Journal of Econometrics 230 (2022) 416–431

Contents lists available at ScienceDirect

Journal of Econometrics

journal homepage: www.elsevier.com/locate/jeconom

Robust post-selection inference of high-dimensional mean regression with heavy-tailed asymmetric or heteroskedastic errors

Dongxiao Han [a], Jian Huang [b], Yuanyuan Lin [c,*], Guohao Shen [c]

[a] School of Statistics and Data Science, LPMC and KLMDASR, Nankai University, Tianjin, China
[b] Department of Statistics and Actuarial Science, University of Iowa, IA, USA
[c] Department of Statistics, The Chinese University of Hong Kong, Shatin, New Territories, Hong Kong, China

ARTICLE INFO

Article history:
Received 11 November 2019
Received in revised form 5 April 2021
Accepted 17 May 2021
Available online 18 June 2021

Keywords:
Confidence interval
Huber loss
Linear model
Post-selection inference

ABSTRACT

We propose a robust post-selection inference method based on the Huber loss for the regression coefficients, when the error distribution is heavy-tailed and asymmetric in a high-dimensional linear model with an intercept term. The asymptotic properties of the resulting estimators are established under mild conditions. We also extend the proposed method to accommodate heteroscedasticity assuming the error terms are symmetric and other suitable conditions. Statistical tests for low-dimensional parameters or individual coefficient in the high-dimensional linear model are also studied. Simulation studies demonstrate desirable properties of the proposed method. An application to a genomic dataset about riboflavin production rate is provided.

1. Introduction

Massive and high-dimensional data have now become commonplace in various scientific disciplines, owing to the fast development in information technologies. There have been many novel statistical methodologies and computational algorithms developed for analyzing high-dimensional data. In particular, regularization methods have been successfully applied in high-dimensional regression problems. Examples include Meinshausen and Bühlmann (2006), Zhao and Yu (2006), Fan and Lv (2008), Huang et al. (2008), Zhang and Huang (2008), Wainwright (2009), Ishwaran et al. (2010), Zhang (2010), Bradic et al. (2011), Liu et al. (2014), among many others. However, variable selection procedures focus on point estimation. Since perfect model recovery may not be delivered by variable selection methods, statistical inference based on the selected model may give inaccurate or even wrong results. Statistical inference, including interval estimation and hypothesis testing with high-dimensional data, is largely untouched until the pioneering works of Zhang and Zhang (2014), van de Geer et al. (2014), Javanmard and Montanari (2014) and Belloni et al. (2015). Other important works include Janková and van de Geer (2015), Cai and Guo (2017), Belloni et al. (2019) and the references therein.

Indeed, there has been ever-increasing interest in developing post-selection inference methods with high-dimensional data in recent years. For high-dimensional linear models with sub-Gaussian errors, post-selection inference based on least squares estimation and novel debiase ideas are studied by Zhang and Zhang (2014), van de Geer et al. (2014), Javanmard and Montanari (2014), etc. Nevertheless, statistical procedures based on least squares methods are sensitive to outliers. The sub-Gaussian assumption is made for technical convenience but may not be realistic in many practical situations

* Corresponding author.
E-mail address: ylin@sta.cuhk.edu.hk (Y. Lin).

https://doi.org/10.1016/j.jeconom.2021.05.006

(Cont, 2001; Wang et al., 2015; Eklund et al., 2016), especially for data with heavy-tailed errors that are common in finance and economics (Fan et al., 2016, 2017). Robust methods based on quantile regression or least absolute deviation (LAD) are studied by Li and Zhu (2008), Zou and Yuan (2008), Wu and Liu (2009), Belloni and Chernozhukov (2011), Wang (2013), Fan et al. (2014), Belloni et al. (2015, 2019), Cheng et al. (2020), among many others. Other than the LAD and the quantile check loss, the Huber loss (Huber, 1964) is an important robust criterion for parameter estimation. Asymptotic properties of the Huber estimators have been studied extensively under the fixed or low-dimensional settings (Huber, 1973; Yohai and Maronna, 1979; Portnoy, 1985; He and Shao, 1996, 2000; Zhou et al., 2018). Recently, novel findings on adaptive robust estimation based on the Huber loss for high-dimensional mean regression are reported by Fan et al. (2017), Loh (2018), Wang et al. (2020), Sun et al. (2020). Specially, in the presence of asymmetric errors, Fan et al. (2017) and Sun et al. (2020) study Huber-type estimators and provide non-asymptotic estimation bounds. Under symmetric or asymmetric errors, data-driven robustification parameter selection can be found in Wang et al. (2020) and Loh (2018), respectively. Confidence intervals for low-dimensional parameters based on the de-sparsified lasso for high-dimensional generalized linear models are studied by Janková and van de Geer (2016). Post-selection inference for high-dimensional linear models based on the weighted Huber loss is considered by Loh (2018). Both works assume a linear model without an intercept term and assumptions on the error distribution, for example, symmetry around zero, are required to establish the asymptotic properties.

Despite these developments, different from least squares estimation for mean regression, the intercept in the linear model cannot be simply removed by centering the response variable with the Huber loss. In this article, we consider a high-dimensional linear model with an intercept term and develop a one-step post-selection inference procedure based on the Huber loss. Our proposed procedure is robust in the sense that the error term can be heavy-tailed and asymmetrically distributed. We also extend proposed method to accommodate heteroscedasticity when the error distribution is symmetric. Numerical studies confirm that our method is robust in various practical situations.

The rest of the article is organized as follows. Section 2.1 presents the model and the proposed inference procedure. Theoretical properties of the proposed estimators are given in Section 2.2. Statistical tests for single or low-dimensional components of the slope parameter vector are developed. Section 3 contains an extension of the proposed method to handle the heteroscedasticity. Supportive simulation results are reported in Section 4 and an application to a genomic dataset is provided in Section 5. A few closing remarks are given in Section 6. All proofs are deferred to the Supplementary Material.

2. Homoscedastic linear model

2.1. Model and estimation method

Consider the linear regression model

$$Y = \mu^* + X^\top \beta^* + \epsilon, \tag{1}$$

where $Y \in \mathbb{R}$ is a response variable, $X \in \mathbb{R}^d$ is a d-dimensional vector of covariates, ϵ is a zero-mean error term independent of X, $\mu^* \in \mathbb{R}$ and $\beta^* \in \mathbb{R}^d$ are the intercept and the slope parameter vector, respectively. The error term can be heavy-tailed and asymmetrically distributed. The observations (X_i, Y_i), $i = 1, \ldots, n$, are independent and identically distributed copies of (X, Y). Throughout the paper, we focus on the high-dimensional setting where d can be of the same order as n or greater than n, depending on the assumption on the design matrix. Our goal is to conduct post-selection inference for each component β_j^*, as well as simultaneous inference for $\beta_G^* := \{\beta_j^* : j \in G\}$, where β_j^* is the jth element of β^* and G is any fixed-dimensional subset of $\{1, \ldots, d\}$. Since our inference method is not based on least squares method, the intercept term μ^* in model (1) cannot be simply removed by centering the response and the predictors.

Owing to the fact that sparse estimators such as the lasso do not have a tractable limiting distribution, statistical inference with high-dimensional data is challenging, especially in the context of large d and relatively small n problems. Write $Z = (1, X^\top)^\top$ and $\theta^* = (\mu^*, \beta^{*\top})^\top$. To pursue robust estimation, we consider the Huber loss function

$$l_\tau(x) = \begin{cases} x^2/2 & \text{if } |x| \le \tau; \\ \tau|x| - \tau^2/2 & \text{if } |x| > \tau. \end{cases}$$

For any given $\tau > 0$, we define

$$\theta_\tau^* = (\mu_\tau^*, \beta_\tau^{*\top})^\top \equiv \underset{\theta \in \mathbb{R}^{d+1}}{\operatorname{argmin}} El_\tau(Y - Z^\top \theta).$$

Since the distribution of the error may not be symmetric, $\theta_\tau^* \neq \theta^*$ in general, indicating that the bias induced by the Huber loss is nonnegligible. Nonetheless, it is shown in Proposition 1 that the asymmetry of the error can only lead to biased estimation of the intercept μ^*, but not the slope parameter vector β^*. Wang et al. (2020) also pointed out this phenomenon under slightly different assumptions. With this view, the Huber loss can provide a leeway to perform post-selection inference for the slope parameters.

To estimate the regression coefficient β_j^*, we first consider minimizing the objective function:

$$l(\beta_j) \equiv \frac{1}{n}\sum_{i=1}^{n} l_\tau(Y_i - \mu_\tau^{init} - X_{i,-j}^\top \beta_{\tau,-j}^{init} - X_{i,j}\beta_j), \tag{2}$$

where μ_τ^{init} and β_τ^{init} are certain initial estimators of μ_τ^* and β^*, $\beta_{\tau,-j}^{init} \equiv \{\beta_{\tau,k}^{init} : k \neq j\}$, and $X_{i,-j} \equiv \{X_{i,k} : k \neq j\}$. However, it is known that the asymptotic normality of the estimator for β_j^* by minimizing (2) cannot be established if the initial estimators μ_τ^{init} and β_τ^{init} are not $n^{1/2}$-consistent. Thus, regularized estimators cannot serve as initial estimators. To tackle this problem, inspired by the ideas of orthogonalization (Neyman, 1959; Zhang and Zhang, 2014; Belloni et al., 2015, 2019) and decorrelated score (Ning and Liu, 2017), we consider the following estimating equation for β_j:

$$\frac{1}{n}\sum_{i=1}^{n}(-X_{i,j} + Z_{i,-(j+1)}^\top \widehat{\gamma}_j)\psi_\tau(Y_i - \mu_\tau^{init} - X_{i,-j}^\top \beta_{\tau,-j}^{init} - X_{i,j}\beta_j) = 0, \tag{3}$$

where $\psi_\tau(x) = dl_\tau(x)/dx$, $\widehat{\gamma}_j$ is a consistent estimator of γ_j^*, and

$$\gamma_j^* \equiv \operatorname*{argmin}_{\gamma_j} E(X_{i,j} - Z_{i,-(j+1)}^\top \gamma_j)^2.$$

It can be easily verified that the estimating Eq. (3) corresponds to the following orthogonality property

$$\frac{\partial}{\partial \eta} E\{(-X_{i,j} + Z_{i,-(j+1)}^\top \gamma_j)\psi_\tau(Y_i - \mu - X_{i,-j}^\top \beta_{-j} - X_{i,j}\beta_j^*)\}|_{\eta=\eta^*} = 0, \tag{4}$$

where $\eta = (\gamma_j^\top, \mu, \beta_{-j}^\top)^\top$, and $\eta^* = (\gamma_j^{*\top}, \mu_\tau^*, \beta_{-j}^{*\top})^\top$. The orthogonal property in (4) ensures that the convergence rate of the estimator of β_j^* derived from (3) will not be affected by the estimation of μ_τ^{init} and $\beta_{\tau,-j}^{init}$, namely, μ_τ^{init} and $\beta_{\tau,-j}^{init}$ are allowed to converge to μ_τ^* and β_{-j}^* at a slower rate than $n^{-1/2}$, for instance, $o(n^{-1/4})$. However, solving (3) directly is numerically inconvenient due to the discontinuity of the indicator function and sign function. Invoking the idea of one-step estimation in Bickel (1975), we define

$$S(\beta_j) \equiv E\{(-X_{i,j} + Z_{i,-(j+1)}^\top \gamma_j^*)\psi_\tau(Y_i - \mu_\tau^* - X_{i,-j}^\top \beta_{\tau,-j}^* - X_{i,j}\beta_j)\}$$

and

$$\dot{S}(\beta_j) \equiv -E\{X_{i,j}(-X_{i,j} + Z_{i,-(j+1)}^\top \gamma_j^*)I(|Y_i - \mu_\tau^* - X_{i,-j}^\top \beta_{\tau,-j}^* - X_{i,j}\beta_j| \le \tau)\}.$$

