Statistica Sinica **34** (2024), 2089-2113
doi:https://doi.org/10.5705/ss.202022.0065

ONE-STEP REGULARIZED ESTIMATOR FOR HIGH-DIMENSIONAL REGRESSION MODELS

Yi Wang*[1,2], Donglin Zeng[3], Yuanjia Wang[4] and Xingwei Tong[5]

[1]*Shanghai University of International Business and Economics,*
[2]*Peking University,* [3]*University of Michigan,*
[4]*Columbia University and* [5]*Beijing Normal University*

Abstract: Statistical inference for high-dimensional regression models is a challenging problem. Existing methods focus on inference for finite-dimensional components of the model parameters. Constructing the parameter estimators and establishing the asymptotic inference are specific to each model. In this study, we treat a high-dimensional model as a special case of a semiparametric model. We propose a general framework for constructing one-step regularized estimators for any smooth functional of high-dimensional parameters, which can be viewed as an extension of the one-step efficient estimator for semiparametric models to an M-estimation in the high-dimensional model setting. We show that the proposed estimator is asymptotically normal under some general regularity conditions. We apply the proposed method to an inference for the coefficients in a high-dimensional lasso regression, and to determine the l^2-norm of the functional coefficients in a high-dimensional additive model, allowing the number of covariates to grow exponentially with the sample size. A simulation study and a microarray data example are presented to demonstrate the performance of the proposed method.

Key words and phrases: Confidence intervals, high-dimension regression, M-estimation, one-step regularized estimators, semiparametric model.

1. Introduction

In high-dimensional regression models, the logarithm of the number of covariates can grow at a polynomial rate as the sample size increases, and many statistical methods have been developed for both model prediction and variable selection. For linear models, a regularized or penalized least-square estimation is widely used to handle high-dimensional covariates. Examples include the least absolute shrinkage and selection operator (lasso) (Tibshirani (1996)), smoothly clipped absolute deviation (SCAD) (Fan and Li (2001)), and minimax concave penalty (MCP) (Zhang (2010)). Theoretical properties on the oracle properties of variable selection are given for lasso estimators in Meinshausen and Bühlmann (2006), Zhao and Yu (2006), and Wainwright (2009), and later established in Fan and Lv (2011) and Bradic, Fan and Wang (2011)

*Corresponding author.

for a general concave penalty. The oracle properties of variable selection are also obtained by van de Geer (2008) for high-dimensional generalized linear models, and by Huang, Horowitz and Wei (2010) for nonparametric additive models (NAMs) in a high-dimensional setting. However, statistical inference for high-dimensional regression models remains a challenging problem, and traditional inference results may not hold for high-dimensional estimators. For example, it has been shown that the lasso estimator is not root-n consistent (Candes and Tao (2007); Zhang and Huang (2008); Bühlmann and van de Geer (2011)). It is also well known that no post-selection estimators are locally regular estimators. Knight and Fu (2000) point out that, even in a low-dimensional case, the asymptotic distribution of the lasso estimator is not normal, and Chatterjee and Lahiri (2010) show that an inference based on bootstrap methods may fail.

A growing number of studies are trying to determine how to obtain a correct inference in high-dimensional regression models. Some methods propos modified bootstrap procedures for inference (Chatterjee and Lahiri (2011); Dezeure, Bühlmann and Zhang (2017)) or focus on conditional inference post-selection (Lockhart et al. (2014); Taylor et al. (2014); Lee et al. (2016); Yang et al. (2016)). Belloni, Chernozhukov and Hansen (2014) introduce post-selection to structural and treatment effects, which they refer to as "double selection", and Kozbur (2020) extend this to an additive model. As an alternative, some methods propose improving existing estimators to yield a regular inference asymptotically. For example, Zhang and Zhang (2014) propose a low-dimensional projection approach to obtain the confidence intervals for finite-dimensional parameters in a high-dimensional linear model. The key idea is to project model-based residuals onto the linear space of the covariates with coefficients that were not of interest for inference, and then to remove this projection from the initial estimators. This procedure, called "debiasing", obtains a new estimator that is locally regular and asymptotically normal. Since then, this idea has been used in many high-dimensional settings to obtain valid confidence intervals for prespecified parameters of interest, with different ways of constructing the projections. van de Geer et al. (2014) study a debiased estimator for high-dimensional generalized linear models with a convex loss function. Ren et al. (2015) extend this idea to a Gaussian graphical model, and Ning and Liu (2017) propose a de-correlated score function, in the same spirit as debiasing, with a Dantzig-type estimator to handle more general likelihood functions with high-dimensional parameters. Other extensions include significance tests for a finite-dimensional subset of the model parameters, under constraints (Yu, Gupta and Kolar (2019)), statistical inference based on post-selection for partial linear models (Fei et al. (2019)), debiased estimators for high-dimensional graph-based linear models (Wang and Loh (2020)), and combining a bootstrap with debiased lasso estimators (Zhang and Cheng (2017)), thus improving the estimation of single component in high-dimensional additive models with a debiased modification (Gregory, Mammen

and Wahl (2016)). Chernozhukov, Hansen and Spindler (2015) introduce an orthogonal estimating equation for inference on a single component of high-dimensional parameters. Chernozhukov et al. (2016) give a general construction of moment functions for the generalized method of moments (GMM). More recently, Chernozhukov et al. (2018) considered a debiased estimation based on a Neyman orthogonal score function for treatment effect estimation. Bach et al. (2020) use an orthogonal score function to obtain confidence bands for a single component in additive models. Lu, Kolar and Liu (2020) combine this idea with kernel estimation, and propose a kernel-sieve hybrid regression estimator. These methods all focus on inference for one or a finite number of coefficients in high-dimensional regression models. Furthermore, the construction of the debiasing methods is specific to each model. However, there is no general guidance on how to obtain asymptotically regular estimators for a finite-dimensional functional of the parameters (finite-dimensional components are special finite-dimensional functionals) in general high-dimensional regression models.

In this study, we fill this gap by providing a general theory and framework for performing an inference for any smooth functionals of the parameters in a high-dimensional regression setting. Specifically, we cast high-dimensional regression models as a special case of general semiparametric models, which allow the parameters to be of infinite dimension. An estimation for a high-dimensional model based on, for instance, a penalized least-squares or likelihood, is essentially a special type of constrained or sieve M-estimation in the semiparametric context, which has been studied extensively (e.g., Geman and Hwang (1982); Newey and Powell (2003); Chen (2007); Chen and Shen (1998); Shen and Wong (1994)). Furthermore, an inference for one particular coefficient in high-dimensional models is equivalent to inference for some smooth functional of the parameters in semiparametric models. From this point of view, we propose a general one-step regularized estimator (OSRE) based on semiparametric efficiency theory, with an extension from likelihood-based estimation to more general M-estimation. The proposed estimator reduces to commonly used debiased estimators under high-dimensional linear models and a decorrelated score function. It is also equivalent to a linear approximation of the Neyman orthogonal score function proposed by Chernozhukov et al. (2018).

The main contribution of this work is that we provide general regularity conditions to show that the proposed estimators have an asymptotically linear expansion, so that the distributions are locally regular and asymptotically normal. This leads to a unified approach for testing a high-dimensional regression model using a one-step regularization. As an additional contribution, using a high-dimensional linear model and a NAM as examples, we show that our proposed estimators lead to correct inference for some functionals of the parameters, for example, the total sum of the squared coefficients in the linear model, and the l^2-norm of one functional component in the additive model, even if the dimension

of the covariates is power-exponential of the sample size. To the best of our knowledge, our study is the first to obtain such results for these models. Similarly, this kind of extension can be extended to other cases, making traditional debiased methods more general.

2. Method

2.1. General M-estimation setup

We assume that the data consist of n independent and identically distributed (i.i.d.) observations, $\boldsymbol{Z}_i^{(n)} = (\boldsymbol{X}_i^{(n)}, Y_i)$, for $i = 1, \ldots, n$, where $\boldsymbol{X}^{(n)}$ denotes p_n-dimensional covariates, Y denotes the outcome of interest, and $\boldsymbol{Z}^{(n)}$ follows a probability measure P^n in $\mathbb{R}^{p_n+1}$. We let $\mathcal{Z}^{(n)}$ be the support of $\boldsymbol{Z}_i^{(n)}$. Here, P^n, $\boldsymbol{X}^{(n)}$ and $\boldsymbol{Z}^{(n)}$ may vary with the sample size n, but to simplify the notation, we write P for P^n, $\boldsymbol{X}$ for $\boldsymbol{X}^{(n)}$, and $\boldsymbol{Z}$ for $\boldsymbol{Z}^{(n)}$ in the remainder of this work. For all high-dimensional regression problems, the main goal is to find a prediction function, $f(\boldsymbol{X})$, for the outcome Y. The true optimal prediction function, denoted by f_{n0}, maximizes the expectation of some objective function indexed by f, denoted as $m(\boldsymbol{Z}, f)$, and is assumed to be unique. That is, $P\{m(\boldsymbol{Z}, f_{n0})\} > P\{m(\boldsymbol{Z}, f)\}$, for all $f(\boldsymbol{X}) \neq f_{n0}(\boldsymbol{X})$, with nonzero probability. For our method, we assume that f_{n0} belongs to a known space $\mathcal{F}_n$ that is a Hilbert space consisting of measurable functions of $\boldsymbol{Z}$ equipped with the inner product $\langle \cdot, \cdot \rangle_n$ and the norm $\| \cdot \|_{(n)}$.

