for Market Data

$\widehat{s}$	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

Since the sample size $n = 464$, we set $\widehat{s} = [n^{4/5}/\log(n^{4/5})] = 28$. The naive error variance estimate equals 0.0938, while the RCV error variance estimate equals 0.1340, an 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 $\widehat{s}$, and clearly shows that the RCV estimate of error variance is stable with different choices of $\widehat{s}$, while the estimate of error variance by the naive method reduces as $\widehat{s}$ increases. This is consistent with our theoretical and simulation results.

Regarding the selected models with $\widehat{s}$ predictors as a correct model and ignoring the approximation errors (if any) due to B-spline, we further employ the Wald's χ^2-test for hypothesis whether $(\gamma_{j1}, \cdots, \gamma_{jd_j})^T$ equals zero, namely whether the j^{th} variable is active in presence of the rest variables. Such Wald's χ^2 statistics offer us a rough picture whether X_j is significant or not. The Wald's χ^2-test with the naive error variance estimate concludes 12 significant predictors at significant level 0.05, while the Wald's χ^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 χ^2-test statistic of those insignificant predictors identified by the Wald's test. Figure 4 clearly shows that the χ^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.

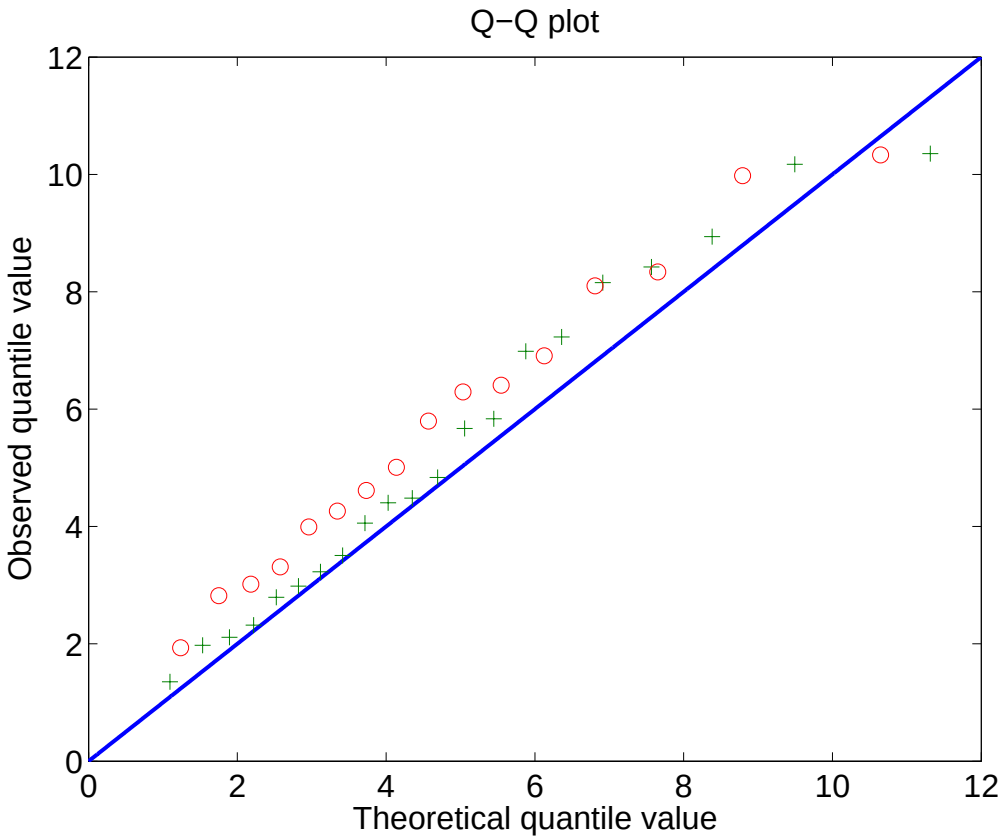

Figure 4: Quantile-quantile plot of χ^2-test values. "o" stands for χ^2-test using naive error variance estimate. "+" stands for χ^2-test using RCV error variance estimate.

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 7 predictors. The corresponding mean squared errors is 0.1207, which is close to the $\widehat{\sigma}^2_{\text{RCV}} = 0.1340$. Note that σ^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 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 $\widehat{\sigma}^2_{\text{RCV}}$.

