On High Dimensional Post-Regularization Prediction Intervals

Shih-Kang Chao* Yang Ning† Han Liu‡

Abstract

This paper considers the construction of prediction intervals for future observations in high dimensional regression models. We propose a new approach to evaluate the uncertainty for estimating the mean parameter based on the widely-used penalization/regularization methods. The proposed method is then applied to construct prediction intervals for sparse linear models as well as sparse additive models. We establish the asymptotic normality of the estimator for the mean parameter and the asymptotic coverage probability of the prediction intervals. The theoretical properties of the proposed methods are verified by extensive simulation studies and real data analysis.

Keyword: Asymptotic normality, Prediction, Lasso, SCAD, Dantzig selector, Sparse additive model, Spline, Linear model

1 Introduction

In many modern research areas including genomics, biomedical imaging, signal processing, epidemiological studies, and high frequency finance, a large amount of variables is often collected, which poses new challenges for statistical analysis. During the past decade, such high dimensional data have been extensively studied, and a variety of penalization/regularization methods is proposed. In particular, for the Lasso estimator (Tibshirani, 1996), the rate of convergence and the variable selection consistency has been studied by Bickel et al. (2009); Zhang (2009); Negahban et al. (2012); Meinshausen and Bühlmann (2006); Zhao and Yu (2006); Wainwright (2009), among others. Moreover, the nonconvex penalized estimators, including MCP (Zhang, 2010a), SCAD (Fan and Li, 2001), and capped-L_1 penalty (Zhang, 2010b), are also proposed. The estimation consistency and oracle properties of the nonconvex estimators are established by Fan et al. (2012b); Fan and Lv (2011); Loh and Wainwright (2013); Wang et al. (2013); Zhang (2013), among others. To relax the

*Ladislaus von Bortkiewicz Chair of Statistics, C.A.S.E. - Center for applied Statistics and Economics, Humboldt-Universität zu Berlin, Unter den Linden 6, 10099, Berlin, Germany; email: shih-kang.chao@hu-berlin.de.

†Department of Operations Research Research and Financial Engineering, Princeton University, Princeton, NJ 08544, USA; e-mail: yning@princeton.edu.

‡Department of Operations Research Research and Financial Engineering, Princeton University, Princeton, NJ 08544, USA; e-mail: hanliu@princeton.edu.

linearity assumption in the linear model, more flexible models such as sparse additive models have been proposed and studied by Meier et al. (2009); Ravikumar et al. (2009); Huang et al. (2010); Fan et al. (2011); Raskutti et al. (2012), among others. Recently, The problem of constructing confidence intervals for the low dimensional component of high dimensional parameters has been considered by Zhang and Zhang (2014); van de Geer et al. (2014); Ning and Liu (2014); Javanmard and Montanari (2013); Belloni et al. (2013).

In regression analysis for low dimensional data, prediction of future observations based on the observed data and its uncertainty assessment are fundamental problems and have been well explained in many (under)graduate-level textbooks; see Faraway (2014) and Graybill (2000). However, the solution to such problems remains largely unknown in high dimensional models. To close this gap, this paper considers how to construct prediction intervals for two fundamental models: (1) sparse linear models and (2) sparse additive models. For the illustration purpose, assume that the data $\{Y_i, \boldsymbol{X}_i\}_{i=1}^n$ are i.i.d. and Y_i given $\boldsymbol{X}_i$ is generated from a linear model $Y_i = \boldsymbol{X}_i^T\boldsymbol{\beta}^* + \varepsilon_i$, where $\varepsilon_i \sim N(0, \sigma^2)$, and $\boldsymbol{\beta}^*$ is a d dimensional unknown parameter. Our goal is to predict the value of Y corresponding to a new covariate $\boldsymbol{X} = \mathbf{x}^*$, when d is much larger than n. As shown in Faraway (2014), the problem of constructing prediction intervals for Y is equivalent to constructing the confidence intervals for the mean parameter $\mathbf{x}^{*\top}\boldsymbol{\beta}^*$.

This article has two main contributions. First, we propose a new calibration method to evaluate the uncertainty for estimating the mean parameter $\mathbf{x}^{*\top}\boldsymbol{\beta}^*$. The proposed method can be applied for a wide class of regularized estimators $\widehat{\boldsymbol{\beta}}$, including nonconvex estimators, provided the convergence rate of $\widehat{\boldsymbol{\beta}}$ to $\boldsymbol{\beta}^*$ is sufficiently fast. In addition, our method can applied to more sophisticated nonparametric models such as sparse additive models. This further extends the scope of our method. Second, we establish the asymptotic normality for an estimator of the mean parameter $\mathbf{x}^{*\top}\boldsymbol{\beta}^*$. We prove this result under very mild conditions, which only needs that $\boldsymbol{\beta}^*$ is sparse, $\boldsymbol{\Sigma} = \mathbb{E}(\boldsymbol{X}_i\boldsymbol{X}_i^T)$ has bounded condition number and $\boldsymbol{X}_i$ satisfies some sub-Gaussian type condition. In particular, we highlight that our method does not need (1) the sparsity of the new covariate $\mathbf{x}^*$; (2) the irrepresentable or the minimal signal strength condition for variable selection consistency; (3) the sparsity of $\boldsymbol{\Sigma}^{-1}$. The similar asymptotic results are also established for the sparse additive models.

Compared with the prediction intervals in low dimensional linear models (Faraway, 2014; Graybill, 2000), the calibration step is new and essential. The reason is that the penalized estimator such as Lasso no longer possess a tractable asymptotic distribution (Knight and Fu, 2000), which makes the prediction intervals considered in Faraway (2014); Graybill (2000) infeasible. In the discussion to Lockhart et al. (2014), Wasserman (2014) briefly described the use of conformal prediction interval (Vovk et al., 2005, 2009). For low dimensional models, Lei et al. (2014) and Lei and Wasserman (2014) extended the conformal inference, and established the properties of the conformal prediction intervals. However, theoretical properties for high dimensional models do not exist for this approach. In high dimensional linear models, a de-biased approach is proposed by Zhang and Zhang (2014); van de Geer et al. (2014); Javanmard and Montanari (2013) to establish the element-wise asymptotic normality of some estimator $\widetilde{\boldsymbol{\beta}}$. Compared to these works, our method has the following three advantages: (1) Their results do not imply the asymptotic normality of $\mathbf{x}^*\widetilde{\boldsymbol{\beta}}$,

unless the new covariate $\mathbf{x}^*$ is sparse which is very restrictive in practice. In contrast, we do not need $\mathbf{x}^*$ to be sparse. (2) To derive the de-biased estimator $\widetilde{\boldsymbol{\beta}}$, one needs to estimate the inverse of the $d \times d$ matrix $\boldsymbol{\Sigma}$, which is computationally intensive for large d. In contrast, our proposed calibration method only requires to estimate a d dimensional vector, which shows our computational advantage. (3) The de-biased estimator has not been theoretically justified for sparse additive models. This papers for the first time provides rigorous theoretical results for confidence and prediction intervals.

The paper is organized as follows. Section 2 presents the prediction intervals for linear models. In Section 3 we consider the prediction intervals for the SParse Additive Model (SPAM). The simulation results and real data analysis are presented in Sections 4 and 5, respectively. Section 6 contains summary and discussions. The proofs are deferred to Appendices.

1.1 Notations

The following notations are adopted throughout this paper. For $v = (v_1, ..., v_d)^T \in \mathbb{R}^d$, and $1 \le q \le \infty$, we define $\|v\|_q = (\sum_{i=1}^d |v_i|^q)^{1/q}$, $\|v\|_0 = |\text{supp}(v)|$, where $\text{supp}(v) = \{j : v_j \neq 0\}$ and $|A|$ is the cardinality of a set A. Denote $\|v\|_\infty = \max_{1 \le i \le d} |v_i|$. For a matrix $\mathbf{M} = (M_{ij})$, let $\|\mathbf{M}\|_2$, $\|\mathbf{M}\|_\infty$ be the spectral, elementwise supreme norms of $\mathbf{M}$. Let $\mathcal{S}^c$ denotes the complement of the set $\mathcal{S}$. We use K, C to denote generic constants independent of n in our paper.

Definition 1.1 (Sub-exponential variable and sub-exponential norm)**.** A random variable X is called sub-exponential if there exists some positive constant K_1 such that $\mathbb{P}(|X| > t) \le \exp(1 - t/K_1)$ for all $t \ge 0$. The sub-exponential norm of X is defined as $\|X\|_{\psi_1} = \sup_{p \ge 1} p^{-1} (\mathbb{E}|X|^p)^{1/p}$.

Definition 1.2 (Sub-Gaussian variable and sub-Gaussian norm)**.** A random variable X is called sub-Gaussian if there exists some positive constant K_2 such that $\mathbb{P}(|X| > t) \le \exp(1 - t^2/K_2^2)$ for all $t \ge 0$. The sub-Gaussian norm of X is defined as $\|X\|_{\psi_2} = \sup_{p \ge 1} p^{-1/2} (\mathbb{E}|X|^p)^{1/p}$.

2 Prediction Intervals for Linear Models

Assume that the response Y_i given the covariate $\boldsymbol{X}_i = (X_{i1}, ..., X_{id})$ is independently generated from the linear model,

$$Y_i = \boldsymbol{X}_i^T \boldsymbol{\beta}^* + \varepsilon_i, \quad \text{where} \quad \varepsilon_i \sim N(0, \sigma^2), \tag{2.1}$$

and $\boldsymbol{\beta}^*$ is a d dimensional unknown parameter. Our goal is to predict the value of Y corresponding to a new covariate $\boldsymbol{X} = \mathbf{x}^*$, and construct prediction intervals for Y.

As $\boldsymbol{\beta}^*$ is unknown, one needs to estimate $\boldsymbol{\beta}^*$ to construct prediction intervals. Let $\ell(\boldsymbol{\beta}) = (2n)^{-1} \|\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}\|_2^2$ denote the least square loss, where $\mathbf{Y} = (Y_1, ..., Y_n)$ and $\mathbf{X} = (\boldsymbol{X}_1, ..., \boldsymbol{X}_n)^T$ is a $n \times d$ design matrix. In high dimensional regime, $\boldsymbol{\beta}^*$ is usually estimated by the following penalized M-estimator

$$\widehat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta} \in \mathbb{R}^d}{\text{argmin}}\, \ell(\boldsymbol{\beta}) + \sum_{j=1}^d p_\lambda(\beta_j),$$

where $\lambda \geq 0$ is a tuning parameter and $p_\lambda(\cdot)$ is a generic penalty function, including the ℓ_1 penalty $p_\lambda(u) = \lambda \cdot |u|$, and

$$\text{SCAD penalty (Fan and Li, 2001):} \quad p_\lambda(u) = \lambda \int_0^{|u|} \Big\{ \mathrm{I}(z \leq \lambda) + \frac{(a\lambda - z)_+}{(a-1)\lambda} \mathrm{I}(z > \lambda) \Big\} dz,$$

$$\text{MCP penalty (Zhang, 2010a):} \quad p_\lambda(u) = \lambda \int_0^{|u|} \Big(1 - \frac{z}{\lambda b}\Big)_+ dz,$$

for some constants a, b. Alternative estimators include the post-Lasso estimator (Belloni and Chernozhukov, 2013) and the Dantzig selector (Candes and Tao, 2007).

To construct prediction intervals, we first consider how to construct confidence intervals for the mean parameter $\mathbb{E}(Y \mid \boldsymbol{X} = \mathbf{x}^*) = \mathbf{x}^{*\top}\boldsymbol{\beta}^*$. While $\mathbf{x}^{*\top}\widehat{\boldsymbol{\beta}}$ is consistent for $\mathbf{x}^{*\top}\boldsymbol{\beta}^*$, the distribution of $\mathbf{x}^{*\top}\widehat{\boldsymbol{\beta}}$ is intractable due to regularization. Hence, a calibration is needed to construct convenient confidence intervals for $\mathbf{x}^{*\top}\boldsymbol{\beta}^*$. Denote $\widehat{\boldsymbol{\Sigma}} = n^{-1}\sum_{i=1}^n \boldsymbol{X}_i \boldsymbol{X}_i^T$. We propose the following estimator of $\mathbf{x}^{*\top}\boldsymbol{\beta}^*$,

$$S_{\mathbf{x}}(\widehat{\boldsymbol{\beta}}) = \mathbf{x}^{*\top}\widehat{\boldsymbol{\beta}} - \frac{1}{n}\sum_{i=1}^n \widehat{\mathbf{w}}^T \boldsymbol{X}_i (Y_i - \boldsymbol{X}_i^T \widehat{\boldsymbol{\beta}}),$$

where $\widehat{\mathbf{w}}$ is obtained from

$$\widehat{\mathbf{w}} = \underset{\mathbf{w}\in\mathbb{R}^d}{\operatorname{argmin}}\, \mathbf{w}^T \widehat{\boldsymbol{\Sigma}} \mathbf{w}, \quad \text{subject to} \quad \big\|\mathbf{x}^* + \widehat{\boldsymbol{\Sigma}}\mathbf{w}\big\|_\infty \leq \lambda', \tag{2.2}$$

where λ' is an additional tuning parameter. The purpose of the calibration is to modify $\mathbf{x}^{*\top}\widehat{\boldsymbol{\beta}}$ such that the estimator $S_{\mathbf{x}}(\widehat{\boldsymbol{\beta}})$ will satisfy the following two properties: (1) $S_{\mathbf{x}}(\widehat{\boldsymbol{\beta}})$ remains consistent for $\mathbf{x}^{*\top}\boldsymbol{\beta}^*$ and (2) $S_{\mathbf{x}}(\boldsymbol{\beta})$ is insensitive to the perturbation of $\boldsymbol{\beta}$. The latter can be understood that the derivative of $S_{\mathbf{x}}(\boldsymbol{\beta})$ with respect to $\boldsymbol{\beta}$, i.e., $\partial S_{\mathbf{x}}(\boldsymbol{\beta})/\partial\boldsymbol{\beta} = \mathbf{x}^* + \widehat{\boldsymbol{\Sigma}}\widehat{\mathbf{w}}$, is sufficiently small, since it is controlled by λ' uniformly as specified in (2.2). Thus, the effect of estimating $\boldsymbol{\beta}$ in $S_{\mathbf{x}}(\boldsymbol{\beta})$ by the regularized estimator $\widehat{\boldsymbol{\beta}}$ is asymptotically ignorable, and this avoids the calculation of the limiting distributions of $\widehat{\boldsymbol{\beta}}$. It would be clear later that such an approach in (2.2) essentially minimizes the width of the confidence interval and meanwhile controls the bias at the level of λ'.

Denote $s^* = \|\boldsymbol{\beta}^*\|_0$. The following Lemma establishes the asymptotic normality of $S_{\mathbf{x}}(\widehat{\boldsymbol{\beta}})$.

Lemma 2.1. Under the linear model assumption (2.1), if $\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_1 = \mathcal{O}_{\mathbb{P}}(s^*\sqrt{\log d/n})$ holds, then we have $n^{1/2}(S_{\mathbf{x}}(\widehat{\boldsymbol{\beta}}) - \mathbf{x}^{*\top}\boldsymbol{\beta}^*) = N + \xi$, where $N \mid \mathbf{X}, \mathbf{x}^* \sim N(0, \sigma^2 \widehat{\mathbf{w}}^T \widehat{\boldsymbol{\Sigma}} \widehat{\mathbf{w}})$, and $|\xi| = \mathcal{O}_{\mathbb{P}}(\lambda' s^* \sqrt{\log d})$.