In high-dimensional regression settings, when p_n is larger than n, estimating f_{n0} is usually performed by maximizing a regularized empirical version of the objective function, which is $\mathbb{P}_n\{m(\boldsymbol{Z}, f)\}$ minus a penalty function of f. Here, $\mathbb{P}_n$ denotes the empirical measure based on n observations. Such an estimation is equivalent to maximizing $\mathbb{P}_n\{m(\boldsymbol{Z}, f)\}$ in a constrained set for f. Hence, we consider the estimation problem in high-dimensional regression problems as a constrained M-estimation, that is,

$$\widehat{f}_n \equiv \max_{f \in \mathcal{F}_{ns}} \mathbb{P}_n m(\boldsymbol{Z}, f),$$

where $\mathcal{F}_{ns}$ is the constrained set in $\mathcal{F}_n$. The resulting estimator, $\widehat{f}_n$, is called the sieve estimator of the M-estimation in the semiparametric context.

As an example, in a linear model, f is a linear combination of $\boldsymbol{X}$ (including a constant) and $m(\boldsymbol{Z}, f) = -(Y - f(\boldsymbol{X}))^2/2$. Moreover, $\mathcal{F}_n$ consists of all linear functions of $\boldsymbol{X}$ in $L^2(P)$ with the same inner product inherited from the $L^2(P)$ space. When the lasso is used for estimation, the constrained set $\mathcal{F}_{ns}$ contains all functions in $\mathcal{F}_n$ with coefficients that have an l_1-norm bounded by a constant. In a generalized linear model, everything is the same, except that $m(\cdot, f)$ is from the log-likelihood function given by the model. In another example of a high-dimensional NAM studied in Huang, Horowitz and Wei (2010), f is a summation

of univariate functions for each variable in $\boldsymbol{X}$, and $\mathcal{F}_n$ is the subspace of such functions in $L^2(P)$. When constructing their estimator, they restrict f to the constrained set $\mathcal{F}_{ns}$, which is a linear space of univariate spline bases in which the coefficients of these bases have a bounded l_1-norm.

2.2. The OSRE

Our goal is to make an inference for a finite-dimensional functional of f_{n0} based on $\widehat{f}_n$, defined as $\theta_{n0} \equiv \mathfrak{F}_n(f_{n0})$. To introduce our proposed one-step regularized approach, we first assume the following conditions:

A.1 Assume that $\mathfrak{F}_n$ has a continuous Hadamard derivative at f_{n0}, which is assumed to be in the interior of $\mathcal{F}_n$, denoted as $\nabla\mathfrak{F}_n(f_{n0})$, and its Hadamard derivative in the direction $v \in \mathcal{F}_n$ is defined as

$$\nabla\mathfrak{F}_n(f_{n0})[v] = \frac{\partial\mathfrak{F}_n(f_{n0}+\tau v)}{\partial\tau}\Big|_{\tau=0}.$$

A.2 Assume that $m(\boldsymbol{Z}, f)$ has a second-order Hadamard derivative at $f_{n0} \in \mathcal{F}_n$, denoted by $\nabla^2 m(\boldsymbol{Z}, f_{n0})$, which is a bounded bilinear operator, defined as

$$P\left\{\nabla^2 m(\boldsymbol{Z}, f_{n0})[h_1, h_2]\right\} = P\left\{\frac{\partial\left[\nabla m(\boldsymbol{Z}, f_{n0}+\tau h_2)[h_1]\right]}{\partial\tau}\Big|_{\tau=0}\right\},$$

for $h_1, h_2 \in \mathcal{F}_n$.

A.3 Define

$$\mathcal{N}_{f_n,\epsilon} = \{g \in \mathcal{F}_n : d_{(n)}(g, f_n) \le \epsilon\}, \text{ for } f_n \in \mathcal{F}_n,$$

which is the neighborhood of f_n. Let $\mathcal{V}_{(n)}$ be the closed linear span of $\{f - f_{n0} : f \in \mathcal{N}_{f_n,\epsilon}\}$. We assume that there exists $h_n^* \in \mathcal{F}_n$ such that

$$P\left\{\nabla^2 m(\boldsymbol{Z}, f_{n0})[h_n^*, v]\right\} = \langle v_n^*, v\rangle_{(n)}, \text{ for all } v \in \mathcal{V}_{(n)}, \tag{2.1}$$

where $v_n^* \in \mathcal{V}_{(n)}$ is the Riesz representor satisfying $\nabla\mathfrak{F}_n(f_{n0})[v] = \langle v_n^*, v\rangle_{(n)}$, for all $v \in \mathcal{V}_{(n)}$, and it exists and is unique , from Condition A.1. Note that h_n^* and v_n^* have the same number of components as the dimension of θ_n. The inner product is the summed inner product between each component pair.

Remark 1. Conditions A.1 and A.2 both require smoothness of the objective functional and functional parameter of interest. Condition A.3 is the key assumption for developing our proposed estimators. From the Riesz representation theorem, $P\{\nabla^2 m(\boldsymbol{Z}, f_{n0})[h_n^*, v]\}$ can be written as $\langle \mathcal{M}_n[h_n^*], v\rangle_n$, for some linear operator $\mathcal{M}_n$. Thus, Condition A.3 is equivalent to the invertibility of $\mathcal{M}_n$, and h_n^* is given as $\mathcal{M}_n^{-1}\nabla\mathfrak{F}(f_{n0})$. The direction h_n^* is an analogue to the least favorable

direction in a semiparametric likelihood inference, where the m-function is the log-likelihood function and $\mathcal{M}_n$ corresponds to the negative information operator. For additional details about the connection between our proposed method and semiparametric models, see Section S1 in the Supplementary Material.

We now introduce the OSRE. Supposing $d_{(n)}(\widehat{f}_n, f_{n0})$ converges to zero in probability, we have

$$\mathfrak{F}_n(\widehat{f}_n) - \mathfrak{F}_n(f_{n0}) = \left\langle v_n^*, \widehat{f}_n - f_{n0} \right\rangle_{(n)} + O_p\left(d^2_{(n)}(\widehat{f}_n, f_{n0})\right). \tag{2.2}$$

Because f_{n0} maximizes $P\{m(\boldsymbol{Z}, f)\}$, we have $P\{\nabla m(\boldsymbol{Z}, f_{n0})[h]\} = 0$ and Condition A.2,

$$\begin{aligned} P\left\{\nabla m(\boldsymbol{Z}, \widehat{f}_n)[h]\right\} &= P\left\{\nabla m(\boldsymbol{Z}, \widehat{f}_n)[h]\right\} - P\left\{\nabla m(\boldsymbol{Z}, f_{n0})[h]\right\} \\ &= P\left\{\nabla^2 m(\boldsymbol{Z}, f_{n0})[h, \widehat{f}_n - f_{n0}]\right\} + O_p\left(d^2_{(n)}(\widehat{f}_n, f_{n0})\right) \end{aligned}$$

for any $h \in \mathcal{F}_n$. In particular, we choose $h = h_n^*$ satisfying (2.1), as given in Condition A.3. Thus, from (2.2), we conclude

$$\mathfrak{F}_n(f_{n0}) = \mathfrak{F}_n(\widehat{f}_n) - P\left\{\nabla m(\boldsymbol{Z}, \widehat{f}_n)[h_n^*]\right\} + O_p\left(d^2_{(n)}(\widehat{f}_n, f_{n0})\right).$$

The last term on the right-hand side of the equation is of order $d^2_{(n)}(\widehat{f}_n, f_{n0})$. Therefore, the second term on the right-hand side, $P\{\nabla m(\boldsymbol{Z}, \widehat{f}_n)[h_n^*]\}$, can be considered as the bias from using $\mathfrak{F}_n(\widehat{f}_n)$ to estimate $\mathfrak{F}_n(f_{n0})$, which may not be negligible in high-dimensional settings. This motivates the construction of the OSRE, as follows: given that $\widehat{h}_n$ is a consistent estimator for h_n^*, our proposed estimator for θ_{n0} is defined as

$$\widetilde{\theta}_n = \widehat{\theta}_n - \mathbb{P}_n\left\{\nabla m(\boldsymbol{Z}, \widehat{f}_n)[\widehat{h}_n]\right\}, \tag{2.3}$$

where $\widehat{\theta}_n = \mathfrak{F}_n(\widehat{f}_n)$ is the plug-in estimator based on $\widehat{f}_n$. Because (2.3) is a one-step update for the initial estimator $\widehat{\theta}_n$, we call the proposed estimator OSRE for θ_{n0}.

Remark 2. The Neyman orthogonal score function in Chernozhukov et al. (2018) requires that the score function ψ satisfies

$$\nabla_\eta P[\psi(Z; \theta_0, \eta_0)[\eta - \eta_0]] = 0,$$

where η is the nuisance parameter, and θ_0 and η_0 are true parameters. Because h_n^* satisfies

$$\nabla^2 P[m(Z; \theta_0, \eta_0)[h_n^*, (\theta - \theta_0, \eta - \eta_0)] = \theta - \theta_0 + O_p(n^{-1/2}),$$

when $\|(\theta-\theta_0,\eta-\eta_0)\|_2^2 = O_p(n^{-1/2})$, we have

$$\nabla_\eta \nabla P[m(Z;\theta_0,\eta_0)[h_n^*,\eta-\eta_0] = 0 + O_p(n^{-1/2}),$$

when $\|(\theta-\theta_0,\eta-\eta_0)\|_2^2 = O_p(n^{-1/2})$. Thus, our proposed method can be viewed as a linear approximation of the Neyman orthogonal score function in the neighborhood of (θ_0,η_0).

2.3. General asymptotic properties for the OSRE

Here, we provide regularity conditions and establish asymptotic results for the proposed OSRE. In addition to Conditions A.1-A.3, we further assume the following:

A.4 The initial estimator, $\widehat{f}_n$, satisfies $d_{(n)}(\widehat{f}_n, f_{n0}) = o_p(n^{-1/4})$.

A.5 There exists an estimator, $\widehat{h}_n$, for h_n^* such that $d_{(n)}(\widehat{h}_n, h_n^*) = o_p(n^{-1/4})$.