It can be verified that $\dot{S}(\beta_j)$ is the derivative of $S(\beta_j)$ with respect to β_j. Let $\epsilon_{i,\tau}^{init} = Y_i - \mu_\tau^{init} - X_i^\top \beta_\tau^{init}$. Instead of solving (3) for β_j, we consider a one-step estimator:

$$\widehat{\beta}_{\tau,j} = \beta_{\tau,j}^{init} + \{\dot{S}(\beta_j^*)\}^{-1}\frac{1}{n}\sum_{i=1}^{n}(X_{i,j} - Z_{i,-(j+1)}^\top \widehat{\gamma}_j)\psi_\tau(\epsilon_{i,\tau}^{init}). \tag{5}$$

Since $\{\dot{S}(\beta_j^*)\}^{-1}$ is unknown, we plug in its empirical counterpart into (5) and obtain the proposed estimator

$$\widehat{\beta}_{\tau,j} = \beta_{\tau,j}^{init} + \frac{\sum_{i=1}^{n}(X_{i,j} - Z_{i,-(j+1)}^\top \widehat{\gamma}_j)\psi_\tau(\epsilon_{i,\tau}^{init})}{\sum_{i=1}^{n} X_{i,j}(X_{i,j} - Z_{i,-(j+1)}^\top \widehat{\gamma}_j) \times \frac{1}{n}\sum_{i=1}^{n} I(|\epsilon_{i,\tau}^{init}| \le \tau)}. \tag{6}$$

For the initial estimator, in view of the popularity and simplicity of the lasso (Tibshirani, 1996), we let the initial estimators μ_τ^{init}, β_τ^{init} be the minimizer of

$$\frac{1}{2n}\sum_{i=1}^{n} l_\tau(Y_i - Z_i^\top \theta) + \lambda_n \|\beta\|_1, \tag{7}$$

where λ_n is a tuning parameter and $\|\beta\|_1 = \sum_{j=1}^{d} |\beta_j|$. Meanwhile, according to van de Geer et al. (2014), one can obtain an appropriate estimator of γ_j^*, denoted by $\widehat{\gamma}_j$, by minimizing

$$W_n(\gamma_j) \equiv \frac{1}{n}\sum_{i=1}^{n}(X_{i,j} - Z_{i,-(j+1)}^\top \gamma_j)^2 + \omega_j \|\gamma_j\|_1, \tag{8}$$

over γ_j, where ω_j is a regularization parameter.

2.2. Theoretical results

We first define the notation needed below. Let $\|\cdot\|_2$ be the Euclidean norm, $\|\cdot\|_0$ be the number of nonzero components of a vector and $\|\cdot\|_\infty$ be the maximal absolute value in the components of a vector. Define $s = 1 + \|\beta^*\|_0$ and $S = E(ZZ^\top)$.

Let S_j^{-1} be the $(j+1)$-th row of the inverse matrix of S. The notation $a \asymp b$ represents that there exist two positive constants e_0 and e_1 such that $e_0 a \le b \le e_1 a$.

We assume the following conditions in Theorems 1–3 and Proposition 1.

(C1) The error ϵ is an absolutely continuous random variable with a cumulative distribution function $F_\epsilon(x)$. There exists a positive constant M_1 such that $E|\epsilon| < M_1$.

(C2) There exist two positive constants m and M such that

$$m \le \inf_{\|\Delta\|_2 \neq 0, \Delta \in R^{d+1}} \frac{\|S^{1/2}\Delta\|_2^2}{\|\Delta\|_2^2} \le \sup_{\|\Delta\|_2 \neq 0, \Delta \in R^{d+1}} \frac{\|S^{1/2}\Delta\|_2^2}{\|\Delta\|_2^2} \le M.$$

(C3) There exists a positive constant A_0 such that, for any $a \in R^{d+1}$ and any $t > 0$,

$$P(|\langle a, z\rangle| \ge A_0 \|a\|_2 t) \le 2e^{-t^2},$$

where $z = S^{-1/2}Z$ and $\langle a, z\rangle = a^\top z$.

(C4) The dimensionality d and s satisfy $s^2 \log(d+1)/n = o(1)$.

(C5) There exist two positive constants N_1 and N_2 such that $\|\theta_\tau^*\|_2 \le N_2$ for all $\tau \ge N_1$. Also, there exists a positive constant L such that $\sup_x f_\epsilon(x) \le L$, where $f_\epsilon(x)$ is the density function of ϵ.

(C6) The regularization parameter ω_j in (8) satisfies $\omega_j \asymp \sqrt{\log(d+1)/n}$. Suppose that $\|S_j^{-1}\|_0 \le s_1$ for some positive integer s_1, and $s_1^3 s^3 \log^3(d+1) = o(n^\alpha)$ for some $\alpha \in (0, 1)$. In addition, we assume Condition (C6)(a) or Condition (C6)(b) below holds:

(C6)(a) Let $\lambda_{\max} \equiv \|(1/n)\sum_{i=1}^n Z_i Z_i^\top - S\|_{sp}$, where $\|A\|_{sp}$ is the spectral norm of a matrix A, i.e., the square root of the largest eigenvalue of $A^\top A$. Assume that $\lambda_{\max} = O_p(\max(\sqrt{d/n}, d/n))$ and $d = O(n)$.

(C6)(b) There exists a positive constant N_3 such that with probability tending to one,

$$\sup_{\substack{\|x\|_0 \le 2n/\log(d+1) \\ \|x\|_2 = 1}} x^\top \left(\frac{1}{n}\sum_{i=1}^n Z_i Z_i^\top\right) x \le N_3.$$

Condition (C1) ensures that $\beta_\tau^* = \beta^*$ for any $\tau > 0$. The bounded first moment condition is needed to prove that the restricted strong convexity condition (Fan et al., 2017) is satisfied by the Huber loss. Many commonly-used distributions such as normal distribution, shifted Chi-square distribution, Student's t-distribution with degrees of freedom greater than 1, satisfy Condition (C1). Conditions (C2)-(C4) are regularity conditions for high-dimensional models (van de Geer et al., 2014; Fan et al., 2017). The boundedness conditions in (C5) are assumed for technical convenience. It follows from Theorem 4.7.1 in Vershynin (2018, page 94) and Condition (C2) that Condition (C6)(a) holds for sub-Gaussian covariates. Condition (C6) part (a) indicates that d can have the same order as n; Condition (C6)(b) holds for Gaussian covariates, and it allows d to grow at an exponential rate of n (Belloni and Chernozhukov, 2011).

Remark 1. When d is fixed, by Theorem 4.7.1 in Vershynin (2018), Markov's inequality and Condition (C2), it follows that sub-Gaussian covariates satisfy $\|(1/n)\sum_{i=1}^n Z_i Z_i^\top - S\|_{sp} = O_p(n^{-1/2})$. Therefore, $\lambda_{\max} = O_p(\max(\sqrt{d/n}, d/n))$ in Condition (C6)(a) is satisfied for fixed d.

The following proposition shows that, for the mean regression under an asymmetrical error distribution, estimation based on the Huber loss function still delivers unbiased estimators for the slope parameters but a biased estimator for the intercept term.

Proposition 1. *Under Conditions (C1) and (C2), for any $\tau > 0$, there exists a constant μ_τ depending on τ, such that $\mu_\tau^* = \mu^* - \mu_\tau$ and $\beta_\tau^* = \beta^*$.*

Remark 2. We note that an independent work (Wang et al., 2020) reports the same result under the assumptions that $E\{l_\tau(\epsilon - \alpha)\}$ has a unique minimizer and $P(|\epsilon - \mu_\tau|) > 0$. We prove Proposition 1 under milder conditions that are easier to verify in practice.

The next theorem establishes the consistency of $\theta_\tau^{init} \equiv (\mu_\tau^{init}, (\beta_\tau^{init})^\top)^\top$.

Theorem 1. *Assume Conditions (C1)-(C5) hold. Then, there exists a positive constant c_0 depending on A_0, m, M, M_1, N_1 and N_2, such that when $\tau \ge c_0$ and $\lambda_n = \kappa_\tau \sqrt{\log(d+1)/n}$, with probability at least $1 - (1+e)/(1+d) - c_1 \exp(-c_2 n)$,*

$$\|\theta_\tau^{init} - \theta_\tau^*\|_2 \le f_{\tau,s}\sqrt{\frac{\log(d+1)}{n}}$$

and

$$\|\theta_\tau^{init} - \theta_\tau^*\|_1 \le 4\sqrt{s} f_{\tau,s}\sqrt{\frac{\log(d+1)}{n}},$$

where e is Euler's number, κ_τ could be any positive constant no less than $\tilde{\kappa}_\tau$ which depends on τ, A_0 and M, c_1 and c_2 are two positive constants depending on A_0, m, M, M_1, N_1 and N_2, and $f_{\tau,s}$ depends on A_0, m, M, M_1, N_1, N_2, κ_τ and s.

The next theorem provides the asymptotic distribution of $\widehat{\beta}_{\tau,j}$, which enables us to construct confidence intervals for β_j^*.

Theorem 2. *Under Conditions (C1)-(C6), for any $\tau \geq c_0$ and $\lambda_n = \kappa_\tau\sqrt{\log(d+1)/n}$, $\sigma_{\tau,j}^{-1} n^{1/2}(\widehat{\beta}_{\tau,j} - \beta_j^*) \to N(0,1)$ in distribution as $n \to \infty$, where κ_τ could be any positive constant no less than $\check{\kappa}_\tau$ which depends on τ, A_0, M and N_3, $\sigma_{\tau,j}^2 = E\{\epsilon_{i,\tau}^2 I(|\epsilon_{i,\tau}| \leq \tau) + \tau^2 I(|\epsilon_{i,\tau}| > \tau|)\}/[E\{(X_{i,j}^{\perp})^2\}P^2(|\epsilon_{i,\tau}| \leq \tau)]$, $X_{i,j}^{\perp} = X_{i,j} - Z_{i,-(j+1)}^{\top}\gamma_j^*$ and $\epsilon_{i,\tau} = Y_i - \mu_\tau^* - X_i^{\top}\beta^*$.*

The asymptotic variance of $n^{1/2}(\widehat{\beta}_{\tau,j} - \beta_j^*)$ can be consistently estimated by

$$\widehat{\sigma}_{\tau,j}^2 = \frac{\sum_{i=1}^n (\epsilon_{i,\tau}^{init})^2 I(|\epsilon_{i,\tau}^{init}| \leq \tau) + \tau^2 I(|\epsilon_{i,\tau}^{init} > \tau|)}{\sum_{i=1}^n X_{i,j}(X_{i,j} - Z_{i,-(j+1)}^{\top}\widehat{\gamma}_j)\{\frac{1}{n}\sum_{i=1}^n I(|\epsilon_{i,\tau}^{init}| \leq \tau)\}^2}.$$

We next present the confidence intervals for β_j^* in the following corollary.