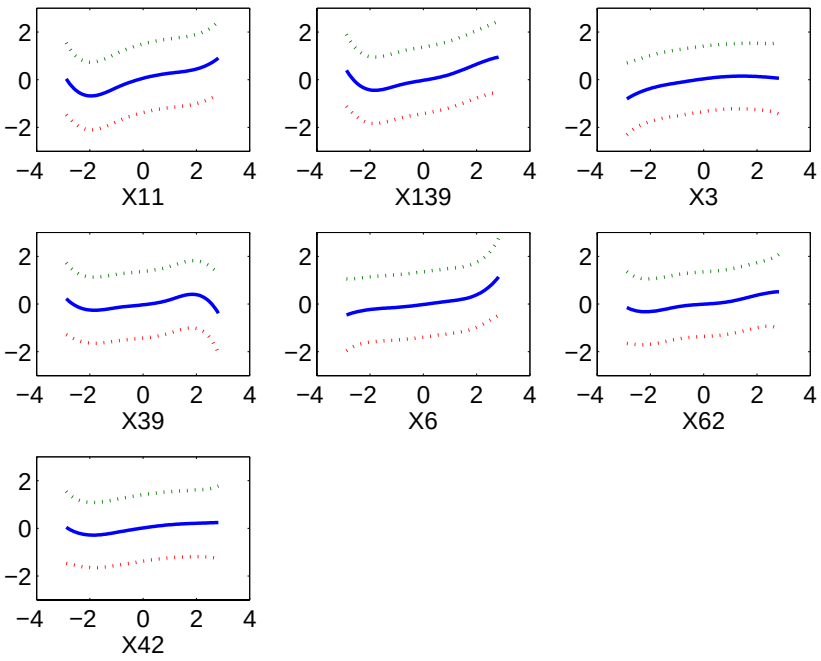

Figure 5: Estimated functions based on 7 variables selected from 28 variables that survive DC-SIS screening by the χ^2-test with the RCV error variance estimator.

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 $\{\mathbf{x}_i, Y_i\}$, $i = 1, \cdots, n$ is an independent and identically distributed random sample from a linear regression model $Y = \mathbf{x}^T\boldsymbol{\beta} + \varepsilon$, the linear predictor $\widehat{Y} = \mathbf{x}^T\widehat{\boldsymbol{\beta}}$, where $\widehat{\boldsymbol{\beta}}$ is the least squares estimate of $\boldsymbol{\beta}$, has prediction error at a new observation $\{\mathbf{x}_*, y_*\}$: $E\{(y_* - \mathbf{x}_*^T\widehat{\boldsymbol{\beta}})^2|\mathbf{X}\} = \sigma^2(1 + \mathbf{x}_*^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{x}_*)$, where σ^2 is the error variance and $\mathbf{X}$ is the corresponding design matrix. This explains why the MSPE is slightly greater than $\widehat{\sigma}^2_{\text{RCV}}$. To further gauge the accuracy of the RCV estimate of σ^2, define weighted prediction error $|y_* - \mathbf{x}_*^T\widehat{\boldsymbol{\beta}}|/\sqrt{1 + \mathbf{x}_*^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{x}_*}$. 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 achieves the benchmark prediction; (b) the modeling biases using the additive models of these seven variables are negligible; (c) $\widehat{\sigma}^2_{\text{RCV}}$ provides a very good estimate for

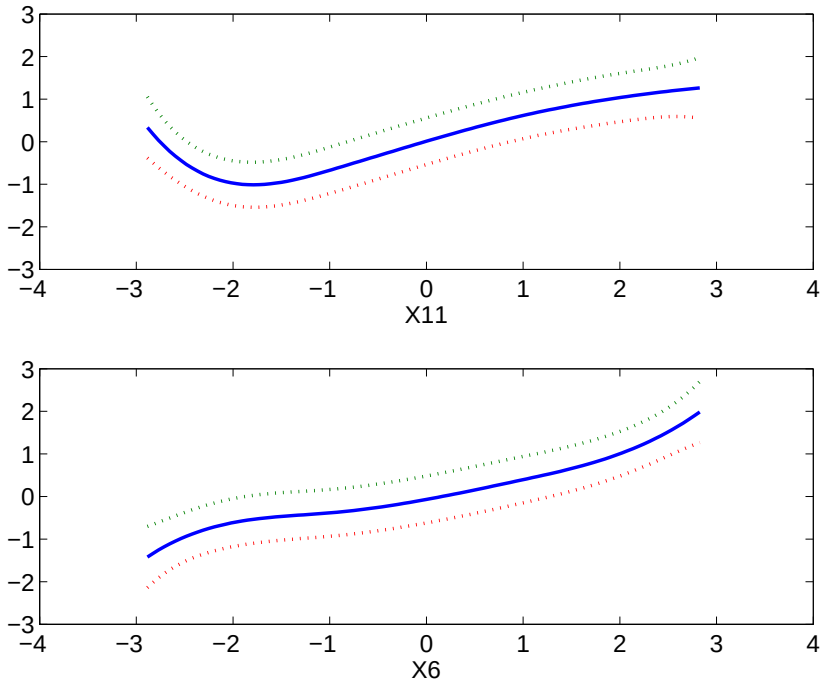

Figure 6: Estimated functions based on 2 variables selected from 28 variables that survive DC-SIS screening by the χ^2-test with the RCV error variance estimator and the Bonferroni adjustment.

σ^2.

Their estimated functions $\widehat{f}_j(x_j)$ 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 due to the fact that 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 $\widehat{f}_{11}(x_{11})$ and $\widehat{f}_6(x_6)$.

4 Discussions

In this paper, 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 $\mathbf{\Psi}$ be the corresponding design matrix of model (2.3). Specifically, $\mathbf{\Psi}$ is a $n \times (pd_n)$ matrix with ith row being $(B_{11}(X_{i1}), \cdots, B_{1d_n}(X_{i1}), B_{21}(X_{i2}), \cdots, B_{pd_n}(X_{ip}))$. Denote by $\mathbf{\Psi}^{(\widehat{\mathcal{D}}^*)}$ the corresponding design matrix of model $\widehat{\mathcal{D}}^*$, and $\mathbf{P}_{\widehat{\mathcal{D}}^*}$ the corresponding projection matrix. That is, $\mathbf{P}_{\widehat{\mathcal{D}}^*} = \mathbf{\Psi}^{(\widehat{\mathcal{D}}^*)}(\mathbf{\Psi}^{(\widehat{\mathcal{D}}^*)^T}\mathbf{\Psi}^{(\widehat{\mathcal{D}}^*)})^{-1}\mathbf{\Psi}^{(\widehat{\mathcal{D}}^*)^T}$. Denote $\mathbf{P}^c_{\widehat{\mathcal{D}}^*} = I_n - \mathbf{P}_{\widehat{\mathcal{D}}^*}$. 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 $\widehat{\mathcal{D}}^* = \{1, 2, \cdots, \widehat{s}\}$, where $\widehat{s} = \left|\widehat{\mathcal{D}}^*\right|$.