A.6 For every $\epsilon, \eta > 0$, there exist $\delta_1, \delta_2 > 0$ such that

$$\lim_n P\left(\sup_{\substack{f_1,f_2\in\mathcal{N}_{f_{n0},\delta_1},\\ h_1,h_2\in\mathcal{N}_{h_n^*,\delta_2}}} \|\mathbb{G}_n\{\nabla m(\boldsymbol{Z},f_1)[h_1]\} - \mathbb{G}_n\{\nabla m(\boldsymbol{Z},f_2)[h_2]\}\|_\infty > \epsilon \right) < \eta,$$

where $\|\boldsymbol{A}\|_\infty = \max_{i,j}|a_{ij}|$ for any matrix $A=(a_{ij})$, $\mathbb{G}_n = n^{1/2}(\mathbb{P}_n - P)$ denotes the empirical process, and $\mathcal{N}_{f,\delta}$ is the δ-neighborhood of f, as defined in Condition A.3.

A.7 When n goes to infinity, $\mathrm{Var}(\nabla m(\boldsymbol{Z},f_{n0})[h_n^*])$ converges to a positive-definite matrix Σ.

Remark 3. Conditions A.4 and A.5 related to the convergence rates for the initial estimator $\widehat{f}_n$ and the estimator $\widehat{h}_n$, respectively. As shown later, these conditions are possible even under high-dimensional settings when p_n is much larger than n. Condition A.6 implies the asymptotically uniform equicontinuity of the empirical process in some neighborhoods of f_{n0} and h_n^*, and holds if some additional function complexity can be established in these neighborhoods.

Theorem 1. *Under Conditions* A.1–A.7, $\widetilde{\theta}_n$ *has an asymptotically linear expansion as*

$$\sqrt{n}(\widetilde{\theta}_n - \theta_{n0}) = -\mathbb{G}_n\{\nabla m(\boldsymbol{Z},f_{n0})[h_n^*]\} + o_p(1).$$

As a result, $\widetilde{\theta}_n$ *is asymptotically regular, and its asymptotic distribution is a multivariate normal distribution with mean zero and covariance matrix* Σ.

Theorem 1 states that the OSRE is asymptotically normal and the variance is

$$\lim_{n\to\infty} \mathrm{Var}\left(\nabla m(\boldsymbol{Z}, f_{n0})[h_n^*]\right).$$

A proof of Theorem 1 is provided in Section S2 of the Supplementary Material. To estimate Σ, it is natural to construct the estimator of the variance as

$$\widehat{\Sigma}_n = n^{-1}\sum_{i=1}^{n}\left(\nabla m(\boldsymbol{Z}_i, \widehat{f}_n)[\widehat{h}_n] - n^{-1}\sum_{i=1}^{n}\nabla m(\boldsymbol{Z}_i, \widehat{f}_n)[\widehat{h}_n]\right)^{\otimes 2}, \tag{2.4}$$

where $\boldsymbol{u}^{\otimes 2} = \boldsymbol{u}\boldsymbol{u}^T$. Our next theorem states that $\widehat{\Sigma}_n$ in (2.4) is a consistent estimator of the variance of the OSRF under the following condition:

A.8 For every $\epsilon, \eta > 0$, there exist $\delta_1, \delta_2 > 0$ such that

$$\lim_n P\left(\sup_{\substack{f_1,f_2\in\mathcal{N}_{f_{n0},\delta_1},\\ h_1,h_2\in\mathcal{N}_{h_n^*,\delta_2}}} \left\|\mathbb{G}_n\left(\nabla m(\boldsymbol{Z}, f_1)[h_1]\right)^{\otimes 2} - \mathbb{G}_n\left(\nabla m(\boldsymbol{Z}, f_2)[h_2]\right)^{\otimes 2}\right\|_\infty > \frac{\epsilon}{\sqrt{n}}\right) < \eta.$$

Theorem 2. *Under Conditions* A.1–A.8, $\widehat{\Sigma}$ *converges to* Σ *in probability.*

The proof is straightforward, because under Condition A.8,

$$(\mathbb{P}_n - P)\left(\nabla m(\boldsymbol{Z}, \widehat{f}_n)[\widehat{h}_n]\right)^{\otimes 2} = (\mathbb{P}_n - P)\left(\nabla m(\boldsymbol{Z}, f_{n0})[h_n^*]\right)^{\otimes 2} + o_p(1).$$

3. Examples

3.1. Example 1: OSRE for a high-dimensional linear model

The first example is from a high-dimensional linear model. Specifically, consider n i.i.d. samples $(\boldsymbol{X}_i, Y_i)$ with $\boldsymbol{X}_i = (X_{i1}, \ldots, X_{ip_n})^T \in \mathbb{R}^{p_n}$, where one X is one, and the other X_{ij} have mean zero for $j > 1$. Moreover, it holds that

$$Y_i = \sum_{j=1}^{p_n} X_{ij}\beta_{nj}^* + \varepsilon_i,\ P[\varepsilon_i|\boldsymbol{X}_i] = 0, \tag{3.1}$$

where $\boldsymbol{\beta}_n^* = (\beta_{n1}^*, \ldots, \beta_{np_n}^*)^T$ is the vector of parameters, and ε_i is a random variable representing the noise in the ith response variable.

A single component of the parameters, say the first coordinate β_{n1}^* is widely used as a "debiased" lasso estimator. Thus, we consider this case in the Supplementary Material. It may also be interesting to consider the total contribution of the covariates, in practice. This is particularly useful when the covariates are obtained from one particular feature domain. Thus, we consider the inference for the sum of the squared $\boldsymbol{\beta}_n^*$, denoted by $\theta_{0n} = \sum_{j=1}^{p_n}\beta_{nj}^{*2}$. We aim

to construct the OSRE of θ_{0n}.

Obviously, $\mathcal{F}_n = \{f(\boldsymbol{x}) = \sum_{j=1}^{p_n} x_j\beta_j\}$ is the functional space. We assume $\boldsymbol{X}$ has mean zero. Because $\theta_{0n} = \sum_{j=1}^{p_n} \beta_{nj}^{*2}$, a simple calculation yields

$$\theta_{n0} = \mathfrak{F}_n(f_{n0}) = P\left[f_{n0}^2(\Sigma^{-1/2}\boldsymbol{X})\right].$$

Then, for all $h_n(\boldsymbol{x}) = \sum_{j=1}^{p_n} x_j\gamma_j$, we have

$$\nabla\mathfrak{F}_n(f_{n0})[h_n] = 2\sum_{j=1}^{p_n} \beta_{nj}^*\gamma_j,$$

which is a continuous linear functional. This verifies Condition A.1. Clearly, Condition A.2 is true. Let $g_{nj}(\boldsymbol{x})$ be a function such that

$$g_{nj}(\boldsymbol{X}) = X_j - \pi(X_j|X_{-j}),$$

where $X_{-j} = (X_1, \ldots, X_{j-1}, X_{j+1}, \ldots, X_{p_n})^T$, and $\pi(X_j|X_{-j})$ is the $L^2(P)$ projection of X_j onto the linear space of X_{-j}. We show that h_n^* in Condition A.3 is

$$h_n^*(\boldsymbol{x}) = -2\sum_{j=1}^{p_n}(Pg_{nj}^2(\boldsymbol{X}))^{-1}g_{nj}(\boldsymbol{x})\beta_{nj}^*. \tag{3.2}$$

To see that h_n^* satisfies (2.1), because $\pi(X_j|X_{-j})$ is the $L^2(P)$ projection of X_j onto the linear span of X_{-j} , we obtain

$$P[(X_j - \pi(X_j|X_{-j}))\pi(X_j|X_{-j})] = 0,$$

and

$$P[(X_k - \pi(X_k|X_{-k}))X_j] = 0 \text{ for all } k \neq j.$$

For any $X_j\gamma_j$,

$$\begin{aligned}
&P\left\{\nabla^2 m(\boldsymbol{Z}, f_{n0})[h_n^*, X_j\gamma_j]\right\}\\
&= 2(Pg_{nj}^2(\boldsymbol{X}))^{-1}P[(X_j - \pi[X_j|X_{-j}])X_j]\beta_{nj}^*\gamma_j\\
&= 2(Pg_{nj}^2(\boldsymbol{X}))^{-1}P[(X_j - \pi[X_j|X_{-j}])^2]\beta_{nj}^*\gamma_j\\
&= 2\beta_{nj}^*\gamma_j.
\end{aligned}$$

Thus, for any $h_n(\boldsymbol{x}) = \sum_{j=1}^{p_n} x_j\gamma_j$,

$$P\left\{\nabla^2 m(\boldsymbol{Z}, f_{n0})[h_n^*, h_n]\right\} = 2\sum_{j=1}^{p_n}\gamma_j\beta_{nj}^* = \nabla\mathfrak{F}_n(f_{n0})[h_n].$$

Thus, h_n^* is a function satisfying (2.1).

Therefore, suppose $\widehat{\boldsymbol{\beta}}_n$ is an initial estimator of $\boldsymbol{\beta}_n^*$, and we can find a proper estimator for h_n^*, denoted by $\widehat{h}_n$. The OSRE for β_{n1}^* is then given as

$$\widetilde{\theta}_n = \sum_{j=1}^{p_n} \widehat{\beta}_{nj}^2 - \frac{1}{n}\sum_{i=1}^{n} \widehat{h}_n(\boldsymbol{X}_i)\left(Y_i - \boldsymbol{X}_i^T\widehat{\boldsymbol{\beta}}_n\right), \tag{3.3}$$

where $\widehat{\beta}_{nj}$ is the jth coordinate of $\widehat{\boldsymbol{\beta}}_n$.

