The Annals of Statistics
2012, Vol. 40, No. 2, 812–831
DOI: 10.1214/12-AOS982

ESTIMATION IN HIGH-DIMENSIONAL LINEAR MODELS WITH DETERMINISTIC DESIGN MATRICES

BY JUN SHAO[1] AND XINWEI DENG

East China Normal University, University of Wisconsin and Virginia Polytechnic Institute and State University

Because of the advance in technologies, modern statistical studies often encounter linear models with the number of explanatory variables much larger than the sample size. Estimation and variable selection in these high-dimensional problems with deterministic design points is very different from those in the case of random covariates, due to the identifiability of the high-dimensional regression parameter vector. We show that a reasonable approach is to focus on the projection of the regression parameter vector onto the linear space generated by the design matrix. In this work, we consider the ridge regression estimator of the projection vector and propose to threshold the ridge regression estimator when the projection vector is sparse in the sense that many of its components are small. The proposed estimator has an explicit form and is easy to use in application. Asymptotic properties such as the consistency of variable selection and estimation and the convergence rate of the prediction mean squared error are established under some sparsity conditions on the projection vector. A simulation study is also conducted to examine the performance of the proposed estimator.

1. Introduction. Consider the following linear model:

$$y_i = \mathbf{x}_i'\boldsymbol{\beta} + \varepsilon_i, \qquad i = 1, \ldots, n, \tag{1}$$

where y_i is an observed response variable, $\mathbf{x}_i$ is a p-dimensional vector of observed covariates or design points associated with y_i, $\boldsymbol{\beta}$ is a p-dimensional vector of unknown parameters and ε_i's are independent and identically distributed unobserved random errors with mean 0 and unknown variance σ^2. The theory of linear models is well established for traditional applications where the dimension p is fixed and the sample size $n > p$. With modern technologies, however, in many biological, medical, social and economical studies, p is comparable with, or much larger than, n, and making valid statistical inference is a great challenge.

In the case of $p < n$, there is a rich literature on variable selection, that is, identifying nonzero components of $\boldsymbol{\beta}$ in (1). For variable selection in the case of $p > n$ and statistical inference afterwards, the development of statistical theory

Received September 2010; revised January 2012.
[1]Supported in part by NSF Grants SES-07-05033 and DMS-10-07454.
MSC2010 subject classifications. Primary 62J07; secondary 62G20, 62J05.
Key words and phrases. Identifiability, projection, ridge regression, sparsity, thresholding, variable selection.

started about a decade ago. Some excellent advances in asymptotic theory have been made recently in situations where p diverges to infinity as the sample size n increases to infinity with the divergence rate $O(n^l)$ for some $l > 0$ (polynomial-type divergence rate) or $O(e^{n^\nu})$ for some $\nu \in (0, 1)$ (ultra-high dimension). See, for example, Fan and Peng (2004), Hunter and Li (2005), Meinshausen and Buhlmann (2006), Zhao and Yu (2006), Zou (2006), Wang, Li and Tsai (2007), Fan and Lv (2008), Zhang and Huang (2008), Meinshausen and Yu (2009), Wang (2009) and a review by Fan and Lv (2010). When $\mathbf{x}_i$'s are random covariates, under some conditions, some variable selection methods have been shown to be selection-consistent in the sense that, with probability tending to 1 as $n \to \infty$, the selected variables are exactly those related to the response, where the probability is with respect to the joint distribution of $(y_i, \mathbf{x}_i)$'s. As Fan and Lv (2008) commented in the end of their stimulating paper, however, no selection-consistency result is available for deterministic $\mathbf{x}_i$'s and many applications, such as biomedical imaging and signal processing, involve deterministic design points. Another example in which $\mathbf{x}_i$ can be treated as deterministic is an analysis conditional on the observed covariates.

Let $\mathbf{X}$ be the matrix whose ith row is $\mathbf{x}_i'$, $i = 1, \ldots, n$. For simplicity, we call $\mathbf{X}$ the design matrix although $\mathbf{x}_i$'s are not necessarily designed points. When $p > n$, a key difference between a random $\mathbf{X}$ and a deterministic design matrix is the identifiability of the regression parameter $\boldsymbol{\beta}$ in (1), caused by the fact that the probabilities under consideration are different. For random $\mathbf{x}_i$'s that are independent and identically distributed and independent of ε_i's, $\boldsymbol{\beta} = [\operatorname{cov}(\mathbf{x}_i)]^{-1} \operatorname{cov}(\mathbf{x}_i, y_i)$. Hence, even when $p > n$, components of $\boldsymbol{\beta}$ can be estimated, and nonzero components of $\boldsymbol{\beta}$ can be identified consistently with respect to the joint probability distribution of $(y_i, \mathbf{x}_i)$'s, under some conditions on $\operatorname{cov}(\mathbf{x}_i)$ and $\operatorname{cov}(\mathbf{x}_i, y_i)$. On the other hand, when the design matrix is deterministic or an analysis conditional on $\mathbf{X}$ is considered, the underlying probability is the probability distribution of $(y_1, \ldots, y_n)$ conditional on $\mathbf{X}$, and $\boldsymbol{\beta}$ is identifiable if and only if it lies in a set having a one-to-one correspondence with $\mathcal{R}(\mathbf{X})$, the linear space spanned by rows of $\mathbf{X}$. Since the dimension of $\mathcal{R}(\mathbf{X})$ is at most n, when $p > n$, $\boldsymbol{\beta}$ is generally not identifiable with respect to the probability distribution of $(y_1, \ldots, y_n)$ conditional on $\mathbf{X}$. Consequently, with deterministic $\mathbf{X}$ and $p > n$, it is not realistic to derive consistent estimators of $\boldsymbol{\beta}$ or consistent variable selection procedures.

Without selection-consistency [as previously described; see definition (7) in Section 4.1], we may still derive consistent estimators of some useful functions of $\boldsymbol{\beta}$ under the p-dimensional linear model given by (1) with deterministic $\mathbf{X}$ and $p > n$. This is the main focus of the current paper. Although $\boldsymbol{\beta}$ is generally not identifiable when $p > n$, we argue in Section 2 that we may not need to estimate the entire vector $\boldsymbol{\beta}$. For statistical analysis, $\boldsymbol{\theta}$, the projection of $\boldsymbol{\beta}$ onto $\mathcal{R}(\mathbf{X})$, is what we are able to estimate, and perhaps the estimation of $\boldsymbol{\theta}$ is sufficient for valid statistical inference.

To estimate $\boldsymbol{\theta}$, we first consider the ridge regression estimator in Section 3. For any linear combination of the ridge regression estimator, we establish the asymptotic convergence rate of its mean squared error. We also obtain the convergence

rate of the expected L_2-norm error for the ridge regression estimator of $\mathbf{X}\boldsymbol{\theta}$. This expected L_2-norm error divided by n is equal to the average prediction mean squared error minus σ^2.

When $\boldsymbol{\theta}$ is sparse in the sense that many of its components are small, we consider in Section 4 a sparse estimator of $\boldsymbol{\theta}$ obtained by thresholding the ridge regression estimator of $\boldsymbol{\theta}$. We show that, with probability tending to 1 at a fast rate, we can eliminate small components of $\boldsymbol{\theta}$ and keep large components of $\boldsymbol{\theta}$, that is, thresholding the ridge regression estimator provides a variable selection procedure, that is, consistent in some sense. This method is computationally much simpler than methods such as the LASSO [Tibshirani (1996)], SCAD [Fan and Li (2001)] and the ENET [Zou and Hastie (2005)], since no numerical minimization is required as the proposed estimator has an explicit form. We show that the convergence rate of the expected L_2-norm error or average prediction mean squared error of the thresholded ridge regression estimator is much faster than that of the ridge regression estimator when $\boldsymbol{\theta}$ is sparse. In particular, the thresholded ridge regression estimator is estimation-consistent (defined in Section 4), but the ridge regression estimator may not be.

Thresholding the ridge regression estimator is closely related to the SIS as shown in Fan and Lv (2008). However, its asymptotic behavior for deterministic $\mathbf{X}$ is different from that for random $\mathbf{X}$, and its consistency also requires very different conditions. For deterministic $\mathbf{X}$ and $p > n$, there does not exist any result on the consistency of the LASSO, SCAD or ENET. When $p < n$, Zhang and Huang (2008) showed that the LASSO is estimation-consistent, but the required conditions are more stringent and complicated than those required for the consistency of the thresholded ridge regression estimator.

Some simulation results are presented in Section 5 to study the estimation and prediction performance of the proposed method, the ridge regression, the LASSO and the ENET. All technical proofs are given in Section 6.

2. Identifiability and projection. We consider model (1) with deterministic design matrix $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)'$, where the dimension of $\mathbf{x}_i$, p, is larger than n. Let $r = r_n$ be the rank of $\mathbf{X}$. From the singular value decomposition,

$$\mathbf{X} = \mathbf{PDQ}', \tag{2}$$

where $\mathbf{P}$ is an $n \times r$ matrix satisfying $\mathbf{P}'\mathbf{P} = \mathbf{I}_r$, $\mathbf{Q}$ is a $p \times r$ matrix satisfying $\mathbf{Q}'\mathbf{Q} = \mathbf{I}_r$, $\mathbf{I}_a$ denotes the identity matrix of order a and $\mathbf{D}$ is an $r \times r$ diagonal matrix of full rank. Let $\mathbf{Q}_\perp$ be a $p \times (p-r)$ matrix such that $\mathbf{Q}'\mathbf{Q}_\perp = \mathbf{0}$ (the matrix of 0's with an appropriate order) and $\mathbf{Q}_\perp'\mathbf{Q}_\perp = \mathbf{I}_{p-r}$. Throughout, we denote the q-dimensional Euclidean space by $\mathcal{R}^q$ for any positive integer q and the subspace of $\mathcal{R}^p$ generated by the rows of $\mathbf{X}$ by $\mathcal{R}(\mathbf{X})$.

We say that $\boldsymbol{\beta}$ in (1) is identifiable if $\boldsymbol{\beta}_1 \in \mathbf{B}$, $\boldsymbol{\beta}_2 \in \mathbf{B}$ and $\mathbf{X}\boldsymbol{\beta}_1 = \mathbf{X}\boldsymbol{\beta}_2$ imply $\boldsymbol{\beta}_1 = \boldsymbol{\beta}_2$, where $\mathbf{B}$ is the parameter space of $\boldsymbol{\beta}$. The following lemma gives a sufficient and necessary condition for the identifiability of $\boldsymbol{\beta}$.