Corollary 1. *Under Conditions (C1)-(C6), for any $\tau \geq c_0$ and $\lambda_n = \kappa_\tau\sqrt{\log(d+1)/n}$, and any $0 < \tilde{\xi} < 1$,*

$$\left|P\{\beta_j^* \in [\widehat{\beta}_{\tau,j} \pm n^{-1/2}\widehat{\sigma}_{\tau,j}\Phi^{-1}(1-\tilde{\xi}/2)]\} - (1-\tilde{\xi})\right| \to 0$$

as $n \to \infty$, where $\Phi(\cdot)$ is the cumulative distribution function of the standard normal distribution and $\Phi^{-1}(\cdot)$ is its inverse function.

Let $\widehat{v}_j = (1/n)\sum_{i=1}^n X_{i,j}(X_{i,j} - Z_{i,-(j+1)}^{\top}\widehat{\gamma}_j)$, $\widehat{\rho}_j = (-\widehat{\gamma}_{j,1}, \ldots, -\widehat{\gamma}_{j,j}, 1, -\widehat{\gamma}_{j,j+1}, \ldots, -\widehat{\gamma}_{j,d})$, $\widehat{\Sigma}_{\tau,j}^{-1} = \widehat{\rho}_j/\{\widehat{v}_j(1/n)\sum_{i=1}^n I(|\epsilon_{i,\tau}^{init}| \leq \tau)\}$, $\widehat{\Sigma}_{\tau,G}^{-1} = \{\widehat{\Sigma}_{\tau,j}^{-1} : j \in G\}$, and $\widehat{S} = (1/n)\sum_{i=1}^n Z_i Z_i^{\top}$.

In addition to Conditions (C1)-(C6), the following condition is needed for simultaneous inference for β_G^*.

(C7) $\max_{j\in G} \|S_j^{-1}\|_0 \leq s_1$ and $\omega_j \asymp \sqrt{\log(d+1)/n}$ uniformly in $j \in G$.

Condition (C7) is a standard assumption in the context of post-selection inference (van de Geer et al., 2014).

Theorem 3. *Under Conditions (C1)-(C7), for any $\tau \geq c_0$ and $\lambda_n = \kappa_\tau\sqrt{\log(d+1)/n}$, and any fixed-dimensional subset $G \subset \{1, \ldots, d\}$, we have*

$$\widehat{\beta}_{\tau,G} - \beta_G^* = \widehat{\Sigma}_{\tau,G}^{-1}\frac{1}{n}\sum_{i=1}^n Z_i\psi_\tau(\epsilon_{i,\tau}) + \widehat{\beta}_{\tau,G}^{rem}, \ \ \|\widehat{\beta}_{\tau,G}^{rem}\|_\infty = o_p(n^{-1/2}),$$

where $\widehat{\beta}_{\tau,G} = \{\widehat{\beta}_{\tau,j} : j \in G\}$, and κ_τ could be any positive constant no less than $\check{\kappa}_\tau$ which depends on τ, A_0, M and N_3.

Define

$$\widehat{\Omega} = \frac{1}{n}\sum_{i=1}^n \{(\epsilon_{i,\tau}^{init})^2 I(|\epsilon_{i,\tau}^{init}| \leq \tau) + \tau^2 I(|\epsilon_{i,\tau}^{init} > \tau|)\}\widehat{\Sigma}_{\tau,G}^{-1}\widehat{S}(\widehat{\Sigma}_{\tau,G}^{-1})^{\top}.$$

By Theorem 3, under $H_{0,G} : \beta_j^* = 0, \ \forall j \in G$, the distribution of $\|\sqrt{n}\widehat{\Omega}^{-1/2}\widehat{\beta}_{\tau,G}\|_2^2$ is asymptotically equal to $\chi^2(|G|)$, where $|G|$ is the cardinality of the set G. Let $u_{\tilde{\xi}}$ be the (1-$\tilde{\xi}$)-quantile of $\chi^2(|G|)$. One may reject $H_{0,G}$ if $\|\sqrt{n}\widehat{\Omega}^{-1/2}\widehat{\beta}_{\tau,G}\|_2^2 > u_{\tilde{\xi}}$.

Remark 3. Janková and van de Geer (2016) developed a similar one-step debiased estimator and pointwise post-selection inference, which requires the error is symmetrically distributed, for high-dimensional homoscedastic linear models. Compared with Janková and van de Geer (2016), our proposed method is valid for a homoscedastic linear model with an intercept term, allowing the error distribution is asymmetric; when the error distribution is symmetric, our method can be extended to accommodate heteroscedasticity.

3. Heteroscedastic linear model

3.1. Model and estimation method

In this section, we consider model (1) when $\epsilon_1, \ldots, \epsilon_n$ are independent but not identically distributed. Similar to Section 2, for any given $\tau > 0$, we define

$$\bar{\theta}_\tau^* = (\bar{\mu}_\tau^*, \bar{\beta}_\tau^{*\top})^{\top} \equiv \underset{\theta\in\mathbb{R}^{d+1}}{\operatorname{argmin}} E\{\sum_{i=1}^n l_\tau(Y_i - Z_i^{\top}\theta)\}.$$

The following assumption is needed for parameter identifiability.

(D1) For any $1 \le i \le n$, the distribution of ϵ_i is symmetric. There exists a positive constant N_4 such that for any $\tau > N_4$, the function $\theta \to E\{\sum_{i=1}^n l_\tau(Y_i - Z_i^\top\theta)\}$ has a unique minimizer $\bar\theta_\tau^*$.

In the presence of heteroscedasticity, the symmetry assumption of the error distributions ensures that the true parameter θ^* in model (1) is a minimizer of $E\{\sum_{i=1}^n l_\tau(Y_i - Z_i^\top\theta)\}$. For parameter identifiability, we also need the assumption that $E\{\sum_{i=1}^n l_\tau(Y_i - Z_i^\top\theta)\}$ has a unique minimizer. The following proposition shows that the target parameter $\bar\theta_\tau^*$ coincides with θ^*.

Proposition 2. *Under Condition (D1), for any $\tau > N_4$, we have $\bar\theta_\tau^* = \theta^*$.*

We extend the method proposed in Section 2 for the inference of β_j^*. To avoid technical complications arising from the heteroscedasticity, a data-splitting technique is employed. Without loss of generality, we assume that n is an even number. Given the observations $(X_1, Y_1), \ldots, (X_n, Y_n)$, the first sub-sample $\{(X_i, Y_i)_{i=1}^{n/2}\}$ are used to construct an initial estimator $\widehat\theta_\tau^{init1}$ of θ^* and an estimator $\widehat\gamma_j^{(1)}$ of γ_j^* according to (7) and (8) respectively. Similarly, $\widehat\theta_\tau^{init2}$ and $\widehat\gamma_j^{(2)}$ can be calculated with the second sub-sample $\{(X_i, Y_i)_{i=n/2+1}^{n}\}$. Define

$$\widehat\beta_{\tau,j}^{(1)} = \beta_{\tau,j}^{init1} + \frac{\sum_{i=n/2+1}^n (X_{i,j} - Z_{i,-(j+1)}^\top \widehat\gamma_j^{(2)})\psi_\tau(\epsilon_{i,\tau}^{init1})}{\sum_{i=n/2+1}^n X_{i,j}(X_{i,j} - Z_{i,-(j+1)}^\top \widehat\gamma_j^{(2)}) \times \frac{2}{n}\sum_{i=n/2+1}^n I(|\epsilon_{i,\tau}^{init1}| \le \tau)}$$

and

$$\widehat\beta_{\tau,j}^{(2)} = \beta_{\tau,j}^{init2} + \frac{\sum_{i=1}^{n/2} (X_{i,j} - Z_{i,-(j+1)}^\top \widehat\gamma_j^{(1)})\psi_\tau(\epsilon_{i,\tau}^{init2})}{\sum_{i=1}^{n/2} X_{i,j}(X_{i,j} - Z_{i,-(j+1)}^\top \widehat\gamma_j^{(1)}) \times \frac{2}{n}\sum_{i=1}^{n/2} I(|\epsilon_{i,\tau}^{init2}| \le \tau)}.$$

where $\epsilon_{i,\tau}^{init1} = Y_i - \mu_\tau^{init1} - X_i^\top\beta_\tau^{init1}$ and $\epsilon_{i,\tau}^{init2} = Y_i - \mu_\tau^{init2} - X_i^\top\beta_\tau^{init2}$. To avoid efficiency loss due to the sample splitting, we propose the following average estimator

$$\widehat\beta_{\tau,j}^{avg} = (\widehat\beta_{\tau,j}^{(1)} + \widehat\beta_{\tau,j}^{(2)})/2.$$

The asymptotic properties of $\widehat\beta_{\tau,j}^{avg}$ are presented in the next subsection.

3.2. Theoretical results

Let

$$\sigma_{n,\tau,j} = \frac{1}{n}\sum_{i=1}^{n/2} \frac{E\{\epsilon_i^2 I(|\epsilon_i| \le \tau) + \tau^2 I(|\epsilon_i > \tau|)\}}{E\{(X_{i,j}^\perp)^2\}\{\frac{2}{n}\sum_{i=1}^{n/2} P^2(|\epsilon_i| \le \tau)\}^2} + \frac{1}{n}\sum_{i=n/2+1}^{n} \frac{E\{\epsilon_i^2 I(|\epsilon_i| \le \tau) + \tau^2 I(|\epsilon_i > \tau|)\}}{E\{(X_{i,j}^\perp)^2\}\{\frac{2}{n}\sum_{i=n/2+1}^{n} P^2(|\epsilon_i| \le \tau)\}^2}.$$

Apart from Conditions (C2)-(C7), additional assumptions are needed.

(D2) There exists a positive constant M_2 such that $\max_{1\le i\le n} E|\epsilon_i| < M_2$.

(D3) There exists a positive constant M_3 such that for any $\tau > N_4$, $\max\{2\sum_{i=1}^{n/2} E\{I(|\epsilon_i| \le \tau)\}/n, 2\sum_{i=n/2+1}^{n} E\{I(|\epsilon_i| \le \tau)\}/n\} > M_3$ and $\sigma_{n,\tau,j} > M_3$.

(D4) Assume that $\max_{1\le i\le n}\sup_x f_{\epsilon_i}(x) \le L$, where $f_{\epsilon_i}(x)$ is the density function of ϵ_i. Moreover, $\|S_j^{-1}\|_0 \le s_1$, $(s_1^2 s + s^2 s_1)\log^2(d+1) = o(n)$ and $\omega_j \asymp \sqrt{\log(d+1)/n}$.

