Journal of the American Statistical Association

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/uasa20

Sparse Reduced Rank Huber Regression in High Dimensions

Kean Ming Tan, Qiang Sun & Daniela Witten

To cite this article: Kean Ming Tan, Qiang Sun & Daniela Witten (2022): Sparse Reduced Rank Huber Regression in High Dimensions, Journal of the American Statistical Association, DOI: 10.1080/01621459.2022.2050243

To link to this article: https://doi.org/10.1080/01621459.2022.2050243

JOURNAL OF THE AMERICAN STATISTICAL ASSOCIATION
2022, VOL. 00, NO. 0, 1–11: Theory and Methods
https://doi.org/10.1080/01621459.2022.2050243

Taylor & Francis
Taylor & Francis Group

Sparse Reduced Rank Huber Regression in High Dimensions

Kean Ming Tan[a], Qiang Sun[b], and Daniela Witten[c]

[a]Department of Statistics, University of Michigan, Ann Arbor, MI; [b]Department of Statistical Sciences, University of Toronto, Toronto, ON, Canada; [c]Departments of Statistics and Biostatistics, University of Washington, Seattle, WA

ABSTRACT

We propose a sparse reduced rank Huber regression for analyzing large and complex high-dimensional data with heavy-tailed random noise. The proposed method is based on a convex relaxation of a rank- and sparsity-constrained nonconvex optimization problem, which is then solved using a block coordinate descent and an alternating direction method of multipliers algorithm. We establish nonasymptotic estimation error bounds under both Frobenius and nuclear norms in the high-dimensional setting. This is a major contribution over existing results in reduced rank regression, which mainly focus on rank selection and prediction consistency. Our theoretical results quantify the tradeoff between heavy-tailedness of the random noise and statistical bias. For random noise with bounded $(1+\delta)$th moment with $\delta \in (0,1)$, the rate of convergence is a function of δ, and is slower than the sub-Gaussian-type deviation bounds; for random noise with bounded second moment, we obtain a rate of convergence as if sub-Gaussian noise were assumed. We illustrate the performance of the proposed method via extensive numerical studies and a data application. Supplementary materials for this article are available online.

ARTICLE HISTORY
Received April 2019
Revised January 2022

KEYWORDS
Convex relaxation; Huber loss; Low rank approximation; Sparsity

1. Introduction

Low rank matrix approximation methods have enjoyed successes in modeling and extracting information from large and complex data across various scientific disciplines. However, large-scale datasets are often accompanied by outliers due to possible measurement error, or because the population exhibits a leptokurtic distribution. As shown in She and Chen (2017), one single outlier can have a devastating effect on low rank matrix estimation. Consequently, nonrobust procedures for low rank matrix estimation could lead to inferior estimates and spurious scientific conclusions. For instance, in the context of financial data, it is evident that asset prices follow heavy-tailed distributions: if the heavy-tailedness is not accounted for in statistically modeling, then the recovery of common market behaviors and asset return forecasting may be jeopardized (Müller, Dacorogna, and Pictet 1998; Cont 2001).

In the context of reduced rank regression, She and Chen (2017) addressed this challenge by explicitly modeling the outliers with a sparse mean shift matrix of parameters. Similar ideas have been considered in the context of robust linear regression (She and Owen 2011) and robust clustering (Liu et al. 2012; Wang et al. 2016). In many statistical applications, the outliers themselves are not of interest. Rather than introducing additional parameters to model the outliers, it is more natural to develop robust statistical methods that are less sensitive to outliers. There is limited work along these lines in low rank matrix approximation problems. In fact, She and Chen (2017) pointed out that in the context of reduced rank regression, directly applying a robust loss function that down-weights the outliers, such as the Huber loss, may result in nontrivial computational and theoretical challenges due to the low rank constraint. So a natural question arises: can we develop a computationally efficient robust sparse low rank matrix approximation procedure that is less sensitive to outliers and yet has sound statistical guarantees?

In this article, we propose a novel method for fitting robust sparse reduced rank regression in the high-dimensional setting. We propose to minimize the Huber loss function subject to both sparsity and rank constraints. This leads to a nonconvex optimization problem, and is thus, computational intractable. To address this challenge, we consider a convex relaxation for both the sparsity and rank constraints, which can be solved efficiently. A similar convex relaxation has also been considered in Chen and Huang (2012) and Richard, Savalle, and Vayatis (2012) under the least squares loss. We note that Bunea, She, and Wegkamp (2012) proposed a group-lasso type penalty with a rank constraint to encourage the solution to be group-wise sparse and low rank under the least squares loss.