Suppose the linear regression model is defined as (3.1). The vector parameter $\boldsymbol{\beta}_n^*$ is sparse, which means we can estimate the initial estimator $\widehat{\boldsymbol{\beta}}_n$ in (3.3) using the lasso method:

$$\widehat{\boldsymbol{\beta}}_n = \underset{\boldsymbol{\beta}\in\mathbb{R}^{p_n}}{\operatorname{argmin}} \left\{ \frac{1}{n}\|\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}\|_2^2 + 2\lambda\|\boldsymbol{\beta}\|_1 \right\}, \tag{3.4}$$

where $\|\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}\|_2^2 = \sum_{i=1}^n (Y_i - \boldsymbol{X}_i^T\boldsymbol{\beta})^2$, $\|\boldsymbol{\beta}\|_1 = \sum_{j=1}^{p_n} |\beta_j|$ is the l_1-norm on $\mathbb{R}^{p_n}$, and $\lambda \geq 0$ is a penalty parameter.

Next, we estimate h_n^* defined by (3.2). Recalling the definition of h_n^*, we first estimate $\pi(X_j|X_j,\ldots,X_{j-1},X_{j+1},\ldots,X_{p_n})$, the projection of X_1 onto the linear space spanned by $X_1,\ldots,X_{p_n}$. The sparsity of the regression parameters implies a finite number of covariates. Note that such sparsity, equivalent to the maximal sparsity level of Σ^{-1}, also appears in other works (van de Geer et al. (2014); Javanmard and Montanari (2014, 2018)). This estimation can be treated as a high-dimensional linear regression problem. Thus, we adopt the lasso to estimate the coefficients.

We estimate the coefficients of $\boldsymbol{X}_{-j}$ for X_j using

$$\widehat{\boldsymbol{\eta}}_j = \underset{\boldsymbol{\eta}\in\mathbb{R}^{p_n-1}}{\operatorname{argmin}} \left\{ \frac{1}{2n}\|\boldsymbol{X}_j - \boldsymbol{X}_{-j}^T\boldsymbol{\eta}\|_2^2 + \widetilde{\lambda}_j\|\eta\|_1 \right\},$$

where $\boldsymbol{X}_j = (X_{1j},\ldots,X_{nj})^T$, $\boldsymbol{X}_{-j}$ is the sub-matrix of $\boldsymbol{X}$ obtained by removing the jth column. With $\widehat{\boldsymbol{\eta}}_j$, we obtain

$$\widehat{g}_{nj}(X) = X_j - \widehat{\pi}(X_j|X_{-j}) = X_j - X_{-j}^T\widehat{\boldsymbol{\eta}}_j.$$

On the other hand, $Pg_{nj}^2(X)$ is estimated by

$$\widehat{\tau}_j^2 = \frac{\|\boldsymbol{X}_j - \boldsymbol{X}_{-j}^T\widehat{\boldsymbol{\eta}}_j\|_2^2}{n} + \lambda\|\widehat{\boldsymbol{\eta}}_j\|_1.$$

Finally, the estimator for h_n^* is given as

$$\widehat{h}_n(X) = 2\sum_{j=1}^{p_n} \frac{\widehat{g}_{nj}(X)\widehat{\beta}_{nj}}{\widehat{\tau}_j^2} = 2\widehat{\boldsymbol{\beta}}_n^T\widehat{T}^{-2}\widehat{\Gamma}X, \tag{3.5}$$

where $\widehat{T}^2 = \mathrm{diag}(\widehat{\tau}_1^2, \ldots, \widehat{\tau}_{p_n}^2)$ and

$$\widehat{\Gamma} = \begin{pmatrix} 1 & -\widehat{\eta}_{1,2} & \ldots & -\widehat{\eta}_{1,p_n} \\ -\widehat{\eta}_{2,1} & 1 & \ldots & -\widehat{\eta}_{2,p_n} \\ \vdots & \vdots & \ddots & \vdots \\ -\widehat{\eta}_{p_n,1} & -\widehat{\eta}_{p_n,2} & \ldots & 1 \end{pmatrix}.$$

From (3.3) to (3.5), the OSRE for θ_{n0} is

$$\widetilde{\theta}_n = \sum_{j=1}^{p_n} \widehat{\beta}_{nj}^2 + \frac{2}{n} \sum_{i=1}^{n} \widehat{\boldsymbol{\beta}}_n^T \widehat{T}^{-2} \widehat{\Gamma} \boldsymbol{X}_i \left(Y_i - \boldsymbol{X}_i^T \widehat{\boldsymbol{\beta}}_n \right),$$

where $\widehat{\beta}_{nj}$ is the jth element of $\widehat{\boldsymbol{\beta}}_n$, and $\widehat{\boldsymbol{\beta}}_n$ is the lasso estimator of $\boldsymbol{\beta}_{n0}$.

To state the asymptotic properties for the OSRE, we need some technical assumptions, Conditions B.1-B.8 which are provided in the Supplementary Material.

Theorem 3. *Suppose that Conditions* B.1-B.8 *hold, and that* $\lambda \asymp \sqrt{\log p_n / n}$ *and* $\widetilde{\lambda}_j \asymp \sqrt{\log p_n / n}$ *uniformly in* j. *Then,* $\widetilde{\theta}_n$ *satisfies*

$$\sqrt{n}(\widetilde{\theta}_n - \theta_{n0}) \xrightarrow{p} N(0, c^2),$$

where c^2 *is defined in Condition* B.8 *in the Supplementary Material.*

A proof of Theorem 3 is given in Section S3 of the Supplementary Material.

3.2. Example 2: OSRE for high-dimensional additive model

Whereas the previous example was a parametric problem, we now examine a high-dimensional NAM. Suppose

$$Y_i = \mu + \sum_{j=1}^{p_n} f_{nj}^*(X_{ij}) + \varepsilon_i, \tag{3.6}$$

where μ is a constant and ε_i is the error term, with mean zero and finite variance σ^2. In this model, the true regression function $f_n^*(\boldsymbol{x}) = \sum_{j=1}^{p_n} f_{nj}^*(x_j)$ belongs to the functional class $\mathcal{F}_n = \{f \in L^2(P) : f(\boldsymbol{x}) = \sum_{j=1}^{p_n} f_j(x_j), P f_j(X_{ij}) = 0\}$, equipped with an inner product

$$\langle f_1, f_2 \rangle_{(n)} = \sum_{j=1}^{p_n} \int f_{1j}(x) f_{2j}(x) dx.$$

For a NAM, we can estimate f_{n0} by maximizing $P\{m(\boldsymbol{X}, Y, f)\}$, with $m(\boldsymbol{X}, Y, f) = -(Y - f(\boldsymbol{X}))^2/2$. We are interested in the contribution of one specific component of $\boldsymbol{X}$, say, X_1. For this purpose, we define the parameter of interest as

$\mathfrak{F}_n(f_n^*) = \int f_{n1}^{*2}(x)dx$ to quantify the contribution of X_1 in terms of predicting Y.

To find h_n^* satisfying Condition A.3, we first use a sequence of basis functions, $\phi_{jk}(x)$, for $k = 1, 2, \ldots$, in the support of X_j (splines, Fourier bases, for instance), so that $f_j(x)$ can be treated as a linear combination of basis functions $\phi_{jk}(x)$, for $k = 1, 2, \ldots$ Let $V_{1k} = \phi_{1k}(X_1)$ and $\boldsymbol{V}_{1,-k} = (V_{11}, \ldots, V_{1,k-1}, V_{1,k+1}, \ldots)$. We show that the solution to the above equation is

$$h_n^*(X) = -\sum_{k=1}^{\infty} u_k \left[V_{1k} - \pi\left(V_{1k}|\boldsymbol{V}_{1,-k}, X_2, X_3, \ldots, X_{p_n}\right)\right],$$

where

$$u_k = 2\left\{P\left[V_{1k} - \pi\left(V_{1k}|\boldsymbol{V}_{1,-k}, X_2, X_3, \ldots, X_{p_n}\right)\right]^2\right\}^{-1} \int f_{n1}^*(t)\phi_{1k}(t)dt.$$

Suppose that $\widehat{f}_n$ is an initial estimator of f_n^*, and that h_n^* is estimated by $\widehat{h}_n$. Then, the OSRE for θ_{n0} is defined as

$$\widetilde{\theta}_n = \int \widehat{f}_n^2(x)dx - \frac{1}{n}\sum_{i=1}^{n} \widehat{h}_n(\boldsymbol{X}_i)\left(Y_i - \bar{Y} - \widehat{f}_n(\boldsymbol{X}_i)\right), \tag{3.7}$$

where $\bar{Y}$ is the average of Y_i.

