Journal of Econometrics 239 (2024) 105459

Contents lists available at ScienceDirect

Journal of Econometrics

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

Retire: Robust expectile regression in high dimensions

Rebeka Man [a], Kean Ming Tan [a,*], Zian Wang [b], Wen-Xin Zhou [c]

[a] Department of Statistics, University of Michigan, Ann Arbor, MI, 48109, USA
[b] Department of Mathematics, University of California, San Diego, La Jolla, CA, 92093, USA
[c] Department of Information and Decision Sciences, University of Illinois at Chicago, Chicago, IL, 60607, USA

ARTICLE INFO

Article history:
Received 16 August 2021
Received in revised form 21 March 2023
Accepted 21 April 2023
Available online 17 June 2023

Keywords:
Expectile regression
Heavy-tailed error
Quantile regression
Robustness
Concave regularization

ABSTRACT

High-dimensional data can often display heterogeneity due to heteroscedastic variance or inhomogeneous covariate effects. Penalized quantile and expectile regression methods offer useful tools to detect heteroscedasticity in high-dimensional data. The former is computationally challenging due to the non-smooth nature of the check loss, and the latter is sensitive to heavy-tailed error distributions. In this paper, we propose and study (penalized) robust expectile regression (`retire`), with a focus on iteratively reweighted ℓ_1-penalization which reduces the estimation bias from ℓ_1-penalization and leads to oracle properties. Theoretically, we establish the statistical properties of the retire estimator under two regimes: (i) low-dimensional regime in which $d \ll n$; (ii) high-dimensional regime in which $s \ll n \ll d$ with s denoting the number of significant predictors. In the high-dimensional setting, we thoroughly analyze the statistical properties of the solution path of iteratively reweighted ℓ_1-penalized `retire` estimation, adapted from the local linear approximation algorithm for folded-concave regularization. Under a mild minimum signal strength condition, we demonstrate that with as few as $\log(\log d)$ iterations, the final iterate of our proposed approach achieves the oracle convergence rate. At each iteration, we solve the weighted ℓ_1-penalized convex program using a semismooth Newton coordinate descent algorithm. Numerical studies demonstrate the promising performance of the proposed procedure in comparison to both non-robust and quantile regression based alternatives.

1. Introduction

Penalized least squares has become a baseline approach for fitting sparse linear models in high dimensions. Its focus is primarily on inferring the conditional mean of the response given a large number of predictors/covariates. In many economic applications, however, more aspects than the mean of the conditional distribution of the response given the covariates are of interest, and that the covariate effects may be inhomogeneous and/or the noise variables exhibit heavy-tailed and asymmetric tails. For instance, in the Job Training Partners Act studied in Abadie et al. (2002), one is more interested in the lower tail than the mean of the conditional distribution of income given predictors such as enrollment in a subsidized training program and demographic variables. To capture heterogeneity in the set of covariates at different locations of the response distribution, methods such as quantile regression (Koenker and Bassett, 1978) and asymmetric least squares regression (expectile regression) (Newey and Powell, 1987) have been widely used. We refer the reader

* Corresponding author.
E-mail addresses: mrebeka@umich.edu (R. Man), keanming@umich.edu (K.M. Tan), ziw105@ucsd.edu (Z. Wang), wez243@ucsd.edu (W.-X. Zhou).

https://doi.org/10.1016/j.jeconom.2023.04.004
0304-4076/

to Koenker and Bassett (1978), Koenker (2005), and Koenker et al. (2017) for a comprehensive overview of quantile regression, and Newey and Powell (1987) and Gu and Zou (2016) for conventional and penalized expectile regressions.

Both quantiles and expectiles are useful descriptors of the tail behavior of a distribution in the same way as the median and mean are related to its central behavior. As shown by Jone (1994), expectiles are exactly quantiles of a transformed version of the original distribution. In fact, the expectile regression can be interpreted as a least squares analogue of regression quantile estimation (Newey and Powell, 1987). Quantile regression is naturally more dominant in the literature due to the fact that expectiles lack an intuitive interpretation while quantiles are the inverse of the distribution function and directly indicate relative frequency. The key advantage of expectile regression is its computational expediency and the asymptotic covariance matrix can be estimated without the need of estimating the conditional density function (nonparametrically). Therefore, it offers a convenient and relatively efficient method of summarizing the conditional response distribution.

Expectile regression has found applications in various fields, including risk analysis (Taylor, 2008; Kuan et al., 2009; Xie et al., 2014; Bellini and Bernardino, 2017; Daouia et al., 2018), as well as the study of determinants of inflation (Busetti et al., 2021) and life expectancy and economic production (Schnabel and Eilers, 2009). In finance applications, the expectile, also known as the expectile-Value at Risk, represents the minimum amount of capital needed to add to a position in order to achieve a specified high gain–loss ratio (Bellini and Bernardino, 2017; Gu and Zou, 2019). In contrast to quantile, the expectile is a coherent measure of risk that is desirable in finance applications (Kuan et al., 2009). While the expected shortfall is another popular coherent risk measure in finance (Acerbie and Tasche, 2002), it is not elicitable (Gneiting, 2011), which poses challenges for its estimation. In fact, the expectile is the only risk measure that is both coherent and elicitable, making it a valuable tool for financial risk management and decision-making (Bellini and Bignozzi, 2015; Ziegel, 2016; Bellini and Bernardino, 2017).

In contrast to quantile regression, expectile regression minimizes a quadratic-type loss and is therefore sensitive to heavy-tailed response distributions. To address this limitation, we propose a robust expectile regression approach that is designed to effectively handle heavy-tailed response distributions in both low- and high-dimensional models. In the latter setting, the number of covariates/regressors, d, is substantially greater than the number of observations, denoted by n. High-dimensional data analysis greatly benefits from the sparsity assumption—only a small number of significant predictors are associated with the response. This motivates the use of various convex and non-convex penalty functions so as to achieve a desirable trade-off between model complexity and statistical accuracy (Bühlmann and van de Geer, 2011; Wainwright, 2019; Fan et al., 2020). The most widely used penalty functions include the ℓ_1/Lasso penalty (Tibshirani, 1996), the smoothly clipped absolute deviation (SCAD) penalty (Fan and Li, 2001), and the minimax concave penalty (MCP) (Zhang, 2010a). Albeit being computationally efficient and statistically (near-)optimal under ℓ_2-errors, the ℓ_1 penalty induces non-negligible estimation bias, which may prevent consistent variable selection. The selected model with a relatively small prediction error tends to include many false positives, unless stringent assumptions are imposed on the design matrix (Zou and Li, 2008; Zhang and Zhang, 2012; Su et al., 2017; Lahiri, 2021).

Folded-concave (non-convex) penalty functions, on the other hand, have been designed to reduce the bias induced by the ℓ_1 penalty. With either the ℓ_2 or a robust loss function, the resulting folded-concave penalized estimator is proven to achieve the oracle property provided the signals are sufficiently strong, i.e., the estimator has the same rate of convergence as that of the oracle estimator obtained by fitting the regression model with true active predictors that are unknown in practice (Fan and Li, 2001; Zou and Li, 2008; Zhang and Zhang, 2012; Loh and Wainwright, 2015; Loh, 2017). Due to non-convexity, directly minimizing the concave penalized loss raises numerical instabilities. Standard gradient-based algorithms are often guaranteed to find a stationary point, while oracle results are primarily derived for the hypothetical global minimum. Zou and Li (2008) proposed a unified algorithm for folded-concave penalized estimation based on local linear approximation (LLA). It relaxes the non-convex optimization problem into a sequence of iteratively reweighted ℓ_1-penalized subproblems. The statistical properties of the final iterate have been studied by Zhang (2010b), Fan et al. (2014), (Fan et al., 2018) and Pan et al. (2021) under different regression models. We refer to Wainwright (2019) and Fan et al. (2020), and the references therein, for a comprehensive introduction of penalized M-estimation based on various convex and folded-concave (non-convex) penalty functions.

For sparse quantile regression (QR) in high dimensions, Belloni and Chernozhukov (2011) studied ℓ_1-penalized quantile regression process, and established the uniform (over a range of quantile levels) rate of convergence. To alleviate the bias induced by the ℓ_1 penalty, Wang et al. (2012) proposed concave penalized quantile regression, and showed that the oracle estimator is a local solution to the resulting optimization problem. Via the one-step LLA algorithm, Fan et al. (2014) proved that the oracle estimator can be obtained (with high probability) as long as the magnitude of true nonzero regression coefficients is at least of order $\sqrt{s\log(d)/n}$. We refer to Wang and He (2022) for a unified analysis of global and local optima of penalized quantile regressions. While quantile regression offers the flexibility to model the conditional response distribution and is robust to outliers, together, the non-differentiability of the check function and the non-convexity of the penalty pose substantial technical and computational challenges. To our knowledge, the theoretical guarantee of the convergence of a computationally efficient algorithm to the oracle QR estimator under the weak minimum signal strength condition, i.e., the true nonzero regression coefficients is at least of order $\sqrt{\log(d)/n}$, remains unclear.

In high-dimensional sparse models, Gu and Zou (2016) considered the penalized expectile regression using both convex and concave penalty functions. Since the expectile loss is convex and twice-differentiable, scalable algorithms, such as the cyclic coordinate decent and proximal gradient descent, can be employed to solve the resulting optimization problem.

Theoretically, the consistency of penalized expectile regression in the high-dimensional regime "$\log(d) \ll n \ll d$" requires *sub-Gaussian* error distributions (Gu and Zou, 2016). This is in strong contrast to penalized QR, the consistency of which requires no moment conditions (Belloni and Chernozhukov, 2011; Wang and He, 2022) although certain regularity conditions on the conditional density function are still needed. This lack of robustness to heavy-tailedness is also observed in numerical studies. Since expectile regression is primarily introduced to explore the tail behavior of the conditional response distribution, its sensitivity to the tails of the error distributions, particularly in the presence of high-dimensional covariates, raises a major concern from a robustness viewpoint.

In this work, we aim to shrink the gap between quantile and expectile regressions, specifically in high dimensions, by proposing a robust expectile regression (`retire`) method that inherits the computational expediency and statistical efficiency of expectile regression and is nearly as robust as quantile regression against heavy-tailed response distributions. The main idea, which is adapted from Sun et al. (2020), is to replace the asymmetric squared loss associated with expectile regression with a Lipschitz and locally quadratic robust alternative, parameterized by a data-dependent parameter to achieve a desirable trade-off between bias and robustness. Under the low-dimensional regime "$d \ll n$", we provide nonasymptotic high probability error bounds, Bahadur representation, and a Berry–Esseen bound (normal approximation) for the retire estimator when the noise variable has bounded variance.

In the high-dimensional sparse setting "$\max\{s, \log(d)\} \ll n \ll d$", we propose an iteratively reweighted ℓ_1-penalized (IRW-ℓ_1) algorithm to obtain the penalized retire estimator, where s denotes the number of significant predictors. The problem boils down to iteratively minimizing convex loss functions (proven to be locally strongly convex with high probability), solvable by (but not limited to) a semismooth Newton coordinate descent type algorithm proposed by Yi and Huang (2017). Theoretically, we provide explicit error bounds (in high probability) for the solution path of IRW-ℓ_1. More specifically, we first obtain the statistical error of the ℓ_1-penalized retire estimator, i.e., the first iterate of the IRW-ℓ_1 algorithm initialized at zero. We then show that the statistical error for the subsequent estimators can be improved sequentially by a δ-fraction at each iteration for some constant $\delta \in (0, 1)$. Under a near necessary and sufficient minimum signal strength condition, we show that the IRW-ℓ_1 algorithm with $\mathcal{O}\{\log(\log d)\}$ iterations delivers an estimator that achieves the oracle rate of convergence with high probability.

The rest of the paper is organized as follows. In Section 2, we briefly revisit the connection and distinction between quantile and expectile regressions. We describe the proposed method in Section 3, where we construct the new loss function and detail the semismooth Newton algorithm to solve the resulting optimization problem. The theoretical properties of the proposed estimator are presented in Section 4. Sections 5 and 6 consist of extensive numerical studies and two data applications, respectively. The proofs of the theoretical results are given in the online supplementary material.

2. Background and problem setup

Let $y \in \mathbb{R}$ be a scalar response variable and $\boldsymbol{x} = (x_1, \ldots, x_d)^{\mathrm{T}} \in \mathbb{R}^d$ be a d-dimensional vector of covariates. The training data $(y_1, \boldsymbol{x}_1), \ldots, (y_n, \boldsymbol{x}_n)$ are independent copies of $(y, \boldsymbol{x})$. Given a location parameter $\tau \in (0, 1)$, we consider the linear model

$$y_i = \boldsymbol{x}_i^{\mathrm{T}} \boldsymbol{\beta}^*(\tau) + \varepsilon_i(\tau), \tag{2.1}$$

where $\boldsymbol{\beta}^*(\tau)$ is the unknown d-dimensional vector of regression coefficients, and $\varepsilon_i(\tau)$'s are independent random noise. Model (2.1) allows the regression coefficients $\boldsymbol{\beta}^*(\tau)$ to vary across different values of τ, and thereby offers insights into the entire conditional distribution of y given $\boldsymbol{x}$. Throughout the paper, we let $x_1 = 1$ so that β_1^* denotes the intercept term. We suppress the dependency of $\boldsymbol{\beta}^*(\tau)$ and $\varepsilon(\tau)$ on τ whenever there is no ambiguity.