Proof. See Appendix A for a detailed proof. □

To apply this Lemma, we need to ensure that (1) the conditional variance of N exists and is bounded and (2) the magnitude of λ' can be sufficiently small, such that $\xi = o_{\mathbb{P}}(1)$. To this end, we focus on the random design. That means $\boldsymbol{X}_1, ..., \boldsymbol{X}_n$ are i.i.d. In addition, we need to normalize the new covariate $\mathbf{x}^*$. Otherwise, $\mathbf{x}^{*\top}\boldsymbol{\beta}^*$ can be arbitrarily large, as $d \to \infty$. Without loss of generality, we assume $\|\mathbf{x}^*\|_2 = 1$.

We first invoke the following lemma to show that $\mathbf{w}^* := -\boldsymbol{\Sigma}^{-1}\mathbf{x}^*$, where $\boldsymbol{\Sigma} = \mathbb{E}(\boldsymbol{X}_i \boldsymbol{X}_i^T)$, is in the feasible set of the problem (2.2).

Lemma 2.2. Assume that $\boldsymbol{X}_1, ..., \boldsymbol{X}_n$ are n i.i.d random variables with mean 0 and variance $\boldsymbol{\Sigma}$. In addition, there exist two constants $C_{\min}$ and $C_{\max}$, such that $0 < C_{\min} \leq \lambda_{\min}(\boldsymbol{\Sigma}) \leq \lambda_{\max}(\boldsymbol{\Sigma}) \leq C_{\max} < \infty$. Assume that $\|\boldsymbol{\Sigma}^{-1/2}\boldsymbol{X}_i\|_{\psi_2} = C$, and $\|\mathbf{x}^*\|_2 = 1$. Then, with probability at least $1 - 2d^{-3}$,

$$\left\|\mathbf{x}^* + \widehat{\boldsymbol{\Sigma}}\mathbf{w}^*\right\|_\infty \leq 2(1 + 2\rho^{1/2}C^2)\sqrt{\frac{C''^{-1}\log d}{n}}, \quad \text{where} \quad \rho = C_{\max}/C_{\min},$$

provided $4C''^{-1}\log d \leq n$, and C'' is a universal constant given in Lemma C.3.

Proof. See Appendix A for a detailed proof. □

By this Lemma, one can take $\lambda' = 2(1 + 2\rho^{1/2}C^2)\sqrt{C''^{-1}\log d/n}$, and the feasible set in (2.2) is non-empty. This implies that (1) the conditional variance of N in Lemma 2.1 exists and (2) ξ in Lemma 2.1 can be $o_{\mathbb{P}}(1)$. Combining Lemma 2.1 and Lemma 2.2, we can establish the main theorem of this section.

Theorem 2.3. Assume that the linear model (2.1), and the conditions in Lemma 2.2 hold. If $\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_1 = \mathcal{O}_{\mathbb{P}}(s^*\sqrt{\log d/n})$ and $s^*\log d = o(n^{1/2})$ hold, then conditioning on the design $\mathbf{X}$ and $\mathbf{x}^*$, $n^{1/2}(S_{\mathbf{x}}(\widehat{\boldsymbol{\beta}}) - \mathbf{x}^{*\top}\boldsymbol{\beta}^*) \rightsquigarrow N(0, \sigma^2\widehat{\mathbf{w}}^T\widehat{\boldsymbol{\Sigma}}\widehat{\mathbf{w}})$. Finally, if $\|\mathbf{w}^*\|_1^2\sqrt{\log d/n} = \mathcal{O}(1)$, then $\widehat{\mathbf{w}}^T\widehat{\boldsymbol{\Sigma}}\widehat{\mathbf{w}} = \mathcal{O}_{\mathbb{P}}(1)$.

Proof. See Appendix A for a detailed proof. □

If $\widehat{\boldsymbol{\beta}}$ corresponds to the Lasso estimator, then the following corollary establishes the nonasymptotic characterization of $S_{\mathbf{x}}(\widehat{\boldsymbol{\beta}})$ conditional on the design variables $\mathbf{X}$ and $\mathbf{x}^*$.

Corollary 2.4. Assume that the linear model (2.1), and the conditions in Lemma 2.2 hold. Then $n^{1/2}(S_{\mathbf{x}}(\widehat{\boldsymbol{\beta}}) - \mathbf{x}^{*\top}\boldsymbol{\beta}^*) = N + \xi$, where $N \mid \mathbf{X}, \mathbf{x}^* \sim N(0, \sigma^2\widehat{\mathbf{w}}^T\widehat{\boldsymbol{\Sigma}}\widehat{\mathbf{w}})$. In addition, assume $\widehat{\Sigma}_{jj} = 1$, for $j = 1, ..., d$,

$$2\sqrt{\frac{C''^{-1}\log d}{n}} \leq 1, \quad \text{and} \quad 32(2C^2 + 1)\rho s^*\sqrt{\frac{C''^{-1}\log d}{n}} \leq 1/2,$$

where $\rho = C_{\max}/C_{\min}$ is the condition number and C'' is a universal constant in Lemma C.3. Then, with

$$\lambda = 4\sigma^2\sqrt{\frac{6\log d}{n}}, \quad \text{and} \quad \lambda' = 2(1 + 2\rho^{1/2}C^2)\sqrt{\frac{C''^{-1}\log d}{n}},$$

we further have ξ satisfies

$$|\xi| \leq 64(1 + 2C^2)\rho\sigma^2\sqrt{\frac{6}{C''}}\frac{s^*\log d}{\sqrt{n}},$$

with probability at least $1 - 4d^{-2} - 2d^{-3}$.

Proof. See Appendix A for a detailed proof. □

Note that Theorem 2.3 and Corollary 2.4 establish the asymptotic normality of $S_{\mathbf{x}}(\widehat{\boldsymbol{\beta}})$ under very mild conditions. In particular, we only require that $\boldsymbol{\beta}^*$ is sparse, $\boldsymbol{\Sigma}$ has bounded condition number and $\boldsymbol{X}_i$ satisfies some sub-Gaussian type condition, which are all routine assumptions for high dimensional models. Our results are different from the asymptotic normality based on the oracle properties (Fan and Li, 2001). The reason is that we do not assume the minimal signal strength of $\boldsymbol{\beta}^*$ is sufficiently large, which is a necessary condition for variable selection consistency. Unlike the confidence interval for β_j based on desparsifying the Lasso estimator in van de Geer et al. (2014), we do not assume the sparsity of $\boldsymbol{\Sigma}^{-1}$, and our Theorem 2.3 holds for a general class of regularized estimators, including the nonconvex penalty. Although the sparsity of $\boldsymbol{\Sigma}^{-1}$ is not needed in Javanmard and Montanari (2013), one cannot directly use the existing methods for confidence intervals for β_j to infer the mean parameter $\mathbf{x}^{*\top}\boldsymbol{\beta}^*$. There are two reasons. First, the error of normal approximation may accumulate very fast for estimating $\mathbf{x}^{*\top}\boldsymbol{\beta}^*$. This is because the desparsifying/debiased estimators of $\boldsymbol{\beta}^*$ in van de Geer et al. (2014); Javanmard and Montanari (2013) are no longer sparse. To control the estimation error of $\mathbf{x}^{*\top}\boldsymbol{\beta}^*$ based on their methods, the sparsity of $\mathbf{x}^*$ may be needed, which is very restrictive in practice. Second, the approaches in van de Geer et al. (2014) and Javanmard and Montanari (2013) are computationally intensive, because they need to estimate the precision matrix $\boldsymbol{\Sigma}^{-1}$, which is either decomposed into d Lasso problems (van de Geer et al., 2014) or d constrained optimization problems (Javanmard and Montanari, 2013). In contrast, we only need to solve one constrained optimization problem (2.2), which demonstrates the computational advantage of our procedure.

As an interesting special case, if $\mathbf{x}^* = \mathbf{e}_j$, where $\mathbf{e}_j$ is a unite vector with 1 on the jth component and 0 otherwise, then the parameter of interest $\mathbf{x}^{*\top}\boldsymbol{\beta}^*$ reduces to β_j and our method and theory implies those in van de Geer et al. (2014); Javanmard and Montanari (2013). However, in general, we do not require $\mathbf{x}^*$ to be sparse. Thus, Theorem 2.3 is a novel contribution in high dimensional inference towards understanding the asymptotic results for a nonsparse linear combination of parameters.

To apply our method, it is crucial to estimate the variance of the model error σ^2. To make our theory hold, it is sufficient to come up with an estimator $\widehat{\sigma}^2$ satisfying $|\widehat{\sigma}^2 - \sigma^2| = o_{\mathbb{P}}(1)$. In Lemma 2.5, we show that the estimator

$$\widehat{\sigma}^2 = n^{-1}\sum_{i=1}^{n}(Y_i - \boldsymbol{X}_i^T\widehat{\boldsymbol{\beta}})^2 \tag{2.3}$$

is applicable. Similar results can be found in Greenshtein and Ritov (2004).

Lemma 2.5 (Convergence of $\widehat{\sigma}$)**.** Under the conditions in Lemma 2.1 and $\sigma^2 > C$ for constant $C > 0$, we have $|\widehat{\sigma}^2 - \sigma^2| = o_{\mathbb{P}}(1)$.

Proof. See Appendix A for a detailed proof. □

An estimator that we apply in our simulation studies is

$$\widehat{\sigma}^2_{\widehat{\lambda}} = n^{-1}\sum_{i=1}^{n}(Y_i - \boldsymbol{X}_i^T\widehat{\boldsymbol{\beta}}_{\widehat{\lambda}})^2, \tag{2.4}$$

where $\widehat{\lambda}$ is estimated from a K-fold cross-validation procedure. Some alternative estimators have been considered in the literature. For instance, Fan et al. (2012a) proposed a two-stage refitted cross-validation estimator, which reduces the effect of overselection of Lasso estimators. In addition, Reid et al. (2014) compared several estimators of σ^2 based on the Lasso procedure by simulations, and they adopted the result of Homrighausen and McDonald (2013) to conclude that if $\|\boldsymbol{\beta}^*\|_1 = o((n/\log n)^{1/4})$, then (2.4) remains consistent. Another popular method to estimate σ^2 is the scaled Lasso proposed by Sun and Zhang (2012). In particular, they developed an iterative algorithm to solve $\boldsymbol{\beta}$ and σ simultaneously. They also proved the consistency of the estimator $\widehat{\sigma}$; see Sun and Zhang (2012) for details.

Combining Theorem 2.3 and Lemma 2.5, the Slutsky's theorem implies that conditioning on the design $\mathbf{X}$ and $\mathbf{x}^*$, we have

$$n^{1/2}(S_{\mathbf{x}}(\widehat{\boldsymbol{\beta}}) - \mathbf{x}^{*\top}\boldsymbol{\beta}^*) \rightsquigarrow N(0, \widehat{\sigma}^2 \widehat{\mathbf{w}}^T \widehat{\boldsymbol{\Sigma}} \widehat{\mathbf{w}}),$$

and therefore the prediction interval with asymptotic coverage probability $(1-\alpha)$ for Y corresponding to $\boldsymbol{X} = \mathbf{x}^*$ is given by $[S_{\mathbf{x}}(\widehat{\boldsymbol{\beta}}) - \Phi^{-1}(1-\alpha/2)V, S_{\mathbf{x}}(\widehat{\boldsymbol{\beta}}) + \Phi^{-1}(1-\alpha/2)V]$, where $\Phi^{-1}(\cdot)$ is the inverse of c.d.f of standard normal variables and

$$V = \widehat{\sigma} + n^{-1/2}(\widehat{\sigma}^2 \widehat{\mathbf{w}}^T \widehat{\boldsymbol{\Sigma}} \widehat{\mathbf{w}})^{1/2}.$$

For reader's convenience, the procedure to construct the high-dimensional prediction intervals for linear models is summarized in Algorithm 1.

Algorithm 1 Calculate the prediction intervals for linear models

Require: : Loss function $\ell(\boldsymbol{\beta}) = (2n)^{-1}\|\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}\|_2^2$, penalty function $P(\cdot)$ and tuning parameters λ and λ', covariates $\mathbf{x}^*$, significance level $0 < \alpha < 1$.

(i) Calculate $\widehat{\boldsymbol{\beta}}$ as $\widehat{\boldsymbol{\beta}} = \operatorname{argmin}_{\boldsymbol{\beta}} \ell(\boldsymbol{\beta}) + P_\lambda(\boldsymbol{\beta})$.

(ii) Estimate $\widehat{\mathbf{w}}$ by

$$\widehat{\mathbf{w}} = \operatorname*{argmin}_{\mathbf{w}\in\mathbb{R}^d} \mathbf{w}^T \widehat{\boldsymbol{\Sigma}} \mathbf{w}, \quad \text{s.t.} \ \left\| \mathbf{x}^* + \frac{1}{n}\sum_{i=1}^n \boldsymbol{X}_i \boldsymbol{X}_i^T \mathbf{w} \right\|_\infty \le \lambda'.$$

(iii) Calculate

$$S_{\mathbf{x}}(\widehat{\boldsymbol{\beta}}) = \mathbf{x}^{*\top}\widehat{\boldsymbol{\beta}} - \frac{1}{n}\sum_{i=1}^n \widehat{\mathbf{w}}^T \boldsymbol{X}_i (Y_i - \boldsymbol{X}_i^T \widehat{\boldsymbol{\beta}}), \quad \text{and} \quad \widehat{\sigma}^2 = n^{-1}\sum_{i=1}^n (Y_i - \boldsymbol{X}_i^T \widehat{\boldsymbol{\beta}})^2.$$

return Prediction intervals $[S_{\mathbf{x}}(\widehat{\boldsymbol{\beta}}) - \Phi^{-1}(1-\alpha/2)V, S_{\mathbf{x}}(\widehat{\boldsymbol{\beta}}) + \Phi^{-1}(1-\alpha/2)V]$, where $V = \widehat{\sigma} + n^{-1/2}(\widehat{\sigma}^2 \widehat{\mathbf{w}}^T \widehat{\boldsymbol{\Sigma}} \widehat{\mathbf{w}})^{1/2}$.