Proof of Lemma 1. Note that

$$\boldsymbol{\varepsilon}^T\mathbf{P}_{\widehat{\mathcal{D}}^*}\boldsymbol{\varepsilon} = \boldsymbol{\varepsilon}^T\mathbf{\Psi}^{(\widehat{\mathcal{D}}^*)}(\mathbf{\Psi}^{(\widehat{\mathcal{D}}^*)^T}\mathbf{\Psi}^{(\widehat{\mathcal{D}}^*)})^{-1}\mathbf{\Psi}^{(\widehat{\mathcal{D}}^*)^T}\boldsymbol{\varepsilon} \leq \lambda_{\min}^{-1}(\mathbf{\Psi}^{(\widehat{\mathcal{D}}^*)^T}\mathbf{\Psi}^{(\widehat{\mathcal{D}}^*)})\left\|\mathbf{\Psi}^{(\widehat{\mathcal{D}}^*)^T}\boldsymbol{\varepsilon}\right\|_2^2, \tag{A.1}$$

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 $\mathbf{\Psi}^{(\widehat{\mathcal{D}}^*)^T}\mathbf{\Psi}^{(\widehat{\mathcal{D}}^*)}$. Note that $\mathbf{\Psi}^{(\widehat{\mathcal{D}}^*)} = (\mathbf{\Psi}_1, \cdots, \mathbf{\Psi}_{\widehat{s}})$ with

$$\mathbf{\Psi}_j = \begin{pmatrix} B_{j1}(X_{j1}) & \cdots & B_{j\,d_n}(X_{j1}) \\ \cdots & \cdots & \cdots \\ B_{j1}(X_{jn}) & \cdots & B_{j\,d_n}(X_{jn}) \end{pmatrix}, \qquad j = 1, \cdots, \widehat{s}. \tag{A.2}$$

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

$$(\frac{1-\delta}{2})^{\widehat{s}-1}\left(\|\mathbf{\Psi}_1\mathbf{b}_1\|_2 + \cdots + \|\mathbf{\Psi}_{\widehat{s}}\mathbf{b}_{\widehat{s}}\|_2\right)^2 \leq \|\mathbf{\Psi}_1\mathbf{b}_1 + \cdots + \mathbf{\Psi}_{\widehat{s}}\mathbf{b}_{\widehat{s}}\|_2^2 = \mathbf{b}^T{\mathbf{\Psi}^{(\widehat{\mathcal{D}}^*)}}^T\mathbf{\Psi}^{(\widehat{\mathcal{D}}^*)}\mathbf{b}. \quad \text{(A.3)}$$

This yields that

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

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

$$(\frac{1-\delta}{2})^{\widehat{s}-1}\left(\sum_{i=1}^{\widehat{s}}\mathbf{b}_i\mathbf{\Psi}_i^T\mathbf{\Psi}_i\mathbf{b}_i\right) \geq (\frac{1-\delta}{2})^{\widehat{s}-1}\left(\sum_{\mathbf{b}_i^T\mathbf{b}_i \neq 0}\lambda_{\min}(\mathbf{\Psi}_i^T\mathbf{\Psi}_i)\,\mathbf{b}_i^T\mathbf{b}_i\right).$$

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

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

Thus,

$$(\frac{1-\delta}{2})^{\widehat{s}-1}\left(\sum_{\mathbf{b}_i^T\mathbf{b}_i \neq 0}\lambda_{\min}(\mathbf{\Psi}_i^T\mathbf{\Psi}_i)\,\mathbf{b}_i^T\mathbf{b}_i\right) \geq C_1(\frac{1-\delta}{2})^{\widehat{s}-1}d_n^{-1}n\sum_{\mathbf{b}_i^T\mathbf{b}_i \neq 0}\mathbf{b}_i^T\mathbf{b}_i = C_1(\frac{1-\delta}{2})^{\widehat{s}-1}d_n^{-1}n. \quad \text{(A.6)}$$

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

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

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

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

By the notation (A.2), we have

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

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

$$\mathrm{E}\,|B_{ij}(X_{ik})\varepsilon_k|^m = \mathrm{E}\,|B_{ij}(X_{ik})|^m \cdot \mathrm{E}\,|\varepsilon_k|^m \le \mathrm{E}\,|B_{ij}(X_{ik})|^m\, \mathrm{E}\,(m!\, a^m \exp\{|\varepsilon_1|/a\}). \tag{A.10}$$

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

$$C_4 m!\, a^m d_n^{-1} \mathrm{E}\,(\exp\{|\varepsilon_1|/a\}) \le \frac{C_4}{2} m!\, (2\, d_n^{-1} b\, a^2) a^{m-2}. \tag{A.11}$$