LEMMA 1. *Under model* (1) *with* $p > r$, $\boldsymbol{\beta}$ *is identifiable if and only if there exists a known function* ϕ *from* $\mathcal{R}^r$ *to* $\mathcal{R}^{p-r}$ *such that*

$$\mathbf{B} = \{\boldsymbol{\beta} : \boldsymbol{\beta} = \mathbf{Q}\boldsymbol{\xi} + \mathbf{Q}_\perp \phi(\boldsymbol{\xi}), \boldsymbol{\xi} \in \mathcal{R}^r\}. \tag{3}$$

Lemma 1 reveals that identifiable $\boldsymbol{\beta}$'s must be in a set having a one-to-one correspondence with $\mathcal{R}(\mathbf{X}) = \{\boldsymbol{\beta} : \boldsymbol{\beta} = \mathbf{Q}\boldsymbol{\xi}, \boldsymbol{\xi} \in \mathcal{R}^r\}$. Since the dimension of the set on the right-hand side of (3) is $r \le n \wedge p$ (the minimum of n and p), $\boldsymbol{\beta}$ is typically not identifiable when $p > n$ and, hence, we are not able to obtain a component-wise consistent estimator of $\boldsymbol{\beta}$. However, we may not need to estimate the entire vector $\boldsymbol{\beta}$, that is, if $\mathbf{X}\boldsymbol{\beta}_1 = \mathbf{X}\boldsymbol{\beta}_2$, we can still estimate parameters related to $\mathbf{X}\boldsymbol{\beta}_1 = \mathbf{X}\boldsymbol{\beta}_2$ and make valid inference without trying to distinguish $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$. Therefore, we consider the projection of $\boldsymbol{\beta}$ onto $\mathcal{R}(\mathbf{X})$, which is what we are able to identify in view of Lemma 1. Define

$$(\mathbf{X}\mathbf{X}')^- = \mathbf{P}\mathbf{D}^{-2}\mathbf{P}',$$

which is $(\mathbf{X}\mathbf{X}')^{-1}$ if $r = n$. The projection of $\boldsymbol{\beta}$ onto $\mathcal{R}(\mathbf{X})$ is

$$\boldsymbol{\theta} = \mathbf{X}'(\mathbf{X}\mathbf{X}')^-\mathbf{X}\boldsymbol{\beta} = \mathbf{Q}\mathbf{Q}'\boldsymbol{\beta}. \tag{4}$$

Note that $\boldsymbol{\theta} \in \mathcal{R}(\mathbf{X})$ and $\boldsymbol{\theta} = \boldsymbol{\beta}$ if and only if $\boldsymbol{\beta} \in \mathcal{R}(\mathbf{X})$. Furthermore, $\mathbf{X}\boldsymbol{\theta} = \mathbf{X}\boldsymbol{\beta}$ and model (1) can be written as

$$y_i = \mathbf{x}_i'\boldsymbol{\theta} + \varepsilon_i, \qquad i = 1, \ldots, n. \tag{5}$$

Thus, estimating $\boldsymbol{\theta}$ is enough for inference about parameters $\mathbf{X}\boldsymbol{\beta} = \mathbf{X}\boldsymbol{\theta}$ and prediction.

The dimension of $\boldsymbol{\theta}$ is still p. When $\boldsymbol{\beta}$ has many zero components, $\boldsymbol{\theta}$ may not have any zero component. However, $\boldsymbol{\theta}$ may have many small components. This can be seen from the L_2-norms of $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$. Since $\boldsymbol{\theta} = \mathbf{Q}\mathbf{Q}'\boldsymbol{\beta}$ and $\mathbf{Q}\mathbf{Q}'$ is a projection matrix, we obtain that $\|\boldsymbol{\theta}\| \le \|\boldsymbol{\beta}\|$, where $\|\cdot\|$ denotes the L_2-norm. This implies that if $\boldsymbol{\beta}$ has many zero components so that the order of $\|\boldsymbol{\beta}\|$ is much smaller than $O(p)$, then the order of $\|\boldsymbol{\theta}\|$ is also much smaller than $O(p)$. Hence, if components of $\boldsymbol{\theta}$ are nonzero, then many of them must be negligible, and $\boldsymbol{\theta}$ can be viewed as a sparse vector. More precise descriptions of this sparsity can be found in conditions (C2) in Section 3 and (C4) in Section 4.

3. The ridge regression estimator of the projection. Since the dimension of $\boldsymbol{\theta}$ in (4) is $p > n$, we consider the ridge regression estimator of $\boldsymbol{\theta}$ [Hoerl and Kennard (1970)] under model (5).

$$\hat{\boldsymbol{\theta}} = (\mathbf{X}'\mathbf{X} + h_n\mathbf{I}_p)^{-1}\mathbf{X}'\mathbf{y},$$

where $\mathbf{y} = (y_1, \ldots, y_n)'$ and $h_n > 0$ is an appropriately chosen regularization parameter. The computation of $\hat{\boldsymbol{\theta}}$ involves only inverting an $n \times n$ matrix. This is because (2) implies that

$$(\mathbf{X}'\mathbf{X} + h_n\mathbf{I}_p)^{-1}\mathbf{X}' = \mathbf{X}'(\mathbf{X}\mathbf{X}' + h_n\mathbf{I}_n)^{-1}, \tag{6}$$

which also implies that the ridge regression estimator $\hat{\boldsymbol{\theta}}$ is always in $\mathcal{R}(\mathbf{X})$. In fact, if $\hat{\boldsymbol{\beta}}$ is the ridge regression estimator of $\boldsymbol{\beta}$ constructed under model (1), then $\hat{\boldsymbol{\theta}} = \mathbf{X}'(\mathbf{X}\mathbf{X}')^{-}\mathbf{X}\hat{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}}$. But $\hat{\boldsymbol{\theta}} = \hat{\boldsymbol{\beta}}$ estimates $\boldsymbol{\theta}$, not the nonidentifiable $\boldsymbol{\beta}$ when $p > n$.

We now study the bias and variance of $\hat{\boldsymbol{\theta}}$ as an estimator of $\boldsymbol{\theta}$, which is essential for establishing asymptotic properties of $\hat{\boldsymbol{\theta}}$. For the matrix $\mathbf{Q}$ given in the singular value decomposition (2), $\boldsymbol{\Gamma} = (\mathbf{Q}\mathbf{Q}_\perp)$ is orthogonal, that is, $\boldsymbol{\Gamma}'\boldsymbol{\Gamma} = \boldsymbol{\Gamma}\boldsymbol{\Gamma}' = \mathbf{I}_p$. Then

$$\begin{aligned}
\operatorname{bias}(\hat{\boldsymbol{\theta}}) &= E(\hat{\boldsymbol{\theta}}) - \boldsymbol{\theta} \\
&= (\mathbf{X}'\mathbf{X} + h_n\mathbf{I}_p)^{-1}\mathbf{X}'\mathbf{X}\boldsymbol{\theta} - \boldsymbol{\theta} \\
&= -(h_n^{-1}\mathbf{X}'\mathbf{X} + \mathbf{I}_p)^{-1}\boldsymbol{\theta} \\
&= -\boldsymbol{\Gamma}(h_n^{-1}\boldsymbol{\Gamma}'\mathbf{X}'\mathbf{X}\boldsymbol{\Gamma} + \mathbf{I}_p)^{-1}\boldsymbol{\Gamma}'\mathbf{Q}\mathbf{Q}'\boldsymbol{\theta} \\
&= -(\mathbf{Q} \quad \mathbf{Q}_\perp)\begin{pmatrix} (h_n^{-1}\mathbf{D}^2 + \mathbf{I}_r)^{-1} & \mathbf{0} \\ \mathbf{0} & \mathbf{I}_{p-r} \end{pmatrix}\begin{pmatrix} \mathbf{Q}' \\ \mathbf{Q}_\perp' \end{pmatrix}\mathbf{Q}\mathbf{Q}'\boldsymbol{\theta} \\
&= -(\mathbf{Q}(h_n^{-1}\mathbf{D}^2 + \mathbf{I}_r)^{-1} \quad \mathbf{Q}_\perp)\begin{pmatrix} \mathbf{Q}'\boldsymbol{\theta} \\ \mathbf{0} \end{pmatrix} \\
&= -\mathbf{Q}(h_n^{-1}\mathbf{D}^2 + \mathbf{I}_r)^{-1}\mathbf{Q}'\boldsymbol{\theta},
\end{aligned}$$

where the fourth equality follows from the fact that $\boldsymbol{\Gamma}$ is orthogonal and $\boldsymbol{\theta} = \mathbf{Q}\mathbf{Q}'\boldsymbol{\beta} = \mathbf{Q}\mathbf{Q}'\boldsymbol{\theta}$. The covariance matrix of $\hat{\boldsymbol{\theta}}$ is given by

$$\begin{aligned}
\operatorname{var}(\hat{\boldsymbol{\theta}}) &= \sigma^2(\mathbf{X}'\mathbf{X} + h_n\mathbf{I}_p)^{-1}\mathbf{X}'\mathbf{X}(\mathbf{X}'\mathbf{X} + h_n\mathbf{I}_p)^{-1} \\
&\le \sigma^2(\mathbf{X}'\mathbf{X} + h_n\mathbf{I}_p)^{-1} \\
&\le \sigma^2 h_n^{-1}\mathbf{I}_p,
\end{aligned}$$

where $\mathbf{A} \le \mathbf{B}$ for nonnegative definite matrices $\mathbf{A}$ and $\mathbf{B}$ means $\mathbf{B} - \mathbf{A}$ is nonnegative definite.

To study the asymptotic properties of $\hat{\boldsymbol{\theta}}$, we consider $n \to \infty$ and $p = p_n$, a function of n. Quantities such as $\boldsymbol{\beta}$, $\mathbf{y}$, $\mathbf{x}_i$, etc., form triangular arrays, but the subscript n is omitted for simplicity. We assume that λ_{1n}, the smallest positive eigenvalue of $\mathbf{X}'\mathbf{X}$, satisfies

(C1) $\qquad \lambda_{1n}^{-1} = O(n^{-\eta}), \qquad \eta \le 1$ and η does not depend on n.