3 Prediction Intervals for Sparse Additive Models (SPAM)

In this section we consider the prediction interval for sparse additive models. Assume that the response Y_i given the covariate $\boldsymbol{X}_i = (X_{i1}, ..., X_{id})$ is independently generated from the following

sparse additive model (Huang et al., 2010; Ravikumar et al., 2009):

$$Y_i = \mu + \sum_{j=1}^{d} f_j(X_{ij}) + \varepsilon_i, \tag{3.1}$$

where X_{ij} takes values in a compact interval $[a, b]$ in which $a < b$ are finite numbers, ε_i are i.i.d. $N(0, \sigma^2)$, and $f_j : \mathbb{R} \to \mathbb{R}$ are unknown functions with identifiability conditions $\mathbb{E}(f_j(X_j)) = 0$ for $1 \le j \le d$. We assume $\mu = 0$, $f_j = 0$ for any $j \in \mathcal{S}^c$ where $\mathcal{S} \subset \{1, 2, ..., d\}$ is the support set with $s^* = |\mathcal{S}| \ll n$. The nonzero f_j is assumed to lie in $\mathcal{F}$ which is the class of functions f on $[0, 1]$ whose αth derivative $f^{(\alpha)}$ exists and satisfies a Lipschitz condition of order γ:

$$|f^{(\alpha)}(s) - f^{(\alpha)}(t)| \le C|s - t|^{\gamma}, \quad \text{for any } s, t \in [a, b],$$

where C is a constant and $k = \alpha + \gamma > 0.5$. Based on the construction of Schumaker (1981) and Stone (1985), each function $f_j \in \mathcal{F}$ can be approximated by an element $f_{nj} \in \mathcal{S}_n$ so that $\|f_{nj} - f_j\|_\infty = \mathcal{O}(m_n^{-k})$ (see p. 150 of Newey (1997)), where

$$f_{nj} = \sum_{k=1}^{m_n} \beta_{jk}^* \phi_k(x),$$

and $\mathcal{S}_n$ is an expanding class of polynomials spanned by a set of normalized B-spline basis $\{\phi_k, 1 \le k \le m_n\}$. Let $\|f\|_2 = [\int_a^b f^2(x)dx]^{1/2}$, $\boldsymbol{\beta}_j^* = (\beta_{j1}^*, ..., \beta_{jm_n}^*)$ and $\boldsymbol{\beta}^* = (\boldsymbol{\beta}_1^{*\top}, ..., \boldsymbol{\beta}_d^{*\top})^T$. For any $x \in \mathbb{R}$, let $\boldsymbol{\Phi}(x) = (\phi_1(x), ..., \phi_{m_n}(x))^T$. The centered basis function is

$$\psi_{jk}(x) = \phi_k(x) - \bar{\phi}_{jk}, \text{ where } \bar{\phi}_{jk} = \frac{1}{n}\sum_{i=1}^{n} \phi_k(X_{ij}),$$

and we define $\boldsymbol{\Psi}_j(x) = \boldsymbol{\Phi}(x) - \bar{\boldsymbol{\Phi}}_j \in \mathbb{R}^{m_n}$ as the centered $\boldsymbol{\Phi}(x)$ at the jth covariate where $\bar{\boldsymbol{\Phi}}_j = n^{-1}\sum_{i=1}^{n} \boldsymbol{\Phi}(X_{ij})$. To define the design matrix, let

$$\boldsymbol{Z}_i = \widetilde{\boldsymbol{Z}}_i - \bar{\boldsymbol{Z}} \in \mathbb{R}^{dm_n}, \text{ where } \widetilde{\boldsymbol{Z}}_i = (\boldsymbol{\Phi}(X_{i1})^T, ..., \boldsymbol{\Phi}(X_{id})^T)^T \text{ and } \bar{\boldsymbol{Z}} = \frac{1}{n}\sum_{i=1}^{n} \widetilde{\boldsymbol{Z}}_i. \tag{3.2}$$

Consequently, $\boldsymbol{Z}_i$ are centered in the empirical mean $\sum_{i=1}^{n} \boldsymbol{Z}_i = 0$. The design matrix is given by $\mathbf{Z} = (\boldsymbol{Z}_1, ..., \boldsymbol{Z}_n)^T \in \mathbb{R}^{n \times dm_n}$. Denote $\widehat{\mathbf{C}} = \mathbf{Z}^T\mathbf{Z}/n$, and $\boldsymbol{Y} = (Y_1, ..., Y_n)^T$.

For the sparse additive model, Huang et al. (2010) estimated $f_j(x)$ by $\widehat{f}_j(x) = \sum_{k=1}^{m_n} \widehat{\beta}_{jk}\psi_{jk}(x)$, where $\widehat{\boldsymbol{\beta}}$ is obtained by the following group Lasso procedure:

$$\widehat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \Big\{ \ell(\boldsymbol{\beta}) + \lambda \sum_{j=1}^{d} \|\boldsymbol{\beta}_j\|_2 \Big\}, \quad \text{where } \ell(\boldsymbol{\beta}) = \frac{1}{2n}\|\boldsymbol{Y} - \mathbf{Z}\boldsymbol{\beta}\|_2^2. \tag{3.3}$$

Here we focus on the following prediction problem. Given a new covariate $\mathbf{x}^*$, we would like to construct confidence intervals for $\mathbf{z}^{*\top}\boldsymbol{\beta}^*$, where $\mathbf{z}^* = (\boldsymbol{\Phi}(x_1^*)^T, ..., \boldsymbol{\Phi}(x_d^*)^T)^T$. In addition, we consider the prediction interval for Y corresponding to $\boldsymbol{X} = \mathbf{x}^*$. Similar to the linear regression, for theoretical development, we assume $\|\mathbf{z}^*\|_2 = 1$.

Following the same rationale as in Section 2, to estimate $\mathbf{z}^{*\top}\boldsymbol{\beta}^*$, we propose the following estimator,

$$S_{\mathbf{x}}(\widehat{\boldsymbol{\beta}}) = \mathbf{z}^{*\top}\widehat{\boldsymbol{\beta}} - \frac{1}{n}\sum_{i=1}^{n}\widehat{\mathbf{w}}^T\boldsymbol{Z}_i(Y_i - \boldsymbol{Z}_i^T\widehat{\boldsymbol{\beta}}),$$

where $\widehat{\mathbf{w}}$ is obtained from the following procedure

$$\widehat{\mathbf{w}} = \underset{\mathbf{w}\in\mathbb{R}^{dm_n}}{\operatorname{argmin}}\, \mathbf{w}^T\widehat{\mathbf{C}}\mathbf{w}, \quad \text{subject to } \left\|\mathbf{z}^* + \widehat{\mathbf{C}}\mathbf{w}\right\|_\infty \le \lambda', \tag{3.4}$$

where λ' is an additional tuning parameter. To establish the theoretical results, we denote the matrix $L_{2,1}$ and $L_{2,\infty}$ norms for a vector $\mathbf{v}\in\mathbb{R}^{dm_n}$ as

$$\|\mathbf{v}\|_{2,1} = \sum_{j=1}^{d}\sqrt{\sum_{k=1}^{m_n} v_{(j-1)m_n+k}^2}, \quad \|\mathbf{v}\|_{2,\infty} = \max_{1\le j\le d}\sqrt{\sum_{k=1}^{m_n} v_{(j-1)m_n+k}^2}. \tag{3.5}$$

The following conditions are assumed.

Assumption 3.1 (Gaussian Errors)**.** The random variables $\varepsilon_1, ..., \varepsilon_n$ are i.i.d $N(0, \sigma^2)$.

Assumption 3.2 (Indentifiability)**.** $\mathbb{E}[f_j(X_j)] = 0$ and $f_j \in \mathcal{F}$, $j = 1, ..., d$.

Assumption 3.3 (Bounded Covariates)**.** For $j = 1, ..., d$, the covariate vector X_j has the support set $[a, b]$ and there exist constants C_1 and C_2 such that the probability density function g_j of X_j satisfies $0 < C_1 \le g_j(x) \le C_2$, for $x \in [a, b]$.

Assumption 3.4 (Restricted Eigenvalue Conditions)**.** There exists a constant τ such that

$$\min\left\{\frac{m_n(\mathbf{v}^T\widehat{\mathbf{C}}\mathbf{v})}{\|\mathbf{v}\|_2^2} : \mathbf{v}\in\mathbb{R}^{dm_n}\backslash\{0\}, \|\mathbf{v}_{\mathcal{S}^c}\|_{2,1} \le 3\|\mathbf{v}_{\mathcal{S}}\|_{2,1}\right\} \ge \tau, \tag{3.6}$$

where $\mathcal{S} = \{1 \le j \le d : f_j \ne 0\}$ is the support set and $s^* = |\mathcal{S}|$.

Assumptions 3.1–3.3 are standard for studying the theoretical results for nonparametric regressions; see Huang et al. (2010). To obtain valid prediction intervals, in Assumption 3.1, we impose the Gaussian assumption for the model error. For simplicity, we make the restricted eigenvalue condition explicitly in Assumption 3.4. Since the Gram matrix $\widehat{C}$ is constructed by the spline basis, the magnitude of its eigenvalues shrinks to 0 as $m_n \to \infty$. Hence, we rescale it by m_n in (3.6). Note that Bickel et al. (2009) showed that the type of Assumption 3.4 is implied by the sparse eigenvalue condition. When s^* is bounded, by Lemma 3 of Huang et al. (2010), the sparse eigenvalue condition for the additive model holds with probability tending to 1 and therefore Assumption 3.4 holds. A rigorous verification can be found in Proposition 4.1 of Lu et al. (2015).

Theorem 3.5. Under Assumptions 3.1, 3.2, 3.3, 3.4, then

$$\sqrt{\frac{n}{m_n}}\left(S_{\mathbf{x}}(\widehat{\boldsymbol{\beta}}) - \mathbf{z}^{*\top}\boldsymbol{\beta}^*\right) = N + \Delta_1 + \Delta_2, \tag{3.7}$$

where $N \mid \mathbf{X}, \mathbf{x}^* \sim N(0, m_n^{-1}\sigma^2 \widehat{\mathbf{w}}^T \widehat{\mathbf{C}} \widehat{\mathbf{w}})$. If $\|\mathbf{w}^*\|_1^2 \sqrt{\log(dm_n)/n} = \mathcal{O}_{\mathbb{P}}(m_n)$ and

$$\lambda = 4\sqrt{2\bar{C}C''^{-1} m_n \frac{\log d}{n}} + 4s^* C_\phi m_n^{1/2-k},$$

then $\widehat{\mathbf{w}}^T \widehat{\mathbf{C}} \widehat{\mathbf{w}} = \mathcal{O}_{\mathbb{P}}(m_n)$ and Δ_1 and Δ_2 satisfy

$$|\Delta_1| \leq 6s^* \tau^{-1} m_n \{4(2C''^{-1} m_n \log d)^{1/2} + 4s^* C_\phi n^{1/2} m_n^{1/2-k}\} \lambda'$$

with probability at least $1 - 2d^{-1}$ and

$$|\Delta_2| = \mathcal{O}_{\mathbb{P}}(s^* n^{1/2} m_n^{-k}),$$

where C'' a universal constant given in Lemma C.3, and C_ϕ is a constant satisfying $\sup_{x \in [a,b]} |f_j(x) - \Phi(x)^T \boldsymbol{\beta}_j^*| \leq C_\phi m_n^{-k}$.

Proof. See Appendix B for a detailed proof. □

Theorem 3.5 is analogous to Lemma 2.1 for linear models. Note that the convergence rate of $S_{\mathbf{x}}(\widehat{\boldsymbol{\beta}})$ in this theorem is slower than that in Lemma 2.1, due to the approximation error by using the sieve spaces. To conclude that the estimator $S_{\mathbf{x}}(\widehat{\boldsymbol{\beta}})$ is asymptotically normal, we need to further control $\boldsymbol{\Delta}_1$. The following lemma serves this purpose.

Lemma 3.6. Assume there exist two constants $C_{\min}$ and $C_{\max}$, such that $0 < m_n^{-1} C_{\min} \leq \lambda_{\min}(\mathbf{C}) \leq \lambda_{\max}(\mathbf{C}) \leq m_n^{-1} C_{\max} < \infty$, where $\mathbf{C} = \mathbb{E}(\widetilde{\mathbf{Z}}^T \widetilde{\mathbf{Z}}/n)$. In addition, assume that $\|\mathbf{C}^{-1/2} \boldsymbol{Z}_i\|_{\psi_2} = C$, for any $i = 1, ..., n$, and $\|\mathbf{z}^*\|_2 = 1$. Then, with probability at least $1 - 2(dm_n)^{-3}$,

$$\left\|\mathbf{z}^* + \widehat{\mathbf{C}}\mathbf{w}^*\right\|_\infty \leq 2(1 + 2\rho^{1/2} C^2) \sqrt{\frac{C''^{-1} \log(m_n d)}{n}}, \text{ where } \rho = C_{\max}/C_{\min},$$

and $\mathbf{w}^* = -\mathbf{C}^{-1}\mathbf{z}^*$, provided $4C''^{-1} \log(dm_n) \leq n$, and C'' is given in Lemma C.3.

Proof. See Appendix B for a detailed proof. □

This Lemma implies that one can take

$$\lambda' = 2(1 + 2\rho^{1/2} C^2) \sqrt{C''^{-1} \frac{\log(m_n d)}{n}}$$

in (3.4), such that $|\Delta_1| = \mathcal{O}_{\mathbb{P}}\big(s^* \sqrt{\log(m_n d)/n} \{m_n^{3/2} \log d + s^* n^{1/2} m_n^{3/2-k}\}\big)$.

In what follows, we discuss the order of m_n which makes both Δ_1 and Δ_2 converge to 0. Notice that the choice of m_n must balance the rate of Δ_1 and Δ_2 as $\Delta_1 \to \infty$ and $\Delta_2 \to 0$ as $m_n \to \infty$ when holding other parameters fixed. Suppose $k > 3/2$ and s^* is fixed. Then, $n^{1/(2k)} \lesssim m_n$ would make Δ_2 converge to 0 in probability. To make Δ_1 converge to 0 in probability, we need $m_n \log d \lesssim n^{1/3}$. Hence, ignoring the logarithmic factor of d, it is required to set $n^{1/(2k)} \lesssim m_n \lesssim n^{1/3}$ to make both Δ_1 and $\Delta_2 \to 0$. By the bias and variance trade-off in Lemma C.1, the optimal number of basis is given by $m_n \asymp n^{1/(2k+1)}$. That is, to perform the inference, we need to *undersmooth* to eliminate the bias.

Similar to the variance estimator (2.3), σ^2 in the additive model can be estimated by

$$\widehat{\sigma}^2 = n^{-1}\sum_{i=1}^{n}(Y_i - \boldsymbol{Z}_i^T\widehat{\boldsymbol{\beta}})^2.$$

The following Lemma establishes the consistency of $\widehat{\sigma}^2$.

Lemma 3.7 (Convergence of $\widehat{\sigma}$)**.** Under the conditions in Theorem 3.5, we have $|\widehat{\sigma}^2 - \sigma^2| = o_{\mathbb{P}}(1)$.

Proof. See Appendix B for a detailed proof. □

For practical applications, we suggest to use the estimator

$$\widehat{\sigma}^2_{\widehat{\lambda}} = n^{-1}\sum_{i=1}^{n}(Y_i - \boldsymbol{Z}_i^T\widehat{\boldsymbol{\beta}}_{\widehat{\lambda}})^2, \tag{3.8}$$

where $\widehat{\lambda}$ is given by the K-fold cross-validation. A direct application of Lemma 3.7 implies that the prediction interval with asymptotic coverage probability $(1-\alpha)$ for Y corresponding to $\mathbf{x}^*$ is given by $[S_{\mathbf{x}}(\widehat{\boldsymbol{\beta}}) - \Phi^{-1}(1-\alpha/2)V, S_{\mathbf{x}}(\widehat{\boldsymbol{\beta}}) + \Phi^{-1}(1-\alpha/2)V]$, where

$$V = \widehat{\sigma} + n^{-1/2}(\widehat{\sigma}^2\widehat{\mathbf{w}}^T\widehat{\mathbf{C}}\widehat{\mathbf{w}})^{1/2}.$$

For reader's convenience, the procedure to construct the prediction interval for sparse additive models is summarized in Algorithm 2. Finally, we comment that if we set $\mathbf{x}^* = (x_1^*, 0, ..., 0)$ and choose $n^{1/(2k)} \lesssim m_n \lesssim n^{1/3}$, then the prediction interval in Algorithm 2 reduces to the confidence interval for the first component of function $f_1(x)$ at $x = x_1^*$. By Theorem 3.5 and Lemmas 3.6, 3.7, the confidence interval for $f_1(x_1^*)$ has the correct coverage probability asymptotically.