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

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

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

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

$$\boldsymbol{\varepsilon}^T\mathbf{P}_{\widehat{\mathcal{D}}^*}\boldsymbol{\varepsilon} \le C_1^{-1}\left(\frac{2}{1-\delta}\right)^{\widehat{s}-1} d_n n^{-1}\left\|\boldsymbol{\Psi}^{(\widehat{\mathcal{D}}^*)^T}\boldsymbol{\varepsilon}\right\|_2^2 \le C_5^2 C_1^{-1}\left(\frac{2}{1-\delta}\right)^{\widehat{s}-1} d_n\log(p\,d_n). \tag{A.12}$$

Due to the independent and identically distributed random errors with mean 0 and variance σ^2, by the *Law of Large Number*, we have

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

Thus, we obtain that

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

Proof of Theorem 1. Note that

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

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

$$\Delta_1 = \{\sum_{j=1}^{\widehat{s}}\mathbf{f}_j(\mathbf{X}_j) - \sum_{j=1}^{\widehat{s}}\mathbf{f}_{nj}(\mathbf{X}_j) + \sum_{j=1}^{\widehat{s}}\mathbf{f}_{nj}(\mathbf{X}_j)\}^T\mathbf{P}^c_{\widehat{\mathcal{D}}^*}\{\sum_{j=1}^{\widehat{s}}\mathbf{f}_j(\mathbf{X}_j) - \sum_{j=1}^{\widehat{s}}\mathbf{f}_{nj}(\mathbf{X}_j) + \sum_{j=1}^{\widehat{s}}\mathbf{f}_{nj}(\mathbf{X}_j)\},$$

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

$$\Delta_{11} = \{\sum_{j=1}^{\widehat{s}}\mathbf{f}_j(\mathbf{X}_j) - \sum_{j=1}^{\widehat{s}}\mathbf{f}_{nj}(\mathbf{X}_j)\}^T\mathbf{P}^c_{\widehat{\mathcal{D}}^*}\{\sum_{j=1}^{\widehat{s}}\mathbf{f}_j(\mathbf{X}_j) - \sum_{j=1}^{\widehat{s}}\mathbf{f}_{nj}(\mathbf{X}_j)\},$$
$$\Delta_{12} = \{\sum_{j=1}^{\widehat{s}}\mathbf{f}_{nj}(\mathbf{X}_j)\}^T\mathbf{P}^c_{\widehat{\mathcal{D}}^*}\{\sum_{j=1}^{\widehat{s}}\mathbf{f}_{nj}(\mathbf{X}_j)\},$$

$$\Delta_{13} = \ 2\{\sum_{j=1}^{\widehat{s}} \mathbf{f}_j(\mathbf{X}_j) - \sum_{j=1}^{\widehat{s}} \mathbf{f}_{nj}(\mathbf{X}_j)\}^T \mathbf{P}^c_{\widehat{\mathcal{D}}^*} \{\sum_{j=1}^{\widehat{s}} \mathbf{f}_{nj}(\mathbf{X}_j)\}.$$

Then $\Delta_1 = \Delta_{11} + \Delta_{12} + \Delta_{13}$. Note that $\mathbf{P}^c_{\widehat{\mathcal{D}}^*}$ is a projection matrix on the complement of the linear space of $\mathbf{\Psi}^{(\widehat{\mathcal{D}}^*)}$, and therefore $\mathbf{P}^c_{\widehat{\mathcal{D}}^*}\{\sum_{j=1}^{\widehat{s}} \mathbf{f}_{nj}(\mathbf{X}_j)\} = 0$. Thus, both Δ_{12} and Δ_{13} equal 0. We next calculate the order of Δ_{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 $\mathbf{P}^c_{\widehat{\mathcal{D}}^*}$ is a projection matrix, its eigenvalues equal either 0 or 1. By the Cauchy-Schwarz inequality and some straightforward calculation, it follows that $\Delta_{11} = O_p(\widehat{s}^2 n d_n^{-2d})$. Therefore $\Delta_1 = O_p(\widehat{s}^2 n d_n^{-2d})$. Under conditions in Theorem 1(i), $O_p(\widehat{s}^2 d_n^{-2d}) = o_p(1)$. As a result, $\Delta_1 = o_p(n)$. Under conditions in Theorem 1(ii), $\widehat{s} = o(n^{(2d-1)/4(2d+1)})$ and therefore $\Delta_1 = o_p(\sqrt{n})$.

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

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

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

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

which is bounded by $C_6^{m-2}\, \mathrm{E}\left(\left|f_j(X_{ji}) - f_{nj}(X_{ji})\right|^2 \ \mathrm{E}\left(\left|\varepsilon_i\right|^m \left| \mathbf{x}_i \right) \right)$ 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 $\mathrm{E}\left\{\exp(A_1 \left|\varepsilon_i\right|) | \mathbf{x}_i\right\}$, it

follows that

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

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

$$\mathrm{P}\left(\max_{1\leq j\leq p}\left|\sum_{i=1}^{n}(f_j(X_{ji}) - f_{nj}(X_{ji}))\,\varepsilon_i\right| > M\right) \leq 2\,p\,\exp\left\{-\frac{1}{2}\frac{M^2}{2\,C_7\,n\,d_n^{-2d} - C_8 M}\right\}. \tag{A.15}$$

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,

$$\Delta_{21} = O_p\left(\widehat{s}\sqrt{\log(p)\;n\,d_n^{-2d}}\right). \tag{A.16}$$