We also need a sparsity condition on $\boldsymbol{\theta}$. From the discussion in the end of Section 2, we conclude that, in terms of the L_2-norm, the sparsity of $\boldsymbol{\beta}$ implies the sparsity of $\boldsymbol{\theta}$. We assume that

(C2) $\qquad \|\boldsymbol{\theta}\| = O(n^{\tau}), \qquad \tau < \eta$ and τ does not depend on n.

If the number of nonzero components of $\boldsymbol{\beta}$ is $O(n^{2\tau})$, and all absolute values of nonzero components of $\boldsymbol{\beta}$ are bounded by a constant M, then (C2) holds since $\|\boldsymbol{\theta}\| \le \|\boldsymbol{\beta}\| \le Mn^{\tau}$.

THEOREM 1. *Assume model* (1) *and conditions* (C1) *and* (C2).

(i) *As* $n \to \infty$, $E(\mathbf{l}'\hat{\boldsymbol{\theta}} - \mathbf{l}'\boldsymbol{\theta})^2 = O(h_n^{-1}) + O(h_n^2 n^{-2(\eta-\tau)})$ *uniformly over* p*-dimensional deterministic vector* $\mathbf{l}$ *with* $\|\mathbf{l}\| = 1$.

(ii) $n^{-1}E\|\mathbf{X}\hat{\boldsymbol{\theta}} - \mathbf{X}\boldsymbol{\theta}\|^2 = O(r_n n^{-1}) + O(h_n^2 n^{-(1+\eta-2\tau)})$.

Note that these results hold without any condition on the dimension p. Theorem 1(i) shows that the mean squared error of $\mathbf{l}'\hat{\boldsymbol{\theta}}$ converges to 0 uniformly in $\mathbf{l}$ if $h_n \to \infty$ and $h_n n^{-(\eta-\tau)} \to 0$. Theorem 1(ii) gives the convergence rate of the expected L_2-norm error $E\|\mathbf{X}\hat{\boldsymbol{\theta}} - \mathbf{X}\boldsymbol{\theta}\|^2$ for estimating $E(\mathbf{y}) = \mathbf{X}\boldsymbol{\theta}$. Since the dimension of $\mathbf{X}\boldsymbol{\theta}$ is n, we say that an estimator $\hat{\boldsymbol{\vartheta}}$ of $\boldsymbol{\theta}$ is L_2-consistent if $n^{-1}E\|\mathbf{X}\hat{\boldsymbol{\vartheta}} - \mathbf{X}\boldsymbol{\theta}\|^2 \to 0$ as $n \to \infty$. Typically, r_n/n does not converge to 0 and, hence, $\mathbf{X}\hat{\boldsymbol{\theta}}$ may not be L_2-consistent.

To elaborate the motivation of using the expected L_2-norm error $E\|\mathbf{X}\hat{\boldsymbol{\vartheta}} - \mathbf{X}\boldsymbol{\theta}\|^2$ as a performance measure for an estimator $\hat{\boldsymbol{\vartheta}}$ of $\boldsymbol{\theta}$, we consider the problem of predicting future y-values on deterministic $\mathbf{X}$. Let $\mathbf{y}_*$ be independent of $\mathbf{y}$ but with the same distribution as $\mathbf{y}$. For deterministic $\mathbf{X}$, it is typical to assess the accuracy of the prediction $\mathbf{X}\hat{\boldsymbol{\vartheta}}$ using the average prediction mean squared error $n^{-1}E\|\mathbf{y}_* - \mathbf{X}\hat{\boldsymbol{\vartheta}}\|^2$. It turns out that

$$n^{-1}E\|\mathbf{y}_* - \mathbf{X}\hat{\boldsymbol{\vartheta}}\|^2 = \sigma^2 + n^{-1}E\|\mathbf{X}\hat{\boldsymbol{\vartheta}} - \mathbf{X}\boldsymbol{\theta}\|^2.$$

Hence, having a small expected L_2-norm error is equivalent to having a small average prediction mean squared error.

4. The thresholded ridge regression estimator. The discussion in the previous section indicates that, although the ridge regression estimator $\hat{\boldsymbol{\theta}}$ is consistent for the estimation of any linear combination of $\boldsymbol{\theta}$, it may not be L_2-consistent, that is, $n^{-1}E\|\mathbf{X}\hat{\boldsymbol{\theta}} - \mathbf{X}\boldsymbol{\theta}\|^2$ may not converge to 0. To achieve L_2-consistency (and good prediction property) under some sparsity conditions on $\boldsymbol{\theta}$, we propose to improve the ridge regression estimator by thresholding.

4.1. *Variable selection.* Let $\mathcal{M}_{\boldsymbol{\beta},0}$ be the set of indices of nonzero components of $\boldsymbol{\beta}$, and let $\widehat{\mathcal{M}}$ be the set of indices of components of $\boldsymbol{\beta}$ selected using a variable selection method. The variable selection method or $\widehat{\mathcal{M}}$ is said to be selection-consistent if and only if

$$\lim_{n\to\infty} P(\widehat{\mathcal{M}} = \mathcal{M}_{\boldsymbol{\beta},0}) = 1. \tag{7}$$

Unlike the case of random $\mathbf{X}$, for deterministic $\mathbf{X}$ with $p > n$, the selection-consistency defined by (7) is generally not achievable because $\boldsymbol{\beta}$ is not identifiable. Some selection-consistency results for the case of $p > n$ and deterministic $\mathbf{X}$ published in the literature are based on very strong and sometimes unrealistic conditions on the design matrix $\mathbf{X}$ to ensure the identifiability of $\boldsymbol{\beta}$. In fact, when $\boldsymbol{\beta}$ is

not identifiable, it is not appropriate to use $\boldsymbol{\beta}$ to describe usefulness of components of $\mathbf{x}_i$, since two different $\boldsymbol{\beta}$ may result in the same responses under model (1). Although components of $\mathbf{x}_i$ corresponding to zero components of $\boldsymbol{\beta}$ are not related to y_i, due to the fact that $\boldsymbol{\beta}$ is unknown and not identifiable, these components of $\mathbf{x}_i$ may still be useful in statistical analysis since we have to use model (5) instead of model (1), that is, $\boldsymbol{\theta}$ instead of $\boldsymbol{\beta}$.

The previous discussion leads to variable selection in terms of the projection vector $\boldsymbol{\theta}$, since any linear combination $\mathbf{l}'\boldsymbol{\beta}$ is estimable if and only if $\mathbf{l}'\boldsymbol{\beta} = \mathbf{l}'\boldsymbol{\theta}$. However, when $\boldsymbol{\beta}$ contains many zero components, $\boldsymbol{\theta}$ may not have any zero component, although many components of $\boldsymbol{\theta}$ may be close to zero. Small but not exactly zero components of $\boldsymbol{\theta}$ do not contribute much in estimation but add variability. Thus, we would like to carry out variable selection in a more general sense as defined by Zhang and Huang (2008), that is, we try to eliminate small components of $\boldsymbol{\theta}$. Condition (C4) stated later may be used to define whether a component of $\boldsymbol{\theta}$ can be treated as small.

We propose to threshold the ridge regression estimator $\hat{\boldsymbol{\theta}}$. Let $\hat{\theta}_j$ be the jth components of $\hat{\boldsymbol{\theta}}$, $j = 1, \ldots, p$. The thresholded ridge regression estimator is defined as $\tilde{\boldsymbol{\theta}}$ whose jth component $\tilde{\theta}_j = \hat{\theta}_j$ if $|\hat{\theta}_j| > a_n$ and $\tilde{\theta}_j = 0$ if $|\hat{\theta}_j| \le a_n$, $j = 1, \ldots, p$, where

$$a_n = C_1 n^{-\alpha}, \qquad 0 < \alpha \le 1/2, C_1 > 0, \tag{8}$$

is the thresholding value with α and C_1 not depending on n. The computation of $\tilde{\boldsymbol{\theta}}$ is easy since it has an explicit form. Thresholding can be viewed as a variable selection procedure; that is, we select components of $\boldsymbol{\theta}$ with indices in $\mathcal{M}_{\hat{\boldsymbol{\theta}},a_n}$, the set of indices of nonzero components of $\tilde{\boldsymbol{\theta}}$. We now study the asymptotic behavior of $\mathcal{M}_{\hat{\boldsymbol{\theta}},a_n}$ under some conditions and appropriate choices of a_n and h_n. A condition on the divergence rate of $p = p_n$ as $n \to \infty$ is

$$p = O(e^{n^\nu}), \qquad 0 < \nu < 1 \text{ and } \nu \text{ does not depend on } n. \tag{C3}$$

If $p = e^{n^\nu}$, it is referred to as the ultra-high dimension [Fan and Lv (2010)].

THEOREM 2. *Assume model* (1) *with normally distributed* ε_i *and conditions* (C1)–(C3). *Let* a_n *be given by* (8) *with* $\alpha < (\eta - \nu - \tau)/3$, $u_n = 1 + (\log\log n)^{-1}$ *and* $h_n = C_2 a_n^{-2}(\log\log n)^3 \log(n \vee p)$, *where* $C_2 > 0$ *is a constant and* $n \vee p$ *is the maximum of* n *and* p. *Then, for any constant* $t > 0$,

$$P(\mathcal{M}_{\boldsymbol{\theta},a_n u_n} \subset \mathcal{M}_{\hat{\boldsymbol{\theta}},a_n} \subset \mathcal{M}_{\boldsymbol{\theta},a_n/u_n}) = 1 - O\big((n \vee p)^{-t}\big), \tag{9}$$

where $\mathcal{M}_{\boldsymbol{\xi},c_n}$ *denotes the set of indices of components of* $\boldsymbol{\xi}$ *whose absolute values are larger than* c_n.

Result (9) shows that, by thresholding $\hat{\boldsymbol{\theta}}$, we can eliminate all components of $\boldsymbol{\theta}$ with absolute values less than a_n/u_n, but keep all components of $\boldsymbol{\theta}$ with absolute

values larger than $a_n u_n$, with probability tending to 1 at the rate of $O((n \vee p)^{-t})$ for any $t > 0$. This rate is at least $O(n^{-t})$ for any $t > 0$ and it is $O(e^{-tn^{\nu}})$ for any $t > 0$ if $\log p$ has exactly the order n^{ν}.