In case when many additive components $f_{nj}^*(\cdot)$ are zeros, Huang, Horowitz and Wei (2010) propose using an adaptive group lasso to estimate f_n^*. Consider a normalized B-spline basis $\{\psi_k, 1 \le k \le m_n\}$ for $\mathcal{B}_n$, where $m_n = K_n + l$, in which $K_n = n^\nu$ with $0 < \nu < 0.5$ is a positive integer. Under suitable smoothness assumptions, f_{nj}^* can be well approximated by functions in $\mathcal{B}_n$. Let $\|\boldsymbol{a}\|_2 = (\sum_{j=1}^{m_n} |a_j|^2)^{1/2}$, $\boldsymbol{\beta}_{nj} = (\beta_{j1}, \ldots, \beta_{jm_n})^T$, $\boldsymbol{\beta}_n = (\boldsymbol{\beta}_{n1}^T, \ldots, \boldsymbol{\beta}_{np_n}^T)^T$, and $\boldsymbol{w}_n = (w_{n1}, \ldots, w_{np_n})^T$ be a given vector of weights. Then, the penalized least squares estimation with a group lasso minimizes

$$L_n(\mu, \boldsymbol{\beta}_n) = \sum_{i=1}^{n}\left[Y_i - \mu - \sum_{j=1}^{p_n}\sum_{k=1}^{m_n}\beta_{jk}\psi_k(X_{ij})\right]^2 + \lambda_{n2}\sum_{j=1}^{p_n} w_{nj}\|\beta_{nj}\|_2,$$

where λ_{n2} is a penalty parameter. In order to make the computation identifiable, we impose the additional constraints that

$$\sum_{i=1}^{n}\sum_{k=1}^{m_n}\beta_{jk}\psi_k(X_{ij}) = 0.$$

The constrained optimization problem is converted to an unconstrained problem by centering the response and the basis functions. Let

$$\phi_{jk}(x) = \psi_k(x) - n^{-1}\sum_{i=1}^{n}\psi_k(X_{ij}).$$

For simplicity, we write $\phi_k(x) = \phi_{jk}(x)$, and assume the mean of Y is zero. Huang, Horowitz and Wei (2010) propose a two-step approach for the estimation and the component selection. First, they define

$$\widetilde{\boldsymbol{\beta}}_n = \underset{\boldsymbol{\beta}_n}{\operatorname{argmin}} \sum_{i=1}^{n} \left[Y_i - \sum_{j=1}^{p_n} \sum_{k=1}^{m_n} \beta_{jk} \phi_k(X_{ij}) \right]^2 + \lambda_{n1} \sum_{j=1}^{p_n} \|\beta_{nj}\|_2.$$

Then they use $\widetilde{\boldsymbol{\beta}}_n$ to obtain the weights by setting

$$w_{nj} = \begin{cases} \|\widetilde{\beta}_{nj}\|_2^{-1}, & \text{if } \|\widetilde{\beta}_{nj}\|_2 > 0 \\ \infty, & \text{if } \|\widetilde{\beta}_{nj}\|_2 = 0. \end{cases}$$

Finally, the adaptive group lasso estimator is $\widehat{\boldsymbol{\beta}}_n = \operatorname{argmin}_{\boldsymbol{\beta}_n} L_n(\boldsymbol{\beta}_n)$. Therefore, the group lasso estimators for f_j are

$$\widehat{f}_{nj}(x) = \sum_{k=1}^{m_n} \widehat{\beta}_{jk} \phi_k(x).$$

To estimate h_n^* in Condition A.3, we first need to estimate the projection $\pi\left(V_{1k} | \boldsymbol{V}_{1,-k}, X_2, X_3, \ldots, X_{p_n}\right)$, which can also be viewed as a high-dimensional NAM. Thus, suppose there exists $\boldsymbol{\eta}_{1k}^* = (\eta_{1,k,1}^*, \ldots, \eta_{1,k,k-1}^*, \eta_{1,k,k+1}^*, \ldots)^T$, and $s_k^{*(n)}(X_{-1}) = \sum_{j=2}^{p_n} s_{kj}^{*(n)}(X_j)$ satisfies

$$V_{1,k} = \sum_{l \neq k} V_{1,l} \eta_{1kl}^* + s_k^{*(n)}(X_{-1}) + \varepsilon_k, P[\varepsilon_k | X_{-1}] = 0.$$

The assumption implies a sparse structure of the projection

$$\pi[V_{1k} | \boldsymbol{V}_{1,-k}, X_2, \ldots, X_{p_n}],$$

which means that only a few of the covariates are correlated with X_1. The sparsity of η_k implies that each function component in the additive model can be represented by a finite number of basis functions, although the number can diverge with the sample size. Thus, with greater sample sizes, we can allow h_n^* to be closer to some arbitrary additive function. This is a common assumption in many high-dimensional lasso settings (van de Geer et al. (2014); Javanmard and Montanari (2014, 2018)).

We then follow Huang, Horowitz and Wei (2010) to apply the group Lasso to estimate $\pi\left(V_{1k} | \boldsymbol{V}_{1,-k}, X_2, X_3, \ldots, X_{p_n}\right)$. For simplicity of notation, we omit n in the subscript of $\boldsymbol{\eta}$ in the following. More specifically, the penalized least squares estimators are given as

$$\widetilde{\boldsymbol{\eta}}_k = \underset{\boldsymbol{\eta}_k}{\operatorname{argmin}} \sum_{i=1}^{n} \left[V_{i1k} - \boldsymbol{V}_{i1,-k}^T \boldsymbol{\eta}_{k1} - \sum_{j=2}^{p_n} \sum_{l=1}^{m_n} \eta_{kjl} \phi_l(X_{ij}) \right]^2 + \widetilde{\lambda}_{k1} \left(\sum_{l \neq k} |\eta_{k1l}| + \sum_{j=2}^{p_n} \|\boldsymbol{\eta}_{kj}\|_2 \right),$$

where $\boldsymbol{\eta}_{kn} = (\boldsymbol{\eta}_{k1}^T, \ldots, \boldsymbol{\eta}_{kj}^T)^T$, $\boldsymbol{\eta}_{k1} = (\eta_{k11}, \ldots, \eta_{k,1,k-1}, \eta_{k,1,k+1}, \ldots, \eta_{k,1,m_n})^T$, $\boldsymbol{\eta}_{kj} = (\eta_{kj1}, \ldots, \eta_{kjm_n})^T$, for $j = 2, \ldots, p_n$, and $\widetilde{\lambda}_{k1}$ is a penalty parameter. The estimates for $\widetilde{\boldsymbol{\eta}}_k$ give the weights by setting

$$\widetilde{w}_{k1l} = \begin{cases} |\widetilde{\eta}_{k1l}|^{-1}, & \text{if } |\widetilde{\eta}_{k1l}| > 0 \\ \infty, & \text{if } |\widetilde{\eta}_{k1l}| = 0, \end{cases}$$

for $l \neq k$, and

$$\widetilde{w}_{kj} = \begin{cases} \|\widetilde{\boldsymbol{\eta}}_{kj}\|_2^{-1}, & \text{if } \|\widetilde{\boldsymbol{\eta}}_{kj}\|_2 > 0 \\ \infty, & \text{if } \|\widetilde{\boldsymbol{\eta}}_{kj}\|_2 = 0, \end{cases}$$

for $j = 2, \ldots, p_n$. Finally, we minimize

$$\widetilde{L}_k(\boldsymbol{\eta}_k) = \sum_{i=1}^{n} \left[V_{i1k} - \boldsymbol{V}_{i1,-k}^T \boldsymbol{\eta}_{k1} - \sum_{j=2}^{p_n} \sum_{l=1}^{m_n} \eta_{kjl} \phi_l(X_{ij}) \right]^2 + \widetilde{\lambda}_{kn2} \left(\sum_{l \neq k} \widetilde{w}_{k1l} |\eta_{k1l}| + \sum_{j=2}^{p_n} \widetilde{w}_{kj} \|\boldsymbol{\eta}_{kj}\|_2 \right).$$

The resulting coefficients of the projection are

$$\widehat{\boldsymbol{\eta}}_k = \underset{\boldsymbol{\eta}_k}{\operatorname{argmin}} \widetilde{L}_k(\boldsymbol{\eta}_k).$$

Now, define

$$\widehat{C} = \begin{pmatrix} 1 & -\widehat{\eta}_{1,1,2} & \cdots & -\widehat{\eta}_{1,1,m_n} & -\widehat{\eta}_{1,2,1} & \cdots & -\widehat{\eta}_{1,2,m_n} & -\widehat{\eta}_{1,3,1} & \cdots & -\widehat{\eta}_{1,p_n,m_n} \\ -\widehat{\eta}_{2,1,1} & 1 & \cdots & -\widehat{\eta}_{2,1,m_n} & -\widehat{\eta}_{2,2,1} & \cdots & -\widehat{\eta}_{2,2,m_n} & -\widehat{\eta}_{2,3,1} & \cdots & -\widehat{\eta}_{2,p_n,m_n} \\ \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ -\widehat{\eta}_{m_n,1,1} & -\widehat{\eta}_{m_n,1,2} & \cdots & 1 & -\widehat{\eta}_{m_n,2,1} & \cdots & -\widehat{\eta}_{m_n,2,m_n} & -\widehat{\eta}_{m_n,3,1} & \cdots & -\widehat{\eta}_{m_n,p_n,m_n} \end{pmatrix}.$$

Let

$$\widehat{T}^2 = \operatorname{diag}(\widehat{\tau}_1^2, \ldots, \widehat{\tau}_{m_n}^2),$$

where

$$\widehat{\tau}_k^2 = \frac{1}{n} \sum_{i=1}^{n} \left(V_{i1k} - \sum_{l \neq k} V_{i1l} \widehat{\eta}_{k1l} - \sum_{j=2}^{p_n} \sum_{l=1}^{m_n} V_{ijl} \widehat{\eta}_{kjl} \right) V_{i1k},$$

and let

$$V = (V_{11}, \ldots, V_{1m_n}, \ldots, V_{p_n1}, \ldots, V_{p_nm_n})^T.$$

We obtain the estimator for h_n^* as

$$\widehat{h}_n(\boldsymbol{X}) = -\widehat{\boldsymbol{\kappa}}^T \widehat{T}^{-2} \widehat{C} V,$$

where $\widehat{\boldsymbol{\kappa}} = (\widehat{\kappa}_1, \ldots, \widehat{\kappa}_{m_n})^T$, and $\widehat{\kappa}_k = 2 \int \widehat{f}_{n1}(x)\phi_k(x)dx$. Therefore, from (3.7), the OSRE of θ_{n0} is

$$\widetilde{\theta}_n = \int \widehat{f}_{n1}^2(x)dx + \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{p_n} \widehat{h}_{nj}(X_{ij}) \left(Y_i - \sum_{j=1}^{p_n} \widehat{f}_{nj}(X_{ij}) \right). \tag{3.8}$$

Under some conditions, we can also prove that the OSRE $\widetilde{\theta}_n$ defined in (3.8) follows an asymptotic normal distribution. In order to obtain the asymptotic properties of $\widehat{\theta}_n$, we require the Conditions C.1–C.10 which are given in the Supplementary Material.