The most natural way to relate the conditional distribution of y given $\boldsymbol{x}$ and the parameter process $\{\boldsymbol{\beta}^*(\tau), \tau \in (0, 1)\}$ is through quantile regression, under the assumption that $F^{-1}_{y_i|\boldsymbol{x}_i}(\tau) = \boldsymbol{x}_i^{\mathrm{T}} \boldsymbol{\beta}^*(\tau)$, or equivalently, $\mathbb{P}\{\varepsilon_i(\tau) \le 0 \,|\boldsymbol{x}_i\} = \tau$. Fitting a conditional quantile model involves minimizing a non-smooth piecewise linear loss function, $\varphi_\tau(u) = u\{\tau - \mathbb{1}(u < 0)\}$, typically recast as a linear program, solvable by the simplex algorithm or interior-point methods. For the latter, Portnoy and Koenker (1997) showed that the average-case computational complexity grows as a cubic function of the dimension d, and thus, is computationally demanding for problems with large dimensions.

Adapted from the concept of quantiles, Newey and Powell (1987) and Efron (1991) separately proposed an alternative class of location measures of a distribution, named the expectile according to the former. The resulting regression methods are referred to as the expectile regression or the asymmetric least squares regression, which are easy to compute and reasonably efficient under normality conditions. We start with some basic notation and facts for expectile regression. Let $Z \in \mathbb{R}$ be a random variable with finite moment, i.e., $\mathbb{E}(|Z|) < \infty$. The τth expectile or τ-mean of Z is defined as

$$e_\tau(Z) := \operatorname*{argmin}_{u \in \mathbb{R}} \mathbb{E}\big\{\eta_\tau(Z - u) - \eta_\tau(Z)\big\}, \qquad \tau \in (0, 1), \tag{2.2}$$

where

$$\eta_\tau(u) = |\tau - \mathbb{1}(u < 0)| \cdot \frac{u^2}{2} = \frac{\tau}{2}\{\max(u, 0)\}^2 + \frac{1 - \tau}{2}\{\max(-u, 0)\}^2 \tag{2.3}$$

is the asymmetric squared/ℓ_2 loss (Newey and Powell, 1987). The quantity $e_\tau(Z)$ is well defined as long as $\mathbb{E}|Z|$ is finite. When $\tau = 1/2$, it can be easily seen that $e_{1/2}(Z) = \mathbb{E}(Z)$. Therefore, expectiles can be viewed as an asymmetric

generalization of the mean, and the term "expectile" stems from a combination of "expectation" and "quantile". Moreover, expectiles are uniquely identified by the first-order condition

$$\tau \cdot \mathbb{E}(Z - e_\tau(Z))_+ = (1-\tau) \cdot \mathbb{E}(Z - e_\tau(Z))_-,$$

where $x_+ = \max(x, 0)$ and $x_- = \max(-x, 0)$. Note also that the τ-expectile of Z defined in (2.2) is equivalent to Efron's ω-mean with $\omega = \tau/(1-\tau)$ (Efron, 1991).

The notion of expectiles is a least squares counterpart of quantiles, and can be viewed as an alternative measure of "locations" of the random variable Z. Respectively, 1/2-expectile and 1/2-quantile correspond to the mean and median, both of which are related to the central behavior. In general, τ-expectile and τ-quantile with τ close to zero and one describe the lower and higher regions of the distribution of Z, respectively. The latter is the point below which $100\tau\%$ of the mass of Z lies, whereas the former specifies the position, say e_τ, such that the average distance from Z below e_τ to e_τ itself is $100\tau\%$ of the average distance between Z and e_τ.

Given independent observations $Z_1, \ldots, Z_n$ from Z, the expectile location estimator is given by $\widehat{e}_\tau = \operatorname{argmin}_{u\in\mathbb{R}} \sum_{i=1}^n \eta_\tau(Z_i - u)$, which is uniquely defined due to the strong convexity of the asymmetric ℓ_2-loss. The expectile estimator $\widehat{e}_\tau$ can also be interpreted as a maximum likelihood estimator of a normal distributed sample with unequal weights given to disturbances of differing signs, with a larger relative weight given to less variable disturbances (Aigner et al., 1976).

Essentially the asymmetric squared loss $\eta_\tau(\cdot)$ is an ℓ_2-version of the check function $\varphi_\tau(\cdot)$ for quantile regression. Given training data from the linear model (2.1) subject to $e_\tau(\varepsilon_i|\boldsymbol{x}_i) = 0$, the expectile regression estimator (Newey and Powell, 1987) is defined as the minimizer of the following convex optimization problem

$$\underset{\boldsymbol{\beta}\in\mathbb{R}^d}{\text{minimize}} \ \frac{1}{n}\sum_{i=1}^n \eta_\tau(y_i - \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}), \tag{2.4}$$

which consistently estimates $\boldsymbol{\beta}^*$ when $d = o(n)$ as $n \to \infty$. In particular, expectile regression with $\tau = 0.5$ reduces to the ordinary least squares regression.

3. Retire: Robust Expectile Regression

3.1. A class of asymmetric robust squared losses

Despite its computational advantage over quantile regression, expectile regression (2.4) is much more sensitive to heavy-tailed distributions due to the squared loss component in (2.3). This lack of robustness is amplified in the presence of high-dimensional covariates, and therefore necessitates the development of a new class of asymmetric loss functions that preserves the robustness of the check loss to a degree.

To this end, we construct a class of asymmetric robust loss functions that is more resistant against heavy-tailed error/response distributions. The main idea is to replace the quadratic component in (2.3) with a Lipschitz and locally strongly convex alternative, typified by the Huber loss (Huber, 1964) that is a hybrid ℓ_1/ℓ_2 function. The proposed loss function, $\ell_\gamma(u)$, contains a tuning parameter $\gamma > 0$ that is to be chosen to achieve a balanced trade-off between the robustification bias and the degree of robustness. At a high level, we focus on the class of loss functions that satisfies Condition 1 below.

Condition 1. Let $\ell_\gamma(u) = \gamma^2\ell(u/\gamma)$ for $u \in \mathbb{R}$, where the function $\ell : \mathbb{R} \mapsto [0, \infty)$ satisfies: (i) $\ell'(0) = 0$ and $|\ell'(u)| \le \min(a_1, |u|)$ for all $u \in \mathbb{R}$; (ii) $\ell''(0) = 1$ and $\ell''(u) \ge a_2$ for all $|u| \le a_3$; and (iii) $|\ell'(u) - u| \le u^2$ for all $u \in \mathbb{R}$, where a_1, a_2, and a_3 are positive constants.

Condition 1 encompasses many commonly used robust loss functions such as the Huber loss $\ell(u) = \min\{u^2/2, |u| - 1/2\}$ (Huber, 1964), pseudo-Huber losses $\ell(u) = \sqrt{1+u^2} - 1$ and $\ell(u) = \log(e^u/2 + e^{-u}/2)$, smoothed Huber losses $\ell(u) = \min\{u^2/2 - |u|^3/6, |u|/2 - 1/6\}$ and $\ell(u) = \min\{u^2/2 - u^4/24, (2\sqrt{2}/3)|u| - 1/2\}$, among other smooth approximations of the Huber loss (Lange, 1990). Consequently, we consider the following asymmetric robust loss

$$L_{\tau,\gamma}(u) := |\tau - \mathbb{1}(u < 0)| \cdot \ell_\gamma(u), \tag{3.1}$$

where $\ell_\gamma(\cdot)$ is subject to Condition 1.

In Section 3.2, we consider the robust expectile regression (retire) estimator based on the robust loss (3.1) in the classical setting that $d < n$. Its statistical properties, both asymptotic and nonasymptotic, will be given in Section 4.1 under the so-called "many regressors" model (Belloni et al., 2015) in which the dimension $d = d_n$ is allowed to grow with n subject to the constraint $d_n = o(n^a)$ for some $0 < a \le 1$. To deal with high-dimensional data for which d can be much larger than n, we propose penalized retire estimators in Section 3.3 with statistical guarantees (under sparsity) provided in Section 4.2.

3.2. Retire estimator in low dimensions

Given a location parameter $\tau \in (0, 1)$, we define the retire estimator (when $d < n$) as

$$\widehat{\boldsymbol{\beta}} = \widehat{\boldsymbol{\beta}}_\gamma = \underset{\boldsymbol{\beta}\in\mathbb{R}^d}{\operatorname{argmin}} \frac{1}{n}\sum_{i=1}^{n} L_{\tau,\gamma}(y_i - \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}), \tag{3.2}$$

where $\gamma > 0$ is a robustification parameter that will be calibrated adaptively from data as we detail in Section 5.1. Numerically, the optimization problem (3.2) can be efficiently solved by either gradient descent or quasi-Newton methods (Nocedal and Wright, 1999), such as the Broyden–Fletcher–Goldfarb–Shanno (BFGS) algorithm that can be implemented as on option of the base function `optim()` in R.

Recall that the population parameter $\boldsymbol{\beta}^*$ is uniquely identified as

$$\boldsymbol{\beta}^* = \underset{\boldsymbol{\beta}\in\mathbb{R}^d}{\operatorname{argmin}}\ \mathbb{E}\{L_{\tau,\infty}(y - \boldsymbol{x}^{\mathrm{T}}\boldsymbol{\beta})\} \ \text{ with } \ L_{\tau,\infty}(u) := |\tau - \mathbb{1}(u < 0)| \cdot u^2/2.$$

On the other hand, $\widehat{\boldsymbol{\beta}}$ can be viewed an M-estimator of the following population parameter

$$\boldsymbol{\beta}^*_\gamma := \underset{\boldsymbol{\beta}\in\mathbb{R}^d}{\operatorname{argmin}}\, \mathbb{E}\{L_{\tau,\gamma}(y - \boldsymbol{x}^{\mathrm{T}}\boldsymbol{\beta})\}.$$

It is worth pointing out that $\boldsymbol{\beta}^*_\gamma$ typically differs from $\boldsymbol{\beta}^*$ for any given $\gamma > 0$. To see this, note that the convexity of the robust loss $L_{\tau,\gamma} : \mathbb{R}^d \to \mathbb{R}$ implies the first-order condition, that is, $\mathbb{E}\{|\tau - \mathbb{1}(y < \boldsymbol{x}^{\mathrm{T}}\boldsymbol{\beta}^*_\gamma)| \cdot \ell'_{\tau,\gamma}(y - \boldsymbol{x}^{\mathrm{T}}\boldsymbol{\beta}^*_\gamma)\boldsymbol{x}\} = \mathbf{0}$. On the other hand, we have $e_\tau(\varepsilon|\boldsymbol{x}) = e_\tau(y - \boldsymbol{x}^{\mathrm{T}}\boldsymbol{\beta}^*|\boldsymbol{x}) = 0$, implying $\mathbb{E}\{|\tau - \mathbb{1}(\varepsilon < 0)| \cdot \varepsilon\boldsymbol{x}\} = \mathbf{0}$. Since the random error ε given $\boldsymbol{x}$ is asymmetric around zero, in general we have

$$\mathbf{0} \neq \mathbb{E}\{|\tau - \mathbb{1}(\varepsilon < 0)| \cdot \ell'_{\tau,\gamma}(\varepsilon)\boldsymbol{x}\} = \mathbb{E}\{|\tau - \mathbb{1}(y < \boldsymbol{x}^{\mathrm{T}}\boldsymbol{\beta}^*)| \cdot \ell'_{\tau,\gamma}(y - \boldsymbol{x}^{\mathrm{T}}\boldsymbol{\beta}^*)\boldsymbol{x}\},$$

which in turn implies that $\boldsymbol{\beta}^* \neq \boldsymbol{\beta}^*_\gamma$. We refer to the difference $\|\boldsymbol{\beta}^*_\gamma - \boldsymbol{\beta}^*\|_2$ as the robustification bias, where $\|\cdot\|_2$ is the ℓ_2-norm. In Section 4.1, we will show that under mild conditions, the robustification bias is of the order $\mathcal{O}(\gamma^{-1})$, and a properly chosen γ balances bias and robustness.

To perform statistical inference on β_j^*'s, we construct normal-based confidence intervals based on the asymptotic theory developed in Section 3.2. To this end, we first introduce some additional notation. Let $\widehat{\varepsilon}_i = y_i - \boldsymbol{x}_i^{\mathrm{T}}\widehat{\boldsymbol{\beta}}$ be the residuals from the fitted model and let $\mathbf{e}_j \in \mathbb{R}^d$ be the canonical basis vector, i.e., the jth entry equals one and all other entries equal zero. Let $\widehat{\mathbf{J}} = n^{-1}\sum_{i=1}^n |\tau - \mathbb{1}(\widehat{\varepsilon}_i < 0)| \cdot \boldsymbol{x}_i\boldsymbol{x}_i^{\mathrm{T}}$. An approximate 95% confidence interval for β_j^* can thus be constructed as

$$\left[\widehat{\beta}_j - 1.96\frac{\widehat{\sigma}(\mathbf{e}_j)}{\sqrt{n}},\ \widehat{\beta}_j - 1.96\frac{\widehat{\sigma}(\mathbf{e}_j)}{\sqrt{n}}\right],$$

where

$$\widehat{\sigma}^2(\mathbf{e}_j) := \mathbf{e}_j^{\mathrm{T}}\widehat{\mathbf{J}}^{-1}\left[\frac{1}{n}\sum_{i=1}^{n}\zeta^2(\widehat{\varepsilon}_i)\boldsymbol{x}_i\boldsymbol{x}_i^{\mathrm{T}}\right]\widehat{\mathbf{J}}^{-1}\mathbf{e}_j,$$

and $\zeta(u) = L'_{\tau,\gamma}(u) = |\tau - \mathbb{1}(u < 0)| \cdot \ell'_\gamma(u)$ is the first-order derivative of $L_{\tau,\gamma}(\cdot)$ given in (3.1).