The next theorem establishes the consistency of $\widehat\theta_\tau^{init1}$ and $\widehat\theta_\tau^{init2}$.

Theorem 4. *Under Conditions (C2)-(C4), (D1) and (D2), there exists a positive constant c_0' depending on A_0, m, M_2, and N_4 such that for $\tau \ge c_0'$ and $\lambda_n = \kappa_\tau'\sqrt{\log(d+1)/n}$, with probability at least $1 - (1+e)/(1+d) - c_1'\exp(-c_2' n)$,*

$$\|\theta_\tau^{init1} - \theta^*\|_2 \le f_{\tau,s}'\sqrt{\frac{\log(d+1)}{n}}, \quad \|\theta_\tau^{init2} - \theta^*\|_2 \le f_{\tau,s}'\sqrt{\frac{\log(d+1)}{n}},$$

and

$$\|\theta_\tau^{init1} - \theta^*\|_1 \le 4\sqrt{s} f_{\tau,s}'\sqrt{\frac{\log(d+1)}{n}}, \quad \|\theta_\tau^{init2} - \theta^*\|_1 \le 4\sqrt{s} f_{\tau,s}'\sqrt{\frac{\log(d+1)}{n}},$$

where κ_τ' is a positive constant depending on τ, A_0 and M, c_1' and c_2' are two positive constants depending on A_0, m, M_2, and N_4, and $f_{\tau,s}'$ depends on A_0, m, M, M_2, N_4, κ_τ' and s.

The asymptotic distribution of $\widehat{\beta}_{\tau,j}^{avg}$ is given in the next theorem.

Theorem 5. *Under Conditions (C2)-(C4) and (D1)-(D4), for any $\tau \geq c_0'$ and $\lambda_n = \kappa_\tau' \sqrt{\log(d+1)/n}$, $\sigma_{n,\tau,j}^{-1} n^{1/2}(\widehat{\beta}_{\tau,j}^{avg} - \beta_j^*) \to N(0,1)$ in distribution as $n \to \infty$.*

The asymptotic variance of $n^{1/2}(\widehat{\beta}_{\tau,j}^{avg} - \beta_j^*)$ can be consistently estimated by

$$\widehat{\sigma}_{n,\tau,j}^2 = \frac{\sum_{i=1}^{n/2}(\epsilon_{i,\tau}^{init2})^2 I(|\epsilon_{i,\tau}^{init2}| \leq \tau) + \tau^2 I(|\epsilon_{i,\tau}^{init2}| > \tau|)}{2\sum_{i=1}^{n/2} X_{i,j}(X_{i,j} - Z_{i,-(j+1)}^\top \widehat{\gamma}_j^{(1)})\{\frac{2}{n}\sum_{i=1}^{n/2} I(|\epsilon_{i,\tau}^{init2}| \leq \tau)\}^2} + \frac{\sum_{i=n/2+1}^{n}(\epsilon_{i,\tau}^{init1})^2 I(|\epsilon_{i,\tau}^{init1}| \leq \tau) + \tau^2 I(|\epsilon_{i,\tau}^{init1}| > \tau|)}{2\sum_{i=n/2+1}^{n} X_{i,j}(X_{i,j} - Z_{i,-(j+1)}^\top \widehat{\gamma}_j^{(2)})\{\frac{2}{n}\sum_{i=n/2+1}^{n} I(|\epsilon_{i,\tau}^{init1}| \leq \tau)\}^2}.$$

We provide the pointwise confidence interval for β_j^* in the next corollary.

Corollary 2. *Under Conditions (C2)-(C4) and (D1)-(D4), for any $\tau \geq c_0'$ and $\lambda_n = \kappa_\tau' \sqrt{\log(d+1)/n}$, and any $0 < \tilde{\xi} < 1$,*

$$\left|P\{\beta_j^* \in [\widehat{\beta}_{\tau,j}^{avg} \pm n^{-1/2}\widehat{\sigma}_{n,\tau,j}\Phi^{-1}(1-\tilde{\xi}/2)]\} - (1-\tilde{\xi})\right| \to 0$$

as $n \to \infty$, where $\Phi(\cdot)$ and $\Phi^{-1}(\cdot)$ are the same as in Corollary 1.

Let $\widehat{v}_j^{(1)} = (2/n)\sum_{i=1}^{n/2} X_{i,j}(X_{i,j} - Z_{i,-(j+1)}^\top \widehat{\gamma}_j^{(1)})$, $\widehat{v}_j^{(2)} = (2/n)\sum_{i=n/2+1}^{n} X_{i,j}(X_{i,j} - Z_{i,-(j+1)}^\top \widehat{\gamma}_j^{(2)})$, $\widehat{\rho}_j^{(1)} = (-\widehat{\gamma}_{j,1}^{(1)}, \ldots, -\widehat{\gamma}_{j,j}^{(1)}, 1, -\widehat{\gamma}_{j,j+1}^{(1)}, \ldots, -\widehat{\gamma}_{j,d}^{(1)})$, $\widehat{\rho}_j^{(2)} = (-\widehat{\gamma}_{j,1}^{(2)}, \ldots, -\widehat{\gamma}_{j,j}^{(2)}, 1, -\widehat{\gamma}_{j,j+1}^{(2)}, \ldots, -\widehat{\gamma}_{j,d}^{(2)})$, $\widehat{\Omega}_{\tau,j}^{(1)} = \widehat{\rho}_j^{(1)}/\{\widehat{v}_j^{(1)}(2/n)\sum_{i=1}^{n/2} I(|\epsilon_{i,\tau}^{init2}| \leq \tau)\}$, $\widehat{\Omega}_{\tau,j}^{(2)} = \widehat{\rho}_j^{(2)}/\{\widehat{v}_j^{(2)}(2/n)\sum_{i=n/2+1}^{n} I(|\epsilon_{i,\tau}^{init1}| \leq \tau)\}$, $\widehat{\Omega}_{\tau,G}^{(1)} = \{\widehat{\Omega}_{\tau,j}^{(1)} : j \in G\}$ and $\widehat{\Omega}_{\tau,G}^{(2)} = \{\widehat{\Omega}_{\tau,j}^{(2)} : j \in G\}$. The next theorem provides theoretical guarantee for simultaneous inference of β_G^*.

Theorem 6. *Under Conditions (C2)-(C4), (C7) and (D1)-(D4), for any $\tau \geq c_0'$ and $\lambda_n = \kappa_\tau' \sqrt{\log(d+1)/n}$, and any fixed-dimensional subset $G \subset \{1, \ldots, d\}$, we have*

$$\widehat{\beta}_{\tau,G}^{avg} - \beta_G^* = \widehat{\Omega}_{\tau,G}^{(1)} \frac{1}{n}\sum_{i=1}^{n/2} Z_i \psi_\tau(\epsilon_i) + \widehat{\Omega}_{\tau,G}^{(2)} \frac{1}{n}\sum_{i=n/2+1}^{n} Z_i \psi_\tau(\epsilon_i) + \widetilde{\beta}_{\tau,G}^{rem}, \ \|\widetilde{\beta}_{\tau,G}^{rem}\|_\infty = o_p(n^{-1/2}),$$

where $\widehat{\beta}_{\tau,G}^{avg} = \{\widehat{\beta}_{\tau,j}^{avg} : j \in G\}$.

Define

$$\widetilde{\Omega} = \frac{1}{n}\sum_{i=1}^{n/2}\{(\epsilon_{i,\tau}^{init2})^2 I(|\epsilon_{i,\tau}^{init2}| \leq \tau) + \tau^2 I(|\epsilon_{i,\tau}^{init2}| > \tau|)\}\widehat{\Omega}_{\tau,G}^{(1)}\widehat{S}(\widehat{\Omega}_{\tau,G}^{(1)})^\top + \frac{1}{n}\sum_{i=n/2+1}^{n}\{(\epsilon_{i,\tau}^{init1})^2 I(|\epsilon_{i,\tau}^{init1}| \leq \tau) + \tau^2 I(|\epsilon_{i,\tau}^{init1}| > \tau|)\}\widehat{\Omega}_{\tau,G}^{(2)}\widehat{S}(\widehat{\Omega}_{\tau,G}^{(2)})^\top.$$

By Theorem 6, under $H_{0,G} : \beta_j^* = 0, \ \forall j \in G$, $\|\sqrt{n}\widetilde{\Omega}^{-1/2}\widehat{\beta}_{\tau,G}^{avg}\|_2^2$ converges to $\chi^2(|G|)$ in distribution as $n \to \infty$, where $|G|$ is the cardinality of the set G. As a result, one may reject $H_{0,G}$ if $\|\sqrt{n}\widetilde{\Omega}^{-1/2}\widehat{\beta}_{\tau,G}^{avg}\|_2^2 > u_{\tilde{\xi}}$, where $u_{\tilde{\xi}}$ is the (1-$\tilde{\xi}$)-quantile of $\chi^2(|G|)$.

4. Simulation

We carry out extensive simulation studies to evaluate the finite-sampler performance of the proposed method.

4.1. Simulated data: various error distributions and relatively strong signal

We first investigate the performance when the signals of the regression parameters are relatively strong. The first 5 components of β^* are set to be 1 and the rest components of β^* are 0 (setting (i)). Similar to Janková (2017), we generate the covariate vector X from $N(0, \Lambda^{-1})$, where $\Lambda_{j,j} = 1$; $\Lambda_{j,k} = 0.5$ if $|j-k| = 1$; $\Lambda_{j,k} = 0.4$ if $|j-k| = 2$; otherwise, $\Lambda_{j,k} = 0$. Five error distributions are tried:

(A) Standard normal distribution, shorted as $N(0,1)$;

(B) Student's t distribution with degrees of freedom 3, shorted as $t(3)$;

(C) Shifted Chi-square distribution with degrees of freedom 8, shorted as $\chi^2(8) - 8$;

(D) Skewed normal distribution with location parameter 0, scale parameter 1 and shape parameter 1, shorted as $SN(0,1,1)$;

(E) Pareto distribution with scale parameter 1 and shape parameter 2; shorted as $P(1,2)$.

Table 1
The average of the coverage probabilities (ACP) of the 95% confidence intervals and the average length (AL) of the 95% confidence intervals over S_∞^1, S_∞^{1c}, and $\lfloor d \rfloor$, respectively.