Theorem 4. *If Conditions* C.1–C.10 *hold, then* $\widetilde{\theta}_n$*, defined by* (3.8) *satisfies*

$$\sqrt{n}(\widetilde{\theta}_n - \theta_{n0}) \xrightarrow{p} N(0, \sigma_\varepsilon^2 c^2).$$

Theorem 4 ensures that the asymptotic distribution of the OSRE $\widetilde{\theta}_n$ defined in (3.8) is normal. A proof for Theorem 4 is given in Section S4 of the Supplementary Material.

To illustrate the universality of our proposed method, we provide additional examples in the Supplementary Material. The OSRE for the coefficient inferences of high-dimensional linear model and high-dimensional logistic regression models are given in Sections S5 and S6, respectively, of the Supplementary Material.

4. Simulation Study

4.1. Simulation with high-dimensional linear models

Our first simulation considers a high-dimensional linear model. In this setting, we generate $p = 500$ covariates consisting of $K \equiv p/q$ groups, each group with q variables. For q variables in the kth group, denoted by $X_{k1}, \ldots, X_{kq}$, are generated as

$$X_{kj} = \frac{(w_{kj} + tu_k)}{1+t}, \quad w_{kj} \sim U(0,1),\ u_k \in U(0,1).$$

In this way, we generate a sequence of blocked covariates. We set $t = 2$, so the correlation between any two X in the same block is $\rho = 0.8$, but they are independent if from different blocks. Given $\boldsymbol{X}$, Y is generated from a linear model with an error term from a standard normal distribution. We vary the block size, q,

from two to four, and choose the coefficients according to the following scenarios:

(a) $q = 2$ and $\boldsymbol{\beta}_0 = (1, 1, 1, 0, \ldots, 0)^T$.

(b) $q = 4$ and $\boldsymbol{\beta}_0 = (1, 1, 1, 0, \ldots, 0)^T$.

(c) $q = 4$ and $\boldsymbol{\beta}_0 = (1, 1, 1, 1, 0, \ldots, 0)^T$.

(d) $q = 4$ and $\boldsymbol{\beta}_0 = (1, 1, 1, 1, 1, 0, \ldots, 0)^T$.

To illustrate our proposed method, we focus on inferences for three parameters, that is, the total effect, given by $\theta_0 = \sum_{j=1}^{p} \beta_j^{*2}$, the coefficient of one important covariate, given by β_1^*, and the coefficient of an unimportant covariate, which is specified as the coefficient of the first zero-coefficient covariate in each scenario. The OSRE for a single coordinate can be found in Section S3 of the Supplementary Material. We consider sample sizes $n = 100$ and 200, and replicate each scenario 500 times in the simulation study.

To calculate the OSRE, the initial estimate for $\boldsymbol{\beta}$ is based on the lasso regression. The tuning parameter is selected as the largest penalty parameter for which the corresponding cross-validation error is within one standard deviation of the minimal error. The estimate for h_n^* is also obtained from a lasso regression with cross-validation, but the tuning parameter is set to be a factor of the cross-validation optimal parameter. The usual variable selection procedures, such as the lasso, tend to balance the bias and variance trade-off, and their goal is to minimize prediction errors. However, to obtain a proper inference, it is necessary to remove the bias so that the asymptotic normality with mean zero is a good approximation. Empirically, we find a factor of 2^{-6} yields the best performance. To examine the inference performance of the proposed method, we calculate confidence intervals for the OSRE, which are constructed using the asymptotic normal distribution in our theorem. For comparison purposes, we also report the coverages from two other methods: the first constructs the confidence intervals using a residual bootstrap (RBS), and the second performs an ad-hoc post-selection inference (PSI) by treating selected variables in the lasso method as the only variable in the regression model. It takes less than a second to compute the OSRE of one single regression coefficient in Scenario (a) with sample size 100 on a laptop computer with an Intel Core i5 processor.

Figure 1 plots a histogram of the OSRE and the plug-in estimators for case (a) with sample size $n = 100$. The dashed curve in the left figure is a normal density function, with the true parameter value θ_0 as the mean (dotted line) and the variance given as the average of the estimated $\widehat{\sigma}_n$. Therefore, this curve serves as a theoretical distribution from our theorem. Figure 1 indicates that the OSRE is close to a normal distribution, and its distribution matches the theoretical one very well. In contrast, the plug-in estimator is severely biased. (The standard error of the dashed line in the right figure is the standard error of the plug-in estimator.)

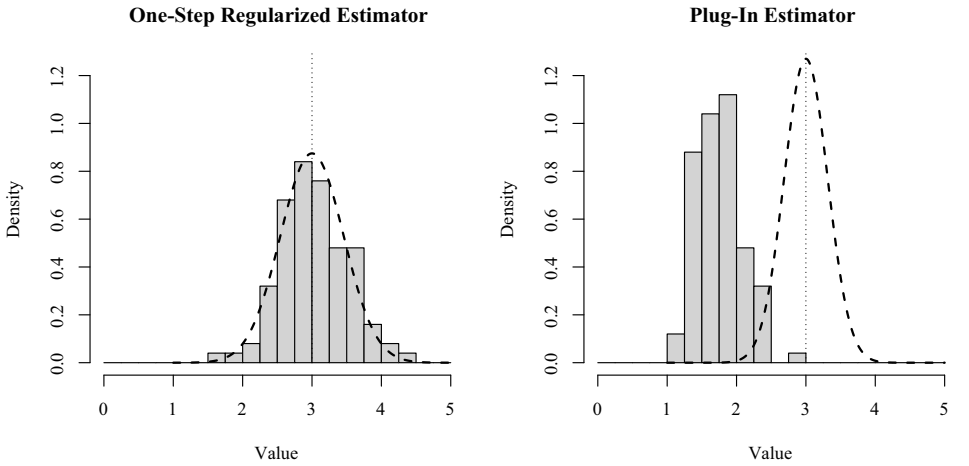

Figure 1. Histograms of the OSRE and plug-in estimators of the total effect for Scenario (a).

In Table 1, we report the simulation results for the bias (Bias), standard error (SE), estimated standard error (ESE), with the coverage probabilities based on $(1-\alpha)$-confidence intervals, where $\alpha = 0.1$ and 0.05; CP95 represents the coverage rates of 95% confidence interval, and CP90 represents those of the 90% confidence interval. As shown in the table, both the RBS and PSI perform poorly in some cases. In contrast, the coverage probabilities of the confidence intervals based on the OSRE are reasonably close to the nominal levels, and the performance is even better when n increases to 200. In addition, the SEs and ESEs are close in our proposed method.

From Table 1, we notice that the post-selection produces similar coverage to that of the OSRE for β_1, but with much shorter confidence intervals. This is because, for this model, the variable selection does not introduce much bias into the estimation of β_1, so debiasing is not necessary for β_1. On the other hand, the extra bias correction based on the empirical data in the OSRE can bring extra variability into the estimation. In the Supplementary Material (S5.1), we demonstrate that if the extra bias correction is known, then the confidence intervals in the ORSE are similar in width to those in the post-selection approach.

4.2. Simulation study with high-dimensional NAM

In this simulation, we generate $\boldsymbol{X}$ in the same way as in the linear model with $q = 5$, except that the correlation ρ with the same block is either 0 (setting $t = 0$) or 0.2 (setting $t = 0.5$). The outcome, Y, is generated from

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

Table 1. Results based on 500 replicates for high-dimensional linear models.

n	Method	Parameter	Bias	SE	ESE	CP95	CP90	Bias	SE	ESE	CP95	CP90
					(a)					(b)		
100	RBS	θ	-1.253	0.297	0.204	0.336	0.292	-0.831	0.293	0.249	0.696	0.638
		β_1	-0.215	0.179	0.179	0.902	0.876	-0.176	0.191	0.201	0.936	0.900
		β_k	0.013	0.043	0.042	0.970	0.944	0.037	0.082	0.094	0.974	0.946
	PSI	θ	0.021	0.319	0.292	0.916	0.868	0.023	0.318	0.273	0.912	0.850
		β_1	-0.026	0.180	0.158	0.904	0.838	-0.023	0.195	0.183	0.938	0.872
		β_k	0.035	0.092	0.023	0.966	0.942	0.069	0.125	0.054	0.960	0.932
	OSRE	θ	-0.020	0.461	0.456	0.934	0.880	0.299	0.530	0.576	0.956	0.890
		β_1	0.016	0.219	0.208	0.930	0.874	0.055	0.257	0.254	0.950	0.894
		β_k	0.082	0.203	0.207	0.930	0.870	0.103	0.236	0.255	0.948	0.888
					(c)					(d)		
	RBS	θ	-0.777	0.286	0.289	0.864	0.794	-1.193	0.372	0.338	0.550	0.476
		β_1	-0.108	0.205	0.210	0.946	0.906	-0.095	0.222	0.215	0.918	0.884
		β_k	-0.000	0.008	0.000	0.998	0.998	0.016	0.052	0.040	0.952	0.934
	PSI	θ	0.169	0.277	0.271	0.934	0.868	0.081	0.383	0.333	0.896	0.832
		β_1	0.011	0.205	0.191	0.934	0.880	-0.001	0.209	0.186	0.916	0.844
		β_k	-0.001	0.014	0.000	0.998	0.998	0.031	0.090	0.023	0.970	0.944
	OSRE	θ	0.583	0.547	0.603	0.892	0.800	0.438	0.694	0.721	0.914	0.852
		β_1	0.078	0.284	0.253	0.918	0.862	0.077	0.273	0.257	0.918	0.864
		β_k	0.003	0.243	0.250	0.958	0.906	0.068	0.250	0.254	0.938	0.890
					(a)					(b)		
200	RBS	θ	-1.004	0.227	0.149	0.494	0.434	-0.701	0.179	0.164	0.824	0.734
		β_1	-0.157	0.133	0.126	0.930	0.884	-0.136	0.132	0.141	0.944	0.912
		β_k	0.010	0.033	0.030	0.964	0.938	0.020	0.050	0.059	0.978	0.958
	PSI	θ	0.013	0.239	0.210	0.924	0.864	-0.011	0.205	0.189	0.928	0.884
		β_1	-0.010	0.129	0.116	0.910	0.858	-0.013	0.134	0.133	0.940	0.898
		β_k	0.025	0.071	0.014	0.956	0.930	0.045	0.085	0.036	0.978	0.952
	OSRE	θ	-0.035	0.414	0.383	0.922	0.866	0.101	0.429	0.465	0.950	0.914
		β_1	0.011	0.176	0.180	0.958	0.912	0.013	0.216	0.215	0.952	0.912
		β_k	0.030	0.175	0.179	0.938	0.902	0.040	0.195	0.215	0.976	0.928
					(c)					(d)		
	RBS	θ	-0.654	0.192	0.176	0.888	0.810	-1.005	0.239	0.199	0.646	0.568
		β_1	-0.098	0.146	0.146	0.940	0.882	-0.081	0.150	0.147	0.956	0.922
		β_k	0.000	0.000	0.000	1.000	1.000	0.004	0.020	0.027	0.982	0.966
	PSI	θ	0.095	0.170	0.176	0.930	0.872	0.026	0.252	0.226	0.924	0.862
		β_1	-0.000	0.146	0.137	0.922	0.866	-0.004	0.142	0.136	0.946	0.892
		β_k	0.000	0.000	0.000	1.000	1.000	0.014	0.052	0.010	0.992	0.972
	OSRE	θ	0.266	0.389	0.431	0.938	0.872	0.152	0.503	0.540	0.964	0.910
		β_1	0.048	0.210	0.214	0.958	0.906	0.006	0.215	0.215	0.950	0.886
		β_k	-0.010	0.204	0.215	0.958	0.918	0.020	0.213	0.215	0.950	0.898