Algorithm 2 Calculate the prediction intervals for sparse additive models

Require: : Loss function $\ell(\boldsymbol{\beta}) = (2n)^{-1}\|\mathbf{Y} - \mathbf{Z}\boldsymbol{\beta}\|_2^2$, where $\mathbf{Z} = (\boldsymbol{Z}_1, ...\boldsymbol{Z}_n)^T$ and $\boldsymbol{Z}_i$ is defined in (3.2). Tuning parameters λ and λ', covariate $\mathbf{x}^*$ and $\mathbf{z}^* = (\boldsymbol{\Phi}(x_1^*)^T, ..., \boldsymbol{\Phi}(x_d^*)^T)^T$, significance level $0 < \alpha < 1$.

(i) Calculate $\widehat{\boldsymbol{\beta}}$ as $\widehat{\boldsymbol{\beta}} = \text{argmin}_{\boldsymbol{\beta}}\left\{\ell(\boldsymbol{\beta}) + \lambda\sum_{j=1}^{d}\|\boldsymbol{\beta}_j\|_2\right\}$.

(ii) Estimate $\widehat{\mathbf{w}}$ by

$$\widehat{\mathbf{w}} = \underset{\mathbf{w}\in\mathbb{R}^{dm_n}}{\text{argmin}}\,\mathbf{w}^T\widehat{\mathbf{C}}\mathbf{w}, \quad \text{subject to } \left\|\mathbf{z}^* + \widehat{\mathbf{C}}\mathbf{w}\right\|_\infty \le \lambda'.$$

(iii) Calculate the estimator

$$S_{\mathbf{x}}(\widehat{\boldsymbol{\beta}}) = \mathbf{z}^{*\top}\widehat{\boldsymbol{\beta}} - \frac{1}{n}\sum_{i=1}^{n}\widehat{\mathbf{w}}^T\boldsymbol{Z}_i(Y_i - \boldsymbol{Z}_i^T\widehat{\boldsymbol{\beta}}), \quad \text{and} \quad \widehat{\sigma}^2 = n^{-1}\sum_{i=1}^{n}(Y_i - \boldsymbol{Z}_i^T\widehat{\boldsymbol{\beta}})^2.$$

return Prediction intervals $[S_{\mathbf{x}}(\widehat{\boldsymbol{\beta}}) - \Phi^{-1}(1-\alpha/2)V, S_{\mathbf{x}}(\widehat{\boldsymbol{\beta}}) + \Phi^{-1}(1-\alpha/2)V]$, where $V = \widehat{\sigma} + n^{-1/2}(\widehat{\sigma}^2\widehat{\mathbf{w}}^T\widehat{\mathbf{C}}\widehat{\mathbf{w}})^{1/2}$.

4 Simulation Results

4.1 Linear regression model

A simulation study is conducted to examine the finite sample performance of the proposed prediction interval. We first consider the coverage ratio of the mean parameter,

$$\mathbf{x}^{*\top}\boldsymbol{\beta}^* \in \left[S_{\mathbf{x}}(\widehat{\boldsymbol{\beta}}) - \Phi^{-1}(1-\alpha/2)\widehat{\sigma}\sqrt{\widehat{\mathbf{w}}^T\widehat{\boldsymbol{\Sigma}}\widehat{\mathbf{w}}/n},\ S_{\mathbf{x}}(\widehat{\boldsymbol{\beta}}) + \Phi^{-1}(1-\alpha/2)\widehat{\sigma}\sqrt{\widehat{\mathbf{w}}^T\widehat{\boldsymbol{\Sigma}}\widehat{\mathbf{w}}/n}\right].$$

We fix the significance level at $\alpha = 0.05$, and consider sample size $n = 100$ and number of covariates $d = 200$. The data are generated from model (2.1). In particular, the error ε_i follows i.i.d. $N(0,1)$ distribution. The covariates $\boldsymbol{X}_i$ are generated from a multivariate Gaussian distribution $N(0, \boldsymbol{\Sigma})$ where $\boldsymbol{\Sigma}_{ij} = 0.6^{|i-j|}$. We assume in the model that the intercept is 0. Furthermore, we set $\beta_j^* = \mu > 0$, for $j = 1, 2, 3$ and $\beta_j^* = 0$ for $j = 4, ..., d$. Hence, $s^* = \|\boldsymbol{\beta}^*\|_0 = 3$. The future covariates $\mathbf{x}^*$ is drawn randomly from $N(0, \boldsymbol{\Sigma})$ and is fixed in the simulation.

We consider the following three methods to estimate $\boldsymbol{\beta}$, namely, Lasso, SCAD and post-Lasso. In the Lasso procedure, we estimate $\widehat{\boldsymbol{\beta}}$ with tuning parameter λ chosen from cross-validation with `cv.glmnet` command in `R` package `glmnet`. For SCAD, we use the `R` package `ncvreg`. Similarly, the tuning parameter is chosen by cross-validation. For post-Lasso, we first fit Lasso with tuning parameter chosen by cross-validation to find the estimator $\widehat{\boldsymbol{\beta}}$ and its support $\widehat{\mathcal{S}} = \text{supp}(\widehat{\boldsymbol{\beta}})$, and then perform ordinary least square estimation on covariates $\boldsymbol{X}_{i,\widehat{\mathcal{S}}}$ restricted to the estimated support. The resulting estimator is called the post-Lasso estimator $\widehat{\boldsymbol{\beta}}_{post}$ (Belloni and Chernozhukov, 2013). In the simulation studies, $S_{\mathbf{x}}(\widehat{\boldsymbol{\beta}}_{\text{Lasso}})$ and $S_{\mathbf{x}}(\widehat{\boldsymbol{\beta}}_{\text{SCAD}})$ are constructed with Lasso and SCAD estimators as in Algorithm 1. For post-Lasso, we consider two possibilities to construct confidence intervals for the mean parameter based on the post-Lasso estimator $\widehat{\boldsymbol{\beta}}_{post}$. The first approach (Post-Lasso 1) is to estimate $\mathbf{x}^{*\top}\boldsymbol{\beta}^*$ by $S_{\mathbf{x}}(\widehat{\boldsymbol{\beta}}_{post}) = \mathbf{x}^{*\top}\widehat{\boldsymbol{\beta}}_{post} - n^{-1}\widehat{\mathbf{w}}^T\sum_{i=1}^n \boldsymbol{X}_i(Y_i - \boldsymbol{X}_i^T\widehat{\boldsymbol{\beta}}_{post})$, where $\widehat{\mathbf{w}}$ is defined in (2.2). The confidence interval is constructed by the same way as that for the Lasso estimator. In the second approach (Post-Lasso 2), we directly estimate the mean parameter by $\mathbf{x}^{*\top}\widehat{\boldsymbol{\beta}}_{post}$, and the confidence interval is constructed based on the ordinary least square regression (Faraway, 2014).

We can select the tuning parameter based on

$$\lambda' = C\lambda_{\max}(\boldsymbol{\Sigma})\sqrt{\frac{\log d}{n}},$$

where C is a constant. We choose C such that λ' satisfies $0.8\|\mathbf{x}^*\|_\infty \le \lambda' \le 0.9\|\mathbf{x}^*\|_\infty$. Hence, for Lasso and Post-Lasso we set $C = 0.045, 0.0475$; for SCAD we set $C = 0.25, 0.26$. To perform the linear programing required for computing $\widehat{\mathbf{w}}$, we use `fastlp` in `R` package `fastclime`. The number of simulation iterations is 500.

Table 1 shows the averaged coverage probability and length of confidence intervals based on the Lasso, SCAD, Post-Lasso 1 and Post-Lasso 2. We find that the Lasso, SCAD, and Post-Lasso 1 all have similar results in terms of coverage probability and length of confidence intervals. This shows that our method for constructing confidence intervals for mean parameters is quite robust

to the initial estimator $\widehat{\boldsymbol{\beta}}$. In addition, when the signal strength μ is weak, say $\mu = 0.2, 0.5$, the naive approach (Post-Lasso 2) by ignoring the uncertainty in the model selection leads to incorrect coverage probability. We also note that the length of confidence intervals for Post-Lasso 2 is much wider than the other methods for $\mu = 0.2$. This is due to the fact that in such weak signal situation the recovered support is often empty, so that the length of the confidence interval is totally determined by the estimated model errors.

Table 1: The averaged coverage probability and length (in parenthesis) of confidence intervals for $\mathbf{x}^{*\top}\boldsymbol{\beta}^*$ under linear regression models. Nominal coverage $1 - \alpha = 0.95$.

Method	C	$\mu = 0.2$	$\mu = 0.5$	$\mu = 1$	$\mu = 2$	$\mu = 3$	$\mu = 4$	$\mu = 5$
Lasso	0.045	0.992	0.992	0.992	0.996	0.996	1.000	1.000
		(0.162)	(0.135)	(0.135)	(0.136)	(0.145)	(0.159)	(0.178)
	0.0475	0.962	0.940	0.938	0.950	0.958	0.966	0.976
		(0.085)	(0.085)	(0.085)	(0.086)	(0.092)	(0.101)	(0.113)
SCAD	0.0625	0.994	0.986	0.998	0.992	0.994	0.994	0.994
		(0.134)	(0.144)	(0.128)	(0.132)	(0.133)	(0.133)	(0.134)
	0.065	0.968	0.936	0.970	0.984	0.986	0.986	0.986
		(0.097)	(0.091)	(0.093)	(0.096)	(0.096)	(0.097)	(0.097)
Post-Lasso 1	0.045	0.984	0.984	0.988	0.990	0.990	0.992	0.998
		(0.156)	(0.135)	(0.135)	(0.136)	(0.145)	(0.159)	(0.178)
	0.0475	0.952	0.914	0.946	0.950	0.962	1.000	0.978
		(0.085)	(0.085)	(0.085)	(0.086)	(0.092)	(0.101)	(0.113)
Post-Lasso 2	-	0.830	0.878	0.920	0.920	0.920	0.954	0.950
		(2.148)	(0.049)	(0.075)	(0.072)	(0.054)	(0.086)	(0.077)

We also consider the coverage ratio of the prediction intervals

$$Y^* \in \left[S_{\mathbf{x}}(\widehat{\boldsymbol{\beta}}) - \Phi^{-1}(1-\alpha/2)\widehat{\sigma}(1+\sqrt{\widehat{\mathbf{w}}^T\widehat{\boldsymbol{\Sigma}}\widehat{\mathbf{w}}/n}),\ S_{\mathbf{x}}(\widehat{\boldsymbol{\beta}}) + \Phi^{-1}(1-\alpha/2)\widehat{\sigma}(1+\sqrt{\widehat{\mathbf{w}}^T\widehat{\boldsymbol{\Sigma}}\widehat{\mathbf{w}}/n}) \right],$$

where $Y^* = \mathbf{x}^{*\top}\boldsymbol{\beta}^* + \varepsilon^*$, where ε^* is independent of $\{\boldsymbol{X}_i\}_{i=1}^n$ and i.i.d. with $\{\varepsilon_i\}_{i=1}^n$. The simulation setting is exactly the same as that for Table 1.

Table 2 shows the averaged coverage probability and length of prediction intervals based on the Lasso, SCAD and Post-Lasso 1. A general tendency is that the length of prediction intervals for Y^* is much larger than that for $\mathbf{x}^{*\top}\boldsymbol{\beta}^*$ shown in Tables 1. Note that when $\alpha = 0.05$, $\widehat{\sigma} \approx 1$ if the estimator is accurate. Thus, it is usual that the size of the prediction interval is greater than 4. For all procedures, when μ increases, the length of the prediction interval are also getting slightly larger. In particular, the prediction intervals constructed by the SCAD estimator is wider than the prediction intervals constructed by the other methods, when μ is large. This results from large variance of model error $\widehat{\sigma}$ estimated by using the SCAD estimator. Indeed, for $\mu = 5$, the median of the estimated variance of model error using the SCAD estimator is 1.947, while using the Lasso estimator this value is 1.159.

Table 2: The averaged coverage probability and length (in parenthesis) of prediction intervals for Y^* under linear regression models. Nominal coverage $1-\alpha = 0.95$.

Method	C	$\mu = 0.2$	$\mu = 0.5$	$\mu = 1$	$\mu = 2$	$\mu = 3$	$\mu = 4$	$\mu = 5$
Lasso	0.045	0.962	0.964	0.964	0.966	0.974	0.978	0.980
		(4.199)	(4.210)	(4.206)	(4.235)	(4.439)	(4.818)	(5.334)
	0.0475	0.962	0.964	0.964	0.966	0.974	0.978	0.980
		(4.194)	(4.208)	(4.204)	(4.232)	(4.436)	(4.815)	(5.331)
SCAD	0.0625	0.960	0.958	0.972	0.970	0.976	0.976	0.976
		(4.124)	(4.198)	(4.431)	(4.850)	(5.667)	(6.622)	(7.627)
	0.065	0.960	0.958	0.972	0.970	0.976	0.976	0.976
		(4.123)	(4.197)	(4.430)	(4.849)	(5.665)	(6.620)	(7.625)
Post-Lasso 1	0.045	0.962	0.962	0.962	0.964	0.972	0.978	0.980
		(4.199)	(4.210)	(4.206)	(4.235)	(4.439)	(4.818)	(5.334)
	0.0475	0.964	0.962	0.962	0.964	0.972	0.978	0.980
		(4.194)	(4.208)	(4.204)	(4.232)	(4.436)	(4.815)	(5.331)

4.2 Sparse additive model

We consider the following simulation scenarios for sparse additive models. The data are generated from the model: $Y_i = f(\boldsymbol{X}_i) + \varepsilon_i$, where $f(\boldsymbol{X}_i) = \sum_{j=1}^{s} f_j(X_{ij})$. The following choices of functions are adopted,

$$
\begin{aligned}
&f_1(t) = 5t; f_2(t) = 3(2t-1)^2;\\
&f_3(t) = 4\sin(2\pi t)/(2-\sin(2\pi t)) + 6\big\{0.1\sin(2\pi t) + 0.2\cos(2\pi t) + 0.3\sin(2\pi t)^2\\
&\qquad + 0.4\cos(2\pi t)^3 + 0.5\sin(2\pi t)^3\big\},
\end{aligned}
$$

with $t \in [0,1]$. The covariate $\boldsymbol{X}_i$ is generated by first simulating $\widetilde{\boldsymbol{X}}_i$ from a multivariate Gaussian distribution $N(0, \boldsymbol{\Sigma})$ where $\boldsymbol{\Sigma}_{ij} = 0.6^{|i-j|}$, and then transforming back to interval $[0,1]$ by $\boldsymbol{X}_{ij} = \Phi(\widetilde{\boldsymbol{X}}_{ij})$ for each $j = 1, .., d$. Therefore, $\boldsymbol{X}_{ij}$ is uniformly distributed and has compact support. The sample size is $n = 100$ and $d = 50$. The B-spline basis is applied and the number of spline basis m_n is $m_n = 6$. The functions are estimated by the group Lasso implemented in `R` package `grpreg` with $\widehat{\lambda}$ chosen by K-fold cross-validation, where $K = 10$. Again, $\widehat{\mathbf{w}}$ is estimated with `fastlp` in `R` package `fastclime`, with λ' chosen by

$$
\lambda' = C\frac{\lambda_{\max}(\widehat{\boldsymbol{C}})}{\lambda_{\min}(\widehat{\boldsymbol{C}})}\sqrt{\frac{\log(6d)}{n}},
$$

where C is a constant. In fact, using Lemma 6.2 of Zhou et al. (1998), equidistant grid and the fact that the marginal distributions of $\boldsymbol{X}_i$ are uniform, we can estimate $\lambda_{\max}(\widehat{\boldsymbol{C}})/\lambda_{\min}(\widehat{\boldsymbol{C}}) = q/c_0^2$, where q is the degree of the spline function and c_0 is an absolute constant. Therefore, we have $\lambda' = C'\sqrt{\log(6d)/n}$, and we choose $C' = 1.225, 1.25$ and 1.3 The future covariates $\mathbf{x}^*$ is drawn randomly from $N(0, \boldsymbol{\Sigma})$ and is fixed in the simulation.