Let q_{n-} and q_{n+} be the numbers of elements in $\mathcal{M}_{\boldsymbol{\theta},a_n u_n}$ and $\mathcal{M}_{\boldsymbol{\theta},a_n/u_n}$, respectively. Then $q_{n-} \le q_{n+}$. Since $u_n \to 1$, it is often true that $q_{n+} - q_{n-} \to 0$ as $n \to \infty$. Then, result (9) implies that

$$P(\mathcal{M}_{\hat{\boldsymbol{\theta}},a_n} = \mathcal{M}_{\boldsymbol{\theta},a_n}) = 1 - O\big((n \vee p)^{-t}\big), \tag{10}$$

which will be referred to as the consistency of $\mathcal{M}_{\hat{\boldsymbol{\theta}},a_n}$. This consistency is weaker than the selection-consistency given by (7), but the latter may not be achieved.

We now consider nonnormal ε_i under model (1), that is, the normality assumption on ε_i is replaced by

$$E(\varepsilon_i^k) < \infty \qquad \text{for an even integer } k \text{ not depending on } n, \tag{M}$$

and condition (C3) is replaced by

$$p = O(n^l), \qquad 1 \le l < k/6 \text{ and } l \text{ does not depend on } n, \tag{C3$'$}$$

while the other conditions, (C1) and (C2), remain the same. When the normality condition is relaxed to the moment condition (M), we cannot handle a dimension at the divergence rate given by (C3), although the polynomial-type divergence rate given by (C3$'$) can still be much larger than n. The integer k in condition (M) has to be sufficiently large so that $3l(t+1)/k < \eta - \tau$, where $t > 0$ is in the convergence rate of $\mathcal{M}_{\hat{\boldsymbol{\theta}},a_n}$.

THEOREM 2A. *Assume model* (1) *and conditions* (M), (C1), (C2) *and* (C3$'$). *For any* $t > 0$, *let* a_n *be given by* (8) *with* $\alpha \le (\eta - \xi - \tau)/3$ *and* $\xi = 3l(t+1)/k$, *and* $u_n = 1 + (\log\log n)^{-1}$. *If* $h_n = C_2 a_n^{-2}(\log\log n)^2 (n \vee p)^{2\xi/(3l)}$, *where* $C_2 > 0$ *is a constant, then result* (9) *holds.*

4.2. *L_2-consistency.* The following result shows that, after the variable selection, the thresholded estimator $\tilde{\boldsymbol{\theta}}$ has asymptotically smaller expected L_2-norm error than $\hat{\boldsymbol{\theta}}$, and it is in fact L_2-consistent, under the following sparsity condition on $\boldsymbol{\theta}$:

$$q_{n+} - q_{n-} \to 0, \qquad q_n/r_n \to 0 \quad \text{and} \quad a_n v_n \to 0, \tag{C4}$$

where

$$v_n = \sum_{j\,:\,|\theta_j| \le a_n} |\theta_j|,$$

θ_j is the jth component of $\boldsymbol{\theta}$, r_n is the rank of $\mathbf{X}$, a_n is given by (8), and q_n, q_{n-} and q_{n+} are, respectively, the numbers of elements in sets $\mathcal{M}_{\boldsymbol{\theta},a_n}$, $\mathcal{M}_{\boldsymbol{\theta},a_n u_n}$ and $\mathcal{M}_{\boldsymbol{\theta},a_n/u_n}$ given by (9).

The last two conditions in (C4) are very similar to condition (2.4) in Zhang and Huang (2008); that is, there exist q_n "large" components of $\boldsymbol{\theta}$ with q_n much smaller than the rank of $\mathbf{X}$, and v_n, the L_1 norm of the "small" components of $\boldsymbol{\theta}$, may diverges to ∞, but at a rate slower than a_n^{-1}.

THEOREM 3. *Assume the conditions in Theorem 2 or 2A. Assume further that (C4) holds and the maximum eigenvalue of* $\mathbf{X}'\mathbf{X}$ *is* $O(n)$. *Then*

$$n^{-1}E\|\mathbf{X}\tilde{\boldsymbol{\theta}} - \mathbf{X}\boldsymbol{\theta}\|^2 = O(q_n n^{-1}) + O(v_n a_n) + O\big(h_n^2 n^{-(1+\eta-2\tau)}\big). \tag{11}$$

Result (11) shows the gain of variable selection by thresholding. The expected L_2-norm error $n^{-1}E\|\mathbf{X}\tilde{\boldsymbol{\theta}} - \mathbf{X}\boldsymbol{\theta}\|^2$ is smaller than $n^{-1}E\|\mathbf{X}\hat{\boldsymbol{\theta}} - \mathbf{X}\boldsymbol{\theta}\|^2$ for sufficiently large n. The former converges to 0 at a certain rate and hence $\tilde{\boldsymbol{\theta}}$ is L_2-consistent, whereas the latter may not converge to 0 when r_n/n does not converge to 0.

If $q_n/n \to 0$, result (11) can also be established with the vector of nonzero components of $\tilde{\boldsymbol{\theta}}$ replaced by the ordinary least squares estimator of the sub-vector of $\boldsymbol{\theta}$ indexed by the set $\mathcal{M}_{\hat{\boldsymbol{\theta}},a_n}$.

4.3. *Tuning parameters.* To apply thresholding, we need to choose the constants C_1 in the thresholding value a_n given by (8) and C_2 in the regularization parameter h_n given in Theorem 2 or 2A. Similar to many other problems, C_1 and C_2 can be viewed as tuning parameters, and there is no optimal way to find their values. Some discussions can be found, for example, in Fan and Lv (2008). It is possible to use a data-driven method to find values of tuning parameters by minimizing the average prediction mean squared error $n^{-1}E\|\mathbf{y}_* - \mathbf{X}\tilde{\boldsymbol{\theta}}\|^2 = \sigma^2 + n^{-1}E\|\mathbf{X}\tilde{\boldsymbol{\theta}} - \mathbf{X}\boldsymbol{\theta}\|^2$.

Let $\psi(C)$ be the average prediction mean squared error when $C = (C_1, C_2)$ is used in a_n and h_n. Since $\psi(C)$ is unknown, we minimize the cross-validation estimator

$$\hat{\psi}(C) = \frac{1}{n}\sum_{i=1}^{n}\big(y_i - \mathbf{x}_i'\tilde{\boldsymbol{\theta}}_{-i}^{(C)}\big)^2,$$

where $\tilde{\boldsymbol{\theta}}_{-i}^{(C)}$ is the thresholded ridge regression estimator of $\boldsymbol{\theta}$ based on the data set with $(y_i, \mathbf{x}_i)$ removed, $i = 1, \ldots, n$. To avoid repeated computation of $\tilde{\boldsymbol{\theta}}_{-i}^{(C)}$, we may use an equivalent formula for $\hat{\psi}(C)$,

$$\hat{\psi}(C) = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{y_i - \mathbf{x}_i'\tilde{\boldsymbol{\theta}}^{(C)}}{1 - w_i(C)}\right)^2, \tag{12}$$

where $w_i(C) = \mathbf{x}_i'(\mathbf{X}'\mathbf{X} + h_n I_p)^{-1}\mathbf{x}_i$ and $\tilde{\boldsymbol{\theta}}^{(C)}$ is the thresholded ridge regression estimator based on the whole data set. This method is applied in the simulation study presented in the next section.

5. Simulation results. With deterministic $\mathbf{X}$ and $p > n$, we examined the L_2-norm errors and the expected L_2-norm errors of the ridge regression estimator, the thresholded ridge regression estimator, and the popular LASSO estimator and ENET estimator (for comparison purpose) in four simulation studies. In the first two simulation studies, the design matrix $\mathbf{X}$ was generated from a multivariate normal distribution but fixed throughout the simulation, which corresponds to analysis conditional on $\mathbf{X}$. In the last two simulation studies, $\mathbf{X}$ is a nearly orthogonal Latin hypercube design or a Latin hypercube design.

5.1. *Simulation study I.* We considered linear model (1) with normally distributed ε_i and $\sigma = 10$. Three sets of sample and variable sizes were considered, $(n, p) = (30, 100)$, $(100, 500)$ and $(200, 2000)$, with increasing ratio p/n. A set of $\mathbf{x}_1, \ldots, \mathbf{x}_n$ were independently generated with $\mathbf{x}_i \sim N(\mathbf{0}, \boldsymbol{\Sigma})$, where the diagonal elements of $\boldsymbol{\Sigma}$ are all equal to 1 and off-diagonal elements of $\boldsymbol{\Sigma}$ are all equal to 0.75. This set of $\mathbf{X}$ was fixed throughout the simulation. The first 20 components of $\boldsymbol{\beta}$ are $1 + 0.1j$ for $j = 1, \ldots, 20$, and the rest of the components of $\boldsymbol{\beta}$ are all equal to 0. The L_2 cumulative proportion plot of the projection vector $\boldsymbol{\theta}$, that is, $\sum_{j=1}^{k} \theta_{(j)}^2 / \|\boldsymbol{\theta}\|^2$, $k = 1, \ldots, p$, is given in Figure 1, where $\theta_{(j)}^2$ is the jth ordered value of $\theta_1^2, \ldots, \theta_p^2$. Although $\boldsymbol{\beta}$ has many zero components, $\boldsymbol{\theta}$ does not have any zero component but many components of $\boldsymbol{\theta}$ are small.

For the thresholded ridge regression estimator, we selected the tuning parameter $C = (C_1, C_2)$ by minimizing $\hat{\psi}(C)$ given by (12). For the ridge regression, LASSO, and ENET estimators, the tuning parameters were selected by a 5-fold cross-validation.

Let $\hat{\boldsymbol{\vartheta}}$ denote the thresholded ridge regression estimator $\tilde{\boldsymbol{\theta}}$, the ridge regression estimator $\hat{\boldsymbol{\theta}}$, the LASSO estimator or the ENET estimator. We independently generated 100 values of $\mathbf{y}$ and obtained 100 values of $n^{-1}\|\mathbf{X}\boldsymbol{\beta} - \mathbf{X}\hat{\boldsymbol{\vartheta}}\|^2$, the L_2-norm error (divided by the sample size). Box plots of 100 values of $n^{-1}\|\mathbf{X}\boldsymbol{\beta} - \mathbf{X}\hat{\boldsymbol{\vartheta}}\|^2$ for four estimation methods are given in Figure 1. The average of 100 values of $n^{-1}\|\mathbf{X}\boldsymbol{\beta} - \mathbf{X}\hat{\boldsymbol{\vartheta}}\|^2$, a simulation approximation to the expected L_2-norm error $n^{-1}E\|\mathbf{X}\boldsymbol{\beta} - \mathbf{X}\hat{\boldsymbol{\vartheta}}\|^2$, is listed in Table 1 for each of the four methods.