3.3. Penalized retire estimator in high dimensions

In this section, we propose the penalized retire estimator for modeling high-dimensional data with $d > n$, obtained by minimizing the robust loss in (3.2) plus a penalty function that induces sparsity on the regression coefficients. As mentioned in Section 1, the non-negligible estimation bias introduced by convex penalties (e.g., the Lasso penalty) can be reduced by folded-concave regularization when the signals are sufficiently strong, that is, the minimum of magnitudes of all nonzero coefficients are away from zero to some extent. The latter, however, is computationally more challenging and unstable due to non-convexity.

Adapted from the local linear approximation algorithm proposed by Zou and Li (2008), we apply an iteratively reweighted ℓ_1-penalized algorithm for fitting sparse robust expectile regression models with the robust loss $L_{\tau,\gamma}(\cdot)$. At each iteration, the penalty weights depend on the previous iterate and the choice of a (folded) concave regularizer satisfying Condition 2 (Zhang and Zhang, 2012) below. Some popular examples include the smoothly clipped absolute deviation (SCAD) penalty (Fan and Li, 2001), the minimax concave penalty (Zhang, 2010a), and the capped-ℓ_1 penalty. We refer the reader to Zhang and Zhang (2012) and Section 4.4 of Fan et al. (2020) for more details.

Condition 2. The penalty function p_λ ($\lambda > 0$) is of the form $p_\lambda(t) = \lambda^2 p_0(t/\lambda)$ for $t \geq 0$, where the function $p_0 : \mathbb{R}_+ \to \mathbb{R}_+$ satisfies: (i) $p_0(\cdot)$ is non-decreasing on $[0, \infty)$ with $p_0(0) = 0$; (ii) $p_0(\cdot)$ is differentiable almost everywhere on $(0, \infty)$ and $\lim_{t\downarrow 0} p'_0(t) = 1$; (iv) $p'_0(t_1) \leq p'_0(t_2)$ for all $t_1 \geq t_2 > 0$.

Let $p_\lambda(\cdot)$ be a prespecified concave regularizer that satisfies Condition 2, and let $p'_\lambda(\cdot)$ be its first-order derivative. Starting at iteration zero an initial estimate $\widehat{\boldsymbol{\beta}}^{(0)}$, we sequentially solve the following weighted ℓ_1-penalized convex optimization problems:

$$\widehat{\boldsymbol{\beta}}^{(t)} \in \underset{\boldsymbol{\beta}\in\mathbb{R}^d}{\text{minimize}} \left\{ \frac{1}{n}\sum_{i=1}^{n} L_{\tau,\gamma}(y_i - \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}) + \sum_{j=2}^{d} p'_\lambda(|\widehat{\beta}_j^{(t-1)}|)|\beta_j| \right\}, \tag{3.3}$$

where $\widehat{\boldsymbol{\beta}}^{(t)} = (\widehat{\beta}_1^{(t)}, \ldots, \widehat{\beta}_d^{(t)})^{\mathrm{T}}$. At each iteration, $\widehat{\boldsymbol{\beta}}^{(t)}$ is a weighted ℓ_1-penalized robust expectile regression estimate, where the weight $p'_\lambda(|\widehat{\beta}_j^{(t-1)}|)|\beta_j|$ can be viewed as a local linear approximation of the concave regularizer $p_\lambda(|\beta_j|)$ around $|\widehat{\beta}_j^{(t-1)}|$. With the trivial initialization $\widehat{\boldsymbol{\beta}}^{(0)} = \mathbf{0}$, the first optimization problem (3.3) (when $t = 1$) reduces to the ℓ_1-penalized robust expectile regression because $p'_\lambda(0) = \lambda$. This iterative procedure outputs a sequence of estimates $\widehat{\boldsymbol{\beta}}^{(1)}, \ldots, \widehat{\boldsymbol{\beta}}^{(T)}$, where the number of iterations T can either be set before running the algorithm or depend on a stopping criterion. Throughout this paper, we refer to the sequence of estimates $\{\widehat{\boldsymbol{\beta}}^{(t)}\}_{t=1,\ldots,T}$ given in (3.3) as the *iteratively reweighted ℓ_1-penalized retire* estimators. We will characterize their statistical properties in Section 4.2, including the theoretical choice of T in order to obtain a statistically optimal estimator.

We now outline a coordinate descent type algorithm, the semismooth Newton coordinate descent (SNCD) algorithm proposed by Yi and Huang (2017), to solve the weighted ℓ_1-penalized convex optimization problem in (3.3). Recall that the key component of the asymmetric loss function $L_{\tau,\gamma}(\cdot)$ is the robust loss $\ell_\gamma(u) = \gamma^2\ell(u/\gamma)$. For convenience, we focus on the Huber loss for which $\ell(u) = u^2/2 \cdot \mathbb{1}(|u| \le 1) + (|u| - 1/2) \cdot \mathbb{1}(|u| > 1)$. The main crux of the SNCD algorithm is to combine the semismooth Newton method and the cyclic coordinate descent algorithm to iteratively update the parameter of interest one at a time via a Newton-type step until convergence. In the following, we provide a brief derivation of the algorithm, and defer the details to Section A of the supplementary material.

Let $L'_{\tau,\gamma}(u)$ and $L''_{\tau,\gamma}(u)$ be the first- and second-order derivatives (with respect to u) of the loss function in (3.1), respectively. For notational convenience, let $\lambda_j^{(t)} = p'_\lambda(|\widehat{\beta}_j^{(t-1)}|)$ be the penalty weights at the tth iteration. Then, the Karush–Kuhn–Tucker conditions for (3.3) read

$$\begin{cases} -\frac{1}{n}\sum_{i=1}^{n} L'_{\tau,\gamma}(y_i - \boldsymbol{x}_i^{\mathrm{T}}\widehat{\boldsymbol{\beta}}) = 0 \quad \text{for } j = 1, \\ -\frac{1}{n}\sum_{i=1}^{n} L'_{\tau,\gamma}(y_i - \boldsymbol{x}_i^{\mathrm{T}}\widehat{\boldsymbol{\beta}})x_{ij} + \lambda_j^{(t)}\widehat{z}_j = 0 \quad \text{for } j = 2, \ldots, d, \\ \widehat{\beta}_j - S(\widehat{\beta}_j + \widehat{z}_j) = 0 \quad \text{for } j = 2, \ldots, d, \end{cases} \tag{3.4}$$

where $\widehat{z}_j \in \partial|\widehat{\beta}_j|$ is a subgradient of the absolute value function, and $S(u) = \mathrm{sign}(u)\max(|u| - 1, 0)$. Finding the optimum to the optimization problem (3.3) is equivalent to solving the system of Eqs. (3.4). The latter can be done iteratively in a cyclic fashion. That is, at each iteration, we update the pair of parameters (β_j, z_j) by solving the corresponding equations in (3.4) while keeping the remaining parameters fixed. Each pair of parameters is updated by a semismooth Newton step, which we detail in Section A of the online supplementary material. The whole procedure is summarized in Algorithm 1.

Algorithm 1 Semismooth Newton Coordinate Descent Algorithm for Solving (3.3) with a Huber Loss.

Input: regularization parameter λ, Huber loss tuning parameter γ, and convergence criterion ϵ.

Initialization: $\widehat{\boldsymbol{\beta}}^0 = \mathbf{0}$.

Iterate: the following until the stopping criterion $\|\widehat{\boldsymbol{\beta}}^k - \widehat{\boldsymbol{\beta}}^{k-1}\|_2 \le \epsilon$ is met, where $\widehat{\boldsymbol{\beta}}^k$ is the value of $\boldsymbol{\beta}$ obtained at the kth iteration.

1. $\widehat{\beta}_1^{k+1} \leftarrow \widehat{\beta}_1^k + \{\sum_{i=1}^n L'_{\tau,\gamma}(y_i - \boldsymbol{x}_i^{\mathrm{T}}\widehat{\boldsymbol{\beta}}^k)\}/\{\sum_{i=1}^n L''_{\tau,\gamma}(y_i - \boldsymbol{x}_i^{\mathrm{T}}\widehat{\boldsymbol{\beta}}^k)\}$.
2. for $j = 2, \ldots, d$, update the pair of parameters (β_j, z_j) as follows:

$$\begin{bmatrix} \widehat{\beta}_j^{k+1} \\ \widehat{z}_j^{k+1} \end{bmatrix} = \begin{cases} \begin{bmatrix} \widehat{\beta}_j^k + \frac{\sum_{i=1}^n L'_{\tau,\gamma}(y_i - \boldsymbol{x}_i^{\mathrm{T}}\widehat{\boldsymbol{\beta}}^k)x_{ij} - \lambda_j^{(t)} \cdot \mathrm{sign}(\widehat{\beta}_j^k + \widehat{z}_j^k)}{\sum_{i=1}^n L''_{\tau,\gamma}(y_i - \boldsymbol{x}_i^{\mathrm{T}}\widehat{\boldsymbol{\beta}}^k)x_{ij}^2} \\ \mathrm{sign}(\widehat{\beta}_j^k + \widehat{z}_j^k) \end{bmatrix}, & \text{if } |\widehat{\beta}_j^k + \widehat{z}_j^k| > 1, \\ \begin{bmatrix} 0 \\ (n\lambda_j^{(t)})^{-1}\sum_{i=1}^n L'_{\tau,\gamma}(y_i - \boldsymbol{x}_i^{\mathrm{T}}\widehat{\boldsymbol{\beta}}^k)x_{ij} + (n\lambda_j^{(t)})^{-1}\widehat{\beta}_j^k\sum_{i=1}^n L''_{\tau,\gamma}(y_i - \boldsymbol{x}_i^{\mathrm{T}}\widehat{\boldsymbol{\beta}}^k)x_{ij}^2 \end{bmatrix}, & \text{if } |\widehat{\beta}_j^k + \widehat{z}_j^k| \le 1. \end{cases}$$

Output: the final iterate $\widehat{\boldsymbol{\beta}}^k$.

For the Huber loss, the first- and second-order derivatives of $L_{\tau,\gamma}(u)$ are

$$L'_{\tau,\gamma}(u) = \begin{cases} -(1-\tau)\gamma, & \text{if } u < -\gamma, \\ (1-\tau)u, & \text{if } -\gamma \le u < 0, \\ \tau u, & \text{if } 0 \le u \le \gamma, \\ \tau\gamma, & \text{if } u > \gamma, \end{cases} \quad \text{and} \quad L''_{\tau,\gamma}(u) = \begin{cases} 1-\tau, & \text{if } -\gamma \le u < 0, \\ \tau, & \text{if } 0 \le u \le \gamma, \\ 0, & \text{otherwise}, \end{cases}$$

respectively. In Algorithm 1, the update $\widehat{\beta}_j^{k+1}$ involves the second-order derivative of the loss function, $\sum_{i=1}^n L''_{\tau,\gamma}(y_i - \boldsymbol{x}_i^{\mathrm{T}}\widehat{\boldsymbol{\beta}}^k)$, in the denominator. For extreme values of τ that are near zero or 1, the denominator may be close to zero, causing

instability. To address this issue, Yi and Huang (2017) implemented a continuity approximation in their R package `hqreg`, and we adopt the same technique to implement Algorithm 1. In particular, if the sum of second-order derivatives is equal to zero or the percentage of residuals with magnitude below γ is less than 5% or n^{-1}, we instead substitute the sum of second-order derivatives by the quantity $\sum_{i=1}^{n}(|y_i - \boldsymbol{x}_i^{\mathrm{T}}\widehat{\boldsymbol{\beta}}^k|)^{-1}\mathbb{1}(|y_i - \boldsymbol{x}_i^{\mathrm{T}}\widehat{\boldsymbol{\beta}}^k| > \gamma)$. Such a continuity approximation works well across all of the numerical settings that we have considered.

4. Theoretical results

We provide an explicit characterization of the estimation error for the retire estimator $\widehat{\boldsymbol{\beta}}$ defined in (3.2) (low-dimensional setting), and the sequence of penalized retire estimators $\{\widehat{\boldsymbol{\beta}}^{(t)}\}_{t=1,\ldots,T}$ defined in (3.3) (high-dimensional setting) in Sections 4.1 and 4.2, respectively. Our proposed estimator relies on the choice of robust loss function in Condition 1. For simplicity, we focus on the Huber loss $\ell(u) = u^2/2 \cdot \mathbb{1}(|u| \le 1) + (|u| - 1/2) \cdot \mathbb{1}(|u| > 1)$ throughout our analysis, i.e., $a_1 = a_2 = a_3 = 1$ in Condition 1, but note that similar results hold for any robust loss that satisfies Condition 1. Throughout the theoretical analysis, we assume that the location measure $\tau \in (0, 1)$ is fixed.

We first defined the empirical loss function and its gradient as

$$\mathcal{R}_n(\boldsymbol{\beta}) = \frac{1}{n}\sum_{i=1}^{n} L_{\tau,\gamma}(y_i - \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}) \quad \text{and} \quad \nabla\mathcal{R}_n(\boldsymbol{\beta}) = -\frac{1}{n}\sum_{i=1}^{n} L'_{\tau,\gamma}(y_i - \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta})\boldsymbol{x}_i,$$

respectively. Moreover, we impose some common conditions on the random covariates $\boldsymbol{x}$ and the random noise ε for both low-and high-dimensional settings. In particular, we assume that the random covariates $\boldsymbol{x} \in \mathbb{R}^d$ are sub-exponential and that the random noise ε is heavy-tailed with finite second moment.