Most of the existing theoretical analysis of reduced rank regression focuses on rank selection consistency and prediction consistency (Bunea, She, and Wegkamp 2011; Mukherjee and Zhu 2011; Bunea, She, and Wegkamp 2012; Chen, Dong, and Chan 2013; Luo and Qi 2017; She 2017). Moreover, as shown by several authors, in order to achieve consistency, the number of covariates and the number of responses need to be much

CONTACT Qiang Sun *qiang.sun@utoronto.ca* Department of Statistical Sciences, University of Toronto, 700 University Avenue, Toronto, ON M5G 1X6, Canada.
Supplementary materials for this article are available online. Please go to *www.tandfonline.com/r/JASA*.

smaller than the sample size (Candes et al. 2011; She 2017). This motivates the use of sparsity penalty to accommodate possible high-dimensional covariates and responses. However, in the high-dimensional setting, nonasymptotic analysis of the estimation error is not well established, even in the context of reduced rank regression.[1] To bridge this gap in the literature, we provide nonasymptotic analysis of the estimation error under both Frobenius and nuclear norms for robust sparse reduced rank regression with the Huber loss.

The Huber loss has a *robustification parameter* that trades bias for robustness. In past work, the robustification parameter is usually fixed using the 95%-efficiency rule (among others, Huber 1964, 1973; Portnoy 1985; Mammen 1989; He and Shao 1996). Therefore, estimators obtained under Huber loss are typically biased. To achieve asymptotic unbiasedness and robustness simultaneously, within the context of robust linear regression, Sun, Zhou, and Fan (2018) showed that the robustification parameter has to adapt to the sample size, dimensionality, and moments of the random noise. Motivated by Sun, Zhou, and Fan (2018), we will establish theoretical results for the proposed method by allowing the robustification parameter to diverge.

The robustness of our proposed estimator is evidenced by its finite sample performance in the presence of heavy-tailed data, that is, data for which high-order moments are not finite. When the sampling distribution is heavy-tailed, there is a higher chance that some data are sampled far away from their mean. We refer to these outlying data as heavy-tailed outliers. Theoretically, we establish nonasymptotic results that quantify the tradeoff between heavy-tailedness of the random noise and statistical bias: for random noise with bounded $(1+\delta)$th moment, the rate of convergence, depending on δ, is slower than the sub-Gaussian-type deviation bounds; for random noise with bounded second moment, we recover results as if sub-Gaussian errors were assumed; and the transition between the two regimes is smooth.

Notation: For any vector $\mathbf{u} = (\mathrm{u}_1, \ldots, \mathrm{u}_p)^{\mathrm{T}} \in \mathbb{R}^p$ and $q \geq 1$, let $||\mathbf{u}||_q = \left(\sum_{j=1}^p |\mathrm{u}_j|^q\right)^{1/q}$ denote the ℓ_q norm. Let $||\mathbf{u}||_0 = \sum_{j=1}^p 1(\mathrm{u}_j \neq 0)$ denote the number of nonzero entries of $\mathbf{u}$, and let $||\mathbf{u}||_\infty = \max_{1\leq j\leq p} |\mathrm{u}_j|$. For any two vectors $\mathbf{u}, \mathbf{v} \in \mathbb{R}^p$, let $\langle \mathbf{u}, \mathbf{v}\rangle = \mathbf{u}^{\mathrm{T}}\mathbf{v}$. Moreover, for two sequences of real numbers $\{a_n\}_{n\geq 1}$ and $\{b_n\}_{n\geq 1}$, $a_n \lesssim b_n$ signifies that $a_n \leq Cb_n$ for some constant $C > 0$ that is independent of n, $a_n \gtrsim b_n$ if $b_n \lesssim a_n$, and $a_n \asymp b_n$ signifies that $a_n \lesssim b_n$ and $b_n \lesssim a_n$. If $\mathbf{A}$ is an $m \times n$ matrix, we use $||\mathbf{A}||_q$ to denote its order-q operator norm, defined by $||\mathbf{A}||_q = \max_{\mathbf{u}\in\mathbb{R}^n} ||\mathbf{A}\mathbf{u}||_q / ||\mathbf{u}||_q$. We define the (p, q)-norm of a $m \times n$ matrix $\mathbf{A}$ as the usual ℓ_q norm of the vector of row-wise ℓ_p norms of $\mathbf{A}$: $\|\mathbf{A}\|_{p,q} \equiv \left\|\left(||\mathbf{A}_{1\cdot}||_p, \ldots, ||\mathbf{A}_{m\cdot}||_p\right)\right\|_q$, where $\mathbf{A}_{j\cdot}$ is the jth row of $\mathbf{A}$. We use $||\mathbf{A}||_* = \sum_{k=1}^{\min\{m,n\}} \lambda_k$ to denote the nuclear norm of $\mathbf{A}$, where λ_k is the kth singular value of $\mathbf{A}$. Let $||\mathbf{A}||_{\mathrm{F}} = \sqrt{\sum_{i=1}^m \sum_{j=1}^n A_{ij}^2}$ be the Frobenius norm of $\mathbf{A}$. Finally, let $\mathrm{vec}(\mathbf{A})$ be the vectorization of the matrix $\mathbf{A}$, obtained by concatenating the columns of $\mathbf{A}$ into a vector.

[1] We note that the prediction error bound can be used to derive an estimation error bound under some further incoherence condition on the design matrix.