5.2. *Simulation study II.* The setting in this study is the same as that in simulation study I except that the values of $\mathbf{x}_i$'s were generated with a $\boldsymbol{\Sigma}$ whose (k, l)th element is equal to $(0.5)^{|k-l|}$ when $|k - l| \le 10$ and 0 when $|k - l| > 10$. The L_2 cumulative proportion plot of $\boldsymbol{\theta}$ and box plots of values of $n^{-1}\|\mathbf{X}\boldsymbol{\beta} - \mathbf{X}\hat{\boldsymbol{\vartheta}}\|^2$ based on 100 simulation runs for four estimation methods are given in Figure 2. The simulation approximations to $n^{-1}E\|\mathbf{X}\boldsymbol{\beta} - \mathbf{X}\hat{\boldsymbol{\vartheta}}\|^2$ are included in Table 1.

5.3. *Simulation study III.* Let NOLH(n, p) denote a nearly orthogonal Latin hypercube design with n rows (runs) and p columns (variables). We considered two sets of n and p. In the first case, $n = 49$, $p = 96$ and $\mathbf{X}$ is an NOLH(49, 96)

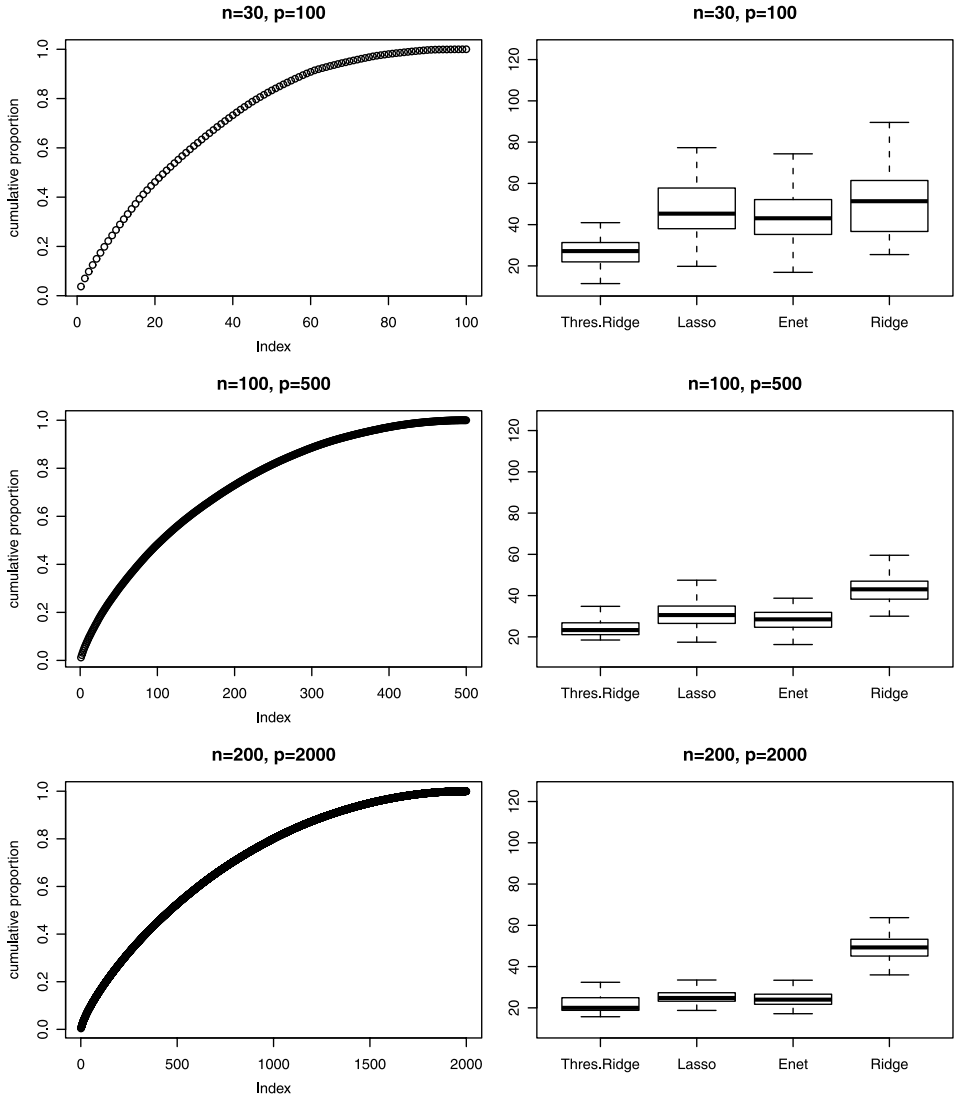

FIG. 1. *Study I*: L_2 *cumulative proportion plot of* $\boldsymbol{\theta}$ *and box plots of* L_2*-norm error for the thresholded ridge regression*, *LASSO*, *ENET and ridge regression*.

constructed by using the orthogonal array-based method in Lin, Mukerjee and Tang (2009). In the second case, $n = 64$, $p = 192$ and $\mathbf{X}$ is an NOLH(64, 192). In both cases, the first 15 components of $\boldsymbol{\beta}$ are equal to $0.2, 0.4, \ldots, 2.8, 3.0$, and the rest components of $\boldsymbol{\beta}$ are equal to 0. The standard deviation of ε_i is 8. The rest of the simulation setting is the same as that in simulation study I. The L_2 cumulative proportion plot of $\boldsymbol{\theta}$ and box plots of values of $n^{-1}\|\mathbf{X}\boldsymbol{\beta} - \mathbf{X}\hat{\boldsymbol{\vartheta}}\|^2$ based on 100

TABLE 1
Simulation approximation to the expected L_2-norm error

Study	n	p	Method			
			Thres. Ridge	**LASSO**	**ENET**	**Ridge**
I	30	100	27.34	48.46	44.56	51.48
	100	500	24.72	32.01	28.46	44.32
	200	2000	21.86	25.37	24.17	49.37
II	30	100	56.50	69.05	70.70	76.05
	100	500	59.35	68.33	64.43	94.06
	200	2000	74.59	85.14	82.35	100.75
III	49	96	61.58	78.40	76.83	85.46
	64	192	54.79	81.54	79.78	78.34
IV	30	100	43.44	55.35	49.29	59.72
	100	500	46.49	56.60	52.83	65.85
	200	2000	48.53	51.78	56.26	71.21

simulation runs for four estimation methods are given in Figure 3. The simulation approximations to $n^{-1}E\|\mathbf{X}\boldsymbol{\beta} - \mathbf{X}\hat{\boldsymbol{\vartheta}}\|^2$ are included in Table 1.

5.4. *Simulation study IV.* The setting in this study is the same as that in simulation study I except that $\mathbf{X}$ is a deterministic Latin hypercube design [McKay, Beckman and Conover (1979)]: each column of $\mathbf{X}$ is a random permutation of n points $6(i/n) - 3$, $i = 1, \ldots, n$, and all columns are generated independently. The L_2 cumulative proportion plot of $\boldsymbol{\theta}$ and box plots of values of $n^{-1}\|\mathbf{X}\boldsymbol{\beta} - \mathbf{X}\hat{\boldsymbol{\vartheta}}\|^2$ based on 100 simulation runs for four estimation methods are given in Figure 4. The simulation approximations to $n^{-1}E\|\mathbf{X}\boldsymbol{\beta} - \mathbf{X}\hat{\boldsymbol{\vartheta}}\|^2$ are included in Table 1.

5.5. *Conclusions based on simulation studies.* From Table 1 and Figures 1–4, we conclude that the thresholded ridge regression estimator is much better than the ridge regression estimator in terms of the L_2-norm error or the expected L_2-norm error, which supports our asymptotic theory, that is, the thresholded ridge regression estimator is L_2-consistent whereas the ridge regression estimator is not. Because the expected L_2-norm error is linearly related to the average prediction mean squared error (Section 3), these results show that thresholding ridge regression has better prediction performance. Except for study III, the LASSO performs worse than the ENET and thresholded ridge regression, but better than the ridge regression, and the ENET performs worse than the thresholded ridge regression, although the difference is small in some cases. Since the ENET uses a combination of L_1- and L_2-penalty, it is not surprising that its performance is between the LASSO and thresholded ridge regression. However, both LASSO and ENET have large variability in simulation study III. It is well known that the LASSO requires