Table 3 shows that the coverage probability and length of confidence intervals for $\mathbf{z}^{*\top}\boldsymbol{\beta}^*$ under the sparse additive models. We find that the confidence intervals are accurate under a variety of scenarios from $f = f_1$ to $f = \sum_{i=1}^{3} f_i$. Similar patterns are observed in Table 4 for prediction intervals.

Table 3: The averaged coverage probability and length (in parenthesis) of confidence intervals for $\mathbf{z}^{*\top}\boldsymbol{\beta}^*$ under sparse additive models. Nominal coverage $1-\alpha = 0.95$.

	$f = f_1$	$f = \sum_{i=1}^{2} f_i$	$f = \sum_{i=1}^{3} f_i$
$C = 1.225$	0.980	0.986	0.982
	(11.662)	(11.109)	(14.041)
$C = 1.25$	0.978	0.988	0.960
	(11.504)	(11.077)	(13.793)
$C = 1.3$	0.974	0.982	0.962
	(10.913)	(10.546)	(13.182)

Table 4: The averaged coverage probability and length (in parenthesis) of prediction intervals for Y^* under sparse additive models. Nominal coverage $1-\alpha = 0.95$. $d = 50$.

	$f = f_1$	$f = \sum_{i=1}^{2} f_i$	$f = \sum_{i=1}^{3} f_i$
$C = 1.225$	0.964	0.978	0.966
	(12.207)	(11.649)	(14.725)
$C = 1.25$	0.964	0.974	0.956
	(12.065)	(11.623)	(14.472)
$C = 1.3$	0.956	0.956	0.938
	(11.493)	(11.123)	(13.899)

5 Real Data Analysis

In this section, we apply our method to two real datasets. In Section 5.1, we construct the prediction interval for the median value of owner-occupied homes with housing-specific variables in the Boston Housing data. In Section 5.2, we illustrate our method to whole-genome regression (WGR) with data containing BMI and genome markers from mice.

5.1 Boston Housing Data

In this section we apply our method to Boston Housing data, which is available in `R` package `MASS`. The data consist of 14 variables with sample size 506. We use the median value of owner-occupied homes as the response variable Y_i and the remaining 13 variables as covariates. To demonstrate

the applicability of our method to high dimensional data, we simulate additional variable $\boldsymbol{X}_i$ of dimension $d-13$, where $d=400$ in the first scenario and $d=600$ in the second scenario. In particular, the simulated covariate follows a joint Gaussian distribution $N(0,\boldsymbol{\Sigma})$, where $\boldsymbol{\Sigma}$ is a $(d-13)\times(d-13)$ matrix with $\boldsymbol{\Sigma}_{ij}=0.6^{|i-j|}$. To examine the validity of prediction, the data set is divided into two parts: the training data with sample size n and testing data with size $506-n$, where $n=100, 200$ and $n=300$. The training data are used to derive the estimators $\widehat{\boldsymbol{\beta}}$ under the linear model assumption, where $\widehat{\boldsymbol{\beta}}$ is given by the Lasso estimator with tuning parameter λ chosen by cross-validation. For each sample $(\widetilde{Y}_i, \widetilde{\boldsymbol{X}}_i)$ in the testing data set, we predict the response given the covariate $\mathbf{x}^*=\widetilde{\boldsymbol{X}}_i$. The prediction interval as described in Algorithm 1 is constructed. We examine whether it covers the true testing sample $\widetilde{Y}_i$. The selection of tuning parameter λ' is similar to that of the simulation study in Section 4.1.

Table 5 shows the coverage probabilities and length of prediction intervals. It is seen that the coverage probabilities are quite robust to the choice of C for a variety of combinations of (n,d).

Table 5: The 95% coverage probability and length (in parenthesis) of the prediction intervals for Boston Housing data. Nominal coverage $1-\alpha=0.95$.

	$\boldsymbol{d=400}$			$\boldsymbol{d=600}$		
	$n=100$	$n=200$	$n=300$	$n=100$	$n=200$	$n=300$
$C=0.00005$	0.9655	1.0000	1.0000	0.9655	1.0000	1.0000
	(47.472)	(51.394)	(63.573)	(47.346)	(53.713)	(62.757)
$C=0.0001$	0.9729	1.0000	1.0000	0.9754	0.9967	1.0000
	(43.352)	(53.671)	(55.754)	(43.362)	(53.475)	(55.582)
$C=0.0005$	0.9286	0.9542	0.9515	0.9286	0.9412	0.9515
	(32.660)	(40.236)	(37.938)	(32.532)	(37.431)	(37.812)
$C=0.005$	0.9409	0.9412	0.9563	0.9384	0.9412	0.9563
	(32.465)	(33.015)	(38.028)	(32.407)	(32.827)	(37.670)

5.2 Prediction with Whole-Genome Regression

Prediction based on genetic markers is important for many applications, such as animal and plant breeding. Such prediction is usually carried out by the whole-genome regression (WGR) proposed by Meuwissen et al. (2001), where the phenotype as a response variable is regressed on thousands of genetic markers. Please see de los Campos et al. (2013) for a detailed review on this approach.

In this section, we apply our method to a dataset with genotype and phenotype for mice from the Wellcome Trust (http://gscan.well.ox.ac.uk). The dataset, built in `R`-package `BGLR`, consists of genotypes and phenotypes of 1814 mice. Each mouse was genotyped at 10346 single nucleotide polymorphisms (SNPs). For more detailed description and the pre-processing of the dataset, please see Pérez and de los Campos (2014). We use the BMI index in the variable `Obesity.BMI` as the response variable, and the SNPs in `mice.X` as the independent variables. To reduce the dimensionality, we first apply the sure screening method in Fan and Lv (2008), which is implemented by the

function `SIS` in `R`-package `SIS`. The results suggest that 146 SNPs are relevant. To demonstrate the performance of our method under high dimensionality, we randomly add in SNPs other than these 146 to make a total number of d SNPs as independent variables. We take $d = 1000$ and 2000. Similar to the Boston Housing data example, we select n training samples and treat the remaining $1814 - n$ samples as validation samples. The size of the training data is chosen as $n = 200, 500$ and 1000.

As in the Boston Housing data example, the training data are used to derive the Lasso estimator $\widehat{\boldsymbol{\beta}}$ under the linear model assumption with tuning parameter λ chosen by the cross-validation. For each sample $(\widetilde{Y}_i, \widetilde{\boldsymbol{X}}_i)$ in the testing dataset, we set $\mathbf{x}^* = \widetilde{\boldsymbol{X}}_i$ and consider the its prediction interval. The selection of tuning parameter λ' is similar to that of the simulation study in Section 4.1.

Table 6 summarizes the results of coverage probability and length of the prediction intervals under various situations. For $d = 1000$, similar to the simulation study, smaller C gives wider prediction interval and larger coverage. Moreover, the coverage and length of prediction intervals increases when the sample size of training data increases. The same phenomenon can be found for the case of $d = 2000$, while maintaining comparable coverage probabilities as those of $d = 1000$ requires wider prediction interval. In conclusion, our prediction intervals have reasonable accuracy and can be used to predict the BMI index based on the high dimensional genomic information.

Table 6: The 95% coverage probability and length (in parenthesis) of the prediction intervals for the mice gene data. Nominal coverage $1 - \alpha = 0.95$.

	$\boldsymbol{d = 1000}$				$\boldsymbol{d = 2000}$		
	$n = 200$	$n = 500$	$n = 1000$		$n = 200$	$n = 500$	$n = 1000$
$C = 0.0008$	0.998	0.998	0.998	$C = 0.004$	0.998	0.998	0.988
	(0.971)	(1.897)	(3.134)		(1.038)	(2.789)	(3.859)
$C = 0.0010$	0.994	1.000	1.000	$C = 0.005$	0.988	0.998	0.996
	(0.480)	(1.313)	(2.280)		(0.478)	(1.505)	(2.530)
$C = 0.0012$	0.926	0.998	0.998	$C = 0.006$	0.912	1.000	1.000
	(0.269)	(0.929)	(1.805)		(0.281)	(0.971)	(1.950)

6 Discussion

In this paper, we propose a general approach to construct confidence intervals for mean parameters and prediction intervals for future observations in high dimensional linear models and sparse additive models. From the practical perspective, the method can be easily implemented by using the existing software packages and our numerical results suggest that the proposed method outperforms the confidence and prediction intervals by using the naive Post-Lasso method (Post-Lasso 2). From the theoretical perspective, we provide theoretical guarantees for our method under very mild conditions for both linear models and sparse additive models.

Similar to the existing inference approaches for β_j (Zhang and Zhang, 2014; van de Geer et al.,

2014; Javanmard and Montanari, 2013), our method involves two tuning parameters, and their choice may influence the results of inference. One future research direction is to develop an objective procedure to determine tuning parameters and meanwhile the asymptotic coverage of confidence and prediction intervals is preserved.

This paper focuses on the post-regularization prediction problems for the linear models and additive models. It is of interest to further extend our methods to more complex regression models, such as generalized linear models and generalized additive models.

Acknowledgement

We thank all participants in the "Inference in high dimensional regression" workshop sponsored by American Institute of Mathematics. We especially thank Professor Richard Samworth, Cun-Hui Zhang and Rina Foygel Barber for their helpful comments. This research is partially supported by the grants NSF IIS1408910, NSF IIS1332109, NIH R01MH102339, NIH R01GM083084, and NIH R01HG06841. Shih-Kang Chao is partially supported by Einstein Foundation Berlin via Berlin Doctoral Program in Economics and Management Science (BDPEMS) and Collaborative Research Center 649.

Appendix

A Proofs for Sparse High Dimensional Linear Model

A.1 Proof of Lemma 2.1

Proof of Lemma 2.1. By definition, we can show that

$$\begin{aligned} S(\widehat{\boldsymbol{\beta}}) - \mathbf{x}^{*T}\boldsymbol{\beta}^* &= \mathbf{x}^{*T}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*) - \widehat{\mathbf{w}}^T \frac{1}{n}\sum_{i=1}^n \boldsymbol{X}_i(Y_i - \boldsymbol{X}_i^T\widehat{\boldsymbol{\beta}}) \\ &= \underbrace{(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)^T \left[\mathbf{x}^* + \frac{1}{n}\sum_{i=1}^n \boldsymbol{X}_i\boldsymbol{X}_i^T\widehat{\mathbf{w}}\right]}_{E_1} - \underbrace{\widehat{\mathbf{w}}^T\left[\frac{1}{n}\sum_{i=1}^n \boldsymbol{X}_i\varepsilon_i\right]}_{E_2}. \end{aligned}$$

By Hölder's inequality,

$$|\sqrt{n}E_1| \le \sqrt{n}\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_1 \underbrace{\|\mathbf{x}^* + \widehat{\boldsymbol{\Sigma}}\widehat{\mathbf{w}}\|_\infty}_{\le \lambda'} = \mathcal{O}_{\mathbb{P}}(s^*\lambda'\sqrt{\log d/n}).$$

For E_2, it is seen that $n^{1/2}E_2 = N$, and $N \mid \mathbf{X} \sim N(0, \sigma^2\widehat{\mathbf{w}}^T\widehat{\boldsymbol{\Sigma}}\widehat{\mathbf{w}})$. this completes the proof. □

A.2 Proof of Lemma 2.2

Proof of Lemma 2.2. Since $\widehat{\boldsymbol{\Sigma}} = n^{-1}\sum_{i=1}^n \boldsymbol{X}_i\boldsymbol{X}_i^T$, we can show that

$$\mathbf{x}^* + \widehat{\boldsymbol{\Sigma}}\mathbf{w}^* = -\frac{1}{n}\sum_{i=1}^n(\boldsymbol{X}_i\boldsymbol{X}_i^T\boldsymbol{\Sigma}^{-1}\mathbf{x}^* - \mathbf{x}^*) = -\frac{1}{n}\sum_{i=1}^n(\boldsymbol{\Sigma}^{1/2}\boldsymbol{U}_i\boldsymbol{U}_i^T\boldsymbol{\Sigma}^{-1/2}\mathbf{x}^* - \mathbf{x}^*),$$

where $\boldsymbol{U}_i = \boldsymbol{\Sigma}^{-1/2}\boldsymbol{X}_i$. Now, we consider the jth component of $\boldsymbol{T}_i = \boldsymbol{\Sigma}^{1/2}\boldsymbol{U}_i\boldsymbol{U}_i^T\boldsymbol{\Sigma}^{-1/2}\mathbf{x}^* - \mathbf{x}^*$. For $\boldsymbol{\Sigma}_{j*}^{1/2}\boldsymbol{U}_i\boldsymbol{U}_i^T\boldsymbol{\Sigma}^{-1/2}\mathbf{x}^*$ with $1 \le j \le d$, we have by Lemma C.2,

$$\|\boldsymbol{\Sigma}_{j*}^{1/2}\boldsymbol{U}_i\boldsymbol{U}_i^T\boldsymbol{\Sigma}^{-1/2}\mathbf{x}^*\|_{\psi_1} \le 2\underbrace{\|\boldsymbol{\Sigma}_{j*}^{1/2}\boldsymbol{U}_i\|_{\psi_2}}_{E_1}\underbrace{\|\boldsymbol{U}_i^T\boldsymbol{\Sigma}^{-1/2}\mathbf{x}^*\|_{\psi_2}}_{E_2}.$$

For E_1, by the definition of ψ_2 norm, we can show that $E_1 \le \|\boldsymbol{\Sigma}_{j*}^{1/2}\|_2\|\boldsymbol{U}_i\|_{\psi_2} \le C_{\max}^{1/2}C$. The similar arguments yield $E_2 \le \|\boldsymbol{\Sigma}^{-1/2}\|_2\|\mathbf{x}^*\|_2\|\boldsymbol{U}_i\|_{\psi_2} \le C_{\min}^{-1/2}C$. Finally, note that $\|\mathbf{x}^*\|_2 = 1$ implies $\|\mathbf{x}^*\|_\infty \le 1$. These together imply that $\|T_{ij}\|_{\psi_1} \le (1 + 2\rho^{1/2}C^2)$. Finally, by Lemma C.3, with $t = 2(1 + 2\rho^{1/2}C^2)\sqrt{C''^{-1}\log(d)/n}$, we obtain

$$\left\|\mathbf{x}^* + \widehat{\boldsymbol{\Sigma}}\mathbf{w}^*\right\|_\infty \le 2(1 + 2\rho^{1/2}C^2)\sqrt{\frac{C''^{-1}\log(d)}{n}},$$

with probability at least $1 - 2d^{-3}$, provided $2\sqrt{C''^{-1}\log(d)/n} \le 1$, where C'' is given in Lemma C.3.