β^* in Setting (i)	Method	ACP			AL		
		S_∞^1	S_∞^{1c}	$\lfloor d \rfloor$	S_∞^1	S_∞^{1c}	$\lfloor d \rfloor$
$n = 200, d = 400, N(0, 1)$	Proposed	0.918	0.981	0.980	0.369	0.416	0.415
	Z. & Z.	0.922	0.950	0.950	0.392	0.402	0.402
$n = 200, d = 400, t(3)$	Proposed	0.918	0.979	0.978	0.453	0.508	0.507
	Z. & Z.	0.921	0.946	0.946	0.490	0.505	0.504
$n = 300, d = 400, N(0, 1)$	Proposed	0.942	0.978	0.977	0.313	0.361	0.360
	Z. & Z.	0.942	0.943	0.943	0.332	0.348	0.347
$n = 300, d = 400, t(3)$	Proposed	0.912	0.978	0.977	0.371	0.426	0.425
	Z. & Z.	0.915	0.950	0.949	0.430	0.449	0.449
$n = 200, d = 800, N(0, 1)$	Proposed	0.906	0.979	0.978	0.394	0.400	0.400
	Z. & Z.	0.904	0.951	0.950	0.384	0.392	0.392
$n = 200, d = 800, t(3)$	Proposed	0.916	0.982	0.981	0.436	0.473	0.473
	Z. & Z.	0.928	0.950	0.949	0.481	0.493	0.493
$n = 300, d = 800, N(0, 1)$	Proposed	0.918	0.981	0.980	0.306	0.343	0.342
	Z. & Z.	0.890	0.952	0.952	0.320	0.332	0.332
$n = 300, d = 800, t(3)$	Proposed	0.922	0.981	0.980	0.367	0.409	0.409
	Z. & Z.	0.920	0.952	0.952	0.378	0.389	0.389

Notes: Z. & Z. stands for the method in Zhang and Zhang (2014).

For cases (A)-(C), we set the intercept $\mu^* = 0$ for comparison with the method by Zhang and Zhang (2014). In cases (D) and (E), for identifiability, the mean of the skewed normal variable and the Pareto variable are shifted to the intercept term. As a result, cases (D) and (E) could be used to examine the performance of the proposed method under a linear model with a non-zero intercept term. Note that the variance of the Pareto distribution in case (E) is infinite. For hypothesis testing, two configurations are considered: $G_1 = \{1, 2, 3, 4, 5\}$ and $G_2 = \{6, 7, 8, 9, 10\}$.

In our numerical studies, similar to many existing works, such as Neykov et al. (2016), Fan et al. (2017) and Sun et al. (2020), the tuning parameter λ_n in (7) is selected by the 10-fold cross validation. For each $j = 1, \ldots, d$, we select the tuning parameter ω_j using 5-fold cross validation. Moreover, the robustification parameter τ is tuned by the criterion that 80% of the predicted errors are in $[-\tau, \tau]$. For a set $A \subset \{2, \ldots, d+1\}$, the average of the empirical coverage probabilities (ACP) of the 95% confidence intervals over the set A is defined as

$$\mathrm{ACP}(A) = \sum_{j \in A} \mathrm{CP}_j / |A|,$$

where CP_j is the empirical coverage probability of the 95% confidence interval for θ_j^*. The average length (AL) of the 95% confidence intervals over the set A can be defined analogously. The results presented below are based on 100 replications with sample sizes $n = 200$ and 300 and dimensionality $d = 400$ and 800. We implement the proposed method with the `hqreg` package in R (https://cran.r-project.org/web/packages/hqreg/index.html), in which a semismooth Newton coordinate descent algorithm (SNCD) is employed to reduce the computational cost per iteration from $O(nd^2)$ to $O(nd)$ compared with the semismooth Newton algorithm (SNA) (Yi and Huang, 2017).

For comparison, we also consider the methods by Zhang and Zhang (2014), Fan et al. (2017) and Sun et al. (2020). Let $S_\infty = \{j | \theta_j^* \neq 0\}$, $S_\infty^c = \{j | \theta_j^* = 0\}$, $S_\infty^1 = S_\infty \setminus \{1\}$, and $S_\infty^{1c} = S_\infty^c \setminus \{1\}$. Table 1 reports the ACP and AL over S_∞^1, S_∞^{1c}, and $\lfloor d \rfloor$ in cases (A) and (B), where $\lfloor d \rfloor = \{2, \ldots, d+1\}$. Tables 2 and 3 present the averaged ℓ_1 and ℓ_2 distances between $\widehat{\beta}_\tau$ and β^*, the empirical probability of selecting the correct model (CM) over replications in the model selection step, and the empirical sizes and powers of the test statistic $\|\sqrt{n}\widehat{\Omega}^{-1/2}\widehat{\beta}_{\tau,G}\|_2^2$ at the significance level of $\tilde{\xi} = 0.05$. The results for cases (C)-(E) are given in Tables 4–6, respectively.

It can be seen from Tables 1–6 that the empirical sizes of our proposed method are close to the nominal level 0.05 and the ACP is reasonable across various settings. The empirical probability of selecting the correct model is close to 1 for most of the cases in Tables 2, 3, 5 and 6, indicating that the true sparse model can be recovered in the model selection step. In addition, Tables 2, 3, 5 and 6 also show that the statistical test based on the proposed statistic $\|\sqrt{n}\widehat{\Omega}^{-1/2}\widehat{\beta}_{\tau,G}\|_2^2$ is powerful. The overall performance of the proposed method gets better when the sample size increases from 200 to 300. For cases (A) and (B), the proposed method is comparable to that of Zhang and Zhang (2014). For cases (C)-(E), especially the heavy-tailed case (E), our method outperforms the method by Zhang and Zhang (2014) in terms of AL and the averaged ℓ_1 and ℓ_2 distance, indicating that our method is more robust than that of Zhang and Zhang (2014). Moreover, we observe that the shrinkage methods by Fan et al. (2017) and Sun et al. (2020) outperform the proposed method in terms of ℓ_1 and ℓ_2 distance. We believe this is due to the fact that post-selection inference methods, such as that of Zhang and Zhang (2014) and van de Geer et al. (2014), focus on post-selection inference, but not the shrinkage or sparsity recovery of the regression coefficients. We also display the histograms of $\sqrt{n}(\widehat{\beta}_{\tau,j} - \beta_j^*)/\widehat{\sigma}_{\tau,j}$, $j = 1, 3, 6, 8$, in Fig. 1, in which $n = 300$

Table 2
The averaged ℓ_1 and ℓ_2 distances between $\widehat{\beta}_\tau$ and β^*, the empirical sizes and powers of $||\sqrt{n}\widehat{\Omega}^{-1/2}\widehat{\beta}_{\tau,G}||_2^2$ at the significant level 0.05, and the empirical probability of selecting the correct model (CM) in the model selection step.

β^* in Setting (i)	Method	Distance		CM	Testing	
		ℓ_1	ℓ_2		Size	Power
$n = 200, d = 400, N(0, 1)$	Proposed	28.289	3.168	1.000	0.040	1.000
	Z. & Z.	32.809	4.244	1.000	–	–
	FLW17	0.962	0.113	1.000	–	–
	SZF19	0.729	0.143	1.000	–	–
$n = 200, d = 400, t(3)$	Proposed	34.651	5.120	1.000	0.040	0.990
	Z. & Z.	41.119	7.660	1.000	–	–
	FLW17	1.182	0.288	1.000	–	–
	SZF19	1.179	0.288	1.000	–	–
$n = 200, d = 800, N(0, 1)$	Proposed	53.402	5.640	1.000	0.080	1.000
	Z. & Z.	63.715	8.013	1.000	–	–
	FLW17	1.231	0.132	1.000	–	–
	SZF19	0.852	0.127	1.000	–	–
$n = 200, d = 800, t(3)$	Proposed	62.685	7.843	1.000	0.030	1.000
	Z. & Z.	78.629	16.218	1.000	–	–
	FLW17	0.973	0.179	1.000	–	–
	SZF19	1.112	0.170	1.000	–	–

Notes: Z. & Z. stands for the method by Zhang and Zhang (2014); FLW17 represents the method by Fan et al. (2017); SZF19 stands for the method by Sun et al. (2020); "-" means not applicable.

Table 3
The averaged ℓ_1 and ℓ_2 distances between $\widehat{\beta}_\tau$ and β^*, the empirical sizes and powers of $||\sqrt{n}\widehat{\Omega}^{-1/2}\widehat{\beta}_{\tau,G}||_2^2$ at the significant level 0.05, and the empirical probability of selecting the correct model (CM) in the model selection step.

β^* in Setting (i)	Method	Distance		CM	Testing	
		ℓ_1	ℓ_2		Size	Power
$n = 300, d = 400, N(0, 1)$	Proposed	24.875	2.431	1.000	0.050	1.000
	Z. & Z.	30.138	5.077	1.000	–	–
	FLW17	0.603	0.070	1.000	–	–
	SZF19	0.843	0.191	1.000	–	–
$n = 300, d = 400, t(3)$	Proposed	29.667	3.523	1.000	0.080	1.000
	Z. & Z.	35.640	5.678	1.000	–	–
	FLW17	0.726	0.141	1.000	–	–
	SZF19	0.924	0.235	1.000	–	–
$n = 300, d = 800, N(0, 1)$	Proposed	46.536	4.264	1.000	0.070	1.000
	Z. & Z.	53.711	5.662	1.000	–	–
	FLW17	0.729	0.080	1.000	–	–
	SZF19	0.837	0.194	1.000	–	–
$n = 300, d = 800, t(3)$	Proposed	55.231	6.082	1.000	0.030	1.000
	Z. & Z.	62.783	7.912	1.000	–	–
	FLW17	1.148	0.119	1.000	–	–
	SZF19	0.785	0.170	1.000	–	–

Notes: Z. & Z. stands for the method by Zhang and Zhang (2014); FLW17 represents the method by Fan et al. (2017); SZF19 stands for the method by Sun et al. (2020); "-" means not applicable.

and $d = 800$. For each j and different error distributions, one can see that the distribution of $\sqrt{n}(\widehat{\beta}_{\tau,j} - \beta_j^*)/\widehat{\sigma}_{\tau,j}$ is similar to the standard normal distribution, which confirms the theory. Similar conclusions can be drawn under other settings.

4.2. Simulated data: weak signals

In the second part, we conduct simulations to check the performance when the signals of the parameters are weak. The first 10 components of β^* are set to be 0.15 and other components of β^* are 0 (setting (ii)). For hypothesis testing, we consider two configurations: $G_3 = \{1, 2, \ldots, 10\}$ and $G_4 = \{11, 12, \ldots, 20\}$. We consider two error distributions: the standard normal distribution, shorted as $N(0, 1)$; Student's t distribution with degrees of freedom 3, shorted as $t(3)$. Other setups are the same as in Section 4.1. The simulation results are presented in Tables 7–9 and Fig. 2. It can be seen that the proposed method performs reasonably well when the signals are relatively weak. And it is more robust than that of Zhang and Zhang (2014). Note that all methods cannot identify the correct model in the model selection step when the

Table 4
The average of the coverage probabilities (ACP) of the 95% confidence intervals and the average length (AL) of the 95% confidence intervals over S_∞^1, S_∞^{1c}, and $\lfloor d \rfloor$, respectively.