Condition 3. Let $\boldsymbol{\Sigma} = \mathbb{E}(\boldsymbol{x}\boldsymbol{x}^{\mathrm{T}})$ be a positive definite matrix with $\lambda_u \ge \lambda_{\max}(\boldsymbol{\Sigma}) \ge \lambda_{\min}(\boldsymbol{\Sigma}) \ge \lambda_l > 0$ and assume that $\lambda_l = 1$ for simplicity. There exists $\nu_0 \ge 1$ such that $\mathbb{P}(|\boldsymbol{u}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1/2}\boldsymbol{x}| \ge \nu_0\|\boldsymbol{u}\|_2 \cdot t) \le e^{-t}$ for all $t \in \mathbb{R}$ and $\boldsymbol{u} \in \mathbb{R}^d$. For notational convenience, let $\sigma_{\boldsymbol{x}}^2 = \max_{1\le j\le d}\sigma_{jj}$, where σ_{jj} is the jth diagonal entry of $\boldsymbol{\Sigma}$.

Condition 4. The random noise ε has a finite second moment, i.e., $\mathbb{E}(\varepsilon^2|\boldsymbol{x}) \le \sigma_\varepsilon^2 < \infty$. Moreover, the conditional τ-expectile of ε satisfies $\mathbb{E}[w_\tau(\varepsilon)\varepsilon|\boldsymbol{x}] = 0$, where $w_\tau(u) := |\tau - \mathbb{1}(u < 0)|$.

4.1. Statistical theory for the retire estimator in (3.2)

In this section, we provide nonasymptotic error bounds for the retire estimator, $\widehat{\boldsymbol{\beta}}$, under the regime in which $n > d$ but d is allowed to diverge. Moreover, we establish a nonasymptotic Bahadur representation for $\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*$, based on which we construct a Berry–Esseen bound for a normal approximation. As mentioned in Section 3.2, the robustification bias $\|\boldsymbol{\beta}_\gamma^* - \boldsymbol{\beta}^*\|_2$ is inevitable due to the asymmetry nature of error term ε. Let $\underline{\tau} = \min(\tau, 1-\tau)$, $\bar{\tau} = \max(\tau, 1-\tau)$, and $A_1 \ge 1$ be a constant satisfying $\mathbb{E}(\boldsymbol{u}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1/2}\boldsymbol{x})^4 \le A_1^4\|\boldsymbol{u}\|_2^4$ for all $\boldsymbol{u} \in \mathbb{R}^d$. The following proposition reveals the fact that the robustification bias scales at the rate γ^{-1}, which decays as γ grows.

Proposition 4.1. *Assume Conditions 1, 3, and 4 hold. Provided that $\gamma \ge 2\sigma_\varepsilon A_1^2\bar{\tau}/\underline{\tau}$, we have*

$$\|\boldsymbol{\Sigma}^{1/2}(\boldsymbol{\beta}_\gamma^* - \boldsymbol{\beta}^*)\|_2 \le 2\gamma^{-1}(\bar{\tau}/\underline{\tau})\sigma_\varepsilon^2.$$

The key to our subsequent analysis for the retire estimator $\widehat{\boldsymbol{\beta}}$ is the strong convexity property of the empirical loss function $\mathcal{R}_n(\cdot)$ uniformly over a local ellipsoid centered at $\boldsymbol{\beta}^*$ with high probability. Let $\kappa_1 = \min_{|u|\le 1}\ell''(u)$ and let $\mathbb{B}_{\boldsymbol{\Sigma}}(r) = \{\boldsymbol{\delta} \in \mathbb{R}^d : \|\boldsymbol{\Sigma}^{1/2}\boldsymbol{\delta}\|_2 \le r\}$ be an ellipsoid. We characterize the strong convexity of $\mathcal{R}_n(\cdot)$ in Lemma 4.1. With the aid of Lemma 4.1, we establish a non-asymptotic error bound for the retire estimator $\widehat{\boldsymbol{\beta}}$ in Theorem 4.1.

Lemma 4.1. *Let (γ, n) satisfy $\gamma \ge 4\sqrt{2}\max\{\sigma_\varepsilon, 2A_1^2 r\}$ and $n \gtrsim (\gamma/r)^2(d+t)$. Under Conditions 1, 3, and 4, with probability at least $1 - e^{-t}$, we have*

$$\langle\nabla\mathcal{R}_n(\boldsymbol{\beta}) - \nabla\mathcal{R}_n(\boldsymbol{\beta}^*), \boldsymbol{\beta} - \boldsymbol{\beta}^*\rangle \ge \frac{1}{2}\kappa_1\underline{\tau}\|\boldsymbol{\Sigma}^{1/2}(\boldsymbol{\beta} - \boldsymbol{\beta}^*)\|_2^2 \quad \textit{uniformly over} \quad \boldsymbol{\beta} \in \boldsymbol{\beta}^* + \mathbb{B}_{\boldsymbol{\Sigma}}(r).$$

Theorem 4.1. *Assume Conditions 1, 3, and 4 hold. For any $t > 0$, the retire estimator $\widehat{\boldsymbol{\beta}}$ in (3.2) with $\gamma = \sigma_\varepsilon\sqrt{n/(d+t)}$ satisfies the bound*

$$\|\boldsymbol{\Sigma}^{1/2}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)\|_2 \le C(\bar{\tau}/\underline{\tau})\kappa_1^{-1}\sigma_\varepsilon\nu_0\sqrt{\frac{d+t}{n}},$$

with probability at least $1 - 2e^{-t}$ as long as $n \gtrsim d + t$, where $C > 0$ is an absolute constant.

Theorem 4.1 shows that under the sub-exponential design with heavy-tailed random errors with bounded second moment, the retire estimator $\widehat{\boldsymbol{\beta}}$ exhibits a sub-Gaussian type deviation bound, provided that the robustification parameter is properly chosen, i.e., $\gamma = \sigma_\varepsilon\sqrt{n/(d+t)}$. In other words, the proposed retire estimator gains robustness to heavy-tailed random noise without compromising statistical accuracy.

Remark 4.1. The choice of $\gamma = \sigma_\varepsilon\sqrt{n/(d+t)}$ in Theorem 4.1 is a reflection of the bias and robustness trade-off for the retire estimator $\widehat{\boldsymbol{\beta}}$. Intuitively, a large γ creates less robustification bias but sacrifices robustness. More specifically, we shall see from the proof of Theorem 4.1 that conditioning on the event $\{\widehat{\boldsymbol{\beta}} \in \boldsymbol{\beta}^* + \mathbb{B}_{\boldsymbol{\Sigma}}(r_{\mathrm{loc}})\}$,

$$\|\boldsymbol{\Sigma}^{1/2}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)\|_2 \lesssim \underbrace{\frac{\sigma_\varepsilon^2}{\gamma}}_{\text{robustification bias}} + \underbrace{\sigma_\varepsilon\sqrt{\frac{d}{n}} + \gamma\frac{d}{n}}_{\text{statistical error}}$$

with high probability. Therefore, we choose $\gamma = \sigma_\varepsilon\sqrt{n/(d+t)}$ to minimize the right-hand side as a function of γ.

Next, we establish nonasymptotic Bahadur representation for the difference $\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*$. To this end, we need slightly stronger conditions on both the random covariates $\boldsymbol{x}$ and the random noise ε. In particular, we require that the random covariate vector $\boldsymbol{x}$ is sub-Gaussian and that the conditional density of the random noise ε is upper bounded. We formalize the above into the following conditions.

Condition 5. There exists $\nu_1 \geq 1$ such that $\mathbb{P}(|\boldsymbol{u}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1/2}\boldsymbol{x}| \geq \nu_1\|\boldsymbol{u}\|_2 t) \leq 2e^{-t^2/2}$ for all $t \in \mathbb{R}$ and $\boldsymbol{u} \in \mathbb{R}^d$.

Condition 6. Let $f_{\varepsilon|\boldsymbol{x}}(\cdot)$ be the conditional density function of the random noise ε. There exists $\bar{f}_{\varepsilon|\boldsymbol{x}} > 0$ such that $\sup_{u\in\mathbb{R}} f_{\varepsilon|\boldsymbol{x}}(u) \leq \bar{f}_{\varepsilon|\boldsymbol{x}}$ almost surely (for all $\boldsymbol{x}$).

Recall that $w_\tau(u) = |\tau - \mathbb{1}(u<0)|$ and that $\zeta(u) = L'_{\tau,\gamma}(u) = w_\tau(u)\ell'_\gamma(u)$. Moreover, let $\mathbf{J} = \mathbb{E}\{w_\tau(\varepsilon)\boldsymbol{x}\boldsymbol{x}^{\mathrm{T}}\}$ be the Hessian matrix. Theorem 4.2 establishes the Bahadur representation of the retire estimator $\widehat{\boldsymbol{\beta}}$. Specifically, we show that the remainder of the Bahadur representation for $\widehat{\boldsymbol{\beta}}$ exhibits sub-exponential tails, which we will use to establish the Berry–Esseen bound for linear projections of $\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*$ in Theorem 4.3.

Theorem 4.2. *Assume Conditions 1, 4, 5, and 6 hold. For any $t > 0$, the retire estimator $\widehat{\boldsymbol{\beta}}$ given in (3.2) with $\gamma = \sigma_\varepsilon\sqrt{n/(d+t)}$ satisfies the following nonasymptotic Bahadur representation*

$$\left\|\boldsymbol{\Sigma}^{-1/2}\mathbf{J}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*) - \frac{1}{n}\sum_{i=1}^n \zeta(\varepsilon_i)\boldsymbol{\Sigma}^{-1/2}\boldsymbol{x}_i\right\|_2 \leq C\sigma_\varepsilon \cdot \frac{d+t}{n} \tag{4.1}$$

with probability at least $1 - 3e^{-t}$ as long as $n \gtrsim d + t$, where $C > 0$ is a constant independent of (n, d) and t.

Theorem 4.3. *Under the same set of conditions as in Theorem 4.2, assume further that $\mathbb{E}(|\varepsilon|^3|\boldsymbol{x}) \leq \nu_3 < \infty$ (almost surely). Then, the retire estimator $\widehat{\boldsymbol{\beta}}$ in (3.2) with $\gamma = \sigma_\varepsilon\sqrt{n/(d+\log n)}$ satisfies*

$$\sup_{\boldsymbol{u}\in\mathbb{R}^d, z\in\mathbb{R}} \left|\mathbb{P}(n^{1/2}\langle \boldsymbol{u}, \widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\rangle \leq \sigma z) - \Phi(z)\right| \lesssim \frac{d+\log n}{\sqrt{n}},$$

where $\sigma^2 = \sigma^2(\boldsymbol{u}) := \boldsymbol{u}^{\mathrm{T}}\mathbf{J}^{-1}\mathbb{E}\{\zeta^2(\varepsilon)\boldsymbol{x}\boldsymbol{x}^{\mathrm{T}}\}\mathbf{J}^{-1}\boldsymbol{u}$ and $\Phi(\cdot)$ is the standard normal cumulative distribution function.

Theorem 4.3 shows that with a diverging parameter $\gamma = \sigma_\varepsilon\sqrt{n/(d+\log n)}$, for any $\boldsymbol{u} \in \mathbb{R}^d$, the linear projection of $\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*$ is asymptotically normal after some standardization as long as (n, d) satisfies the scaling condition $d = o(\sqrt{n})$.

4.2. Statistical theory for the iteratively reweighted ℓ_1-penalized retire estimator

In this section, we analyze the sequence of estimators $\{\widehat{\boldsymbol{\beta}}^{(t)}\}_{t=1}^T$ obtained in (3.3) under the high-dimensional regime in which $d > n$. Throughout the theoretical analysis, we assume that the regression parameter $\boldsymbol{\beta}^* \in \mathbb{R}^d$ in model (2.1) is exactly sparse, i.e., $\boldsymbol{\beta}^*$ has s non-zero coordinates. Let $\mathcal{S} = \{1 \leq j \leq d : \beta_j^* \neq 0\}$ be the active set of $\boldsymbol{\beta}^*$ with cardinality $|\mathcal{S}| = s$. Recall that $\underline{\tau} = \min(\tau, 1-\tau)$, $\kappa_1 = \min_{|u|\leq 1}\ell''(u)$ and $A_1 > 0$ is a constant that satisfies $\mathbb{E}(\boldsymbol{u}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1/2}\boldsymbol{x})^4 \leq A_1^4\|\boldsymbol{u}\|_2^4$ for all $\boldsymbol{u} \in \mathbb{R}^d$, where $\boldsymbol{x}$ satisfies Condition 3. Similar to the low-dimensional setting, the key to our high-dimensional analysis is an event $\mathcal{E}_{\mathrm{rsc}}$ that characterizes the local restricted strong convexity property of the empirical loss function $\mathcal{R}_n(\cdot)$ over the intersection of an ℓ_1-cone and a local ℓ_2-ball centered at $\boldsymbol{\beta}^*$ (Loh and Wainwright, 2015). Lemma 4.2 below shows that the event $\mathcal{E}_{\mathrm{rsc}}$ occurs with high probability for suitably chosen parameters.