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

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

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

$$\mathbf{Y}^T\left(I_n - \boldsymbol{\Psi}^{(\widehat{\mathcal{D}}^*)}(\boldsymbol{\Psi}^{(\widehat{\mathcal{D}}^*)^T}\boldsymbol{\Psi}^{(\widehat{\mathcal{D}}^*)})^{-1}\boldsymbol{\Psi}^{(\widehat{\mathcal{D}}^*)^T}\right)\mathbf{Y}$$

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

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

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

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

$$
\frac{\widehat{\sigma}^2_{\widehat{\mathcal{D}}^*}}{(1 - \widehat{\gamma}_n^2)} = \frac{1}{n - \widehat{s}\, d_n}\, \boldsymbol{\varepsilon}^T \boldsymbol{\varepsilon} + O_p(\widehat{s}^2\, d_n^{-2d}) + O_p(\sqrt{\log(p)}\, \widehat{s}\, n^{-1/2}\, d_n^{-d}) + O_p(\frac{\Delta_{22}}{n}). \quad \text{(A.17)}
$$

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

$$
\frac{\widehat{\sigma}^2_{\widehat{\mathcal{D}}^*}}{1 - \widehat{\gamma}_n^2} \xrightarrow{\text{p}} \sigma^2. \quad \text{(A.18)}
$$

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

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

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.

A.2 Proof of Theorem 2.

Define events $\mathcal{A}_{n1} = \{\mathcal{D}^* \subset \widehat{\mathcal{D}}_1^*\}$, $\mathcal{A}_{n2} = \{\mathcal{D}^* \subset \widehat{\mathcal{D}}_2^*\}$ 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 $\boldsymbol{\Psi}^{(\widehat{\mathcal{D}}_1^*)}$ be the design matrix corresponding to $\widehat{\mathcal{D}}_1^*$, $\mathbf{P}_{\widehat{\mathcal{D}}_1^*} = \boldsymbol{\Psi}^{(\widehat{\mathcal{D}}_1^*)} (\boldsymbol{\Psi}^{(\widehat{\mathcal{D}}_1^*)^T} \boldsymbol{\Psi}^{(\widehat{\mathcal{D}}_1^*)})^{-1} \boldsymbol{\Psi}^{(\widehat{\mathcal{D}}_1^*)^T}$,

and $\mathbf{P}^c_{\widehat{\mathcal{D}}_1^*} = I - \mathbf{P}_{\widehat{\mathcal{D}}_1^*}$. Note that $\mathbf{P}^c_{\widehat{\mathcal{D}}_1^*}\left(\sum_{j=1}^{\widehat{s}_1}\mathbf{f}_{nj}(\mathbf{X}_j^{(2)})\right) = 0$. Thus,

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

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

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

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

$$\mathrm{E}\left(\boldsymbol{\varepsilon}^{(2)^T}\mathbf{P}_{\widehat{\mathcal{D}}_1^*}\boldsymbol{\varepsilon}^{(2)} - \sigma^2\mathbf{tr}(\mathbf{P}_{\widehat{\mathcal{D}}_1^*}) \,\Big|\, \mathbf{X}^{(2)}_{\widehat{\mathcal{D}}_1^*}\right) = 0.$$

We now calculate its variance

$$\mathrm{Var}\left(\boldsymbol{\varepsilon}^{(2)^T}\mathbf{P}_{\widehat{\mathcal{D}}_1^*}\boldsymbol{\varepsilon}^{(2)} - \sigma^2\mathbf{tr}(\mathbf{P}_{\widehat{\mathcal{D}}_1^*}) \,\Big|\, \mathbf{X}^{(2)}_{\widehat{\mathcal{D}}_1^*}\right) = \mathrm{E}\left(\left(\boldsymbol{\varepsilon}^{(2)^T}\mathbf{P}_{\widehat{\mathcal{D}}_1^*}\boldsymbol{\varepsilon}^{(2)}\right)^2 \,\Big|\, \mathbf{X}^{(2)}_{\widehat{\mathcal{D}}_1^*}\right) - \sigma^4\mathbf{tr}^2\mathbf{P}_{\widehat{\mathcal{D}}_1^*}. \quad \text{(A.20)}$$

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

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

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

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

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

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

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

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

Note that

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

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

$$\begin{aligned} \mathrm{Var}\left(\boldsymbol{\varepsilon}^{(2)T} \mathbf{P}_{\widehat{\mathcal{D}}_1^*} \boldsymbol{\varepsilon}^{(2)} - \sigma^2 \mathrm{tr}(\mathbf{P}_{\widehat{\mathcal{D}}_1^*}) \,\middle|\, \mathbf{X}^{(2)}_{\widehat{\mathcal{D}}_1^*} \right) &\leq (\mathrm{E}\varepsilon_1^4 + \sigma^4)\,\mathrm{tr}(\mathbf{P}_{\widehat{\mathcal{D}}_1^*}) - \frac{2\sigma^4}{n}\,\mathrm{tr}^2(\mathbf{P}_{\widehat{\mathcal{D}}_1^*}) \\ &\leq (\mathrm{E}\varepsilon_1^4 + \sigma^4)\widehat{s}_1\, d_n. \end{aligned}$$

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

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

Therefore, we have that

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

Similarly, it follows that

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

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

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

This completes the proof of Theorem 2.