β^* in Setting (i)	Method	ACP			AL		
		S_∞^1	S_∞^{1c}	$\lfloor d \rfloor$	S_∞^1	S_∞^{1c}	$\lfloor d \rfloor$
$n = 200, d = 400, \chi^2(8) - 8$	Proposed	0.910	0.975	0.974	1.349	1.516	1.514
	Z. & Z.	0.902	0.952	0.951	1.572	1.615	1.614
$n = 300, d = 400, \chi^2(8) - 8$	Proposed	0.914	0.974	0.974	1.189	1.366	1.363
	Z. & Z.	0.910	0.943	0.942	1.330	1.388	1.388
$n = 200, d = 800, \chi^2(8) - 8$	Proposed	0.880	0.977	0.977	1.388	1.536	1.535
	Z. & Z.	0.880	0.951	0.951	1.544	1.578	1.578
$n = 300, d = 800, \chi^2(8) - 8$	Proposed	0.902	0.976	0.976	1.169	1.309	1.308
	Z. & Z.	0.904	0.952	0.952	1.297	1.339	1.339
$n = 200, d = 400, SN(0, 0, 1)$	Proposed	0.896	0.970	0.969	0.277	0.311	0.311
	Z. & Z.	0.939	0.950	0.950	0.488	0.501	0.501
$n = 300, d = 400, SN(0, 0, 1)$	Proposed	0.858	0.970	0.968	0.279	0.313	0.312
	Z. & Z.	0.913	0.952	0.952	0.369	0.379	0.379
$n = 200, d = 800, SN(0, 0, 1)$	Proposed	0.892	0.968	0.968	0.292	0.318	0.318
	Z. & Z.	0.920	0.953	0.953	0.350	0.358	0.358
$n = 300, d = 800, SN(0, 0, 1)$	Proposed	0.852	0.961	0.961	0.220	0.247	0.247
	Z. & Z.	0.906	0.953	0.953	0.268	0.277	0.277
$n = 200, d = 400, P(1, 2)$	Proposed	0.894	0.962	0.961	0.304	0.341	0.341
	Z. & Z.	0.931	0.952	0.952	0.972	1.000	1.000
$n = 300, d = 400, P(1, 2)$	Proposed	0.892	0.962	0.961	0.279	0.317	0.317
	Z. & Z.	0.924	0.952	0.951	1.350	1.408	1.407
$n = 200, d = 800, P(1, 2)$	Proposed	0.880	0.969	0.969	0.352	0.381	0.381
	Z. & Z.	0.904	0.952	0.952	0.784	0.803	0.803
$n = 300, d = 800, P(1, 2)$	Proposed	0.890	0.962	0.962	0.250	0.279	0.279
	Z. & Z.	0.926	0.950	0.950	0.731	0.757	0.757

Notes: Z. & Z. stands for method by Zhang and Zhang (2014).

signals are weak, but post-selection methods are still able to carry out valid statistical inference. Similar to Section 4.1, the methods by Fan et al. (2017) and Sun et al. (2020) give smaller ℓ_1 and ℓ_2 distance than our method.

4.3. Simulated data: heteroscedastic errors

In the third part, we generate data from a heteroscedastic linear model (case F), where the error follows a hybrid distribution, i.e., half of the errors follow $N(0, 1)$ and another half follow $t(3)$. Since a sample-splitting technique is employed, we set the sample size $n = 400$ and 600. Other setups are the same as in Section 4.1. The simulation results are summarized in Table 10 and Fig. 3, from which one can see that the proposed method performs well in the presence of heteroscedasticity. Similar conclusions to those in Section 4.1 can be drawn in the comparison with the methods by Zhang and Zhang (2014), Fan et al. (2017) and Sun et al. (2020).

5. Application

We apply our method to analyze a genomic dataset concerning the riboflavin (vitamin B_2) production rate. This dataset has been analyzed in van de Geer et al. (2014), Javanmard and Montanari (2014) and Janková and van de Geer (2016). A total of 71 samples of genetically engineered mutants of bacillus subtilis are included in the analysis. The response variable is the logarithm of the riboflavin production rate. In addition, 4088 covariates which measure the logarithm of the expression level of 4088 genes are treated as predictors. The selection methods of the tuning parameters τ, ω_j, and λ_n are the same as those in the simulation studies. Our goal is to select predictors that are associated with the riboflavin production rate.

Following Janková and van de Geer (2016), we first conduct variable screening to reduce the dimensionality to a moderate scale. We choose the top 300 covariates deemed most relevant for subsequent linear modeling. Table 11 presents the estimates, and the 95% confidence intervals for the coefficients of the significant covariates. With the proposed method, it can be seen from Table 11 that two genes, XLYA_at and YCKE_at, are selected and both are positively correlated with the riboflavin production rate. In addition, the p-value of $\|\sqrt{n}\widehat{\Omega}^{-1/2}\widehat{\beta}_{\tau,G}\|_2^2$ for testing $H_{0,G}$ with $G = \{\text{XLYA\_at, YCKE\_at}\}$ is 3.559×10^{-6}, which also suggests the significance of these two covariates. For comparison, the method by Zhang and Zhang (2014) identifies 7 genes including YXLD_at, YXLE_at, YEEI_at, YCGO_at, YSFE_at, YCKE_at, and HISI_at with non-zero coefficients.

Table 5
The averaged ℓ_1 and ℓ_2 distances between $\widehat{\beta}_\tau$ and β^*, the empirical sizes and powers of $||\sqrt{n}\widehat{\Omega}^{-1/2}\widehat{\beta}_{\tau,G}||_2^2$ at the significant level 0.05, and the empirical probability of selecting the correct model (CM) in the model selection step.

β^* in Setting (i)	Method	Distance		CM	Testing	
		ℓ_1	ℓ_2		Size	Power
$n = 200, d = 400, \chi^2(8) - 8$	Proposed	108.566	46.539	1.000	0.090	1.000
	Z. & Z.	130.598	67.327	1.000	–	–
	FLW17	6.661	2.287	1.000	–	–
	SZF19	5.228	1.978	1.000	–	–
$n = 200, d = 800, \chi^2(8) - 8$	Proposed	211.536	88.276	1.000	0.030	1.000
	Z. & Z.	256.712	129.922	1.000	–	–
	FLW17	3.316	2.983	1.000	–	–
	SZF19	3.319	2.990	1.000	–	–
$n = 200, d = 400, SN(0, 0, 1)$	Proposed	22.697	2.046	1.000	0.110	1.000
	Z. & Z.	27.231	2.924	1.000	–	–
	FLW17	0.940	0.111	1.000	–	–
	SZF19	0.717	0.143	1.000	–	–
$n = 200, d = 800, SN(0, 0, 1)$	Proposed	46.316	4.232	1.000	0.100	1.000
	Z. & Z.	56.357	9.509	0.990	–	–
	FLW17	0.749	0.144	1.000	–	–
	SZF19	0.797	0.124	1.000	–	–
$n = 200, d = 400, P(1, 2)$	Proposed	26.199	2.720	1.000	0.060	1.000
	Z. & Z.	79.392	37.029	0.970	–	–
	FLW17	4.722	2.156	1.000	–	–
	SZF19	4.088	1.968	1.000	–	–
$n = 200, d = 800, P(1, 2)$	Proposed	55.655	6.143	1.000	0.070	1.000
	Z. & Z.	130.047	40.530	1.000	–	–
	FLW17	4.566	1.532	1.000	–	–
	SZF19	5.406	1.689	1.000	–	–

Notes: Z. & Z. stands for the method by Zhang and Zhang (2014); FLW17 represents the method by Fan et al. (2017); SZF19 stands for the method by Sun et al. (2020); "-" means not applicable.

Table 6
The averaged ℓ_1 and ℓ_2 distances between $\widehat{\beta}_\tau$ and β^*, the empirical sizes and powers of $||\sqrt{n}\widehat{\Omega}^{-1/2}\widehat{\beta}_{\tau,G}||_2^2$ at the significant level 0.05, and the empirical probability of selecting the correct model (CM) in the model selection step.

β^* in Setting (i)	Method	Distance		CM	Testing	
		ℓ_1	ℓ_2		Size	Power
$n = 300, d = 400, \chi^2(8) - 8$	Proposed	98.105	38.003	1.000	0.050	1.000
	Z. & Z.	117.968	61.962	1.000	–	–
	FLW17	2.345	1.287	1.000	–	–
	SZF19	2.353	1.276	1.000	–	–
$n = 300, d = 800, \chi^2(8) - 8$	Proposed	184.822	67.213	1.000	0.070	1.000
	Z. & Z.	217.143	92.569	1.000	–	–
	FLW17	2.681	1.425	1.000	–	–
	SZF19	2.959	2.394	1.000	–	–
$n = 300, d = 400, SN(0, 0, 1)$	Proposed	22.977	2.089	1.000	0.050	1.000
	Z. & Z.	29.470	4.828	0.990	–	–
	FLW17	0.987	0.120	1.000	–	–
	SZF19	1.055	0.306	1.000	–	–
$n = 300, d = 800, SN(0, 0, 1)$	Proposed	38.065	2.855	1.000	0.060	1.000
	Z. & Z.	44.718	3.931	1.000	–	–
	FLW17	0.736	0.078	1.000	–	–
	SZF19	0.675	0.129	1.000	–	–
$n = 300, d = 400, P(1, 2)$	Proposed	28.761	3.281	1.000	0.010	1.000
	Z. & Z.	64.441	19.843	0.990	–	–
	FLW17	2.499	0.805	1.000	–	–
	SZF19	2.010	0.874	1.000	–	–
$n = 300, d = 800, P(1, 2)$	Proposed	42.876	3.627	1.000	0.050	1.000
	Z. & Z.	122.914	43.598	1.000	–	–
	FLW17	2.724	1.441	1.000	–	–
	SZF19	3.157	1.466	1.000	–	–

Notes: Z. & Z. stands for the method by Zhang and Zhang (2014); FLW17 represents the method by Fan et al. (2017); SZF19 stands for the method by Sun et al. (2020); "-" means not applicable.

Fig. 1. Histograms of $\sqrt{n}(\widehat{\beta}_{\tau,j} - \beta_j^*)/\widehat{\sigma}_{\tau,j}$, $j = 1, 3, 6, 8$ under setting (i) with $n = 300, d = 800$. Different rows correspond to different error distributions in cases (A)-(E). The red curve is the density function of the standard normal distribution. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Table 7
The average of the coverage probabilities (ACP) of the 95% confidence intervals and the average length (AL) of the 95% confidence intervals over S_∞^1, S_∞^{1c}, and $\lfloor d \rfloor$, respectively.