Definition 4.1. Given radii parameters $r, L > 0$ and a curvature parameter $\kappa > 0$, define the event

$$\mathcal{E}_{\mathrm{rsc}}(r, L, \kappa) = \left\{ \inf_{\boldsymbol{\beta} \in \boldsymbol{\beta}^* + \mathbb{B}(r) \cap \mathbb{C}(L)} \frac{\langle \nabla \mathcal{R}_n(\boldsymbol{\beta}) - \nabla \mathcal{R}_n(\boldsymbol{\beta}^*), \boldsymbol{\beta} - \boldsymbol{\beta}^* \rangle}{\|\boldsymbol{\beta} - \boldsymbol{\beta}^*\|_2^2} \geq \kappa \right\},$$

where $\mathbb{B}(r) = \{\boldsymbol{\delta} \in \mathbb{R}^d : \|\boldsymbol{\delta}\|_2 \leq r\}$ is an ℓ_2-ball with radius r, and $\mathbb{C}(L) = \{\boldsymbol{\delta} : \|\boldsymbol{\delta}\|_1 \leq L\|\boldsymbol{\delta}\|_2\}$ is an ℓ_1-cone.

Lemma 4.2. *Let the radii parameters (r, L) and the robustification parameter γ satisfy*

$$\gamma \geq 4\sqrt{2}\lambda_u \max\{\sigma_\varepsilon, 2A_1^2 r\} \qquad \text{and} \qquad n \gtrsim (\sigma_{\boldsymbol{x}} \nu_0 \gamma / r)^2 (L^2 \log d + t).$$

Then, under Conditions 1, 3, and 4, event $\mathcal{E}_{\mathrm{rsc}}(r, L, \kappa)$ with $\kappa = \kappa_1 \underline{\tau}/2$ occurs with probability at least $1 - e^{-t}$.

Under the local restricted strong convexity, in Theorem 4.4, we provide an upper bound on the estimation error of $\widehat{\boldsymbol{\beta}}^{(1)}$, i.e., the ℓ_1-penalized retire estimator.

Theorem 4.4. *Assume Conditions 1, 3, and 4 hold. Then, the ℓ_1-penalized retire estimator $\widehat{\boldsymbol{\beta}}^{(1)}$ with $\gamma = \sigma_\varepsilon \sqrt{n/(\log d + t)}$ and $\lambda \asymp \sqrt{(\log d + t)/n}$ satisfies the bounds*

$$\|\widehat{\boldsymbol{\beta}}^{(1)} - \boldsymbol{\beta}^*\|_2 \leq 3(\kappa_1 \underline{\tau})^{-1} s^{1/2} \lambda \qquad \text{and} \qquad \|\widehat{\boldsymbol{\beta}}^{(1)} - \boldsymbol{\beta}^*\|_1 \leq 12(\kappa_1 \underline{\tau})^{-1} s\lambda,$$

with probability as least $1 - 3e^{-t}$.

Theorem 4.4 shows that with an appropriate choice of the tuning parameters γ and λ, the ℓ_1-penalized robust expectile regression satisfies exponential deviation bounds with near-optimal convergence rate as if sub-Gaussian random noise were assumed (Gu and Zou, 2016).

Remark 4.2. Condition 4 can be further relaxed to accommodate heavy-tailed random error with finite $(1 + \phi)$ moment with $0 < \phi < 1$. Specifically, it can be shown that under the ℓ_2 norm, the estimation error of the ℓ_1-penalized Huber regression estimator takes the form $s^{1/2}\{\log(d)/n\}^{\min\{\phi/(1+\phi), 1/2\}}$ (Sun et al., 2020; Tan et al., 2023). Similar results can be obtained for the proposed ℓ_1-penalized retire estimator and we leave it for future work.

Remark 4.3. Throughout this section, we assume that the underlying regression parameter $\boldsymbol{\beta}^* \in \mathbb{R}^d$ is exactly sparse. In this case, iteratively reweighted ℓ_1-penalization helps reduce the estimation bias from ℓ_1-penalization as signal strengthens. For weakly sparse vectors $\boldsymbol{\beta}^*$ satisfying $\sum_{j=1}^d |\beta_j^*|^q \leq R_q$ for some $0 < q \leq 1$ and $R_q > 0$, Fan et al. (2017) showed that the convergence rate (under ℓ_2-norm) of the ℓ_1-penalized adaptive Huber estimator with a suitably chosen robustification parameter is of order $\mathcal{O}(\sigma\sqrt{R_q}\{\log(d)/n\}^{1/2-q/4})$. Using the same argument, the results in Theorem 4.4 can be directly extended to the weakly sparse case where $\boldsymbol{\beta}^*$ belongs to an L_q-ball for some $0 < q \leq 1$. For recovering weakly sparse signals, folded-concave penalization no longer improves upon ℓ_1-penalization, and therefore we will not provide details on such an extension.

Next, we establish the statistical properties for the entire sequence of estimators $\widehat{\boldsymbol{\beta}}^{(1)}, \widehat{\boldsymbol{\beta}}^{(2)}, \ldots, \widehat{\boldsymbol{\beta}}^{(T)}$ obtained from solving the convex optimization problem (3.3) iteratively. Let $\|\boldsymbol{\beta}_{\mathcal{S}}^*\|_{\min} = \min_{j \in \mathcal{S}} |\beta_j^*|$ be the smallest (in absolute value) non-zero regression coefficient. Under a beta-min condition, we show that the estimation error of $\widehat{\boldsymbol{\beta}}^{(1)}$ stated in Theorem 4.4 can be refined. More generally, given the previous iterate $\widehat{\boldsymbol{\beta}}^{(T-1)}$, the estimation error of the subsequent estimator, $\widehat{\boldsymbol{\beta}}^{(T)}$, can be improved by a δ-fraction for some constant $\delta \in (0, 1)$.

Theorem 4.5. *Let $p_0(\cdot)$ be a penalty function satisfying Condition 2. Under Conditions 1, 3 and 4, assume there exist some constants $a_1 > a_0 > 0$ such that*

$$a_0 > \sqrt{5}/(\kappa_1 \underline{\tau}), \qquad p_0'(a_0) > 0, \qquad p_0'(a_1) = 0.$$

Assume further the minimum signal strength condition $\|\boldsymbol{\beta}_{\mathcal{S}}^*\|_{\min} \geq (a_0 + a_1)\lambda$ and the sample size requirement $n \gtrsim s \log d + t$. Picking $\gamma \asymp \sigma_\varepsilon \sqrt{n/(s + \log d + t)}$ and $\lambda \asymp \sigma_\varepsilon \sqrt{(\log d + t)/n}$, we have

$$\|\widehat{\boldsymbol{\beta}}^{(T)} - \boldsymbol{\beta}^*\|_2 \lesssim \delta^{T-1} \sigma_\varepsilon \sqrt{\frac{s(\log d + t)}{n}} + \frac{\sigma_\varepsilon}{1 - \delta} \sqrt{\frac{s + \log d + t}{n}},$$

with probability at least $1 - 4e^{-t}$. Furthermore, setting $T \gtrsim \frac{\log\{\log(d) + t\}}{\log(1/\delta)}$, we have

$$\|\widehat{\boldsymbol{\beta}}^{(T)} - \boldsymbol{\beta}^*\|_2 \lesssim \sigma_\varepsilon \sqrt{\frac{s + \log d + t}{n}} \tag{4.2}$$

$$\text{and} \quad \|\widehat{\boldsymbol{\beta}}^{(T)} - \boldsymbol{\beta}^*\|_1 \lesssim \sigma_\varepsilon s^{1/2} \sqrt{\frac{s + \log d + t}{n}} \tag{4.3}$$

with probability at least $1 - 4e^{-t}$, where $\delta = \sqrt{5}/(a_0 \kappa_1 \underline{\tau}) < 1$.

Theorem 4.5 shows that under the beta-min condition $\|\boldsymbol{\beta}^*_{\mathcal{S}}\|_{\min} \gtrsim \sqrt{\log(d)/n}$, the iteratively reweighted ℓ_1-penalized retire estimator $\widehat{\boldsymbol{\beta}}^{(T)}$ with $T \asymp \log\{\log(d)\}$ achieves the near-oracle convergence rate, i.e., the convergence rate of the oracle estimator that has access to the true support of $\boldsymbol{\beta}^*$. This is also known as the weak oracle property. Picking $t = \log d$, we see that iteratively reweighted ℓ_1-penalization refines the statistical rate from $\sqrt{s\log(d)/n}$ for $\widehat{\boldsymbol{\beta}}^{(1)}$ to $\sqrt{(s+\log d)/n}$ for $\widehat{\boldsymbol{\beta}}^{(T)}$.

Remark 4.4. Theorem 4.5 reveals the so-called *weak oracle property* in the sense that the regularized estimator $\widehat{\boldsymbol{\beta}}^{(T)}$ enjoys the same convergence rate as the oracle estimator defined by regressing only on the significant predictors. To obtain such a result, the required minimum signal strength $\|\boldsymbol{\beta}^*_{\mathcal{S}}\|_{\min} \gtrsim \sqrt{\log(d)/n}$ is almost necessary and sufficient. To see this, consider the linear model $y_i = \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}^* + \varepsilon_i$ with $\varepsilon_i \sim N(0,\sigma^2)$ independent of $\boldsymbol{x}_i$, and define the parameter space $\Omega_{s,a} = \{\boldsymbol{\beta} \in \mathbb{R}^d : \|\boldsymbol{\beta}\|_0 \le s, \min_{j:\beta_j\neq 0}|\beta_j| \ge a\}$ for $a > 0$. Under the assumption that the design matrix $\mathbb{X} = (\boldsymbol{x}_1,\ldots,\boldsymbol{x}_n)^{\mathrm{T}} \in \mathbb{R}^{n\times d}$ satisfies a restricted isometry property and has normalized columns, Ndaoud (2019) derived the following sharp lower bounds for the minimax risk $\psi(s,a) := \inf_{\widehat{\boldsymbol{\beta}}}\sup_{\boldsymbol{\beta}^*\in\Omega_{s,a}}\mathbb{E}\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_2^2$: for any $\epsilon \in (0,1)$,

$$\psi(s,a) \ge \{1+o(1)\}\frac{2\sigma^2 s\log(ed/s)}{n} \quad \text{for any } a \le (1-\epsilon)\sigma\sqrt{\frac{2\log(ed/s)}{n}}$$
$$\text{and} \quad \psi(s,a) \ge \{1+o(1)\}\frac{\sigma^2 s}{n} \quad \text{for any } a \ge (1+\epsilon)\sigma\sqrt{\frac{2\log(ed/s)}{n}},$$

where the limit corresponds to $s/d \to 0$ and $s\log(ed/s)/n \to 0$. The minimax rate $2\sigma^2 s\log(ed/s)/n$ is attainable by both Lasso and Slope (Bellec et al., 2018), while the oracle rate $\sigma^2 s/n$ can only be achieved when the magnitude of the minimum signal is of order $\sigma\sqrt{\log(d/s)/n}$. The beta-min condition imposed in Theorem 4.5 is thus (nearly) necessary and sufficient, and is the weakest possible within constant factors.

Under a stronger beta-min condition $\|\boldsymbol{\beta}^*_{\mathcal{S}}\|_{\min} \gtrsim \sqrt{s\log(d)/n}$, Gu and Zou (2016) showed that with high probability, the IRW-ℓ_1 expectile regression estimator (initialized by zero) coincides with the oracle estimator after three iterations. This is known as the *strong oracle property*. Based on the more refined analysis by Pan et al. (2021), we conjecture that the IRW-ℓ_1 retire estimator $\widehat{\boldsymbol{\beta}}^{(T)}$ with $T \asymp \log(s \vee \log d)$ achieves the strong oracle property provided $\|\boldsymbol{\beta}^*_{\mathcal{S}}\|_{\min} \gtrsim \sqrt{\log(d)/n}$ without the $\sqrt{s}$-factor.

5. Numerical studies

5.1. Simulated data

We evaluate the performance of the proposed IRW-ℓ_1-penalized `retire` estimator *via* extensive numerical studies. We implement the ℓ_1-penalized `retire` and the IRW-ℓ_1-penalized `retire` using SCAD-based weights with $T = 3$, which we compare to three other competitive methods: (i) ℓ_1-penalized Huber regression (`huber`); (ii) ℓ_1-penalized asymmetric least squares regression (`sales`) proposed by Gu and Zou (2016), and (iii) ℓ_1-penalized quantile regression (`qr`) implemented via the `R` package `rqPen` (Sherwood and Maidman, 2020). To assess the performance across different methods, we report the estimation error under the ℓ_2-norm, i.e., $\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_2$, the true positive rate (TPR), and the false positive rate (FPR). Here, TPR is defined as the proportion of the number of correctly identified non-zeros and the false positive rate is calculated as the proportion of the number of incorrectly identified nonzeros.

Note that `huber` and `sales` are special cases of `retire` by taking $\tau = 0.5$ and $\gamma \to \infty$, respectively. Thus, both `huber` and `sales` can be implemented via Algorithm 1. For all methods, the sparsity inducing tuning parameter λ is selected via ten-fold cross-validation. Specifically, for methods `retire`, `huber`, and `sales`, we select the largest tuning parameter that yields a value of the asymmetric least squares loss that is less than the minimum of the asymmetric least squares loss plus one standard error. For `qr`, we use the default cross validation function in `R` package `rqPen` to select the largest tuning parameter that yields a value of its corresponding loss function that is the minimum of the quantile loss function.