References

[1] Bach, F. R. (2008), "Consistency of the group lasso and multiple kernel learning," *The Journal of Machine Learning Research*, 9: 1179-1225.

[2] Bühlmann, P. and Van de Geer, S. (2011). *Statistics for High-Dimensional Data*, Springer, Berlin.

[3] De Boor, C. (1978), *A Practical Guide to Splines,* Vol. 27, New York: Springer-Verlag.

[4] Donoho, D. (2000), "High-dimensional data analysis: The curses and blessings of dimensionality," *AMS Math Challenges Lecture*, 1-32.

[5] 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.

[6] 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 (Statistical Methodology)*, 74(1), 37-65.

[7] 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.

[8] Fan, J. and Li, R. (2006), "Statistical Challenges with High Dimensionality: Feature Selection in Knowledge Discovery," *Proceedings of the International Congress of Mathematicians* (M. Sanz-Sole, J. Soria, J.L. Varona, J. Verdera, eds.), Vol. III, European Mathematical Society, Zurich, 595-622.

[9] 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.

[10] Fan, J. and Lv, J. (2010), "A selective overview of variable selection in high dimensional feature space", *Statistica Sinica*, 20(1), 101.

[11] Friedman, J. and Stuetzle, W. (1981), "Projection pursuit regression." *Journal of the American statistical Association*, 76, 817-823.

[12] 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.

[13] Huang, J., Horowitz, J. L. and Wei, F. (2010), "Variable selection in nonparametric additive models," *Annals of Statistics*, 38, 2282 - 2313.

[14] Li, R., Zhong, W. and Zhu, L. (2012), "Feature screening via distance correlation learning," *Journal of the American Statistical Association*, 107, 1129-1139.

[15] Lin, Y. and Zhang, H. H. (2006), "Component selection and smoothing in multivariate nonparametric regressionm," *The Annals of Statistics*, 34, 2272-2297.

[16] Meier, L., Van de Geer, S., and Bühlmann, P. (2009), "High-dimensional additive modeling," *The Annals of Statistics*, 37, 3779-3821.

[17] Ravikumar, P., Lafferty, J., Liu, H. and Wasserman, L. (2009), "Sparse additive models," *Journal of the Royal Statistical Society: Series B*, 71, 1009-1030.

[18] Stone, C. J. (1985), "Additive regression and other nonparametric models," *The Annals of Statistics*, 689-705.

[19] Van Der Vaart, A. W. and Wellner, J. A. (1996), *Weak Convergence and Empirical Processes*, Springer.

[20] Wang, H. (2009), "Forward regression for ultra-high dimensional variable screening," *Journal of the American Statistical Association*, 104(488).

[21] Xue, L., (2009) "Consistent variable selection in additive models," *Statistica Sinica*, 19, 1281 - 1296.

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

Supplemental Material

In order to give a rigorous proof of (A.3), we extend Lemmas 1 and 3 of Stone (1985) for the centered second moments to the second moments in this supplemental material. It is worth pointing out that (A.3) is a direct application of Lemma S.5 below. Notations for constants in this supplement may have different meaning from those in the main text.

Lemma S.1. Let $V_j = h(X_j)$, $j = 1, \cdots, p$ be random variables and have finite second moment. Then it follows that

$$\mathrm{E}\Big(\sum_{j=1}^{k} V_j\Big)^2 \geq (\frac{1-\delta}{2})^{k-1}\Big(\sum_{j=1}^{k}\sqrt{\mathrm{E}V_j^2}\Big)^2, \tag{S.1}$$

where $\delta = (1 - b^2B^{-2}\zeta)^{1/2}$, for some constant $0 < \zeta < 1$.

Let SD stands for standard deviation. Lemma 1 of Stone (1985) shows that

$$\mathrm{SD}\Big(\sum_{j=1}^{k} V_j\Big) \geq (\frac{1-\delta^*}{2})^{(k-1)/2}\Big(\sum_{j=1}^{k}\mathrm{SD}(V_j)\Big), \tag{S.2}$$

where $\delta^* = (1 - bB^{-1})^{1/2} < 1$. Thus, Lemma S.1 is an extension of Lemma 1 of Stone (1985) in the sense that SD^2 (i.e. the second centered moment) is replaced by the second moment. This extension is needed because we cannot assume that the average of elements in every column of $\mathbf{\Psi}$ equals 0 due to the construction of B-splines bases.

Proof of Lemma S.1. Similar to the proof of Lemma 1 in Stone (1985), we will prove Lemma S.1 by induction. When $k = 1$, (S.1) is valid. Suppose that (S.1) holds for k. We will show that the inequality holds for $k+1$. Let $\tau_k^2 = \mathrm{E}(\sum_{j=1}^{k} V_j)^2$ and $m_k^2 = \mathrm{E}V_k^2$. If $\tau_k^2 = 0$, then $m_1^2 = \cdots = m_k^2 = 0$ and (S.1) is valid. If $m_{k+1}^2 = 0$, (S.1) follows from that for k. Next consider the case that $\tau_k^2 > 0$ and $m_{k+1}^2 > 0$. Set $X = \mathbf{1}^T\mathbf{V} = \mathbf{1}^T(V_1, \cdots, V_k)^T$ and $Y = V_{k+1}$. Consider the optimization problem

$$\min_{\beta} \mathrm{E}(Y - \beta X)^2 = \mathrm{E}Y^2(1 - \frac{(\mathrm{E}XY)^2}{\mathrm{E}X^2\mathrm{E}Y^2}) = m_{k+1}^2(1 - \eta^2), \tag{S.3}$$

where $\eta = \mathrm{E}XY/\sqrt{\mathrm{E}X^2\mathrm{E}Y^2}$. By Condition (C3), $b^2 < b < g_{V_1,\cdots,V_{k+1}}(v_1, \cdots, v_{k+1}) < B$, when