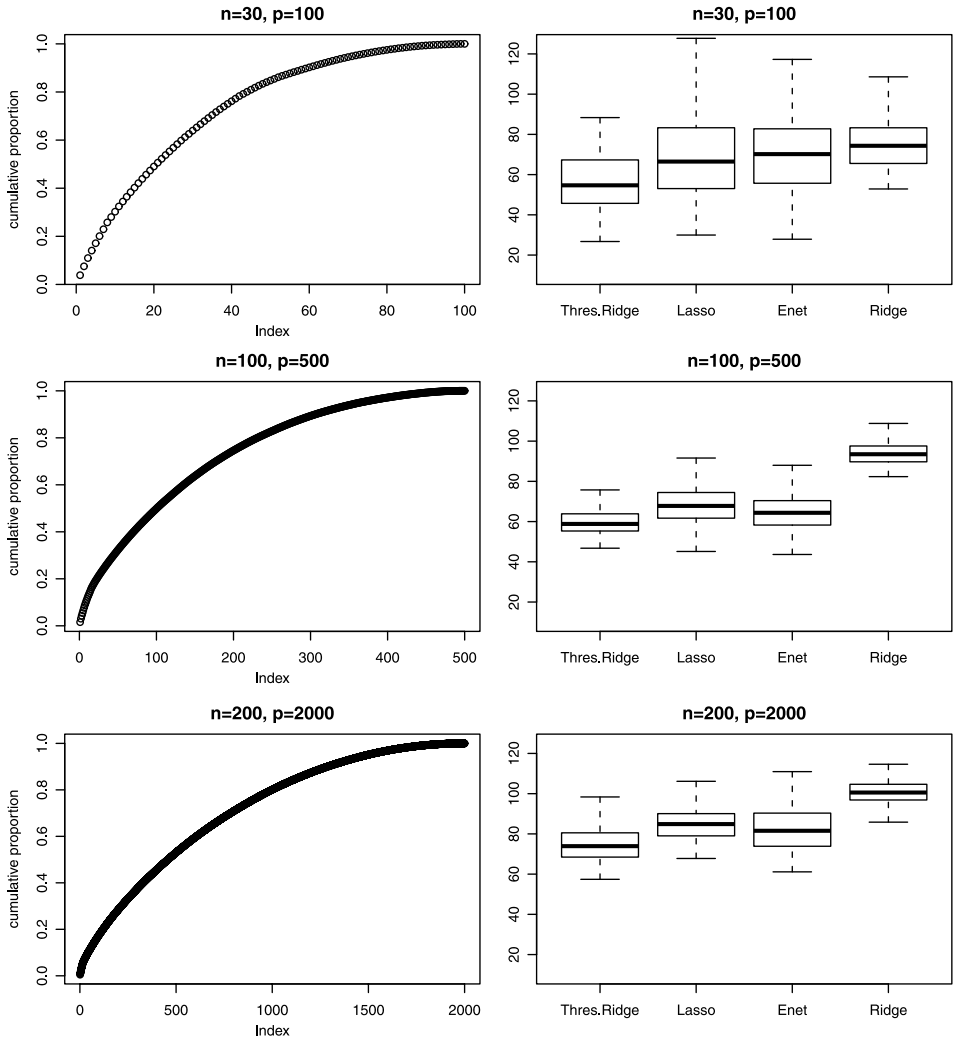

FIG. 2. *Study II*: L_2 *cumulative proportion plot of* $\boldsymbol{\theta}$ *and box plots of* L_2*-norm error for the thresholded ridge regression*, *LASSO*, *ENET and ridge regression*.

more stringent conditions on the design matrix $\mathbf{X}$ [e.g., Zhao and Yu (2006)]. The nearly orthogonal Latin hypercube design in simulation study III may not satisfy these conditions, which results in the poor performance of the LASSO. This also applies to the ENET, since it uses L_1-penalty. Furthermore, no result for the L_2-consistency of LASSO or ENET is available in the situation of deterministic $\mathbf{X}$ and $p > n$.

In terms of the computation, the thresholded ridge regression is much simpler than the LASSO or ENET, especially when p is very large. Because of the iden-

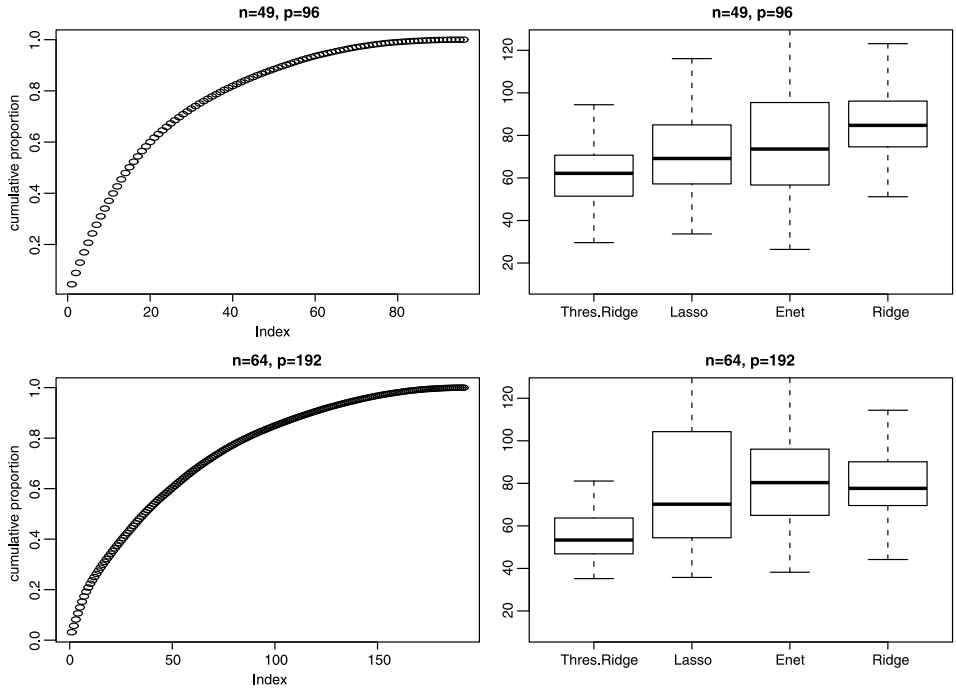

FIG. 3. *Study III*: L_2 *cumulative proportion plot of* $\boldsymbol{\theta}$ *and box plots of* L_2*-norm error for the thresholded ridge regression*, *LASSO*, *ENET and ridge regression*.

tity (6), the computation complexity of the thersholded ridge regression estimator does not increase as p increases.

6. Proofs.

PROOF OF LEMMA 1. Suppose that (3) holds. Let $\boldsymbol{\beta}_j \in \mathbf{B}$, $j = 1, 2$. Then there are $\boldsymbol{\xi}_j \in \mathcal{R}^r$ such that $\boldsymbol{\beta}_j = \mathbf{Q}\boldsymbol{\xi}_j + \mathbf{Q}_\perp \phi(\boldsymbol{\xi}_j)$, $j = 1, 2$. If $\mathbf{X}\boldsymbol{\beta}_1 = \mathbf{X}\boldsymbol{\beta}_2$, then, by (2), $\mathbf{PD}\boldsymbol{\xi}_1 = \mathbf{PD}\boldsymbol{\xi}_2$ and, thus, $\boldsymbol{\xi}_1 = \boldsymbol{\xi}_2$, which implies $\boldsymbol{\beta}_1 = \boldsymbol{\beta}_2$. This shows that the parameter $\boldsymbol{\beta}$ in (1) is identifiable.

Suppose now that $\mathbf{B}$ is not of the form (3). Then, there exist $\boldsymbol{\xi} \in \mathcal{R}^r$, $\boldsymbol{\zeta}_j \in \mathcal{R}^{p-r}$, $j = 1, 2$, $\boldsymbol{\zeta}_1 \neq \boldsymbol{\zeta}_2$ and $\boldsymbol{\beta}_j = \mathbf{Q}\boldsymbol{\xi} + \mathbf{Q}_\perp \boldsymbol{\zeta}_j \in \mathbf{B}$. Then $\boldsymbol{\beta}_1 \neq \boldsymbol{\beta}_2$, but $\mathbf{X}\boldsymbol{\beta}_1 = \mathbf{PD}\boldsymbol{\xi} = \mathbf{X}\boldsymbol{\beta}_2$. This shows that $\boldsymbol{\beta}$ in (1) is not identifiable. □

PROOF OF THEOREM 1.

(i) From Section 3, $\operatorname{bias}(\hat{\boldsymbol{\theta}}) = -\mathbf{Q}(h_n^{-1}\mathbf{D}^2 + \mathbf{I}_r)^{-1}\mathbf{Q}'\boldsymbol{\theta}$. From the facts that $\mathbf{Q}'\mathbf{Q} = \mathbf{I}_r$, $\mathbf{D}^2$ contains positive eigenvalues of $\mathbf{X}'\mathbf{X}$, and $(h_n^{-1}\mathbf{D}^2 + \mathbf{I}_r)^{-1} \leq \frac{h_n/\lambda_{1n}}{1+h_n/\lambda_{1n}}\mathbf{I}_r$, we obtain that $\|\operatorname{bias}(\hat{\boldsymbol{\theta}})\| \leq \|\boldsymbol{\theta}\|(h_n/\lambda_{1n})$. Hence, by (C1) and (C2), $[\mathbf{l}'\operatorname{bias}(\hat{\boldsymbol{\theta}})]^2 \leq \|\operatorname{bias}(\hat{\boldsymbol{\theta}})\|^2 = O(h_n^2 n^{-2(\eta-\tau)})$ uniformly over $\mathbf{l}$ with $\|\mathbf{l}\| = 1$. Also,

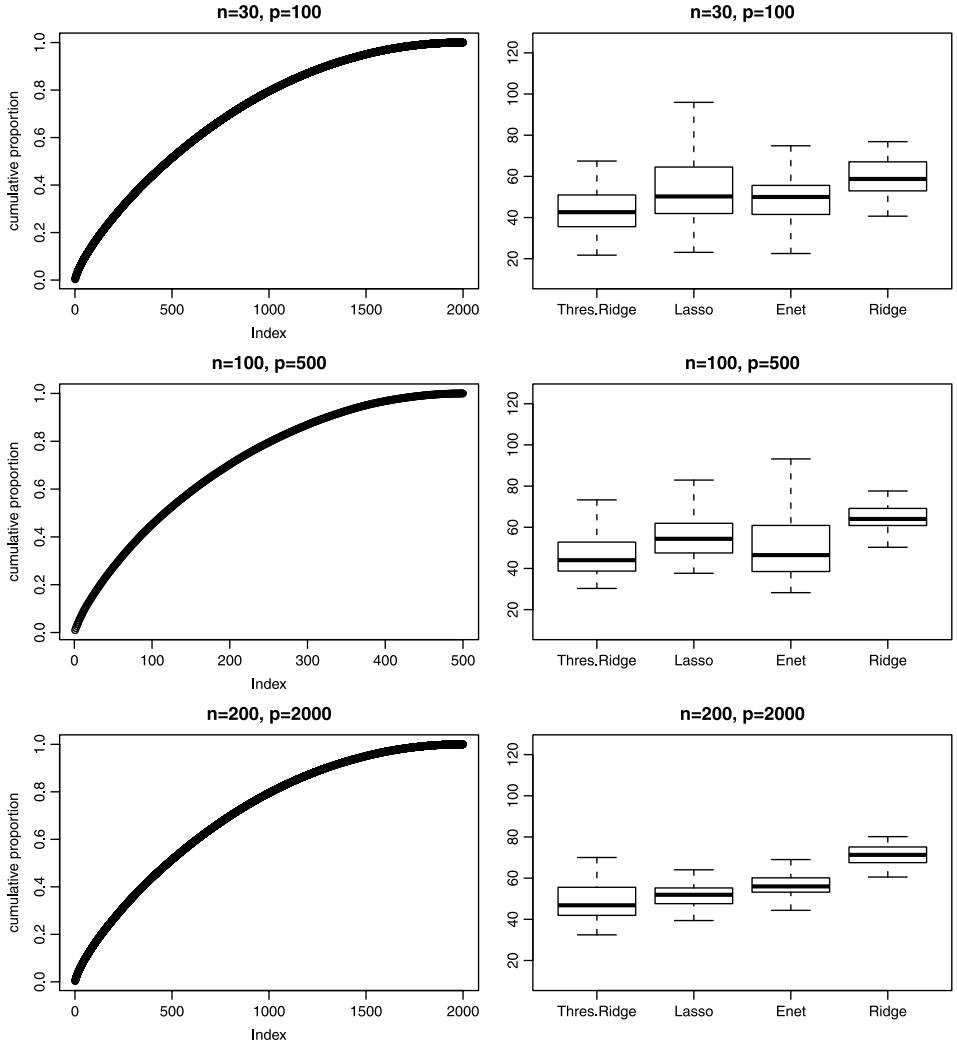

FIG. 4. *Study IV*: L_2 *cumulative proportion plot of* $\boldsymbol{\theta}$ *and box plots of* L_2*-norm error for the thresholded ridge regression*, *LASSO*, *ENET and ridge regression*.