Both `huber` and ℓ_1-penalized `retire` require tuning an additional robustness parameter γ. We propose to select γ using a heuristic tuning method that involves updating γ at the beginning of each iteration in Algorithm 1. Let $r_i^k = y_i - \boldsymbol{x}_i^{\mathrm{T}}\widehat{\boldsymbol{\beta}}^{k-1}, i = 1,\ldots,n$ be the residuals, where $\widehat{\boldsymbol{\beta}}^{k-1}$ is obtained from the $(k-1)$th iteration of Algorithm 1. Let $\widetilde{r}_i^k = (1-\tau)r_i^k\mathbb{1}_{r_i^k\le 0} + \tau r_i^k\mathbb{1}_{r_i^k>0}$ be the asymmetric residuals, and let $\widetilde{\mathbf{r}}^k = (\widetilde{r}_1^k,\ldots,\widetilde{r}_n^k)^{\mathrm{T}}$. We define $\mathrm{mad}(\widetilde{\mathbf{r}}^k) = \{\Phi^{-1}(0.75)\}^{-1}\mathrm{median}(|\widetilde{\mathbf{r}}^k - \mathrm{median}(\widetilde{\mathbf{r}}^k)|)$ as the median absolute deviation of the asymmetric residuals, adjusted by a factor $\Phi^{-1}(0.75)$. We start with setting $\gamma = \sqrt{n/\log(np)}$. At the kth iteration of Algorithm 1, we update the robustification parameter by

$$\gamma^k = \mathrm{mad}(\widetilde{\mathbf{r}}^k)\cdot\sqrt{\frac{n}{\log(np)}}. \tag{5.1}$$

Throughout our numerical studies, we have found that γ chosen using the above heuristic approach works well across different scenarios.

For all of the numerical studies, we generate the covariates $\boldsymbol{x}_i$ from a multivariate normal distribution $N(\mathbf{0}, \boldsymbol{\Sigma} = (\sigma_{jk})_{1\le j,k\le d})$ with $\sigma_{jk} = 0.5^{|j-k|}$. We then generate the response variable y_i from one of the following three models:

Table 1
Homoscedastic model (5.2) with Gaussian noise ($\epsilon \sim N(0, 2)$) and $t_{2.1}$ noise ($\epsilon \sim t_{2.1}$). Estimation error under ℓ_2-norm (and its standard deviation), true positive rate (TPR) and false positive rate (FPR), averaged over 100 repetitions, are reported.

Noise	Method	$n = 400, d = 200$			$n = 400, d = 500$		
		ℓ_2 error	TPR	FPR	ℓ_2 error	TPR	FPR
Gaussian	ℓ_1 retire	0.577 (0.009)	1.000 (0.000)	0.026 (0.002)	0.615 (0.009)	1.000 (0.000)	0.013 (0.001)
	IRW retire (SCAD)	**0.258** (0.006)	1.000 (0.000)	0.011 (0.001)	**0.251** (0.005)	1.000 (0.000)	0.005 (0.001)
	ℓ_1 huber	0.577 (0.009)	1.000 (0.000)	0.026 (0.002)	0.615 (0.009)	1.000 (0.000)	0.013 (0.001)
	ℓ_1 sales	0.577 (0.009)	1.000 (0.000)	0.026 (0.002)	0.614 (0.009)	1.000 (0.000)	0.013 (0.001)
	ℓ_1 qr	0.604 (0.010)	1.000 (0.000)	0.159 (0.008)	0.681 (0.010)	1.000 (0.000)	0.085 (0.005)
$t_{2.1}$	ℓ_1 retire	1.307 (0.039)	0.994 (0.003)	0.006 (0.001)	1.328 (0.040)	0.995 (0.002)	0.003 (0.000)
	IRW retire (SCAD)	**0.780** (0.052)	0.982 (0.005)	0.000 (0.000)	**0.788** (0.052)	0.983 (0.005)	0.000 (0.000)
	ℓ_1 huber	1.307 (0.039)	0.994 (0.003)	0.006 (0.001)	1.328 (0.040)	0.995 (0.002)	0.003 (0.000)
	ℓ_1 sales	1.424 (0.046)	0.990 (0.003)	0.012 (0.002)	1.460 (0.046)	0.987 (0.004)	0.005 (0.001)
	ℓ_1 qr	0.505 (0.010)	1.000 (0.000)	0.142 (0.009)	0.563 (0.010)	1.000 (0.000)	0.078 (0.004)

1. Homoscedastic model:

$$y_i = \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}^* + \epsilon_i, \tag{5.2}$$

2. Quantile heteroscedastic model:

$$y_i = \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}^* + (0.5|x_{id}| + 0.5)\{\epsilon_i - F_{\epsilon_i}^{-1}(\tau)\}, \tag{5.3}$$

3. Expectile heteroscedastic model:

$$y_i = \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}^* + (0.5|x_{id}| + 0.5)\{\epsilon_i - e_\tau(\epsilon_i)\}, \tag{5.4}$$

where ϵ_i is the random noise, $F_{\epsilon_i}^{-1}(\cdot)$ denotes the inverse cumulative distribution function of ϵ_i, and $e_\tau(\epsilon_i)$ denotes the inverse of the expectile function of ϵ_i. Note that under Gaussian and t-distributed noises, the two models (5.4) and (5.3) are the same for $\tau = 0.5$. We set the regression coefficient vector $\boldsymbol{\beta}^* = (\beta_1^*, \beta_2^*, \ldots, \beta_d^*)^{\mathrm{T}}$ as $\beta_1^* = 2$ (intercept), $\beta_j^* = \{1.8, 1.6, 1.4, 1.2, 1, -1, -1.2, -1.4, -1.6, -1.8\}$ for $j = 2, 4, \ldots, 20$, and zero otherwise. The random noise is generated from either a Gaussian distribution, $N(0, 2)$, or a t distribution with 2.1 degrees of freedom. For the heteroscedastic models, we consider two quantile/expectile levels $\tau = \{0.5, 0.8\}$. The results, averaged over 100 repetitions, are reported in Tables 1–4 for the moderate- ($n = 400,\ d = 200$) and high-dimensional ($n = 400,\ d = 500$) settings.

Table 1 contains results ($\tau = 0.5$) under the homoscedastic model with normally and t-distributed noise. For Gaussian noise, the four ℓ_1-penalized estimators have similar performance, and both the estimation error and FPR of IRW `retire` (with SCAD) are notably reduced. Under the $t_{2.1}$ noise, we see that `retire` gains considerable advantage over `sales` in both estimation and model selection accuracy, suggesting that the proposed estimator gains robustness without compromising statistical accuracy.

Tables 2 and 3 show results under the quantile heteroscedastic model with the Gaussian and $t_{2.1}$ noise, respectively. Two quantile levels $\tau = \{0.5, 0.8\}$ are considered. We see that `huber` and ℓ_1-penalized `retire` have the same performance when $\tau = 0.5$ since they are equivalent for the case when $\tau = 0.5$. Moreover, IRW `retire` has the lowest estimation error among all methods under the Gaussian noise. When $\tau = 0.8$, the performance of `huber` deteriorates since `huber` implicitly assumes $\tau = 0.5$ and there is a non-negligible bias when $\tau = 0.8$. Finally, from Table 4 under the expectile heteroscedastic model, we see that the proposed estimator has an even lower estimation error than that of the `qr`.

We want to point out that in general, under the $t_{2.1}$ noise, the quantile regression method `qr` has an advantage because the quantile loss is more robust to outliers than all of the other methods. While `qr` exhibits an advantage in terms of estimation error, it is not as computationally efficient as `retire`, which we will show in Section 5.2. In summary, the numerical studies confirm IRW `retire` as a robust alternative to its least squares counterpart `sales` and as a computationally efficient surrogate for the penalized quantile regression approach.

5.2. Timing comparison

In this section, we show using additional numerical studies that the proposed ℓ_1-penalized `retire` estimator has a significant computational advantage over the ℓ_1-penalized `qr`. We implement `retire` and `qr` using the `R` packages `adaHuber` (Pan and Zhou, 2022) and `rqPen`, respectively. For both methods, their corresponding sparsity regularization parameter is selected from a sequence of 50 λ-values via ten-fold cross-validation. The robustification parameter γ for `retire` is selected using the data adaptive procedure described in Section 5.1.

We generate the data from the homoscedastic model (5.2) with the same setup as in Section 5.1. Results, averaged over 100 independent data sets, for $n = d/2$ and $d = \{100, 200, 300, 400, 500\}$ are summarized in Fig. 1. The curves in panels (a) and (c) of Fig. 1 represent the estimation error (under ℓ_2 norm) as a function of the dimension d, and the curves in panels (b) and (d) of Fig. 1 represent the computational time (in seconds) as a function of the dimension d.

Table 2
Heteroscedastic model (5.3) with Gaussian noise ($\epsilon \sim N(0, 2)$) and quantile levels $\tau = \{0.5, 0.8\}$.

τ	Method	$n = 400, d = 200$			$n = 400, d = 500$		
		ℓ_2 error	TPR	FPR	ℓ_2 error	TPR	FPR
0.5	ℓ_1 retire	0.570 (0.009)	1.000 (0.000)	0.021 (0.002)	0.597 (0.009)	1.000 (0.000)	0.011 (0.001)
	IRW retire (SCAD)	**0.235** (0.006)	1.000 (0.000)	0.006 (0.001)	**0.230** (0.006)	1.000 (0.000)	0.005 (0.001)
	ℓ_1 huber	0.570 (0.009)	1.000 (0.000)	0.021 (0.002)	0.597 (0.009)	1.000 (0.000)	0.011 (0.001)
	ℓ_1 sales	0.575 (0.009)	1.000 (0.000)	0.021 (0.002)	0.599 (0.009)	1.000 (0.000)	0.011 (0.001)
	ℓ_1 qr	0.498 (0.008)	1.000 (0.000)	0.146 (0.007)	0.562 (0.008)	1.000 (0.000)	0.086 (0.004)
0.8	ℓ_1 retire	0.581 (0.008)	1.000 (0.000)	0.061 (0.005)	0.624 (0.011)	1.000 (0.000)	0.064 (0.005)
	IRW retire (SCAD)	**0.448** (0.012)	1.000 (0.000)	0.040 (0.004)	**0.588** (0.025)	1.000 (0.000)	0.047 (0.004)
	ℓ_1 huber	1.210 (0.008)	1.000 (0.000)	0.019 (0.002)	1.233 (0.008)	1.000 (0.000)	0.009 (0.001)
	ℓ_1 sales	0.636 (0.008)	1.000 (0.000)	0.051 (0.004)	0.661 (0.010)	1.000 (0.000)	0.047 (0.004)
	ℓ_1 qr	0.574 (0.009)	1.000 (0.000)	0.138 (0.007)	0.639 (0.011)	1.000 (0.000)	0.073 (0.004)

Table 3
Heteroscedastic model (5.3) with $t_{2.1}$ noise ($\epsilon \sim t_{2.1}$) and quantile levels $\tau = \{0.5, 0.8\}$.

τ	Method	$n = 400, d = 200$			$n = 400, d = 500$		
		ℓ_2 error	TPR	FPR	ℓ_2 error	TPR	FPR
0.5	ℓ_1 retire	1.222 (0.039)	0.995 (0.003)	0.006 (0.001)	1.275 (0.042)	0.995 (0.003)	0.003 (0.000)
	IRW retire (SCAD)	**0.663** (0.051)	0.988 (0.004)	0.000 (0.000)	**0.728** (0.055)	0.979 (0.005)	0.000 (0.000)
	ℓ_1 huber	1.222 (0.039)	0.995 (0.003)	0.006 (0.001)	1.275 (0.042)	0.995 (0.003)	0.003 (0.000)
	ℓ_1 sales	1.351 (0.051)	0.995 (0.003)	0.011 (0.003)	1.399 (0.047)	0.989 (0.003)	0.004 (0.000)
	ℓ_1 qr	0.420 (0.008)	1.000 (0.000)	0.150 (0.008)	0.473 (0.008)	1.000 (0.000)	0.075 (0.004)
0.8	ℓ_1 retire	1.052 (0.053)	0.991 (0.004)	0.015 (0.002)	1.065 (0.060)	0.987 (0.006)	0.009 (0.001)
	IRW retire (SCAD)	**0.498** (0.045)	0.983 (0.006)	0.003 (0.001)	**0.487** (0.056)	0.985 (0.006)	0.002 (0.000)
	ℓ_1 huber	1.556 (0.032)	0.996 (0.002)	0.007 (0.001)	1.593 (0.034)	0.995 (0.003)	0.003 (0.000)
	ℓ_1 sales	1.464 (0.102)	0.979 (0.006)	0.039 (0.004)	1.415 (0.060)	0.983 (0.005)	0.026 (0.002)
	ℓ_1 qr	0.630 (0.013)	1.000 (0.000)	0.132 (0.007)	0.683 (0.014)	1.000 (0.000)	0.070 (0.003)

Table 4
Heteroscedastic model (5.4) with Gaussian noise ($\epsilon \sim N(0, 2)$) and $t_{2.1}$ noise ($\epsilon \sim t_{2.1}$), under the τ-expectile $= 0.8$.