where we choose $f_1(x) = 8x$, $f_2(x) = 3(2x-1)^2$, $f_3(x) = 4\sin(2\pi t)/(2-\sin(2\pi t))$, $f_4(x) = 6(0.1\sin(2\pi t)+0.2\cos(2\pi t)+0.3\sin(2\pi t)^2+0.4\cos(2\pi t)^3+0.5\sin(2\pi t)^3)$, $f_5(x) = \cdots = f_p = 0$, and $\varepsilon \sim N(0, \sigma^2)$. These functions cover both linear and nonlinear patterns. Furthermore, we subtract each f_j from its average value to make the model identifiable when including an intercept. For illustration purposes, we focus on an inference for the total effect of the linear part as

$\int f_1(x)^2 dx$, and the total effect of the nonlinear covariate X_4 as $\int f_4(x)^2 dx$. The signal-to-noise ratio is defined as $sd(f)/sd(\varepsilon)$. Because the standard derivation of the error term is chosen as $\sigma = 2$, the signal-to-noise ratios for f_1 and f_4 are 0.86 and 1.54, respectively. We consider the cases where $p = 500$ and sample size $n = 200$ and 400.

In the simulation, to calculate the OSRE in our method, we first use the third-degree B-spline with six evenly distributed knots to approximate all f_j. We also investigate the choice of three knots and using the sixth degree of splines for the simulation setting with zero within-block correlation and $n = 200$. The results show that the method is fairly robust for these different choices. We obtain the initial estimates for f_j based on these splines and using the adaptive group lasso method proposed by Huang, Horowitz and Wei (2010). In particular, the adaptive group lasso is calculated using the algorithm proposed by Yuan and Lin (2006). Because p is larger than n, we use the BIC (Schwarz (1978)) to select the penalty parameter, as suggested by Huang, Horowitz and Wei (2010). We use the adaptive group lasso with BIC for each basis function of X_1 to estimate the coefficient $\boldsymbol{\eta}_{kn}$, for $k = 1, \ldots, m_n$, and construct the OSRE based on (3.8). It takes an average of 3.12 seconds to run one data set with a sample size of 100. We compare our method with the ad-hoc post-selection method, because the residual bootstrapping method is computationally intensive. The latter treats a spline approximation as a standard linear regression model after the important components are identified.

Table 2 shows the relative bias (Bias), standard errors (SEs), estimated standard errors (ESEs), and coverage rates of the OSRE and ad-hoc (PSI) methods based on 500 replicates. The OSRE and PSI do not perform well when the sample size is 100, so we omit the results in the table. The SEs and ESEs are close to each other in our proposed method, whereaas the ESEs are smaller than the SEs for the ad-hoc method. The simulation shows similar results for the case with and without correlations between the covariates. For the linear and nonlinear components, the coverage probabilities of the OSREs are reasonably close to the nominal levels when the sample size $n = 200$. In contrast, the ad-hoc method gives all cases a coverage probability lower than $(1 - \alpha)$. Both methods seem to work well when the sample size $n = 400$. We also test the performance of the two methods on the zero-influence function. Both methods show a slight overestimation of the coverage probabilities.

We also conduct a simulation study for a high-dimensional generalized linear model and a partial linear model to compare our proposed method with the ad-hoc method (PSI). These simulation results can be found in Section S6 and S7, respectively, of the Supplementary Material.

Table 2. Results based on 500 replications for high-dimensional nonparametric additive models.

Parameter	n	ρ	Method	Bias	SE	ESE	CP90	CP95
$\int f_1^2(x)dx$	200	0	PSI	-0.002	0.905	0.643	0.772	0.858
			OSRE	0.087	1.051	1.081	0.922	0.984
		0.2	PSI	-0.027	0.644	0.474	0.734	0.796
			OSRE	0.007	0.656	0.63	0.846	0.888
	400	0	PSI	0.012	0.523	0.474	0.894	0.954
			OSRE	0.053	0.669	0.744	0.922	0.968
		0.2	PSI	0.013	0.448	0.371	0.886	0.936
			OSRE	0.010	0.540	0.533	0.912	0.962
$\int f_4^2(x)dx$	200	0	PSI	-0.009	1.243	0.854	0.770	0.830
			OSRE	0.059	1.335	1.413	0.946	0.974
		0.2	PSI	-0.028	1.322	0.831	0.742	0.810
			OSRE	0.034	1.344	1.371	0.934	0.976
	400	0	PSI	-0.001	0.664	0.627	0.900	0.956
			OSRE	0.027	0.809	0.936	0.944	0.982
		0.2	PSI	0.016	0.773	0.642	0.868	0.934
			OSRE	0.032	0.978	1.025	0.918	0.964

5. Data Example

In this section, we apply the proposed method to study the association between genes and a particular gene called TRIM32, which has been found to cause Bardet-Biedl syndrome (Chiang et al. (2006)). We use the expression data for an eQTL experiment on rat's eyes reported by Scheetz et al. (2006). In this study, the eye tissue of 120 120-week-old male rats were selected for an Affymetrix expression microarray analysis. Over 31,000 different probe sets were recorded in the Affymetrix Rat Genome 230 2.0 Array. The intensity values are normalized using the robust multi-chip averaging (RMA) method (Irizarry et al. (2003)). The gene expression levels are analyzed on a logarithmic scale. Because many of the probes in the Affymetrix Rat Genome 230 2.0 Array are not expressed in the eye tissue, and initial screening using correlation shows that most probe sets have very low correlation with TRIM32, we select the 500 probe sets that have the highest correlation with TRIM32 in this analysis. We further exclude one sample (GSM130600), because its expression values are extreme. Our final data set has sample size $n = 119$ and 500 covariates.

We fit both a linear model and an additive model to this data to test whether any significant linear or nonlinear association exists between any gene and TRIM32. All covariates are standardized by their ranges, so the values are between zero and one. Linear model fitting is the same as in the first simulation study, where the penalty parameter for the lasso estimation is based on cross-validation. The OSRE for each regression coefficient in the linear model

Table 3. Parameter inference in the microarray data analysis.

Probe	OSRE	Standard Error	p-value
1384035_at	8.34e-06	6.29e-05	0.148
1368136_at	5.52e-06	1.66e-05	<0.001
1398370_at	4.04e-05	8.09e-05	<0.001
1376261_at	4.89e-06	1.23e-04	0.665
1379982_at	9.28e-06	7.24e-05	0.162
1367777_at	8.45e-05	4.68e-04	0.049
1368228_at	9.13e-06	3.61e-05	0.006
1380137_at	1.18e-06	1.18e-05	0.274
1384139_at	8.95e-06	7.60e-05	0.199
1379971_at	1.65e-05	4.29e-05	<0.001
1388491_at	1.18e-05	2.85e-05	<0.001
1375642_at	8.62e-06	3.96e-05	0.018
1369414_at	1.88e-05	1.54e-04	0.183

is then calculated as in the simulation study, and its variance is estimated using the proposed method. To fit the NAM, we use cubic splines with six evenly distributed knots in $[0, 1]$ to estimate each of the additive components. To test the importance of each covariate, we calculate the OSRE for the summary of each functional component as $\int f_k^2(x)dx$, for $k = 1, \ldots, p$. In the estimation, the tuning parameter is chosen using the BIC.

To conserve space, in Table 3, we show only the estimated parameters of the NAM and their associated p-values computed based on normal distributions. The results for the linear model are given in Supplementary Material. Table 3 shows that 13 important genes are selected by the additive model. Only gene 1367777_at is shown to be significantly associated with TRIM32 in both the linear model and the additive model.

To gain further insight into how these selected genes are associated with TRIM32 in the two models, we plot locally weighted scatterplot smoothing estimates for the significant variables from the additive model (in the Supplementary Material). The plot indicates that both 1368228_at and 1379971_at have nonlinear associations with TRIM32. We also observe a clear linear relationship between TRIM32 and 1367777_at, the only gene that is significant in both models.