β^* in Setting (ii)	Method	ACP			AL		
		S_∞^1	S_∞^{1c}	$\lfloor d \rfloor$	S_∞^1	S_∞^{1c}	$\lfloor d \rfloor$
$n = 200, d = 400, N(0, 1)$	Proposed	0.928	0.975	0.975	0.350	0.392	0.392
	Z. & Z.	0.935	0.952	0.952	0.402	0.412	0.412
$n = 300, d = 400, N(0, 1)$	Proposed	0.936	0.973	0.973	0.295	0.340	0.339
	Z. & Z.	0.954	0.956	0.956	0.339	0.353	0.353
$n = 200, d = 800, N(0, 1)$	Proposed	0.928	0.977	0.977	0.367	0.389	0.389
	Z. & Z.	0.943	0.954	0.954	0.391	0.400	0.400
$n = 300, d = 800, N(0, 1)$	Proposed	0.928	0.978	0.978	0.300	0.335	0.335
	Z. & Z.	0.920	0.953	0.953	0.329	0.339	0.339
$n = 200, d = 400, t(3)$	Proposed	0.938	0.974	0.974	0.377	0.423	0.422
	Z. & Z.	0.941	0.952	0.952	0.490	0.505	0.504
$n = 300, d = 400, t(3)$	Proposed	0.912	0.971	0.970	0.307	0.352	0.351
	Z. & Z.	0.950	0.951	0.951	0.405	0.422	0.421
$n = 200, d = 800, t(3)$	Proposed	0.902	0.975	0.974	0.365	0.383	0.383
	Z. & Z.	0.930	0.954	0.953	0.482	0.493	0.493
$n = 300, d = 800, t(3)$	Proposed	0.938	0.975	0.975	0.302	0.339	0.339
	Z. & Z.	0.940	0.953	0.953	0.387	0.400	0.400

Notes: Z. & Z. stands for the method by Zhang and Zhang (2014).

Table 8
The averaged ℓ_1 and ℓ_2 distances between $\widehat{\beta}_\tau$ and β^*, the empirical sizes and powers of $||\sqrt{n}\widehat{\Omega}^{-1/2}\widehat{\beta}_{\tau,G}||_2^2$ at the significant level 0.05, and the empirical probability of selecting the correct model (CM) in the model selection step.

β^* in Setting (ii)	Method	Distance		CM	Testing	
		ℓ_1	ℓ_2		Size	Power
$n = 200, d = 400, N(0, 1)$	Proposed	27.638	3.038	0.080	0.090	1.000
	Z. & Z.	33.236	4.369	0.000	–	–
	FLW17	1.342	0.133	0.040	–	–
	SZF19	1.695	0.151	0.060	–	–
$n = 200, d = 400, t(3)$	Proposed	29.796	3.526	0.030	0.060	1.000
	Z. & Z.	40.814	6.812	0.000	–	–
	FLW17	1.255	0.155	0.000	–	–
	SZF19	1.346	0.159	0.000	–	–
$n = 200, d = 800, N(0, 1)$	Proposed	53.722	5.700	0.030	0.070	1.000
	Z. & Z.	64.052	8.078	0.000	–	–
	FLW17	1.062	0.130	0.000	–	–
	SZF19	1.252	0.127	0.030	–	–
$n = 200, d = 800, t(3)$	Proposed	53.643	5.720	0.060	0.050	1.000
	Z. & Z.	79.432	13.774	0.000	–	–
	FLW17	1.626	0.219	0.000	–	–
	SZF19	1.444	0.202	0.000	–	–

Notes: Z. & Z. stands for the method by Zhang and Zhang (2014); FLW17 represents the method by Fan et al. (2017); SZF19 stands for the method by Sun et al. (2020); "-" means not applicable.

6. Concluding remarks

In this article, we study one-step post-selection inference with the Huber loss for a high-dimensional linear model with an intercept. When the errors are identically distributed, our proposed method allows for asymmetric and heavy-tailed error distributions. As suggested by an anonymous reviewer, we further extend our method to accommodate heteroscedasticity when the error distribution is assumed symmetric. We develop the asymptotic properties of the proposed estimators. Statistical tests are studied for low-dimensional components of the slope parameter vector. The simulation results show that the proposed method works well for various practical situations. An application to a genomic dataset on riboflavin (vitamin B2) production rate is provided to illustrate our method.

The main theorems we established requires a sub-Gaussian assumption on the covariate. It appears to be nontrivial to generalize our main results to heavy-tailed predictors in high dimensions. Some additional conditions on the moment of the covariates may be needed. In addition, to pursue a more robust method in both covariate and the error, one may replace the Neyman least squares projection by a Huber loss. Nevertheless, theoretical justifications become more challenging. Moreover, as pointed by an anonymous reviewer, it would be interesting to consider a smoothed version of the Huber loss. Finally, in the numerical work, we select τ using the criterion that 80% of the predicted errors are in

Table 9
The averaged ℓ_1 and ℓ_2 distance between $\widehat{\beta}_\tau$ and β^*, the empirical sizes and powers of $||\sqrt{n}\widehat{\Omega}^{-1/2}\widehat{\beta}_{\tau,G}||_2^2$ at the significant level 0.05, and the empirical probability of selecting the correct model (CM) in the model selection step.

β^* in Setting (ii)	Method	Distance		CM	Testing	
		ℓ_1	ℓ_2		Size	Power
$n = 300, d = 400, N(0, 1)$	Proposed	24.457	2.363	0.250	0.060	1.000
	Z. & Z.	28.376	3.164	0.100	–	–
	FLW17	0.953	0.098	0.120	–	–
	SZF19	0.915	0.104	0.060	–	–
$n = 300, d = 400, t(3)$	Proposed	25.509	2.592	0.160	0.050	1.000
	Z. & Z.	33.898	4.628	0.020	–	–
	FLW17	1.333	0.129	0.070	–	–
	SZF19	2.212	0.184	0.110	–	–
$n = 300, d = 800, N(0, 1)$	Proposed	46.755	4.304	0.190	0.050	1.000
	Z. & Z.	54.744	5.883	0.050	–	–
	FLW17	1.063	0.101	0.080	–	–
	SZF19	1.028	0.101	0.080	–	–
$n = 300, d = 800, t(3)$	Proposed	48.330	4.595	0.190	0.040	1.000
	Z. & Z.	64.749	8.423	0.060	–	–
	FLW17	1.554	0.139	0.110	–	–
	SZF19	1.350	0.129	0.090	–	–

Notes: Z. & Z. stands for the method in Zhang and Zhang (2014); FLW17 represents the method in Fan et al. (2017); SZF19 stands for the method in Sun et al. (2020); "-" means not applicable.

Table 10
The average ℓ_1 and ℓ_2 distances between $\widehat{\beta}_\tau$ and β^*, the average of the coverage probabilities (ACP) of the 95% confidence intervals, the average length (AL) of the 95% confidence intervals over S_∞^1, S_∞^{1c}, and $\lfloor d \rfloor$, the averaged and the empirical sizes and powers of $||\sqrt{n}\widehat{\Omega}^{-1/2}\widehat{\beta}_{\tau,G}||_2^2$ at the significant level 0.05, and the empirical probability of selecting the correct model (CM) in the model selection step.

Setting (i) of β^*	Method	Distance		ACP			AL			CM	Testing	
		ℓ_1	ℓ_2	S_∞^1	S_∞^{1c}	$\lfloor d \rfloor$	S_∞^1	S_∞^{1c}	$\lfloor d \rfloor$		Size	Power
$n = 400, d = 400,$ NID	Proposed	24.107	2.308	0.890	0.977	0.976	0.309	0.347	0.347	1.000	0.070	1.000
	Z. & Z.	34.964	4.870	0.931	0.951	0.950	0.406	0.429	0.429	1.000	–	–
	FLW17	0.907	0.098	–	–	–	–	–	–	1.000	–	–
	SZF19	1.266	0.424	–	–	–	–	–	–	1.000	–	–
$n = 600, d = 400,$ NID	Proposed	21.263	1.789	0.896	0.973	0.972	0.258	0.296	0.296	1.000	0.030	1.000
	Z. & Z.	31.226	3.974	0.952	0.949	0.949	0.352	0.383	0.383	1.000	–	–
	FLW17	0.670	0.081	–	–	–	–	–	–	1.000	–	–
	SZF19	1.463	0.546	–	–	–	–	–	–	1.000	–	–
$n = 400, d = 800,$ NID	Proposed	44.715	3.948	0.860	0.979	0.978	0.321	0.328	0.328	1.000	0.040	1.000
	Z. & Z.	69.918	11.402	0.945	0.953	0.953	0.423	0.442	0.442	0.990	–	–
	FLW17	0.712	0.126	–	–	–	–	–	–	1.000	–	–
	SZF19	0.945	0.239	–	–	–	–	–	–	1.000	–	–
$n = 600, d = 800,$ NID	Proposed	40.652	3.247	0.870	0.978	0.978	0.260	0.292	0.292	1.000	0.020	1.000
	Z. & Z.	70.022	13.478	0.925	0.940	0.940	1.026	1.092	1.091	0.960	–	–
	FLW17	0.849	0.094	–	–	–	–	–	–	1.000	–	–
	SZF19	1.282	0.430	–	–	–	–	–	–	1.000	–	–

Notes: Z. & Z. stands for the method in Zhang and Zhang (2014); FLW17 represents the method in Fan et al. (2017); SZF19 stands for the method in Sun et al. (2020); NID means non identically-distributed errors in case (F); "-" means not applicable.

Table 11
Estimates (Est) and the 95% confidence intervals (CI) for the coefficients of those significant predictors.

Method	Gene	Est	CI
Proposed	XLYA_at	0.276	[0.088,0.464]
	YCKE_at	0.254	[0.014,0.494]
Z. and Z.	YXLD_at	−0.406	[−0.804,−0.009]
	YXLE_at	−0.510	[−0.928,−0.091]
	YEEI_at	0.994	[0.174,1.81]
	YCGO_at	−0.331	[−0.591,−0.071]
	YSFE_at	0.569	[0.005,1.133]
	YCKE_at	0.370	[0.021,0.719]
	HISI_at	0.670	[0.115,1.225]

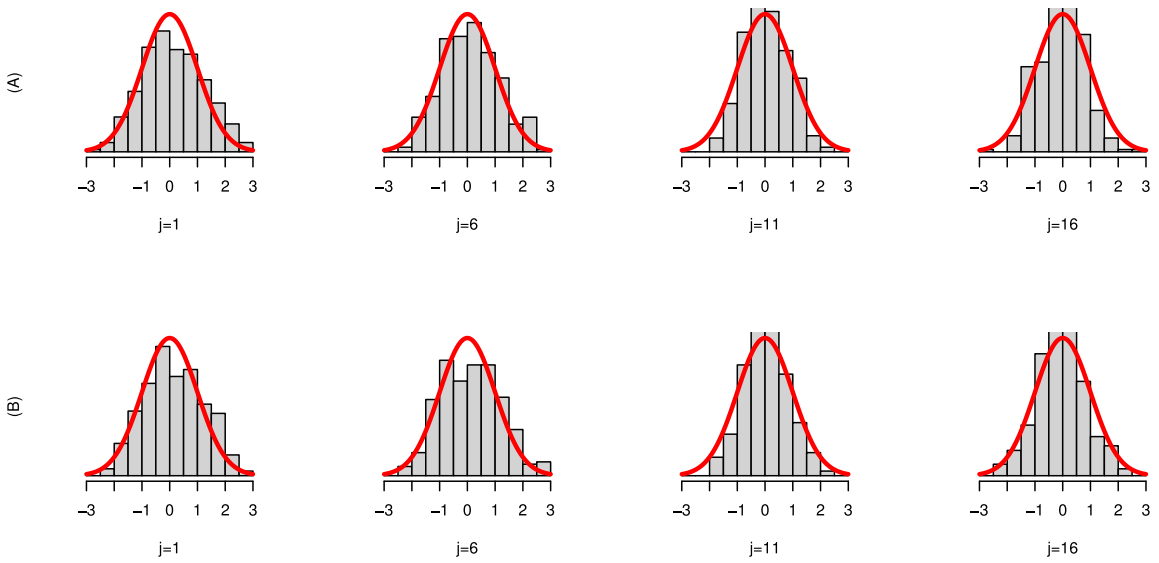

Fig. 2. Histograms of $\sqrt{n}(\widehat{\beta}_{\tau,j} - \beta_j^*)/\widehat{\sigma}_{\tau,j}$, $j = 1, 6, 11, 16$ under setting (ii) with $n = 300, d = 800$. Different rows correspond to different error distributions in cases (A)-(B). The red curve is the density function of the standard normal distribution. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