Noise	Method	$n = 400, d = 200$			$n = 400, d = 500$		
		ℓ_2 error	TPR	FPR	ℓ_2 error	TPR	FPR
Gaussian	ℓ_1 retire	0.534 (0.008)	1.000 (0.000)	0.063 (0.004)	0.557 (0.011)	1.000 (0.000)	0.066 (0.005)
	IRW retire (SCAD)	**0.353** (0.012)	1.000 (0.000)	0.042 (0.004)	**0.501** (0.025)	1.000 (0.000)	0.050 (0.005)
	ℓ_1 huber	0.898 (0.008)	1.000 (0.000)	0.020 (0.002)	0.924 (0.008)	1.000 (0.000)	0.010 (0.001)
	ℓ_1 sales	0.538 (0.009)	1.000 (0.000)	0.058 (0.004)	0.548 (0.010)	1.000 (0.000)	0.052 (0.004)
	ℓ_1 qr	0.671 (0.009)	1.000 (0.000)	0.147 (0.007)	0.716 (0.012)	1.000 (0.000)	0.074 (0.004)
$t_{2.1}$	ℓ_1 retire	1.055 (0.053)	0.991 (0.004)	0.015 (0.002)	1.068 (0.060)	0.987 (0.006)	0.009 (0.001)
	IRW retire (SCAD)	**0.487** (0.045)	0.983 (0.006)	0.004 (0.001)	**0.472** (0.057)	0.985 (0.006)	0.002 (0.000)
	ℓ_1 huber	1.535 (0.032)	0.996 (0.002)	0.007 (0.001)	1.573 (0.035)	0.995 (0.003)	0.003 (0.000)
	ℓ_1 sales	1.470 (0.102)	0.979 (0.006)	0.039 (0.004)	1.420 (0.060)	0.985 (0.005)	0.026 (0.002)
	ℓ_1 qr	0.638 (0.013)	1.000 (0.000)	0.135 (0.007)	0.688 (0.014)	1.000 (0.000)	0.069 (0.003)

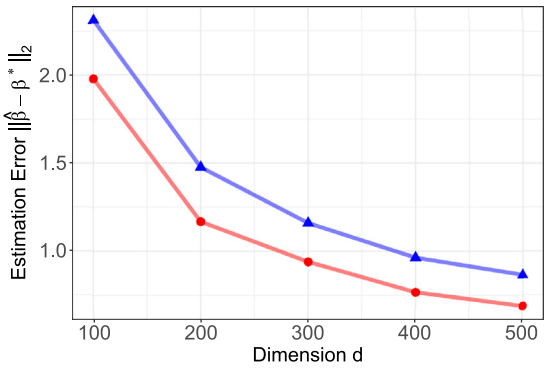
Under the Gaussian random noise, $\epsilon \sim N(0, 2)$, the ℓ_1-penalized `retire` has slightly lower estimation error than ℓ_1-penalized `qr`, and both estimation errors decrease as n and d grow. On the other hand, the ℓ_1-penalized `qr` performs better under the $t_{2.1}$ noise since the quantile loss is more robust to outliers than that of the Huber-type loss. Computationally, the ℓ_1-penalized `retire`, implemented via the `adaHuber` package, exhibits a significant improvement over the ℓ_1-penalized `qr`, implemented via the `rqPen` package, especially when d is large.

6. Data application

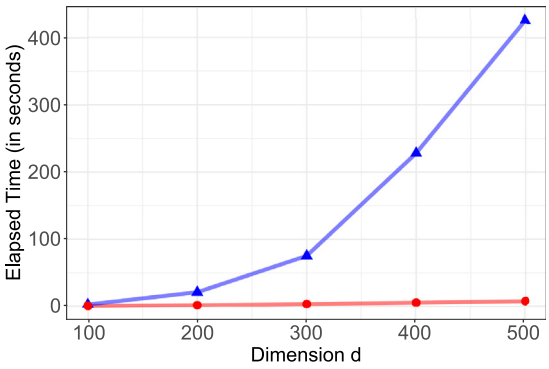
6.1. Job Training Partners Act data

We analyze the Job Training Partners Act (JTPA) data, previously studied in Abadie et al. (2002), using the retire estimator proposed in Section 3.2. The JTPA began funding federal training programs in 1983, and its largest component Title II supports training for the economically disadvantaged. Specifically, applicants who faced "barriers to employment", the most common of which were high-school dropout status and long periods of unemployment, were typically considered eligible for JTPA training. The services offered as a part of training included classroom training, basic education, on-the-job training, job search assistance, and probationary employment.

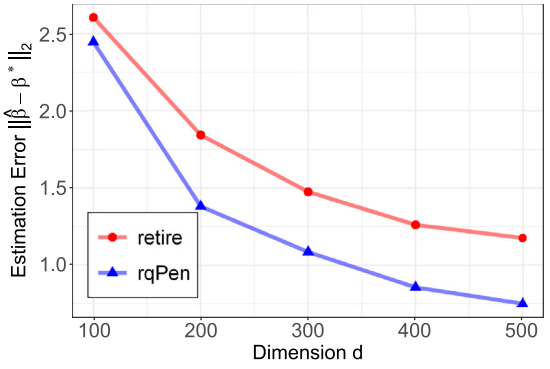
In this data set, applicants who applied for training evaluation between November 1987 and September 1989 were randomly selected to enroll for the JTPA training program. Of the 6102 adult women in the study, 4088 were offered



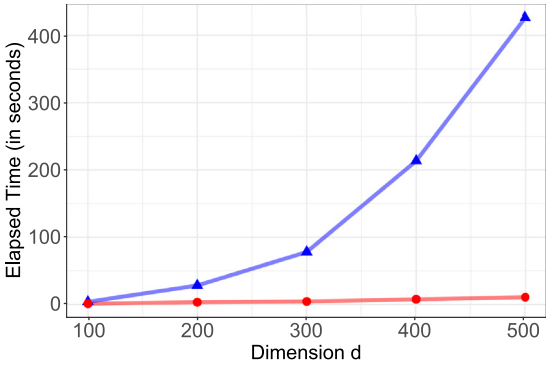
(a) Estimation error for model (5.2) with $N(0, 2)$ error.



(b) Elapsed time for model (5.2) with $N(0, 2)$ error.



(c) Estimation error for model (5.2) with $t_{2.1}$ error.



(d) Elapsed time for model (5.2) with $t_{2.1}$ error.

Fig. 1. Estimation error and elapsedtime (in seconds) under model (5.2) with $N(0, 2)$ and $t_{2.1}$ random noise and $\tau = 0.5$, averaged over 100 data sets for two different methods: (i) the ℓ_1-penalized `retire` implemented using the `R` package `adaHuber`; (ii) the ℓ_1-penalized `qr` implemented using the `R` package `rqPen`. The sample size n is set to equal $n = d/2$.

training and 2722 enrolled in the JTPA services, and of the 5102 adult men in the study, 3399 were offered training and 2136 enrolled in the services. The goal is to assess the effect of subsidized training program on earnings. Motivated by Abadie et al. (2002), we use the 30-month earnings data collected from the Title II JTPA training evaluation study as the response variable. Moreover, we adjust for the following covariates: individual's sex (male = 1, female = 0), whether or not the individual graduated high school or obtained a GED (yes = 1, no = 0), whether or not the individual worked less than 13 weeks in the 12 months preceding random assignment (yes = 1, no = 0), whether or not the individual is black (yes = 1, no = 0), whether or not the individual is Hispanic (yes = 1, no = 0), and marriage status (married = 1, not married = 0). We study the conditional distribution of 30-month earnings at different expectile levels $\tau = \{0.1, 0.5, 0.9\}$. Our proposed method involves robustification parameter γ, which we select using the tuning method described in Section 5.1.

The regression coefficients and their associated 95% confidence intervals are shown in Table 5. We find that covariates with positive regression coefficients for all quantile levels are enrollment for JTPA services, individual's sex, high school graduation or GED status, and marriage status. Black, hispanic, and worked less than 13 weeks in the past year had negative regression coefficients. The regression coefficients varied across the three different expectile levels we considered. The positive regression coefficients increase as the τ level increases and the negative regression coefficients decrease as the τ level increases. That is, for the lower expectile level of 30-month earnings, the covariates have a smaller in magnitude effect on the individual's earnings compared to the higher expectile level. The regression coefficient for enrollment in JTPA services was 1685.34, 2637.57, and 2714.57 at $\tau = \{0.1, 0.5, 0.9\}$, respectively. The τ-expectile of 30-month earnings for $\tau = \{0.1, 0.5, 0.9\}$ is 5068.02, 15815.29, and 32754.89 dollars, respectively. Compared to the expectile at the given τ, the effect of subsidized training was larger for lower expectile levels. Notably, if an individual is a male, conditional on other covariates, their 30-month earnings increase by 5005 dollars for $\tau = 0.5$ and increase by 10,311 dollars for $\tau = 0.9$. From the confidence intervals, we see that all variables are statistically significant except Hispanic.

Table 5
Regression coefficients (and their associated 95% confidence intervals) for the retire estimator.

Variable	$\tau = 0.1$	$\tau = 0.5$	$\tau = 0.9$
Enrolled in services	1685.34 (1401.03, 1969.65)	2637.57 (2079.74, 3195.40)	2714.57 (1766.01, 3663.13)
Male	1706.87 (1435.04, 1978.69)	5005.12 (4449.07, 5561.17)	10310.62 (9338.91, 11282.34)
High school or GED	1477.19 (1218.33, 1736.06)	3656.13 (3140.12, 4172.14)	5718.62 (4803.60, 6633.63)
Black	−580.04 (−917.86, −242.21)	−1567.03 (−2265.51, −868.56)	−2459.81 (−3686.14, −1233.48)
Hispanic	−130.72 (−588.11, 326.66)	−669.76 (−1626.83, 287.32)	−1495.33 (−3306.12, 315.46)
Married	1268.30 (933.66, 1602.94)	3343.63 (2668.95, 4018.30)	4518.43 (3376.92, 5659.93)
Worked less than 13 wks	−3677.98 (−3957.24, −3398.72)	−6879.14 (−7438.20, −6320.08)	−8206.16 (−9151.81, −7260.50)

6.2. Childhood malnutrition data

We apply the IRW ℓ_1-penalized retire estimator with SCAD-based weights to the childhood malnutrition data,. This data set is previously studied in Belloni et al. (2019) and Koenker (2011). The data are collected from the Demographic and Health Surveys (DHS) conducted regularly in more than 75 countries. Similar to Belloni et al. (2019), in this analysis, we will focus on data collected from India, with a total sample size of 37,623. The children studied are between the ages of zero and five.

The goal is to study the conditional distribution of children's height in India given the following covariates: the child's age, months of breastfeeding, mother's body mass index, mother's age, mother's education in years, partner's (father's) education in years, number of dead children in family, and multiple categorical variables including but are not limited to child's sex, child's birth order, mother's employment status, family's wealth (whether they are in poorest, poorer, middle, or richer bracket), electricity, television, refrigerator, bicycle, motorcycle, and car. Additionally, interactions between the following variables were considered: child's age, months of breastfeeding, child's sex, whether or not the child was a twin, mother's BMI, mother's age, mother's years of education, father's years of education, mother's employment status, and mother's residence. There are a total of 75 covariates: 30 individual variables and 45 two-way interactions.

We aim to study the conditional distribution of children's height at different expectile levels $\tau = \{0.1, 0.5, 0.9\}$. Our proposed method involves two tuning parameters γ and λ. The choice of robustification parameter γ was determined by theoretic guidance via a tuning method described in Section 5.1. The choice of sparsity tuning parameter λ is selected using a ten-fold cross validation where we select the largest tuning parameter that yields a value of the asymmetric least squares loss that is less than the minimum of the asymmetric least squares loss plus one standard error. For fair comparison, we apply the same sparsity tuning parameter across the three expectile levels. This is achieved by taking the maximum of the sparsity tuning parameters selected using a ten-fold cross-validation for the three different expectile levels. The selected tuning parameter takes value $\lambda = 0.035$.

The regression coefficients that are non-zero for at least one value of τ are shown in Table 6. There are a total of 38 non-zero coefficients. The regression coefficients for months of breastfeeding vary across the three different expectile levels we consider. At $\tau = 0.1$, the coefficient is 0.445, while at $\tau = \{0.5, 0.9\}$, the coefficients are 0.397 and 0.378 respectively. That is, for lower expectile level of child's height, months of breastfeeding plays a more important role to ensure that the child is not malnourished compared to higher expectile levels.

Other variables of interest are electricity, television, and motorcycle. For $\tau = 0.1$ and $\tau = 0.9$, the regression coefficients are zero, suggesting that access to these resources plays less of a role in a child's height at extreme expectile levels since access becomes a given for $\tau = 0.9$ and vice versa. For $\tau = 0.5$, the coefficients for electricity, television, and motorcycle are 0.647, 0.367, and 0.587 respectively, suggesting that these resources are important.

7. Discussion

In this study, we focused on robust estimation and inference for expectile regression in two scenarios: the low-dimensional setting where $d \ll n$, and the high-dimensional sparse setting where $s \ll n \ll d$. For the latter, we developed a robust penalized expectile regression method through iterative reweighted ℓ_1-penalization and established non-asymptotic high probability bounds. Performing statistical inference in high dimensions is much more challenging than in low dimensions due to the lack of a tractable limiting distribution of the penalized estimator when $d \gg n$. In recent years, there has been a rich development of debiased and de-sparsified procedures for penalized regression. These methods lead to estimators with asymptotically normal distributions, as demonstrated by Zhang and Zhang (2014), van de Geer et al. (2014), Javanmard and Montanari (2014) and Ning and Liu (2017), among others. However, a complete overview of these methods is beyond the scope of this study.

To conduct statistical inference for β_j^* $(2 \le j \le d)$, the jth coordinate of β^*, we consider the score function

$$S_n(\beta_j, \boldsymbol{\beta}_{-j}, \boldsymbol{v}) = \frac{1}{n}\sum_{i=1}^{n} L'_{\tau,\gamma}(y_i - x_{i,j}\beta_j - \boldsymbol{x}_{i,-j}^{\mathrm{T}}\boldsymbol{\beta}_{-j})(x_{i,j} - \boldsymbol{x}_{i,-j}^{\mathrm{T}}\boldsymbol{v}),$$

Table 6
Non-zero regression coefficients for IRW-ℓ_1-penalized `retire` with SCAD-based weights across three expectile levels $\tau = \{0.1, 0.5, 0.9\}$ for the childhood malnutrition data.