6. Conclusion

We have proposed an OSRE for rigorous inference for low-dimensional functionals of high-dimensional parameters. A key component of the OSRE is to solve for h_n^* by inverting the Hessian operator given by the objective function, which is closely related to the information operator when the objective function is a log-likelihood function. For the latter situation, our OSRE reduces to a one-step

efficient estimator in semiparametric models. When the initial estimators $\widehat{f}_n$ and $\widehat{h}_n$ satisfy $n^{-1/4}$ consistency, we have shown that the OSRE is root-n consistent and asymptotically normal. We have applied our method to study the inferences for the parameters in both high-dimensional linear models and high-dimensional additive models. This is the first time such a result has been established for the additive model. Our numerical results suggest that the proposed method works well, even when the sample size is relatively small.

For high-dimensional inference, sample splitting (or cross-fitting) is widely used to construct a estimator. This technique is particularly useful to de-correlate the estimators between the parameter of interest and the nuisance parameters. We can extend the cross-fitting techniques to facilitate the proof of asymptotic equicontinuity conditions in the regularity conditions. Sample splitting can lose efficiency, because it is based on partial data. In such cases, salvage methods include cross-validation-type sample splitting and the estimator average. We will further examine the performance of such techniques in the OSRE methods.

As stated earlier, a key point for the OSRE is the need to estimate h_n^*. Although we can obtain its expression in the models considered here, h_n^* often does not have an explicit expression. Thus, it is difficult to generalize the OSRE to more complicated models. However, in semiparametric inference, the profile likelihood function can be used to approximate the least favorable submodel. For example, the tangent vector of the profile likelihood function is the efficient score function. Hence, one potential direction for future research is to devise a similar profile m-function, without explicitly estimating h_n^*.

We have only considered the dimensionality of variables to be p_n, and assumed the coefficients are nonzero for a much smaller list of variables. However, if we embed p_n-dimensional functions into an infinite-dimensional function space, we can allow assumptions that are even more flexible. For example, nocoefficients are zeros, but they decay at a certain rate. This may lead to an even more general framework for OSREs.

Supplementary Material

The online Supplementary Material provides some technical conditions for the theorems, details of proofs, the connection between OSRE the semiparametric model and additional data examples.

Acknowledgments

Dr. Yi Wang and Xingwei Tong are partially supported by the NSFC and the China Scholarship Council.

References

Bach, P., Klaassen, S., Kueck, J. and Spindler, M. (2020). Uniform inference in high-dimensional generalized additive models. *arXiv:2004.01623*.

Belloni, A., Chernozhukov, V. and Hansen, C. (2014). High-dimensional methods and inference on structural and treatment effects. *Journal of Economic Perspectives* **28**, 29–50.

Bradic, J., Fan, J. and Wang, W. (2011). Penalized composite quasi-likelihood for ultrahigh dimensional variable selection. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)* **73**, 325–349.

Bühlmann, P. and van de Geer, S. (2011). *Statistics for High-Dimensional Data: Methods, Theory and Applications*. Springer Science & Business Media.

Candes, E. and Tao, T. (2007). The Dantzig selector: Statistical estimation when p is much larger than n. *The Annals of Statistics* **35**, 2313–2351.

Chatterjee, A. and Lahiri, S. N. (2010). Asymptotic properties of the residual bootstrap for Lasso estimators. *Proceedings of the American Mathematical Society* **138**, 4497–4509.

Chatterjee, A. and Lahiri, S. N. (2011). Bootstrapping Lasso estimators. *Journal of the American Statistical Association* **106**, 608–625.

Chen, X. (2007). Large sample sieve estimation of semi-nonparametric models. *Handbook of Econometrics* **6**, 5549–5632.

Chen, X. and Shen, X. (1998). Sieve extremum estimates for weakly dependent data. *Econometrica* **66**, 289–314.

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W. et al. (2018). Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal* **21**, C1–C68.

Chernozhukov, V., Escanciano, J. C., Ichimura, H., Newey, W. K. and Robins, J. M. (2016). Locally robust semiparametric estimation. *arXiv:1608.00033*.

Chernozhukov, V., Hansen, C. and Spindler, M. (2015). Valid post-selection and post-regularization inference: An elementary, general approach. *Annual Review of Economics* **7**, 649–688.

Chiang, A. P., Beck, J. S., Yen, H.-J., Tayeh, M. K., Scheetz, T. E., Swiderski, R. E. et al. (2006). Homozygosity mapping with SNP arrays identifies TRIM32, an E3 ubiquitin ligase, as a Bardet–Biedl syndrome gene (BBS11). *Proceedings of the National Academy of Sciences* **103**, 6287–6292.

Dezeure, R., Bühlmann, P. and Zhang, C.-H. (2017). High-dimensional simultaneous inference with the bootstrap. *Test* **26**, 685–719.

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. (2011). Nonconcave penalized likelihood with NP-dimensionality. *IEEE Transactions on Information Theory* **57**, 5467–5484.

Fei, Z., Zhu, J., Banerjee, M. and Li, Y. (2019). Drawing inferences for high-dimensional linear models: A selection-assisted partial regression and smoothing approach. *Biometrics* **75**, 551–561.

Geman, S. and Hwang, C.-R. (1982). Nonparametric maximum likelihood estimation by the method of sieves. *The Annals of Statistics* **10**, 401–414.

Gregory, K., Mammen, E. and Wahl, M. (2016). Statistical inference in sparse high-dimensional additive models. *arXiv:1603.07632*.

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

Irizarry, R. A., Hobbs, B., Collin, F., Beazer-Barclay, Y. D., Antonellis, K. J., Scherf, U. et al. (2003). Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics* **4**, 249–264.

Javanmard, A. and Montanari, A. (2014). Hypothesis testing in high-dimensional regression under the Gaussian random design model: Asymptotic theory. *IEEE Transactions on Information Theory* **60**, 6522–6554.

Javanmard, A. and Montanari, A. (2018). Debiasing the Lasso: Optimal sample size for Gaussian designs. *The Annals of Statistics* **46**, 2593–2622.

Knight, K. and Fu, W. (2000). Asymptotics for Lasso-type estimators. *The Annals of Statistics* **28**, 1356–1378.

Kozbur, D. (2020). Inference in additively separable models with a high-dimensional set of conditioning variables. *Journal of Business & Economic Statistics* **39**, 984–1000.

Lee, J. D., Sun, D. L., Sun, Y. and Taylor, J. E. (2016). Exact post-selection inference, with application to the Lasso. *The Annals of Statistics* **44**, 907–927.

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

Lu, J., Kolar, M. and Liu, H. (2020). Kernel meets sieve: Post-regularization confidence bands for sparse additive model. *Journal of the American Statistical Association* **115**, 2084–2099.

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

Newey, W. K. and Powell, J. L. (2003). Instrumental variable estimation of nonparametric models. *Econometrica* **71**, 1565–1578.

Ning, Y. and Liu, H. et al. (2017). A general theory of hypothesis tests and confidence regions for sparse high dimensional models. *The Annals of Statistics* **45**, 158–195.

Ren, Z., Sun, T., Zhang, C.-H. and Zhou, H. H. (2015). Asymptotic normality and optimalities in estimation of large Gaussian graphical models. *The Annals of Statistics* **43**, 991–1026.

Scheetz, T. E., Kim, K.-Y. A., Swiderski, R. E., Philp, A. R., Braun, T. A., Knudtson, K. L. et al. (2006). Regulation of gene expression in the mammalian eye and its relevance to eye disease. *Proceedings of the National Academy of Sciences* **103**, 14429–14434.

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics* **6**, 461–464.

Shen, X. and Wong, W. H. (1994). Convergence rate of sieve estimates. *The Annals of Statistics* **22**,580–615.

Taylor, J., Lockhart, R., Tibshirani, R. J. and Tibshirani, R. (2014). Exact post-selection inference for forward stepwise and least angle regression. *arXiv:1401.3889*.

Tibshirani, R. (1996). Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society. Series B (Methodological)* **58**, 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.

van de Geer, S. A. (2008). High-dimensional generalized linear models and the Lasso. *The Annals of Statistics* **36**, 614–645.

Wainwright, M. J. (2009). Sharp thresholds for high-dimensional and noisy sparsity recovery using ℓ_1-constrained quadratic programming (Lasso). *IEEE Transactions on Information Theory* **55**, 2183–2202.

Wang, D. and Loh, P.-L. (2020). Adaptive estimation and statistical inference for high-dimensional graph-based linear models. *arXiv:2001.10679*.

Yang, F., Barber, R. F., Jain, P. and Lafferty, J. (2016). Selective inference for group-sparse linear models. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, 2477–2485. Curran Associates Inc., Barcelona.

Yu, M., Gupta, V. and Kolar, M. (2019). Constrained high dimensional statistical inference. *arXiv:1911.07319*.

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

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

Zhang, C.-H. and Huang, J. (2008). The sparsity and bias of the LASSO selection in high-dimensional linear regression. *The Annals of Statistics* **36**, 1567–1594.

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, X. and Cheng, G. (2017). Simultaneous inference for high-dimensional linear models. *Journal of the American Statistical Association* **112**, 757–768.

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

Yi Wang

The School of Statistics and Information, Shanghai University of International Business and Economics, Shanghai 201620, China.

Beijing International Center for Mathematical Research, Peking University, Beijing 100871, China.

E-mail: wangyi@bicmr.pku.edu.cn

Donglin Zeng

Department of Biostatistics, University of Michigan, Ann Arbor, MI 48109, USA.

E-mail: dzeng@umich.edu

Yuanjia Wang

Department of Biostatistics, Mailman School of Public Health, Columbia University, New York, NY 10032, USA.

E-mail: yw2016@cumc.columbia.edu

Xingwei Tong

School of Statistics, Beijing Normal University, Beijing 100875, China.

E-mail: xweitong@bnu.edu.cn

(Received March 2022; accepted March 2023)