□

A.3 Proof of Theorem 2.3

Proof of Theorem 2.3. By Lemma 2.1 with $\lambda' \asymp \sqrt{\log d/n}$, we have $|\xi| = \mathcal{O}_{\mathbb{P}}(s^*\log d/n^{1/2})$, which is $o_{\mathbb{P}}(1)$ under our assumption. This proves that conditioning on the design $\mathbf{X}$ and $\mathbf{x}^*$, $n^{1/2}(S_{\mathbf{x}}(\widehat{\boldsymbol{\beta}}) - \mathbf{x}^{*\top}\boldsymbol{\beta}^*) \rightsquigarrow N(0, \sigma^2\widehat{\mathbf{w}}^T\widehat{\boldsymbol{\Sigma}}\widehat{\mathbf{w}})$. Finally, we would like to show that the asymptotic variance satisfies $\widehat{\mathbf{w}}^T\widehat{\boldsymbol{\Sigma}}\widehat{\mathbf{w}} = \mathcal{O}_{\mathbb{P}}(1)$ under the random design. To this end, note that $\widehat{\mathbf{w}}^T\widehat{\boldsymbol{\Sigma}}\widehat{\mathbf{w}} \le \mathbf{w}^{*\top}\widehat{\boldsymbol{\Sigma}}\mathbf{w}^*$, by the definition of Dantzig selector and Lemma 2.2. In addition

$$\mathbf{w}^{*\top}\widehat{\boldsymbol{\Sigma}}\mathbf{w}^* \le \mathbf{w}^{*\top}\boldsymbol{\Sigma}\mathbf{w}^* + \|\mathbf{w}^*\|_1^2\|\widehat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}\|_{\max}.$$

For the first term in the right hand side,

$$\mathbf{w}^{*\top}\boldsymbol{\Sigma}\mathbf{w}^* \le \|\mathbf{w}^*\|_2^2\|\boldsymbol{\Sigma}\|_2 \le \|\mathbf{x}^*\|_2^2\|\boldsymbol{\Sigma}^{-1}\|_2^2\|\boldsymbol{\Sigma}\|_2 = \mathcal{O}(1).$$

Then, we consider the second term. Note that by the proof of Lemma A.2, we have $\|\widehat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}\|_{\max} = \mathcal{O}_{\mathbb{P}}(\sqrt{\log d/n})$. By assumption $\|\mathbf{w}^*\|_1^2\sqrt{\log d/n} = \mathcal{O}(1)$, the second term satisfies $\|\mathbf{w}^*\|_1^2\|\widehat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}\|_{\max} = \mathcal{O}_{\mathbb{P}}(1)$. Thus, $\mathbf{w}^{*\top}\widehat{\boldsymbol{\Sigma}}\mathbf{w}^* = \mathcal{O}_{\mathbb{P}}(1)$, which completes the proof.

□

A.4 Proof of Corollary 2.4

We denote

$$\kappa = \min\Big\{ s^* \frac{\mathbf{v}^T \widehat{\boldsymbol{\Sigma}} \mathbf{v}}{\|\mathbf{v}_{\mathcal{S}}\|_1^2} : \mathbf{v} \in \mathbb{R}^d \backslash \{0\}, \|\mathbf{v}_{\mathcal{S}^c}\|_1 \le 3\|\mathbf{v}_{\mathcal{S}}\|_1 \Big\},$$

as the compatibility factor (Ye and Zhang, 2010), where $\mathcal{S} = \mathrm{supp}(\boldsymbol{\beta}^*)$ is the support set of $\boldsymbol{\beta}^*$

Proof of Corollary 2.4. To prove this corollary, we need the nonasymptotic bound for E_1 in the proof of Lemma 2.1. Then, we invoke the following lemma.

Lemma A.1. Assume $\widehat{\Sigma}_{jj} = 1$, for $j = 1, ..., d$. Then, for the Lasso estimator $\widehat{\boldsymbol{\beta}}$, with probability at least $1 - 2d^{-2}$, we have $\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_1 \le 4\lambda s^*/\kappa$, provided $\lambda \ge 4\sigma^2\sqrt{6(\log d)/n}$.

Proof. The proof follows by combining Theorem 6.1 and Lemma 6.2 in Bühlmann and van de Geer (2011). □

This Lemma shows that with probability at least $1 - 2d^{-2}$, $|\sqrt{n}E_1| \le 4\sqrt{n}\lambda\lambda' s^*/\kappa$, provided $\lambda \ge 4\sigma^2\sqrt{6(\log d)/n}$. Then, we need the next lemma, which shows that the compatibility factor κ can be bounded from below with high probability.

Lemma A.2. Assume that $2\sqrt{C''^{-1}\log(d)/n} \le 1$ and $32(2C^2+1)\rho s^*\sqrt{C''^{-1}\log(d)/n} \le 1/2$, where $\rho = C_{\max}/C_{\min}$. Under the conditions in Lemma 2.2, we have $\kappa \ge C_{\min}/2$, with probability at least $1 - 2d^{-2}$.

Proof. See Appendix A for a detailed proof. □

The nonasymptotic bound in the theorem is then proved by combining Lemmas 2.2 and A.2. □

A.5 Proof of Lemma 2.5

Proof of Lemma 2.5. We note that

$$\widehat{\sigma}^2 - \sigma^2 = \Big(\frac{1}{n}\sum_{i=1}^n \varepsilon_i^2 - \sigma^2\Big) + \widehat{\boldsymbol{\Delta}}^T\widehat{\boldsymbol{\Sigma}}\widehat{\boldsymbol{\Delta}} - 2\widehat{\boldsymbol{\Delta}}^T\frac{1}{n}\sum_{i=1}^n \varepsilon_i \boldsymbol{X}_i, \tag{A.1}$$

where $\widehat{\boldsymbol{\Delta}} = \widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*$. By the law of large numbers, $|n^{-1}\sum_{i=1}^n \varepsilon_i^2 - \sigma^2| = \mathcal{O}_{\mathbb{P}}(n^{-1})$. For the second term of (A.1), Theorem 7.2 of Bickel et al. (2009) implies $\widehat{\boldsymbol{\Delta}}^T\widehat{\boldsymbol{\Sigma}}\widehat{\boldsymbol{\Delta}} \le Cs^*\log d/n$, for some constant C, with high probability. Finally, by Lemma A.1 and Lemma C.3, the last term of (A.1) satisfies

$$\Big|\widehat{\boldsymbol{\Delta}}^T\frac{1}{n}\sum_{i=1}^n \varepsilon_i \boldsymbol{X}_i\Big| \le \|\widehat{\boldsymbol{\Delta}}\|_1 \cdot \Big\|\frac{1}{n}\sum_{i=1}^n \varepsilon_i \boldsymbol{X}_i\Big\|_\infty \le C\frac{s\log d}{n},$$

for some constant C with high probability. Combining these results with equation (A.1), we have in probability $|\widehat{\sigma}^2 - \sigma^2| \le C\sqrt{\frac{1}{n}} \vee \frac{s\log d}{n}$, for some sufficiently large constant C. □

A.6 Proof of Lemma A.2

Proof of Lemma A.2. By the definition of κ and the fact that $\|\mathbf{v}_{\mathcal{S}}\|_1 \le s^{*1/2}\|\mathbf{v}_{\mathcal{S}}\|_2 \le s^{*1/2}\|\mathbf{v}\|_2$, we have

$$\kappa \ge \min\Big\{\frac{\mathbf{v}^T\widehat{\boldsymbol{\Sigma}}\mathbf{v}}{\|\mathbf{v}\|_2^2} : \mathbf{v}\in\mathbb{R}^d\backslash\{0\}, \|\mathbf{v}_{\mathcal{S}^c}\|_1 \le 3\|\mathbf{v}_{\mathcal{S}}\|_1\Big\}.$$

It is easily seen that

$$\frac{\mathbf{v}^T\widehat{\boldsymbol{\Sigma}}\mathbf{v}}{\|\mathbf{v}\|_2^2} = \frac{\mathbf{v}^T\boldsymbol{\Sigma}\mathbf{v}}{\|\mathbf{v}\|_2^2} + \frac{\mathbf{v}^T(\widehat{\boldsymbol{\Sigma}}-\boldsymbol{\Sigma})\mathbf{v}}{\|\mathbf{v}\|_2^2} \ge C_{\min} - \frac{\|\mathbf{v}\|_1^2\|\widehat{\boldsymbol{\Sigma}}-\boldsymbol{\Sigma}\|_{\max}}{\|\mathbf{v}\|_2^2}.$$

In addition, by $\|\mathbf{v}\|_1^2 \le 16\|\mathbf{v}_{\mathcal{S}}\|_1^2 \le 16s^*\|\mathbf{v}\|_2^2$, we can show that

$$\frac{\mathbf{v}^T\widehat{\boldsymbol{\Sigma}}\mathbf{v}}{\|\mathbf{v}\|_2^2} \ge C_{\min} - 16s^*\|\widehat{\boldsymbol{\Sigma}}-\boldsymbol{\Sigma}\|_{\max}.$$

Since $\widehat{\boldsymbol{\Sigma}} = n^{-1}\sum_{i=1}^n \boldsymbol{\Sigma}^{1/2}(\boldsymbol{\Sigma}^{-1/2}\boldsymbol{X}_i)(\boldsymbol{\Sigma}^{-1/2}\boldsymbol{X}_i)^T\boldsymbol{\Sigma}^{1/2}$, similar to the proof of Lemma 2.2, for any $j = 1,...,d$, we have $\|\boldsymbol{\Sigma}_{j*}^{1/2}(\boldsymbol{\Sigma}^{-1/2}\boldsymbol{X}_i)\|_{\psi_2} \le \|\boldsymbol{\Sigma}_{j*}^{1/2}\|_2\|\boldsymbol{\Sigma}^{-1/2}\boldsymbol{X}_i\|_{\psi_2} \le C_{\max}^{1/2}C$. Thus, we obtain $\|(\widehat{\boldsymbol{\Sigma}}-\boldsymbol{\Sigma})_{jk}\|_{\psi_1} \le \|\widehat{\boldsymbol{\Sigma}}_{jk}\|_{\psi_1} + C_{\max} \le 2C_{\max}C^2 + C_{\max}$, for any $j,k = 1,...,d$. By the union bound inequality and the Bernstein inequality in Lemma C.3, we obtain

$$\|\widehat{\boldsymbol{\Sigma}}-\boldsymbol{\Sigma}\|_{\max} \le 2(2C_{\max}C^2 + C_{\max})\sqrt{\frac{C''^{-1}\log(d)}{n}},$$

with probability at least $1-2d^{-2}$, provided $2\sqrt{C''^{-1}\log(d)/n} \le 1$, where C'' is given in Lemma C.3. Hence, with $32(2C_{\max}C^2 + C_{\max})s^*\sqrt{C''^{-1}\log(d)/n} \le C_{\min}/2$, we derive $\mathbf{v}^T\widehat{\boldsymbol{\Sigma}}\mathbf{v}/\|\mathbf{v}\|_2^2 \ge C_{\min}/2$. This completes the proof.

□

B Proofs for the Sparse High Dimensional Additive Model

B.1 Proof of Theorem 3.5

To prove Theorem 3.5, we first need the following two Lemmas.

Lemma B.1. Denote $\widehat{\boldsymbol{\Delta}} = \widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*$. Then, we have

$$\big(\lambda - \|\nabla\ell(\boldsymbol{\beta}^*)\|_{2,\infty}\big)\|\widehat{\boldsymbol{\Delta}}_{\mathcal{S}^c}\|_{2,1} \le \big(\lambda + \|\nabla\ell(\boldsymbol{\beta}^*)\|_{2,\infty}\big)\|\widehat{\boldsymbol{\Delta}}_{\mathcal{S}}\|_{2,1}.$$

It implies that, $\|\widehat{\boldsymbol{\Delta}}_{\mathcal{S}^c}\|_{2,1} \le 3\|\widehat{\boldsymbol{\Delta}}_{\mathcal{S}}\|_{2,1}$ with probability at least $1-2d^{-1}$ when $\lambda \ge 4(2\bar{C}C''^{-1}m_n\log d/n)^{1/2} + 4s^*C_\phi m_n^{1/2-k}$, where C_ϕ satisfies $\sup_{x\in[a,b]}|f_j(x)-\Phi(x)^T\boldsymbol{\beta}_j^*| \le C_\phi m_n^{-k}$ and $\bar{C}>0$ satisfies $|\psi_k| \le \bar{C}$.

Proof. For brevity, denote $\nabla\ell(\boldsymbol{\beta}^*) = \frac{1}{n}\sum_{i=1}^n \boldsymbol{Z}_i(Y_i - \boldsymbol{Z}_i^T\boldsymbol{\beta}^*)$. Note that $\boldsymbol{\beta}_j^* = 0$ if $j\in\mathcal{S}^c$. Let $D(\boldsymbol{\beta}_1,\boldsymbol{\beta}) = (\boldsymbol{\beta}_1-\boldsymbol{\beta})^T\widehat{\mathbf{C}}(\boldsymbol{\beta}_1-\boldsymbol{\beta})$ denote the symmetrized Bregman divergence. Denote $\widehat{\boldsymbol{\Delta}}_j = \widehat{\boldsymbol{\beta}}_j - \boldsymbol{\beta}_j^*$. Thus

$$\begin{aligned} D(\widehat{\boldsymbol{\beta}},\boldsymbol{\beta}^*) &= \widehat{\boldsymbol{\Delta}}^T\{\nabla\ell(\boldsymbol{\beta}^*+\widehat{\boldsymbol{\Delta}}) - \nabla\ell(\boldsymbol{\beta}^*)\} \\ &= \sum_{j\in\mathcal{S}^c}\widehat{\boldsymbol{\beta}}_j^T\nabla_j\ell(\boldsymbol{\beta}^*+\widehat{\boldsymbol{\Delta}}) + \sum_{j\in\mathcal{S}}\widehat{\boldsymbol{\Delta}}_j\nabla_j L(\boldsymbol{\beta}^*+\widehat{\boldsymbol{\Delta}}) - \widehat{\boldsymbol{\Delta}}^T\ell(\boldsymbol{\beta}^*). \end{aligned}$$

Furthermore, by the KKT condition,

$$\begin{cases} \nabla_j \ell(\widehat{\boldsymbol{\beta}}) = -\lambda \frac{\widehat{\boldsymbol{\beta}}_j}{\|\widehat{\boldsymbol{\beta}}_j\|_2}, & \text{if } \|\widehat{\boldsymbol{\beta}}_j\|_2 \neq 0, \\ \|\nabla_j \ell(\widehat{\boldsymbol{\beta}})\|_2 \leq \lambda, & \text{if } \|\widehat{\boldsymbol{\beta}}_j\|_2 = 0. \end{cases} \tag{B.1}$$