Variable	$\tau = 0.1$	$\tau = 0.5$	$\tau = 0.9$	Variable	$\tau = 0.1$	$\tau = 0.5$	$\tau = 0.9$
cage (child age)	0.650	0.686	0.728	breastfeeding*medu	−0.002	−0.001	−0.001
breastfeeding	0.445	0.397	0.378	breastfeeding*edupartner	0.001	0.000	0.000
cbirthorder1	0.012	0.703	0.000	breastfeeding*munemployed	0.000	0.000	−0.002
deadchildren	0.000	−0.345	0.000	breastfeeding*mresidence	0.000	0.000	−0.011
electricity	0.000	0.647	0.000	csex*mbmi	−0.011	−0.015	0.000
television	0.000	0.367	0.000	csex*mage	−0.039	−0.034	−0.040
motorcycle	0.000	0.587	0.000	ctwin*mage	−0.035	−0.032	0.000
cage*breastfeeding	−0.010	−0.008	−0.008	mbmi*mage	0.000	0.002	0.002
cage*csex	0.000	0.001	0.000	mbmi*medu	0.002	0.001	0.004
cage*ctwin	0.000	0.000	−0.028	mbmi*edupartner	0.000	−0.002	0.000
cage*mbmi	0.002	0.001	0.001	mbmi*munemployed	0.000	0.013	0.019
cage*mage	0.001	0.002	0.002	mage*medu	0.002	0.000	−0.001
cage*medu	0.005	0.003	0.002	mage*edupartner	0.002	0.003	0.001
cage*edupartner	0.001	0.001	0.001	mage*munemployed	0.006	0.006	0.004
cage*munemployed	−0.006	−0.007	−0.007	mage*mresidence	−0.010	0.000	0.000
cage*mresidence	0.000	0.003	0.000	medu*edupartner	0.002	0.003	0.005
breastfeeding*csex	−0.005	−0.007	−0.009	medu*munemployed	−0.009	−0.031	−0.050
breastfeeding*mbmi	0.002	0.002	0.002	edupartner*munemployed	−0.017	−0.030	−0.028
breastfeeding*mage	−0.002	−0.002	−0.003	edupartner*mresidence	0.000	−0.019	−0.015

where $\boldsymbol{\beta}_{-j}$ and $\boldsymbol{x}_{i,-j}$ are, respectively, the subvectors of $\boldsymbol{\beta} \in \mathbb{R}^d$ and $\boldsymbol{x}_i \in \mathbb{R}^d$ with the jth element removed. Let $\widehat{\boldsymbol{\beta}}^{\text{init}}$ be an initial penalized estimator of $\boldsymbol{\beta}^*$ as described in Section 3.3, and denote by $\widehat{\boldsymbol{v}} \in \operatorname{argmin}_{\boldsymbol{v}\in\mathbb{R}^{p-1}}\{(2n)^{-1}\sum_{i=1}^{n}(x_{i,j} - \boldsymbol{x}_{i,-j}^{\mathrm{T}}\boldsymbol{v})^2 + \lambda_v\|\boldsymbol{v}\|_1\}$ a Lasso-type estimator of $\boldsymbol{v}^* := \operatorname{argmin}_{\boldsymbol{v}\in\mathbb{R}^{d-1}}\mathbb{E}(x_{i,j} - \boldsymbol{x}_{i,-j}^{\mathrm{T}}\boldsymbol{v})^2$, the linear projection vector of the regressor of interest $x_{i,j}$ on the remaining covariates $\boldsymbol{x}_{i,-j}$. Assuming that $\boldsymbol{v}^*$ is s_v-sparse, we conjecture that with a properly chosen robustification parameter γ, the "oracle" score $\sqrt{n}\,S_n(\beta_j^*, \widehat{\boldsymbol{\beta}}_{-j}^{\text{init}}, \widehat{\boldsymbol{v}})$ converges in distribution to a centered normal distribution as n and d diverge, provided that $\max\{s, s_v\}\log(d) = o(\sqrt{n})$. Motivated by the classical one-step construction (Bickel, 1975), which aims to improve an initial estimator that is consistent but not efficient, we propose a debiased estimator

$$\widehat{\beta}_j = \widehat{\beta}_j^{\text{init}} - S_n(\widehat{\beta}_j^{\text{init}}, \widehat{\boldsymbol{\beta}}_{-j}^{\text{init}}, \widehat{\boldsymbol{v}}) \Big/ \partial_b S_n(b, \widehat{\boldsymbol{\beta}}_{-j}^{\text{init}}, \widehat{\boldsymbol{v}})|_{b=\widehat{\beta}_j^{\text{init}}}.$$

This estimator is conjectured to follow an asymptotically normal distribution. With a consistent estimate of its asymptotic variance, a Wald-type confidence interval can be constructed. However, a rigorous theoretical investigation of the debiased estimator and the accompanying asymptotic variance estimation problem requires a significant amount of future work, which we leave for future studies.

Acknowledgments

We thank the Editor, Associate Editor, and three anonymous reviewers for their insightful comments that help improve the previous version of the manuscript. K. M. Tan was supported by National Science Foundation, USA Grants DMS-2113356 and CAREER DMS-2238428. W.-X. Zhou acknowledges the support of the National Science Foundation, USA Grant DMS-2113409.

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.jeconom.2023.04.004.

References

Abadie, A., Angrist, J., Imbens, G., 2002. Instrumental variables estimates of the effect of subsidized training on the quantiles of trainee earnings. Econometrica 70 (1), 91–117.
Acerbie, C., Tasche, D., 2002. On the coherence of expected shortfall. J. Bank. Financ. 26, 1487–1503.
Aigner, D.J., Amemiya, T., Poirier, D.J., 1976. On the estimation of production frontiers: Maximum likelihood estimation of the parameters of a discontinuous density function. Internat. Econom. Rev. 17, 377–396.
Bellec, P.C., Lecué, G., Tsybakov, A.B., 2018. Slope meets Lasso: Improved oracle bounds and optimality. Ann. Statist. 46, 3603–3642.
Bellini, F., Bernardino, E.D., 2017. Risk management with expectiles. Eur. J. Finance 23 (6), 487–506.
Bellini, F., Bignozzi, V., 2015. On elicitable risk measures. Quant. Finance 15, 725–733.
Belloni, A., Chernozhukov, V., 2011. ℓ_1-Penalized quantile regression in high-dimensional sparse models. Ann. Statist. 39, 82–130.
Belloni, A., Chernozhukov, V., Chetverikov, D., Kato, K., 2015. Some new asymptotic theory for least squares series: Pointwise and uniform results. J. Econometrics 186, 345–366.

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.
Bühlmann, P., van de Geer, S., 2011. Statistics for High-Dimensional Data: Methods, Theory and Applications. Springer, Heidelberg.
Busetti, F., Caivano, M., Monache, D.D., 2021. Domestic and global determinants of inflation: Evidence from expectile regression. Oxf. Bull. Econ. Stat. 83, 982–1001.
Daouia, A., Girard, S., Stupfler, G., 2018. Estimation of tail risk based on extreme expectiles. J. R. Stat. Soc. Ser. B Stat. Methodol. 80, 263–292.
Efron, B., 1991. Regression percentiles using asymmetric squared error loss. Statist. Sinica 1, 93–125.
Fan, J., Li, R., 2001. Variable selection via nonconcave regularized likelihood and its oracle properties. J. Amer. Statist. Assoc. 96, 1348–1360.
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., Li, R., Zhang, C.-H., Zou, H., 2020. Statistical Foundations of Data Science. CRC Press, Boca Raton.
Fan, J., Liu, H., Sun, Q., Zhang, T., 2018. I-LAMM for sparse learning: Simultaneous control of algorithmic complexity and statistical error. Ann. Statist. 46, 814–841.
Fan, J., Xue, L., Zou, H., 2014. Strong oracle optimality of folded concave regularized estimation. Ann. Statist. 42, 819–849.
Gneiting, T., 2011. Making and evaluating point forecasts. J. Amer. Statist. Assoc. 106, 746–762.
Gu, Y., Zou, H., 2016. High-dimensional generalizations of asymmetric least squares regression and their applications. Ann. Statist. 44, 2661–2694.
Gu, Y., Zou, H., 2019. Aggregated expectile regression by exponential weighting. Statist. Sinica 29 (2), 671–692.
Huber, P.J., 1964. Robust estimation of a location parameter. Ann. Math. Stat. 35, 73–101.
Javanmard, A., Montanari, A., 2014. Confidence intervals and hypothesis testing for high-dimensional regression. J. Mach. Learn. Res. 15, 2869–2909.
Jone, M.C., 1994. Expectiles and M-quantiles are quantiles. Statist. Probab. Lett. 20, 149–153.
Koenker, R., 2005. Quantile Regression. Cambridge University Press, Cambridge.
Koenker, R., 2011. Additive models for quantile regression: Model selection and confidence bandaids. Braz. J. Probab. Stat. 25, 239–262.
Koenker, R., Bassett, G., 1978. Regression quantiles. Econometrica 46, 33–50.
Koenker, R., Chernozhukov, V., He, X., Peng, L., 2017. Handbook of Quantile Regression. CRC Press, Boca Raton, FL.
Kuan, C., Yeh, J., Hsu, Y., 2009. Assessing value at risk with CARE, the conditional autoregressive expectile models. J. Econ. 150, 261–270.
Lahiri, S.N., 2021. Necessary and sufficient conditions for variable selection consistency of the LASSO in high dimensions. Ann. Statist. 49, 820–844.
Lange, K., 1990. Convergence of image reconstruction algorithms with Gibbs smoothing. IEEE Trans. Med. Imaging 9, 439–446.
Loh, P., 2017. Statistical consistency and asymptotic normality for high-dimensional robust M-estimators. Ann. Statist. 45, 866–896.
Loh, P.-L., Wainwright, M.J., 2015. Regularized M-estimators with nonconvexity: Statistical and algorithmic theory for local optima. J. Mach. Learn. Res. 16, 559–616.
Ndaoud, M., 2019. Interplay of minimax estimation and minimax support recovery under sparsity. In: Proceedings of Machine Learning Research, Vol. 98. pp. 647–668.
Newey, W.K., Powell, J.L., 1987. Asymmetric least squares estimation and testing. Econometrica 55, 819–849.
Ning, Y., Liu, H., 2017. A general theory of hypothesis tests and confidence regions for sparse high dimensional models. Ann. Statist. 45, 158–195.
Nocedal, J., Wright, S.J., 1999. Numerical Optimization. Springer, New York.
Pan, X., Sun, Q., Zhou, W.-X., 2021. Iteratively reweighted ℓ_1-penalized robust regression. Electron. J. Stat. 15, 3287–3348.
Pan, X., Zhou, W.-X., 2022. Package `adaHuber`, version 1.1. Reference manual. https://cran.r-project.org/web/packages/adaHuber/adaHuber.pdf.
Portnoy, S., Koenker, R., 1997. The Gaussian hare and the Laplacian tortoise: Computability of squared-error versus absolute-error estimators. Statist. Sci. 12, 279–300.
Schnabel, S.K., Eilers, P.H.C., 2009. An analysis of life expectancy and economic production using expectile frontier zones. Demogr. Res. 21, 109–134.
Sherwood, B., Maidman, A., 2020. Package `rqPen`, version 2.2.2. Reference manual. https://cran.r-project.org/web/packages/rqPen/rqPen.pdf.
Su, W., Bogdan, M., Candés, E., 2017. False discoveries occur early on the Lasso path. Ann. Statist. 45, 2133–2150.
Sun, Q., Zhou, W.-X., Fan, J., 2020. Adaptive Huber regression. J. Amer. Statist. Assoc. 115, 254–265.
Tan, K.M., Sun, Q., Witten, D., 2023. Sparse reduced rank Huber regression in high dimensions. J. Amer. Statist. Assoc. http://dx.doi.org/10.1080/01621459.2022.2050243, (in press).
Taylor, J.W., 2008. Estimating value at risk and expected shortfall using expectiles. J. Financ. Econ. 6, 231–252.
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., Dezeure, R., 2014. On asymptotically optimal confidence regions and tests for high-dimensional models. Ann. Statist. 42, 1166–1202.
Wainwright, M.J., 2019. High-Dimensional Statistics: A Non-Asymptotic Viewpoint. Cambridge University Press, Cambridge.
Wang, L., He, X., 2022. Analysis of global and local optima of regularized quantile regression in high dimensions: A subgradient approach. Econom. Theory 1–45. http://dx.doi.org/10.1017/S0266466622000421.
Wang, L., Wu, Y., Li, R., 2012. Quantile regression for analyzing heterogeneity in ultra-high dimension. J. Amer. Statist. Assoc. 107, 214–222.
Xie, S., Zhou, Y., Wan, A.T.K., 2014. A varying-coefficient expectile model for estimating value at risk. J. Bus. Econ. Stat. 32, 576–592.
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 (3), 547–557.
Zhang, C.-H., 2010a. Nearly unbiased variable selection under minimax concave penalty. Ann. Statist. 38, 894–942.
Zhang, T., 2010b. Analysis of multi-stage convex relaxation for sparse regularization. J. Mach. Learn. Res. 11, 1081–1107.
Zhang, C.-H., Zhang, T., 2012. A general theory of concave regularization for high-dimensional sparse estimation problems. Statist. Sci. 27, 576–593.
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.
Ziegel, J.F., 2016. Coherence and elicitability. Math. Finance 26, 901–918.
Zou, H., Li, R., 2008. One-step sparse estimates in nonconcave penalized likelihood models. Ann. Statist. 36, 1509–1533.