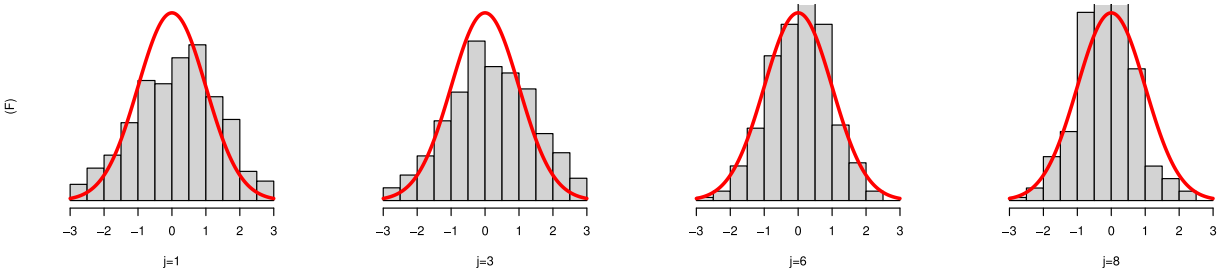

Fig. 3. Histograms of $\sqrt{n}(\widehat{\beta}_{\tau,j} - \beta_j^*)/\widehat{\sigma}_{\tau,j}$, $j = 1, 3, 6, 8$ under setting (i) with $n = 600, d = 800$ and non identically-distributed errors. The red curve is the density function of the standard normal distribution. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

$[-\tau, \tau]$, which leads to satisfactory performance in our simulation studies. However, a rigorous data-driven selector that works in our problem with theoretical guarantees has yet to be found. It would be interesting to consider this problem in the future.

Acknowledgments

The authors wish to thank the Editor, the Associate Editor and the anonymous reviewers for their professional review and constructive comments that lead to significant improvements in the paper. The work of Dongxiao Han is supported by the Fundamental Research Funds for the Central Universities, China, Nankai University, 9920200110. The work of Jian Huang is partially supported by the U.S. National Science Foundation grant DMS-1916199. The work of Yuanyuan Lin is partially supported by the Hong Kong Research Grants Council, HKSAR (Grant No. 14306219 and 14306620), the National Natural Science Foundation of China (Grant No. 11961028) and Direct Grants for Research, The Chinese University of Hong Kong, HKSAR.

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.jeconom.2021.05.006.

References

Belloni, A., Chernozhukov, V., 2011. ℓ_1-penalized quantile regression in high-dimensional sparse models. Ann. Statist. 39, 82–130.

Belloni, A., Chernozhukov, V., Kato, K., 2015. Uniform post-selection inference for least absolute deviation regression and other Z-estimation problems. Biometrika 102, 77–94.
Belloni, A., Chernozhukov, V., Kato, K., 2019. Valid post-selection inference in high-dimensional approximately sparse quantile regression models. J. Amer. Statist. Assoc. 114, 749–758.
Bickel, P.J., 1975. One-step Huber estimates in the linear model. J. Amer. Statist. Assoc. 70, 428–434.
Bradic, J., Fan, J., Wang, W., 2011. Penalized composite quasi-likelihood for ultrahigh dimensional variable selection. J. R. Stat. Soc. Ser. B Stat. Methodol. 73, 325–349.
Cai, T.T., Guo, Z., 2017. Confidence intervals for high-dimensional linear regression: Minimax rates and adaptivity. Ann. Statist. 45, 615–646.
Cheng, C., Feng, X., Huang, J., Liu, X., 2020. Regularized projection score estimation of treatment effects in high-dimensional quantile regression. Statist. Sinica http://dx.doi.org/10.5705/ss.202019.0247.
Cont, R., 2001. Empirical properties of asset returns: Stylized facts and statistical issues. Quant. Finance 1, 223–236.
Eklund, A., Nichols, T.E., Knutsson, H., 2016. Cluster failure: Why fMRI inferences for spatial extent have inflated false-positive rates. Proc. Natl. Acad. Sci. 113, 7900–7905.
Fan, J., Fan, Y., Barut, E., 2014. Adaptive robust variable selection. Ann. Statist. 42, 324–351.
Fan, J., Li, Q., Wang, Y., 2017. Estimation of high dimensional mean regression in the absence of symmetry and light tail assumptions. J. R. Stat. Soc. Ser. B Stat. Methodol. 79, 247–265.
Fan, J., Lv, J., 2008. Sure independence screening for ultrahigh dimensional feature space (with discussion). J. R. Stat. Soc. Ser. B Stat. Methodol. 70, 849–911.
Fan, J., Wang, W., Zhu, Z., 2016. A shrinkage principle for heavy-tailed data: High-dimensional robust low-rank matrix recovery. Available at arXiv:1603.08315.
He, X., Shao, Q.-M., 1996. A general Bahadur representation of M-estimators and its application to linear regression with nonstochastic designs. Ann. Statist. 24, 2608–2630.
He, X., Shao, Q.-M., 2000. On parameters of increasing dimensions. J. Multivariate Anal. 73, 120–135.
Huang, J., Horowitz, J.L., Ma, S., 2008. Asymptotic properties of bridge estimators in sparse high-dimensional regression models. Ann. Statist. 36, 587–613.
Huber, P.J., 1964. Robust estimation of a location parameter. Ann. Math. Stat. 35, 73–101.
Huber, P.J., 1973. Robust regression: asymptotics, conjectures and Monte Carlo. Ann. Statist. 1, 799–821.
Ishwaran, H., Kogalur, U.B., Gorodeski, E.Z., Minn, A.J., Lauer, M.S., 2010. High-dimensional variable selection for survival data. J. Amer. Statist. Assoc. 105, 205–217.
Janková, J., 2017. Asymptotic Inference in Sparse High-Dimensional Models (Doctoral dissertation). ETH Zurich.
Janková, J., van de Geer, S., 2015. Confidence intervals for high-dimensional inverse covariance estimation. Electron. J. Stat. 9, 1205–1229.
Janková, J., van de Geer, S., 2016. Confidence regions for high-dimensional generalized linear models under sparsity. Available at arXiv:1610.01353.
Javanmard, A., Montanari, A., 2014. Confidence intervals and hypothesis testing for high-dimensional regression. J. Mach. Learn. Res. 15, 2869–2909.
Li, Y., Zhu, J., 2008. ℓ_1-norm quantile regression. J. Comput. Graph. Statist. 17, 163–185.
Liu, J., Li, R., Wu, R., 2014. Feature selection for varying coefficient models with ultrahigh-dimensional covariates. J. Amer. Statist. Assoc. 109, 266–274.
Loh, P.L., 2018. Scale calibration for high-dimensional robust regression. Available at arXiv:1811.02096.
Meinshausen, N., Bühlmann, P., 2006. High-dimensional graphs and variable selection with the lasso. Ann. Statist. 34, 1436–1462.
Neykov, M., Liu, J.S., Cai, T., 2016. ℓ_1-Regularized least squares for support recovery of high dimensional single index models with Gaussian designs. J. Mach. Learn. Res. 17, 1–37.
Neyman, J., 1959. Optimal asymptotic tests of composite hypotheses. In: Grenander, U. (Ed.), Probability and Statistics., The Harold Cramer Volume. John Wiley and Sons, New York.
Ning, Y., Liu, H., 2017. A general theory of hypothesis tests and confidence regions for sparse high dimensional models. Ann. Statist. 45, 158–195.
Portnoy, S., 1985. Asymptotic behavior of M estimators of p regression parameters when p^2/n is large; II. Normal approximation. Ann. Statist. 13, 1403–1417.
Sun, Q., Zhou, W.X., Fan, J., 2020. Adaptive huber regression. J. Amer. Statist. Assoc. 115, 254–265.
Tibshirani, R., 1996. Regression shrinkage and selection via the lasso. J. R. Stat. Soc. Ser. B Stat. Methodol. 58, 267–288.
van de Geer, S., Bühlmann, P., Ritov, Y.A., Dezeure, R., 2014. On asymptotically optimal confidence regions and tests for high-dimensional models. Ann. Statist. 42, 1166–1202.
Vershynin, R., 2018. High-Dimensional Probability: An Introduction with Applications in Data Science, Vol. 47. Cambridge University Press, Cambridge.
Wainwright, M.J., 2009. Sharp thresholds for high-dimensional and noisy sparsity recovery using ℓ_1-constrained quadratic programming (Lasso). IEEE Trans. Inform. Theory 55, 2183–2202.
Wang, L., 2013. The ℓ_1 penalized LAD estimator for high dimensional linear regression. J. Multivariate Anal. 120, 135–151.
Wang, L., Peng, B., Li, R., 2015. A high-dimensional nonparametric multivariate test for mean vector. J. Amer. Statist. Assoc. 110, 1658–1669.
Wang, L., Zheng, C., Zhou, W., Zhou, W.X., 2020. A new principle for tuning-free Huber regression. Statist. Sinica http://dx.doi.org/10.5705/ss.202019.0045.
Wu, Y., Liu, Y., 2009. Variable selection in quantile regression. Statist. Sinica 19, 801–817.
Yi, C., Huang, J., 2017. Semismooth Newton coordinate descent algorithm for elastic-net penalized Huber loss regression and quantile regression. J. Comput. Graph. Statist. 26, 547–557.
Yohai, V.J., Maronna, R.A., 1979. Asymptotic behavior of M-estimators for the linear model. Ann. Statist. 7, 258–268.
Zhang, C.H., 2010. Nearly unbiased variable selection under minimax concave penalty. Ann. Statist. 38, 894–942.
Zhang, C.H., Huang, J., 2008. The sparsity and bias of the lasso selection in high-dimensional linear regression. Ann. Statist. 36, 1567–1594.
Zhang, C.H., Zhang, S.S., 2014. Confidence intervals for low dimensional parameters in high dimensional linear models. J. R. Stat. Soc. Ser. B Stat. Methodol. 76, 217–242.
Zhao, P., Yu, B., 2006. On model selection consistency of lasso. J. Mach. Learn. Res. 7, 2541–2563.
Zhou, W.X., Bose, K., Fan, J., Liu, H., 2018. A new perspective on robust M-estimation: Finite sample theory and applications to dependence-adjusted multiple testing. Ann. Statist. 46, 1904–1931.
Zou, H., Yuan, M., 2008. Composite quantile regression and the oracle model selection theory. Ann. Statist. 36, 1108–1126.