We have

$$\begin{aligned} D(\widehat{\boldsymbol{\beta}}, \boldsymbol{\beta}^*) &\leq -\sum_{j \in \mathcal{S}^c} \lambda \|\widehat{\boldsymbol{\beta}}_j\|_2 + \sum_{j \in \mathcal{S}} \|\widehat{\boldsymbol{\Delta}}_j\|_2 \lambda + \sum_{j=1}^{d} \|\widehat{\boldsymbol{\Delta}}_j\|_2 \|\nabla_j \ell(\boldsymbol{\beta}^*)\|_2 \\ &\leq -\lambda \|\widehat{\boldsymbol{\Delta}}_{\mathcal{S}^c}\|_{2,1} + \lambda \|\widehat{\boldsymbol{\Delta}}_{\mathcal{S}}\|_1 + \|\widehat{\boldsymbol{\Delta}}\|_{2,1} \|\nabla \ell(\boldsymbol{\beta}^*)\|_{2,\infty} \\ &= \big(\lambda + \|\nabla \ell(\boldsymbol{\beta}^*)\|_{2,\infty}\big) \|\widehat{\boldsymbol{\Delta}}_{\mathcal{S}}\|_{2,1} - \big(\lambda - \|\nabla \ell(\boldsymbol{\beta}^*)\|_{2,\infty}\big) \|\widehat{\boldsymbol{\Delta}}_{\mathcal{S}^c}\|_{2,1}. \end{aligned}$$

Since $D(\widehat{\boldsymbol{\beta}}, \boldsymbol{\beta}^*) \geq 0$, we have

$$\big(\lambda - \|\nabla \ell(\boldsymbol{\beta}^*)\|_{2,\infty}\big) \|\widehat{\boldsymbol{\Delta}}_{\mathcal{S}^c}\|_{2,1} \leq \big(\lambda + \|\nabla \ell(\boldsymbol{\beta}^*)\|_{2,\infty}\big) \|\widehat{\boldsymbol{\Delta}}_{\mathcal{S}}\|_{2,1}.$$

Next we bound the $\|\nabla \ell(\boldsymbol{\beta}^*)\|_{2,\infty}$,

$$\begin{aligned} \|\nabla \ell(\boldsymbol{\beta}^*)\|_{2,\infty} &= \left\| \frac{1}{n} \sum_{i=1}^{n} \boldsymbol{Z}_i (Y_i - \boldsymbol{Z}_i^T \boldsymbol{\beta}^*) \right\|_{2,\infty} \\ &= \underbrace{\left\| \frac{1}{n} \sum_{i=1}^{n} \boldsymbol{Z}_i \varepsilon_i \right\|_{2,\infty}}_{E_1} + \underbrace{\left\| \frac{1}{n} \sum_{i=1}^{n} \boldsymbol{Z}_i \Big(\sum_{j=1}^{d} f_j(X_{ij}) - \boldsymbol{Z}_i^T \boldsymbol{\beta}^* \Big) \right\|_{2,\infty}}_{E_2}. \end{aligned}$$

Note that $\|n^{-1} \varepsilon_i \boldsymbol{Z}_{ij}\|_2 \leq \sqrt{m_n} \|n^{-1} \varepsilon_i \boldsymbol{Z}_{ij}\|_\infty$. For any $k = 1, ..., m_n$, let $\Delta_{jk} = n^{-1} \varepsilon_i \psi_k(X_{ij})$. Since $\|\varepsilon_i\|_{\psi_1} \leq \|\varepsilon_i\|_{\psi_2} \leq C_\varepsilon$ and $|\psi_k| \leq 2\bar{C}$, for some $\bar{C} > 0$ by the assumption on the basis function. It therefore follows by Lemma C.3 and union bound that $E_1 \leq 2(2\bar{C} C''^{-1} m_n \log d/n)^{1/2}$ with probability at least $1 - 2d^{-1}$.

On the other hand,

$$\left\| \frac{1}{n} \sum_{i=1}^{n} \boldsymbol{Z}_i \Big(\sum_{j=1}^{d} f_j(X_{ij}) - \boldsymbol{Z}_i^T \boldsymbol{\beta}^* \Big) \right\|_{2,\infty} \leq (m_n)^{1/2} \left\| \frac{1}{n} \sum_{i=1}^{n} \boldsymbol{Z}_i \Big(\sum_{j=1}^{d} f_j(X_{ij}) - \boldsymbol{Z}_i^T \boldsymbol{\beta}^* \Big) \right\|_{\infty},$$

and we have

$$\begin{aligned} \left\| \frac{1}{n} \sum_{i=1}^{n} \boldsymbol{Z}_i \Big(\sum_{j=1}^{d} f_j(X_{ij}) - \boldsymbol{Z}_i^T \boldsymbol{\beta}^* \Big) \right\|_{\infty} &\leq 2n^{-1} \sum_{i=1}^{n} \left| \sum_{j \in \mathcal{S}} \big\{ f_j(X_{ij}) - \widetilde{\boldsymbol{Z}}_{ij}^T \boldsymbol{\beta}_j^* \big\} \right| + \Big\| \underbrace{\frac{1}{n} \sum_{i=1}^{n} \boldsymbol{Z}_i (\bar{\boldsymbol{Z}}^T \boldsymbol{\beta}^*)}_{=0} \Big\|_{\infty} \\ &\leq 2 s^* C_\phi m_n^{-k}. \end{aligned}$$

where $\sup_{x \in [a,b]} |f_j(x) - \Phi(x)^T \boldsymbol{\beta}_j^*| \leq C_\phi m_n^{-k}$. Hence, $|E_2| \leq 2 s^* C_\phi m_n^{1/2-k}$. This completes the proof. □

In the following we derive the rate of convergence for the estimator $\widehat{\boldsymbol{\beta}}$.

Theorem B.2. Under Assumption 3.4 and Lemma B.1, with probability at least $1 - 2d^{-1}$,

$$\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_{2,1} \leq 6\tau^{-1} s^* m_n \lambda. \tag{B.2}$$

Proof. Let $\widehat{\boldsymbol{\Delta}} = \widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*$. Note that

$$D(\widehat{\boldsymbol{\beta}}, \boldsymbol{\beta}^*) \leq \big(\lambda + \|\nabla \ell(\boldsymbol{\beta}^*)\|_{2,\infty}\big)\|\widehat{\boldsymbol{\Delta}}_{\mathcal{S}}\|_{2,1} \leq \frac{3\lambda}{2}\|\widehat{\boldsymbol{\Delta}}_{\mathcal{S}}\|_{2,1}.$$

By (3.6), we have

$$D(\widehat{\boldsymbol{\beta}}, \boldsymbol{\beta}^*) \geq \tau m_n^{-1}\|\widehat{\boldsymbol{\Delta}}\|_{2,2}^2 \geq \tau m_n^{-1}\|\widehat{\boldsymbol{\Delta}}_{\mathcal{S}}\|_{2,2}^2 \geq \tau m_n^{-1} s^{*-1}\|\widehat{\boldsymbol{\Delta}}_{\mathcal{S}}\|_{2,1}^2.$$

This implies that

$$\|\widehat{\boldsymbol{\Delta}}_{\mathcal{S}}\|_{2,1} \leq \frac{3 s^* m_n \lambda}{2\tau}.$$

Together with $\|\widehat{\boldsymbol{\Delta}}_{\mathcal{S}^c}\|_{2,1} \leq 3\|\widehat{\boldsymbol{\Delta}}_{\mathcal{S}}\|_{2,1}$ in Lemma B.1, we have

$$\|\widehat{\boldsymbol{\Delta}}\|_{2,1} = \|\widehat{\boldsymbol{\Delta}}_{\mathcal{S}}\|_{2,1} + \|\widehat{\boldsymbol{\Delta}}_{\mathcal{S}^c}\|_{2,1} \leq 4\|\widehat{\boldsymbol{\Delta}}_{\mathcal{S}}\|_{2,1} \leq \frac{6 s^* m_n \lambda}{\tau}.$$

This completes the proof of (B.2). □

Equipped with Lemma B.1 and Lemma B.2, we can now prove Theorem 3.5.

Proof of Theorem 3.5. It follows from the definition of $S_{\mathbf{x}}(\widehat{\boldsymbol{\beta}})$ that

$$\begin{aligned}
&S_{\mathbf{x}}(\widehat{\boldsymbol{\beta}}) - \mathbf{z}^{*\top}\boldsymbol{\beta}^* \\
&= \mathbf{z}^{*\top}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*) - \widehat{\mathbf{w}}^T \frac{1}{n}\sum_{i=1}^n \boldsymbol{Z}_i(Y_i - \boldsymbol{Z}_i^T\widehat{\boldsymbol{\beta}}) \\
&= \underbrace{\left(\mathbf{z}^* + \widehat{\mathbf{w}}^T n^{-1}\sum_{i=1}^n \boldsymbol{Z}_i\boldsymbol{Z}_i^T\right)^T(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)}_{E_1} - \underbrace{\widehat{\mathbf{w}}^T n^{-1}\sum_{i=1}^n \boldsymbol{Z}_i\left[\sum_{j=1}^p f_j(X_{ij}) - \boldsymbol{Z}_i^T\boldsymbol{\beta}^*\right]}_{E_2} \\
&- \widehat{\mathbf{w}}^T n^{-1}\sum_{i=1}^n \boldsymbol{Z}_i\varepsilon_i.
\end{aligned}$$

Observe that $\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_1 \leq \sqrt{m_n}\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_{2,1}$, from Theorem B.2, we can bound E_1 as

$$|E_1| \leq \|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_1\lambda' \leq \sqrt{m_n}\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_{2,1}\lambda' \leq \sqrt{m_n}6s^*\tau^{-1}m_n\lambda'(4(2\bar{C}C''^{-1}m_n\log d/n)^{1/2} + 4s^*C_\phi m_n^{1/2-k})$$

with probability at least $1 - 2d^{-1}$ by Lemmas B.1 and B.2. Recall that $\boldsymbol{Z}_i = \widetilde{\boldsymbol{Z}}_i - \bar{\boldsymbol{Z}}$, and

$$E_2 = \underbrace{n^{-1}\sum_{i=1}^n \widehat{\mathbf{w}}^T\boldsymbol{Z}_i\left[\sum_{j=1}^p f_j(X_{ij}) - \widetilde{\boldsymbol{Z}}_i^T\boldsymbol{\beta}^*\right]}_{E_{21}} + \underbrace{n^{-1}\sum_{i=1}^n \widehat{\mathbf{w}}^T\boldsymbol{Z}_i\bar{\boldsymbol{Z}}^T\boldsymbol{\beta}^*}_{E_{22}}.$$

First, we note that if $f_j = 0$, the corresponding coefficients in $\boldsymbol{\beta}^*$ are also 0. Therefore, by $\sup_{x\in[a,b]} |f_j(x) - \Phi(x)^T\boldsymbol{\beta}_j^*| \le C_\phi m_n^{-1}$,

$$\left|\sum_{j=1}^{p} f_j(X_{ij}) - \widetilde{\boldsymbol{Z}}_i^T\boldsymbol{\beta}^*\right| = \sum_{j\in\mathcal{S}} |f_j(X_{ij}) - \boldsymbol{\Phi}(X_{ij})^T\boldsymbol{\beta}_j^*| = \mathcal{O}(s^* m_n^{-k}). \tag{B.3}$$

To deal with E_{21}, Hölder's inequality implies that

$$\begin{aligned} |E_{21}| &\le \sqrt{n^{-1}\sum_{i=1}^{n}(\widehat{\mathbf{w}}^T\boldsymbol{Z}_i)^2}\sqrt{n^{-1}\sum_{i=1}^{n}\left(\sum_{j=1}^{p} f_j(X_{ij}) - \widetilde{\boldsymbol{Z}}_i^T\boldsymbol{\beta}^*\right)^2} \\ &= \sqrt{\widehat{\mathbf{w}}^T\widehat{\mathbf{C}}\widehat{\mathbf{w}}}\sqrt{n^{-1}\sum_{i=1}^{n}\left(\sum_{j=1}^{p} f_j(X_{ij}) - \widetilde{\boldsymbol{Z}}_i^T\boldsymbol{\beta}^*\right)^2}. \end{aligned} \tag{B.4}$$

To find the order of $\widehat{\mathbf{w}}^T\widehat{\mathbf{C}}\widehat{\mathbf{w}}$, first note that $\widehat{\mathbf{w}}^T\widehat{\mathbf{C}}\widehat{\mathbf{w}} \le \mathbf{w}^{*\top}\widehat{\mathbf{C}}\mathbf{w}^*$, by the definition of Dantzig selector and Lemma 3.6. In addition

$$\mathbf{w}^{*\top}\widehat{\mathbf{C}}\mathbf{w}^* \le \mathbf{w}^{*\top}\mathbf{C}\mathbf{w}^* + \|\mathbf{w}^*\|_1^2\|\widehat{\mathbf{C}} - \mathbf{C}\|_{\max}.$$

For the first term in the right hand side,

$$\mathbf{w}^{*\top}\mathbf{C}\mathbf{w}^* \le \|\mathbf{w}^*\|_2^2\|\mathbf{C}\|_2 \le \|\mathbf{x}^*\|_2^2\|\mathbf{C}^{-1}\|_2^2\|\mathbf{C}\|_2 = \mathcal{O}(m_n).$$

For the second term, note that similar to the proof of Lemma A.2, we have $\|\widehat{\mathbf{C}} - \mathbf{C}\|_{\max} = \mathcal{O}_{\mathbb{P}}(\sqrt{\log dm_n/n})$. Therefore, by assumption on $\|\mathbf{w}^*\|_1$:

$$\|\mathbf{w}^*\|_1^2\|\widehat{\mathbf{C}} - \mathbf{C}\|_{\max} = \mathcal{O}_{\mathbb{P}}\left(\|\mathbf{w}^*\|_1^2\sqrt{\frac{\log dm_n}{n}}\right) = \mathcal{O}_{\mathbb{P}}(m_n).$$

Therefore, we conclude that $\sqrt{\mathbf{w}^{*\top}\widehat{\mathbf{C}}\mathbf{w}^*} = \mathcal{O}_{\mathbb{P}}(m_n^{1/2})$.

Using (B.3), we can bound the second term of (B.4) by $\mathcal{O}_{\mathbb{P}}(s^* m_n^{-k})$. To sum up,

$$|E_{21}| = \mathcal{O}_{\mathbb{P}}(s^* m_n^{1/2-k}).$$

For E_{22}, the centeredness of $\boldsymbol{Z}_i$ yields that

$$E_{22} = (\bar{\boldsymbol{Z}}^T\boldsymbol{\beta}^*)\widehat{\mathbf{w}}^T\left(n^{-1}\sum_{i=1}^{n}\boldsymbol{Z}_i\right) = 0. \tag{B.5}$$

To sum up, $|E_2| = \mathcal{O}_{\mathbb{P}}(s^* m_n^{1/2-k})$.

The asymptotic distribution therefore depends on $N = \sqrt{\frac{n}{m_n}}\widehat{\mathbf{w}}^T n^{-1}\sum_{i=1}^{n}\boldsymbol{Z}_i\varepsilon_i$. Conditioning on the design matrix $\mathbf{X}$ and $\mathbf{x}^*$, we have $N \sim N(0, m_n^{-1}\sigma^2\widehat{\mathbf{w}}^T\widehat{\mathbf{C}}\widehat{\mathbf{w}})$.