from Section 3, $\operatorname{var}(\hat{\boldsymbol{\theta}}) \le \sigma^2 h_n^{-1} \mathbf{I}_p$. Hence, $\mathbf{l}' \operatorname{var}(\hat{\boldsymbol{\theta}})\mathbf{l} = O(h_n^{-1})$ uniformly over $\mathbf{l}$ with $\|\mathbf{l}\| = 1$. Then, the result follows from $E(\mathbf{l}'\hat{\boldsymbol{\theta}} - \mathbf{l}'\boldsymbol{\theta})^2 = \mathbf{l}' \operatorname{var}(\boldsymbol{\theta})\mathbf{l} + [\mathbf{l}' \operatorname{bias}(\boldsymbol{\theta})]^2$.

(ii) Note that $E\|\mathbf{X}\hat{\boldsymbol{\theta}} - \mathbf{X}\boldsymbol{\theta}\|^2 = \operatorname{trace}[\mathbf{X} \operatorname{var}(\hat{\boldsymbol{\theta}})\mathbf{X}'] + \|\mathbf{X} \operatorname{bias}(\hat{\boldsymbol{\theta}})\|^2$. From the proof of (i),

$$\begin{aligned}
\mathbf{X} \operatorname{var}(\hat{\boldsymbol{\theta}})\mathbf{X}' &\le \sigma^2 \mathbf{X}(\mathbf{X}'\mathbf{X} + h_n \mathbf{I}_p)^{-1}\mathbf{X}' \\
&= \sigma^2 \mathbf{P}\mathbf{D}(\mathbf{D}^2 + h_n \mathbf{I}_r)^{-1}\mathbf{D}\mathbf{P}' \\
&\le \sigma^2 \mathbf{P}\mathbf{P}',
\end{aligned}$$

since $\mathbf{D}(\mathbf{D}^2+h_n\mathbf{I}_r)^{-1}\mathbf{D}$ is a diagonal matrix whose diagonal elements are bounded by 1. Hence, $\operatorname{trace}[\mathbf{X}\operatorname{var}(\hat{\boldsymbol{\theta}})\mathbf{X}'] \le \sigma^2\operatorname{trace}(\mathbf{PP}') = \sigma^2 r_n$. Also,

$$\|\mathbf{X}\operatorname{bias}(\hat{\boldsymbol{\theta}})\|^2 = \boldsymbol{\theta}'\mathbf{Q}(h_n^{-1}\mathbf{D}^2+\mathbf{I}_r)^{-1}\mathbf{D}^2(h_n^{-1}\mathbf{D}^2+\mathbf{I}_r)^{-1}\mathbf{Q}'\boldsymbol{\theta} \le h_n^2\lambda_{1n}^{-1}\|\boldsymbol{\theta}\|^2,$$

which is $O(h_n^2 n^{-(\eta-2\tau)})$ by (C1) and (C2). This completes the proof. □

PROOF OF THEOREM 2. From the proof of Theorem 1,

$$\operatorname{bias}(\hat{\theta}_j) = O(\|\boldsymbol{\theta}\|h_n/\lambda_{1n}) = O(h_n/n^{\eta-\tau})$$

uniformly in $j = 1,\dots,p$. For sufficiently large n, $\log\log n > 0$. With $h_n = C_2 a_n^{-2}(\log\log n)^3\log(n\vee p)$ and condition (C3),

$$\frac{h_n}{n^{\eta-\tau}(u_n-1)a_n} = \frac{C_2(\log\log n)^4\log(n\vee p)}{n^{\eta-\tau}a_n^3} \le \frac{c_1(\log\log n)^4}{n^{\eta-\nu-\tau-3\alpha}}$$

for some constant $c_1 > 0$ and, hence, $|\operatorname{bias}(\hat{\theta}_j)|/[(u_n-1)a_n] \to 0$ uniformly in j when $\alpha < (\eta-\nu-\tau)/3$. Since $\operatorname{var}(\hat{\theta}_j) = O(h_n^{-1})$, there is a constant $c_0 > 0$ such that

$$\frac{|\operatorname{bias}(\hat{\theta}_j)| - (u_n-1)a_n}{[\operatorname{var}(\hat{\theta}_j)]^{1/2}} \le -\sqrt{2}c_0\sqrt{h_n}a_n/(\log\log n).$$

Let Φ be the standard normal distribution function. From (1) with normally distributed ε_i,

$$\begin{aligned} P\big(|\hat{\theta}_j-\theta_j| > (u_n-1)a_n\big) &\le 2\Phi\bigg(\frac{|\operatorname{bias}(\hat{\theta}_j)| - (u_n-1)a_n}{[\operatorname{var}(\hat{\theta}_j)]^{1/2}}\bigg) \\ &\le 2\Phi\big(-\sqrt{2}c_0\sqrt{h_n}a_n/(\log\log n)\big) \\ &\le \exp\{-c_0^2 h_n a_n^2/(\log\log n)^2\}, \end{aligned}$$

for sufficiently large n, where the last inequality follows from $2\Phi(-x) \le e^{-x^2/2}$ for $x \ge 2$ and the fact that $h_n a_n^2/(\log\log n)^2 = C_2\log\log n\log(n\vee p) \to \infty$. Using the same argument, we also obtain that

$$P\big(|\hat{\theta}_j-\theta_j| > (1-u_n^{-1})a_n\big) \le \exp\{-c_0^2 h_n a_n^2/(\log\log n)^2\}$$

for sufficiently large n. Let $t > 0$ be given. For sufficiently large n, $c_0^2 C_2\log\log n - 1 > t$ and, hence,

$$\begin{aligned} P(\mathcal{M}_{\boldsymbol{\theta},a_n u_n} \subset \mathcal{M}_{\hat{\boldsymbol{\theta}},a_n}) &\ge 1 - P\bigg(\bigcup_{j:\,|\theta_j|>u_n a_n}\{|\hat{\theta}_j| \le a_n\}\bigg) \\ &\ge 1 - P\bigg(\bigcup_{j:\,|\theta_j|>u_n a_n}\{|\hat{\theta}_j-\theta_j| > (u_n-1)a_n\}\bigg) \end{aligned}$$

$$\begin{aligned}
&\geq 1 - \sum_{j=1}^{p} P\big(|\hat{\theta}_j - \theta_j| > (u_n - 1)a_n\big) \\
&\geq 1 - p\exp\{-c_0^2 h_n a_n^2/(\log\log n)^2\} \\
&\geq 1 - (n \vee p)^{-t}.
\end{aligned}$$

Similarly, for any $t > 0$,

$$\begin{aligned}
P(\mathcal{M}_{\hat{\boldsymbol{\theta}},a_n} \subset \mathcal{M}_{\boldsymbol{\theta},a_n/u_n}) &\geq P\Bigg(\bigcap_{j\,:\,|\theta_j|\leq a_n/u_n} \{|\hat{\theta}_j| \leq a_n\}\Bigg) \\
&\geq 1 - P\Bigg(\bigcup_{j\,:\,|\theta_j|\leq a_n/u_n} \{|\hat{\theta}_j - \theta_j| > (1 - u_n^{-1})a_n\}\Bigg) \\
&\geq 1 - p\exp\{-c_0^2 h_n a_n^2/(\log\log n)^2\} \\
&\geq 1 - (n \vee p)^{-t}
\end{aligned}$$

for sufficiently large n. This completes the proof. □

PROOF OF THEOREM 2A. From the proof of Theorem 1, we still have $\text{bias}(\hat{\theta}_j) = O(h_n/n^{\eta-\tau})$ uniformly in $j = 1, \ldots, p$. Let ζ_j be the jth component of $(\mathbf{X}'\mathbf{X} + h_n\mathbf{I}_p)^{-1}\sum_{i=1}^{n}\mathbf{x}_i(y_i - \mathbf{x}_i'\boldsymbol{\theta})$. Then, for $u_n = 1 + (\log\log n)^{-1}$,

$$\begin{aligned}
P\big(|\hat{\theta}_j - \theta_j| > (u_n - 1)a_n\big) &\leq \frac{E(\hat{\theta}_j - \theta_j)^k}{[(u_n - 1)a_n]^k} \\
&= O\Bigg(\frac{|\,\text{bias}(\hat{\theta}_j)|^k + E(\zeta_j^k)}{[(u_n - 1)a_n]^k}\Bigg) \\
&= O\Bigg(\frac{h_n^k(\log\log n)^k}{n^{k(\eta-\tau)}a_n^k}\Bigg) + O\Bigg(\frac{(\log\log n)^k}{h_n^{k/2}a_n^k}\Bigg),
\end{aligned}$$

where the last equality follows from $E(\zeta_j^k) = O(h_n^{-k/2})$ [Whittle (1960), Theorem 2]. Similarly,

$$P\big(|\hat{\theta}_j - \theta_j| > (1 - u_n^{-1})a_n\big) = O\Bigg(\frac{h_n^k(\log\log n)^k}{n^{k(\eta-\tau)}a_n^k}\Bigg) + O\Bigg(\frac{(\log\log n)^k}{h_n^{k/2}a_n^k}\Bigg).$$