$b < 1$, and $g_{\mathbf{V}}(\mathbf{v}) < B, g_{V_{k+1}}(v_{k+1}) < B$. On the other hand, we also have

$$\begin{aligned}
&\min_{\beta} \mathrm{E}(Y - \beta X)^2 \\
&= \min_{\beta} \int (Y - \beta X)^2 g_{V_1,\cdots,V_{k+1}}(v_1, \cdots, v_{k+1})\, dv_1 \cdots dv_{k+1} \\
&> b^2 B^{-2} \min_{\beta} \int \int (Y - \beta X)^2 g_{\mathbf{V}}(\mathbf{v}) g_{V_{k+1}}(v_{k+1})\, d\mathbf{v} dv_{k+1} \\
&\geq b^2 B^{-2} m_{k+1}^2 (1 - \frac{(\mathrm{E}X)^2(\mathrm{E}Y)^2}{\mathrm{E}X^2\mathrm{E}Y^2}).
\end{aligned} \tag{S.4}$$

Because $0 \leq (\mathrm{E}X)^2(\mathrm{E}Y)^2/(\mathrm{E}X^2\mathrm{E}Y^2) \leq 1$, there exists a positive constant $\zeta_0 < 1$ such that

$$1 - \frac{(\mathrm{E}X)^2(\mathrm{E}Y)^2}{\mathrm{E}X^2\mathrm{E}Y^2} \geq \zeta_0 > 0.$$

since both X and Y are not degenerated. Thus, $\eta^2 \leq 1 - b^2 B^{-2} \zeta_0$ and hence $\eta \geq -\delta$. Consequently,

$$\begin{aligned}
\tau_{k+1}^2 &= \tau_k^2 + 2\eta\tau_k m_{k+1} + m_{k+1}^2 \\
&= (\frac{1+\eta}{2} + \frac{1-\eta}{2})(\tau_k^2 + m_{k+1}^2) + 2\eta\tau_k m_{k+1} \\
&\geq \frac{1+\eta}{2}(\tau_k^2 + m_{k+1}^2) + \frac{1-\eta}{2} 2\tau_k m_{k+1} + 2\eta\tau_k m_{k+1} \\
&= \frac{1+\eta}{2}(\tau_k + m_{k+1})^2 \\
&\geq \frac{1-\delta}{2} \left((\frac{1-\delta}{2})^{(k-1)/2} \left(\sum_{j=1}^{k} m_j \right) + m_{k+1} \right)^2 \\
&\geq (\frac{1-\delta}{2})^k (\sum_{j=1}^{k+1} m_j)^2
\end{aligned} \tag{S.5}$$

Thus, (S.1) holds for $k+1$. By induction the result holds for any integer k. If m_k^2 is bounded, when k goes to infinity, the inequality still holds.

The following two lemmas due to Stone (1985) will be used in the proof of Lemma S.5.

Lemma S.2. Let $\mathbf{X}_1, \cdots, \mathbf{X}_n$ be independent and identically distributed samples from a population satisfying Condition (C3). P_n is the empirical distribution of $\mathbf{X}_1, \cdots, \mathbf{X}_n$. Suppose $d_n^{2d} = o(n^{1-a})$

for some $a > 0$, and let $\varepsilon > 0$. Then except on an event with probability approaching to zero,

$$|\mathrm{P}_n(\mathbf{X}_j \in I_{n\nu}) - \mathrm{P}(\mathbf{X}_j \in I_{n\nu})| \le \varepsilon \mathrm{P}(\mathbf{X}_j \in I_{n\nu}), \quad \text{for } 1 \le j \le p, 1 \le \nu \le d_n,$$
$$|\mathrm{P}_n(\mathbf{X}_{j_1} \in I_{n\nu_1}, \mathbf{X}_{j_2} \in I_{n\nu_2}) - \mathrm{P}(\mathbf{X}_{j_1} \in I_{n\nu_1}, \mathbf{X}_{j_2} \in I_{n\nu_2})| \le \varepsilon \mathrm{P}(\mathbf{X}_{j_1} \in I_{n\nu_1}, \mathbf{X}_{j_2} \in I_{n\nu_2}),$$
$$\text{for } 1 \le j_1, j_2 \le p, j_1 \ne j_2, 1 \le \nu_1, \nu_2 \le d_n.$$

Proof. This indeed is Lemma 2 of Stone (1985).