□

B.2 Proof of Lemma 3.6

Proof of Lemma 3.6. Since $\widehat{\mathbf{C}} = n^{-1}\sum_{i=1}^{n} \boldsymbol{Z}_i \boldsymbol{Z}_i^T$, we can show that

$$\mathbf{z}^* + \widehat{\mathbf{C}}\mathbf{w}^* = -\frac{1}{n}\sum_{i=1}^{n}(\boldsymbol{Z}_i\boldsymbol{Z}_i^T\mathbf{C}^{-1}\mathbf{z}^* - \mathbf{z}^*) = -\frac{1}{n}\sum_{i=1}^{n}(\mathbf{C}^{1/2}\boldsymbol{U}_i\boldsymbol{U}_i^T\mathbf{C}^{-1/2}\mathbf{z}^* - \mathbf{z}^*),$$

where $\boldsymbol{U}_i = \mathbf{C}^{-1/2}\boldsymbol{Z}_i$. Now, we consider the jth component of $\boldsymbol{T}_i = \mathbf{C}^{1/2}\boldsymbol{U}_i\boldsymbol{U}_i^T\mathbf{C}^{-1/2}\mathbf{z}^* - \mathbf{z}^*$. For $\mathbf{C}_{j*}^{1/2}\boldsymbol{U}_i\boldsymbol{U}_i^T\mathbf{C}^{-1/2}\mathbf{z}^*$ with $1 \le j \le d$, we have by Lemma C.2,

$$\|\mathbf{C}_{j*}^{1/2}\boldsymbol{U}_i\boldsymbol{U}_i^T\mathbf{C}^{-1/2}\mathbf{z}^*\|_{\psi_1} \le 2\underbrace{\|\mathbf{C}_{j*}^{1/2}\boldsymbol{U}_i\|_{\psi_2}}_{E_1}\underbrace{\|\boldsymbol{U}_i^T\mathbf{C}^{-1/2}\mathbf{z}^*\|_{\psi_2}}_{E_2}.$$

For E_1, by the definition of ψ_2 norm, we can show that $E_1 \le \|\mathbf{C}_{j*}^{1/2}\|_2\|\boldsymbol{U}_i\|_{\psi_2} \le m_n^{1/2}C_{\max}^{1/2}C$. The similar arguments yield $E_2 \le \|\mathbf{C}^{-1/2}\|_2\|\mathbf{z}^*\|_2\|\boldsymbol{U}_i\|_{\psi_2} \le m_n^{-1/2}C_{\min}^{-1/2}C$. Finally, note that $\|\mathbf{z}^*\|_2 = 1$ implies $\|\mathbf{z}^*\|_\infty \le 1$. These together imply that $\|T_{ij}\|_{\psi_1} \le (1 + 2\rho^{1/2}C^2)$. Finally, by Lemma C.3, with $t = 2(1+2\rho^{1/2}C^2)\sqrt{C''^{-1}\log(m_n d)/n}$, we obtain

$$\left\|\mathbf{z}^* + \widehat{\mathbf{C}}\mathbf{w}^*\right\|_\infty \le 2(1+2\rho^{1/2}C^2)\sqrt{\frac{C''^{-1}\log(m_n d)}{n}},$$

with probability at least $1 - 2(dm_n)^{-3}$, provided $2\sqrt{C''^{-1}\log(dm_n)/n} \le 1$, where C'' is given in Lemma C.3. □

B.3 Proof of Lemma 3.7

Proof of Lemma 3.7. Observe that

$$\widehat{\sigma}^2 - \sigma^2 = \underbrace{\frac{1}{n}\sum_{i=1}^{n}(\varepsilon_i^2 - \sigma^2)}_{E_1} + \underbrace{\frac{1}{n}\sum_{i=1}^{n}\left(\sum_{j=1}^{p} f_j(X_{ij}) - \boldsymbol{Z}_i^T\widehat{\boldsymbol{\beta}}\right)^2}_{E_2} - \underbrace{2\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i\left(\sum_{j=1}^{p} f_j(X_{ij}) - \boldsymbol{Z}_i^T\widehat{\boldsymbol{\beta}}\right)}_{E_3}. \tag{B.6}$$

E_1: By Assumption 3.2 and the Bernstein inequality in Lemma C.3, the first term in (B.6) is of order $\mathcal{O}_{\mathbb{P}}(\sqrt{\frac{\log n}{n}}) = o_{\mathbb{P}}(1)$.

E_2: To bound the second term, since $\boldsymbol{Z}_i = \widetilde{\boldsymbol{Z}}_i - \bar{\boldsymbol{Z}}$, note that by the inequality $(a+b)^2 \le 2a^2 + 2b^2$ for all $a, b \in \mathbb{R}$,

$$\begin{aligned} E_2 &= \frac{1}{n}\sum_{i=1}^{n}\left(\sum_{j=1}^{p} f_j(X_{ij}) - \boldsymbol{Z}_i^T\boldsymbol{\beta}^* + \boldsymbol{Z}_i^T(\boldsymbol{\beta}^* - \widehat{\boldsymbol{\beta}})\right)^2 \\ &\le \underbrace{\frac{2}{n}\sum_{i=1}^{n}\left(\sum_{j=1}^{p} f_j(X_{ij}) - \widetilde{\boldsymbol{Z}}_i^T\boldsymbol{\beta}^* + \bar{\boldsymbol{Z}}^T\boldsymbol{\beta}^*\right)^2}_{\Delta_{21}} + \underbrace{\frac{2}{n}\sum_{i=1}^{n}\left(\boldsymbol{Z}_i^T(\boldsymbol{\beta}^* - \widehat{\boldsymbol{\beta}})\right)^2}_{\Delta_{22}}, \end{aligned} \tag{B.7}$$

For Δ_{21}, by the fact that $f_j \neq 0$ for $j \in \mathcal{S}$ and $|\mathcal{S}| = s^* < \infty$, $|\sum_{j=1}^p f_j(X_{ij}) - \widetilde{\boldsymbol{Z}}_i^T \boldsymbol{\beta}^*| = \mathcal{O}(s^* m_n^{-k})$. Moreover, in the proof for Lemma C.1 in Huang et al. (2010), it is shown that $\bar{\boldsymbol{Z}}^T \boldsymbol{\beta}^* = \mathcal{O}_{\mathbb{P}}(s^* m_n^{-k} + s^* \sqrt{m_n/n})$. Hence, $\Delta_{21} = \mathcal{O}_{\mathbb{P}}(s^{*2} m_n^{-2k} + s^{*2} m_n/n)$.

For Δ_{22}, note that we can express $\Delta_{22} = |\widehat{\boldsymbol{\Delta}}^T \widehat{\mathbf{C}} \widehat{\boldsymbol{\Delta}}|$, where $\widehat{\boldsymbol{\Delta}} = \widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*$. By Theorem B.2, $\Delta_{22} \leq 9 s^* m_n \lambda^2 / \tau = o_{\mathbb{P}}(1)$.

E_3: To control the last term, note that

$$\left[\frac{1}{n}\sum_{i=1}^n \varepsilon_i \left(\sum_{j=1}^p f_j(X_{ij}) - \boldsymbol{Z}_i^T \widehat{\boldsymbol{\beta}}\right)\right] \leq \frac{1}{n}\sum_{i=1}^n \varepsilon_i^2 \cdot \frac{1}{n}\sum_{i=1}^n \left(\sum_{j=1}^p f_j(X_{ij}) - \boldsymbol{Z}_i^T \widehat{\boldsymbol{\beta}}\right)^2.$$

The first term on the right hand side converges in probability to a constant by the law of large numbers. The second term agrees with E_2 and is of order $o_{\mathbb{P}}(1)$. Thus, we have $E_3 = o_{\mathbb{P}}(1)$.

Combining E_1, E_2 and E_3, we get the desired result. □

C Supplementary Lemmas

The first three lemmas are corresponding to Section 3. Let

$$S_{nj}^0 = \left\{ f_{nj} : f_{nj}(x) = \sum_{k=1}^{m_n} b_{jk} \psi_k(x), (\beta_{j1}, ..., \beta_{jm_n}) \in \mathbb{R}^{m_n} \right\},$$

where ψ_k is defined in Section 3. The following Lemma is proved in Huang et al. (2010).

Lemma C.1 (Lemma 1 of Huang et al. (2010))**.** Under Assumptions 3.2 and 3.3, for any $f \in \mathcal{F}$, there exists $f_n \in S_{nj}^0$ satisfying

$$\|f_n - f\|_2 = \mathcal{O}_{\mathbb{P}}(m_n^{-k} + m_n^{1/2} n^{-1/2}).$$

Lemma C.2. Assume that X and Y are sub-Gaussian. Then $\|XY\|_{\psi_1} \leq 2\|X\|_{\psi_2}\|Y\|_{\psi_2}$

Lemma C.3 (Bernstein Inequality)**.** Let $X_1, ..., X_n$ be independent mean 0 sub-exponential random variables and let $K = \max_i \|X_i\|_{\psi_1}$. The for any $t > 0$, we have

$$\mathbb{P}_{\boldsymbol{\beta}^*}\left(\frac{1}{n}\left|\sum_{i=1}^n X_i\right| \geq t\right) \leq 2\exp\left[-C'' \min\left(\frac{t^2}{K^2}, \frac{t}{K}\right) n\right],$$

where $C'' > 0$ is a universal constant.

References

Belloni, A. and Chernozhukov, V. (2013). Least squares after model selection in high-dimensional sparse models. *Bernoulli* **19** 521–547.

Belloni, A., Chernozhukov, V. and Wei, Y. (2013). Honest confidence regions for logistic regression with a large number of controls. *arXiv preprint arXiv:1304.3969* .

Bickel, P. J., Ritov, Y. and Tsybakov, A. B. (2009). Simultaneous analysis of lasso and dantzig selector. *The Annals of Statistics* **37** 1705–1732.

Bühlmann, P. and van de Geer, S. (2011). *Statistics for high-dimensional data: methods, theory and applications*. Springer.

Candes, E. and Tao, T. (2007). The Dantzig selector: statistical estimation when p is much larger than n. *The Annals of Statistics* 2313–2351.

de los Campos, G., Hickey, J. M., Pong-Wong, R., Daetwyler, H. D. and Calus, M. P. L. (2013). Whole-genome regression and prediction methods applied to plant and animal breeding. *Genetics* **193** 327–345.

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**.

Fan, J., Guo, S. and Hao, N. (2012a). Variance estimation using refitted cross-validation in ultrahigh dimensional regression. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **74** 37–65.

Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association* **96** 1348–1360.

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

Fan, J. and Lv, J. (2011). Nonconcave penalized likelihood with np-dimensionality. *Information Theory, IEEE Transactions on* **57** 5467–5484.

Fan, J., Xue, L. and Zou, H. (2012b). Strong oracle optimality of folded concave penalized estimation. *arXiv preprint arXiv:1210.5992* .

Faraway, J. J. (2014). *Linear models with R*. CRC Press.

Graybill, F. A. (2000). *Theory and application of the linear model*. Cengage Learning.

Greenshtein, E. and Ritov, Y. (2004). Persistence in high-dimensional linear predictor selection and the virtue of overparametrization. *Bernoulli* **10** 971–988.

Homrighausen, D. and McDonald, D. J. (2013). The lasso, persistence, and cross-validation. In *Proceedings of the 30th International Conference on Machine Learning*.

Huang, J., Horowitz, J. L. and Wei, F. (2010). Variable selection in nonparametric additive models. *The Annals of Statistics* **38** 2282–2313.

Javanmard, A. and Montanari, A. (2013). Confidence intervals and hypothesis testing for high-dimensional regression. *arXiv preprint arXiv:1306.3171* .

Knight, K. and Fu, W. (2000). Asymptotics for lasso-type estimators. *Annals of Statistics* 1356–1378.

Lei, J., Robins, J. and Wasserman, L. (2014). Distribution-free prediction sets. *Journal of the American Statistical Association* **108** 278–287.

Lei, J. and Wasserman, L. (2014). Distribution-free prediction bands for non- parametric regression. *Journal of the Royal Statistical Society: Series B* **76** 71–96.

Lockhart, R., Taylor, J., Tibshirani, R. J. and Tibshirani, R. (2014). A significance test for the lasso. *The Annals of Statistics* **42** 413–468.

Loh, P.-L. and Wainwright, M. J. (2013). Regularized m-estimators with nonconvexity: Statistical and algorithmic theory for local optima. *arXiv preprint arXiv:1305.2436* .

Lu, J., Kolar, M. and Liu, H. (2015). Post-regularization confidence bands for high dimensional nonparametric models with local sparsity. *arXiv preprint arXiv: 1503.02978* .

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

Meinshausen, N. and Bühlmann, P. (2006). High-dimensional graphs and variable selection with the lasso. *The Annals of Statistics* **34** 1436–1462.

Meuwissen, T. H., Hayes, B. J. and Goddard, M. E. (2001). Prediction of total genetic value using genome-wide dense marker maps. *Genetics* **157** 1819–1829.

Negahban, S. N., Ravikumar, P., Wainwright, M. J. and Yu, B. (2012). A unified framework for high-dimensional analysis of M-estimators with decomposable regularizers. *Statistical Science* **27** 538–557.

Newey, W. K. (1997). Convergence rates and asymptotic normality for series estimators. *Journal of Econometrics* **79** 147–168.

Ning, Y. and Liu, H. (2014). A general theory of hypothesis tests and confidence regions for sparse high dimensional models. *arXiv preprint arXiv:1412.8765* .

Pérez, P. and de los Campos, G. (2014). Genome-wide regression and prediction with the bglr statistical package. *Genetics* **198** 483–495.

Raskutti, G., Wainwright, M. J. and Yu, B. (2012). Minimax-optimal rates for sparse additive models over kernel classes via convex programming. *The Journal of Machine Learning Research* **13** 389–427.

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

Reid, S., Tibshirani, R. and Friedman, J. (2014). A study of error variance estimation in lasso regression. *arXiv preprint arXiv:1311.5274* .

Schumaker, L. (1981). *Spline Functions: Basic Theory*. Wiley, New York.

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

Sun, T. and Zhang, C.-H. (2012). Scaled sparse linear regression. *Biometrika* **99** 879–898.

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

van de Geer, S., Bühlmann, P., Ritov, Y. and Dezeure, R. (2014). On asymptotically optimal confidence regions and tests for high-dimensional models. *The Annals of Statistics* **42** 1166–1202.

Vovk, V., Gammerman, A. and Shafer, G. (2005). *Algorithmic Learning in a Random World*. Springer.

Vovk, V., Nouretdinov, I. and Gammerman, A. (2009). On-line predictive linear regression. *The Annals of Statistics* **37** 1566–1590.

Wainwright, M. J. (2009). Sharp thresholds for high-dimensional and noisy sparsity recovery using¡ formula formulatype=. *Information Theory, IEEE Transactions on* **55** 2183–2202.

Wang, L., Kim, Y. and Li, R. (2013). Calibrating nonconvex penalized regression in ultra-high dimension. *Annals of Statistics* **41** 2505–2536.

Wasserman, L. (2014). Discussion: "a significance test for the lasso". *The Annals of Statistics* **42** 501–508.

Ye, F. and Zhang, C.-H. (2010). Rate minimaxity of the lasso and dantzig selector for the lq loss in lr balls. *The Journal of Machine Learning Research* **9999** 3519–3540.

Zhang, C.-H. (2010a). Nearly unbiased variable selection under minimax concave penalty. *The Annals of Statistics* **38** 894–942.

Zhang, C.-H. and Zhang, S. S. (2014). Confidence intervals for low-dimensional parameters in high-dimensional linear models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **76** 217–242.

Zhang, T. (2009). Some sharp performance bounds for least squares regression with L_1 regularization. *Annals of Statistics* **37** 2109–2144.

Zhang, T. (2010b). Analysis of multi-stage convex relaxation for sparse regularization. *The Journal of Machine Learning Research* **11** 1081–1107.

Zhang, T. (2013). Multi-stage convex relaxation for feature selection. *Bernoulli* **19** 2277–2293.

Zhao, P. and Yu, B. (2006). On model selection consistency of lasso. *The Journal of Machine Learning Research* **7** 2541–2563.

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