Using $h_n = C_2 a_n^{-2}(\log\log n)^2(n \vee p)^{2\xi/(3l)}$, we obtain that

$$\begin{aligned}
P(\mathcal{M}_{\hat{\boldsymbol{\theta}},a_n} \subset \mathcal{M}_{\boldsymbol{\theta},a_n/u_n}) &\geq 1 - \sum_{j=1}^{p} P\big(|\hat{\theta}_j - \theta_j| > (1 - u_n^{-1})a_n\big) \\
&= 1 - O\Bigg(\frac{ph_n^k(\log\log n)^k}{n^{k(\eta-\tau)}a_n^k}\Bigg) - O\Bigg(\frac{p(\log\log n)^k}{h_n^{k/2}a_n^k}\Bigg)
\end{aligned}$$

$$
\begin{aligned}
&= 1 - O\biggl(\frac{p(n \vee p)^{2k\xi/(3l)}(\log\log n)^{3k}}{n^{k(\eta-\tau)}a_n^{3k}}\biggr) \\
&\quad - O\biggl(\frac{p}{(n \vee p)^{\xi k/(3l)}}\biggr) \\
&= 1 - O\biggl(\frac{p(n \vee p)^{k\xi/l}(\log\log n)^{3k}}{(n \vee p)^{(t+1)}n^{k(\eta-3\alpha-\tau)}}\biggr) \\
&\quad - O\biggl(\frac{p}{(n \vee p)^{(t+1)}}\biggr) \\
&= 1 - O\biggl(\frac{n^{k\xi}(\log\log n)^{3k}}{(n \vee p)^t n^{k(\eta-3\alpha-\tau)}}\biggr) - O\biggl(\frac{1}{(n \vee p)^t}\biggr) \\
&= 1 - o\bigl((n \vee p)^{-t}\bigr) - O\bigl((n \vee p)^{-t}\bigr) \\
&= 1 - O\bigl((n \vee p)^{-t}\bigr),
\end{aligned}
$$

since $k\xi/(3l) = t + 1$ and $\alpha \le (\eta - \xi - \tau)/3$. Similarly,

$$
P(\mathcal{M}_{\boldsymbol{\theta},a_n u_n} \subset \mathcal{M}_{\hat{\boldsymbol{\theta}},a_n}) \ge 1 - O\bigl((n \vee p)^{-t}\bigr).
$$

Hence, result (9) follows. □

PROOF OF THEOREM 3. Let $A_n = \{\mathcal{M}_{\hat{\boldsymbol{\theta}},a_n} = \mathcal{M}_{\boldsymbol{\theta},a_n}\}$ and A_n^c be its complement. On the set A_n, the number of nonzero components of $\tilde{\boldsymbol{\theta}}$ is the same as q_n. Let $\boldsymbol{\theta}_1$ be $\boldsymbol{\theta}$ with its components smaller than a_n in absolute value set to 0. Under condition (C4) and the condition that $\mathbf{X}'\mathbf{X}$ has a maximum eigenvalue bounded by cn for a constant c,

$$
\begin{aligned}
n^{-1}\|\mathbf{X}\boldsymbol{\theta}_1 - \mathbf{X}\boldsymbol{\theta}\|^2 &\le c\|\boldsymbol{\theta}_1 - \boldsymbol{\theta}\|^2 \\
&= c \sum_{j\,:\,|\theta_j| \le a_n} \theta_j^2 \\
&\le c a_n \sum_{j\,:\,|\theta_j| \le a_n} |\theta_j| \\
&= O(v_n a_n).
\end{aligned}
$$

Hence,

$$
\begin{aligned}
n^{-1}E\|\mathbf{X}\tilde{\boldsymbol{\theta}} - \mathbf{X}\boldsymbol{\theta}\|^2 &\le 2n^{-1}(E\|\mathbf{X}\tilde{\boldsymbol{\theta}} - \mathbf{X}\boldsymbol{\theta}_1\|^2 + \|\mathbf{X}\boldsymbol{\theta}_1 - \mathbf{X}\boldsymbol{\theta}\|^2) \\
&= 2n^{-1}E\|\mathbf{X}\tilde{\boldsymbol{\theta}} - \mathbf{X}\boldsymbol{\theta}_1\|^2 + O(v_n a_n).
\end{aligned}
$$

Then, it remains to show that

$$
n^{-1}E\|\mathbf{X}\tilde{\boldsymbol{\theta}} - \mathbf{X}\boldsymbol{\theta}_1\|^2 = O(q_n n^{-1}) + O(v_n a_n) + O\bigl(h_n^2 n^{-(1+\eta-2\tau)}\bigr). \tag{13}
$$

Following the proof of Theorem 1 we obtain that

$$n^{-1}E[\|\mathbf{X}\tilde{\boldsymbol{\theta}} - \mathbf{X}\boldsymbol{\theta}_1\|^2 I_{A_n}] = O(q_n n^{-1}) + O\big(h_n^2 n^{-(1+\eta-2\tau)}\big),$$

where I_A is the indicator of the set A. From

$$\|\mathbf{X}\tilde{\boldsymbol{\theta}} - \mathbf{X}\boldsymbol{\theta}_1\|^2 I_{A_n^c} \le 2\|\mathbf{X}\tilde{\boldsymbol{\theta}} - \mathbf{X}\hat{\boldsymbol{\theta}}\|^2 I_{A_n^c} + 2\|\mathbf{X}\hat{\boldsymbol{\theta}} - \mathbf{X}\boldsymbol{\theta}_1\|^2 I_{A_n^c}$$

and Theorem 1, result (13) follows if we can show that

$$n^{-1}E\|\mathbf{X}\tilde{\boldsymbol{\theta}} - \mathbf{X}\hat{\boldsymbol{\theta}}\|^2 I_{A_n^c} = o\big(q_n n^{-1} \vee h_n^2 n^{-(1+\eta-2\tau)}\big).$$

Since

$$\|\mathbf{X}\tilde{\boldsymbol{\theta}} - \mathbf{X}\hat{\boldsymbol{\theta}}\|^2 = (\tilde{\boldsymbol{\theta}} - \hat{\boldsymbol{\theta}})'\mathbf{X}'\mathbf{X}(\tilde{\boldsymbol{\theta}} - \hat{\boldsymbol{\theta}}) \le O(n)\|\tilde{\boldsymbol{\theta}} - \hat{\boldsymbol{\theta}}\|^2 \le O(a_n^2 pn),$$

the result follows from $P(A_n^c) = O((n \vee p)^{-t})$ for any $t > 0$ according to Theorem 2 or 2A. This completes the proof. □

Acknowledgments. The authors would like to thank a referee and an Associate Editor for their helpful comments and suggestions.

REFERENCES

FAN, J. and LI, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *J. Amer. Statist. Assoc.* **96** 1348–1360. MR1946581

FAN, J. and LV, J. (2008). Sure independence screening for ultrahigh dimensional feature space. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **70** 849–911. MR2530322

FAN, J. and LV, J. (2010). A selective overview of variable selection in high dimensional feature space. *Statist. Sinica* **20** 101–148. MR2640659

FAN, J. and PENG, H. (2004). Nonconcave penalized likelihood with a diverging number of parameters. *Ann. Statist.* **32** 928–961. MR2065194

HOERL, A. E. and KENNARD, R. W. (1970). Ridge regression, biased estimation for nonorthogonal problems. *Technometrics* **12** 55–67.

HUNTER, D. R. and LI, R. (2005). Variable selection using MM algorithms. *Ann. Statist.* **33** 1617–1642. MR2166557

LIN, C. D., MUKERJEE, R. and TANG, B. (2009). Construction of orthogonal and nearly orthogonal Latin hypercubes. *Biometrika* **96** 243–247. MR2482150

MCKAY, M. D., BECKMAN, R. J. and CONOVER, W. J. (1979). A comparison of three methods for selecting values of input variables in the analysis of output from a computer code. *Technometrics* **21** 239–245. MR0533252

MEINSHAUSEN, N. and BÜHLMANN, P. (2006). High-dimensional graphs and variable selection with the lasso. *Ann. Statist.* **34** 1436–1462. MR2278363

MEINSHAUSEN, N. and YU, B. (2009). Lasso-type recovery of sparse representations for high-dimensional data. *Ann. Statist.* **37** 246–270. MR2488351

TIBSHIRANI, R. (1996). Regression shrinkage and selection via the lasso. *J. Roy. Statist. Soc. Ser. B* **58** 267–288. MR1379242

WANG, H. (2009). Forward regression for ultra-high dimensional variable screening. *J. Amer. Statist. Assoc.* **104** 1512–1524. MR2750576

WANG, H., LI, R. and TSAI, C.-L. (2007). Tuning parameter selectors for the smoothly clipped absolute deviation method. *Biometrika* **94** 553–568. MR2410008

WHITTLE, P. (1960). Bounds for the moments of linear and quadratic forms in independent variables. *Theory Probab. Appl.* **5** 302–305.

ZHANG, C.-H. and HUANG, J. (2008). The sparsity and bias of the LASSO selection in high-dimensional linear regression. *Ann. Statist.* **36** 1567–1594. MR2435448

ZHAO, P. and YU, B. (2006). On model selection consistency of Lasso. *J. Mach. Learn. Res.* **7** 2541–2563. MR2274449

ZOU, H. (2006). The adaptive lasso and its oracle properties. *J. Amer. Statist. Assoc.* **101** 1418–1429. MR2279469

ZOU, H. and HASTIE, T. (2005). Regularization and variable selection via the elastic net. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **67** 301–320. MR2137327

SCHOOL OF FINANCE AND STATISTICS
EAST CHINA NORMAL UNIVERSITY
500 DONGCHUAN RD.
SHANGHAI, 200241
CHINA
AND
DEPARTMENT OF STATISTICS
UNIVERSITY OF WISCONSIN
1300 UNIVERSITY AVE.
MADISON, WISCONSIN 53706
USA
E-MAIL: shao@stat.wisc.edu

DEPARTMENT OF STATISTICS
VIRGINIA POLYTECHNIC INSTITUTE
AND STATE UNIVERSITY
211 HUTCHESON HALL
BLACKSBURG, VIRGINIA 24061
USA
E-MAIL: xdeng@vt.edu