Lemma S.3. Denote I to be an interval of finite positive length d_I, and (T, U) to be a pair of I-valued random variables with an absolutely continuous distribution. Let E_n be the expectation operator corresponding to the empirical distribution based on a random sample of size n. Suppose that the marginal density of T and U are bounded below by β/d_I on I, where $\beta > 0$. For $t > 0$, except on an event with probability at most $d_k \exp\{-2nt^2\}$, the following inequalities hold for all polynomials Q and R of degree k:

$$\begin{aligned}
|\mathrm{E}_n Q(T) - \mathrm{E}Q(T)| &\le t\beta^{-1/2} c_k \sqrt{\mathrm{E}Q^2(T)}; \\
\left|\mathrm{E}_n Q^2(T) - \mathrm{E}Q^2(T)\right| &\le t\beta^{-1} c_k^2 \, \mathrm{E}Q^2(T); \\
|\mathrm{E}_n R(U) - \mathrm{E}R(U)| &\le t\beta^{-1/2} c_k \sqrt{\mathrm{E}R^2(U)}; \\
|\mathrm{E}_n Q(T)R(U) - \mathrm{E}Q(T)R(U)| &\le t(t+3)\beta^{-1} c_k^2 \sqrt{\mathrm{E}Q^2(t)\,\mathrm{E}R^2(U)}.
\end{aligned}$$

Here d_k is a positive constant only depending on k.

Proof. This is Lemma 12 of Stone (1985).

Lemma S.4. Let (Ω, P) be a probability space, and $A_1, \cdots, A_\Lambda$ be a finite partition of Ω such that $p_\lambda = \mathrm{P}(A_\lambda) > 0$ for $1 \le \lambda \le \Lambda$. Let V_1 and V_2 be a pair of random variables with finite second moments and E be the expectation operator. Let (Ω', P') be another probability space. Define E', A_i', V_j' as before. For any small fixed constant $\varepsilon > 0$, there exists a $\delta > 0$ such that $|\mathrm{P}'(A_\lambda') - \mathrm{P}(A_\lambda)| \le \delta \mathrm{P}(A_\lambda)$ and $\left|\mathrm{E}'(V_1'V_2'|A_\lambda') - \mathrm{E}(V_1V_2|A_\lambda)\right| \le \delta\sqrt{\mathrm{E}(V_1^2|A_\lambda)\mathrm{E}(V_2^2|A_\lambda)}$, for $1 \le$

$\lambda \leq \Lambda$. Then it follows that

$$\left|\mathrm{E}'(V_1'V_2') - \mathrm{E}(V_1V_2)\right| \leq \varepsilon\sqrt{\mathrm{E}V_1^2\mathrm{E}V_2^2}. \tag{S.6}$$

Proof. Observe that $\mathrm{E}(V_1V_2) = \sum_\lambda p_\lambda \mathrm{E}(V_1V_2|A_\lambda)$, $\mathrm{E}V_1^2 = \sum_\lambda p_\lambda \mathrm{E}(V_1^2|A_\lambda)$. It follows by using Cauchy-Schwarz's inequality twice that

$$\mathrm{E}(V_1V_2) \leq \sum_\lambda p_\lambda \sqrt{\mathrm{E}(V_1^2|A_\lambda)\mathrm{E}(V_2^2|A_\lambda)} \leq \sqrt{\sum_\lambda p_\lambda \mathrm{E}(V_1^2|A_\lambda)}\sqrt{\sum_\lambda p_\lambda \mathrm{E}(V_2^2|A_\lambda)}.$$

Thus, (S.6) holds.

Lemma S.5. Suppose that Condition (C3) holds and let $\delta_1 \in (\delta, 1)$. Let E_n be the expectation corresponding to the empirical distribution based on a random sample of size n. Then except on an event with probability tending to zero, it follows that

$$\mathrm{E}_n\Big(\sum_{j=1}^k V_j\Big)^2 \geq \left(\frac{1-\delta_1}{2}\right)^{k-1}\Big(\sum_{j=1}^k \sqrt{\mathrm{E}_n V_j^2}\Big)^2. \tag{S.7}$$

Proof. Choose constant $0 < \varepsilon < \{(1-\delta)/2\}^{k-1}$. By Lemmas S.2, S.3 and S.4, it follows that, except on an event with probability approaching to zero,

$$|\mathrm{E}_n(V_iV_j) - \mathrm{E}(V_iV_j)| \leq \varepsilon\sqrt{\mathrm{E}V_i^2\mathrm{E}V_j^2}, \quad \text{for } 1 \leq i, j \leq k.$$

When $i = j$, it follows that

$$(1-\varepsilon)\mathrm{E}V_i^2 \leq \mathrm{E}_n V_j^2 \leq (1+\varepsilon)\mathrm{E}V_j^2, \quad \text{for } 1 \leq j \leq k. \tag{S.8}$$

Furthermore, we obtain that

$$\mathrm{E}_n(\sum_j V_j)^2 = \mathrm{E}(\sum_j V_j)^2 + E_n(\sum_j V_j)^2 - \mathrm{E}(\sum_j V_j)^2$$

$$\geq \mathrm{E}(\sum_j V_j)^2 - \varepsilon \mathrm{E}(\sum_j V_j)^2$$
$$\geq \mathrm{E}(\sum_j V_j)^2 - \varepsilon(\sum_j \sqrt{\mathrm{E}V_j^2})^2.$$

The last inequality hold by Minkowski's inequality. By Lemma S.1 and (S.8), it follows that

$$\mathrm{E}_n(\sum_j V_j)^2 \geq \left((\frac{1-\delta_1}{2})^{k-1} - \varepsilon \right) (\sum_j \sqrt{\mathrm{E}V_j^2})^2 \geq \left((\frac{1-\delta_1}{2})^{k-1} - \varepsilon \right) (1+\varepsilon)^{-1} (\sum_j \sqrt{\mathrm{E}_n V_j^2})^2.$$

Since ε can be made arbitrarily small, (S.7) holds.