2. Robust Sparse Reduced Rank Regression

2.1. Formulation

Suppose we observe n independent samples of q-dimensional response variables and p-dimensional covariates. Let $\mathbf{Y} \in \mathbb{R}^{n\times q}$ be the observed response and let $\mathbf{X} \in \mathbb{R}^{n\times p}$ be the observed covariates. We consider the matrix regression model

$$\mathbf{Y} = \mathbf{X}\mathbf{A}^* + \mathbf{E}, \tag{1}$$

where $\mathbf{A}^* \in \mathbb{R}^{p\times q}$ is the underlying regression coefficient matrix and $\mathbf{E} \in \mathbb{R}^{n\times q}$ is an error matrix. Each row of $\mathbf{E}$ is an independent mean-zero and potentially heavy-tailed random noise vector.

Reduced rank regression seeks to characterize the relationships between $\mathbf{Y}$ and $\mathbf{X}$ in a parsimonious way by restricting the rank of $\mathbf{A}^*$ (Izenman 1975). An estimator of $\mathbf{A}^*$ can be obtained by solving the optimization problem

$$\underset{\mathbf{A}\in\mathbb{R}^{p\times q}}{\text{minimize}}\ \mathrm{tr}\left\{(\mathbf{Y}-\mathbf{X}\mathbf{A})^{\mathrm{T}}(\mathbf{Y}-\mathbf{X}\mathbf{A})\right\}, \quad \text{subject to } \mathrm{rank}(\mathbf{A}) \leq r, \tag{2}$$

where r is typically much smaller than $\min\{n, p, q\}$. Due to the rank constraint on $\mathbf{A}$, (2) is nonconvex: nonetheless, the global solution of (2) has a closed form solution (Izenman 1975).

It is well-known that squared error loss is sensitive to outliers or heavy-tailed random error (Huber 1973). To address this issue, it is natural to substitute the squared error loss with a loss function that is robust against outliers. We propose to estimate $\mathbf{A}^*$ under the Huber loss function, formally defined as follows.

Definition 1 (Huber Loss and Robustification Parameter). The Huber loss $\ell_\tau(\cdot)$ is defined as

$$\ell_\tau(z) = \begin{cases} \frac{1}{2}z^2, & \text{if } |z| \leq \tau, \\ \tau|z| - \frac{1}{2}\tau^2, & \text{if } |z| > \tau, \end{cases}$$

where $\tau > 0$ is referred to as the *robustification parameter* that trades bias for robustness.

The Huber loss function blends the squared error loss ($|z| \leq \tau$) and the absolute deviation loss ($|z| > \tau$), as determined by the robustification parameter τ. Compared to the squared error loss, large values of z are down-weighted under the Huber loss, thereby resulting in robustness. Generally, an estimator obtained from minimizing the Huber loss is biased. The robustification parameter τ quantifies the tradeoff between bias and robustness: a smaller value of τ introduces more bias but also encourages the estimator to be more robust to outliers. We will provide guidelines for selecting τ based on the sample size and the dimensions of $\mathbf{A}^*$ in later sections. Throughout the article, for $\mathbf{M} \in \mathbb{R}^{p\times q}$, we write $\ell_\tau(\mathbf{M}) = \sum_{i=1}^p \sum_{j=1}^q \ell_\tau(M_{ij})$ for notational convenience.

In the high-dimensional setting in which $n < p$ or $n < q$, it is theoretically challenging to estimate $\mathbf{A}^*$ accurately with only the low rank assumption. To address this challenge, Chen, Chan, and Stenseth (2012) and Chen and Huang (2012) proposed methods for simultaneous dimension reduction and variable selection. In particular, they decomposed $\mathbf{A}^*$ into the product of its singular vectors, and imposed sparsity-inducing penalty on

the left and right singular vectors. Thus, their proposed methods involve solving optimization problems with nonconvex objective.

Given that the goal is to estimate $\mathbf{A}^*$ rather than its singular vectors, we propose to estimate $\mathbf{A}^*$ directly. Under the Huber loss, a robust and sparse estimate of $\mathbf{A}^*$ can be obtained by solving the optimization problem:

$$\underset{\mathbf{A}\in\mathbb{R}^{p\times q}}{\text{minimize}} \left\{\frac{1}{n}\ell_\tau\left(\mathbf{Y}-\mathbf{XA}\right)\right\}, \quad \text{subject to } \text{rank}(\mathbf{A}) \le r \text{ and } \text{card}(\mathbf{A}) \le k, \tag{3}$$

where $\text{card}(\mathbf{A})$ is the number of nonzero elements in $\mathbf{A}$. Optimization problem (3) is nonconvex due to the rank and cardinality constraints on $\mathbf{A}$. We instead propose to estimate $\mathbf{A}^*$ by solving the following convex relaxation:

$$\underset{\mathbf{A}\in\mathbb{R}^{p\times q}}{\text{minimize}} \left\{\frac{1}{n}\ell_\tau\left(\mathbf{Y}-\mathbf{XA}\right)+\lambda\left(||\mathbf{A}||_*+\gamma||\mathbf{A}||_{1,1}\right)\right\}, \tag{4}$$

where λ and γ are nonnegative tuning parameters, $||\cdot||_*$ is the nuclear norm that encourages the solution to be low rank, and $||\cdot||_{1,1}$ is the entry-wise ℓ_1-norm that encourages the solution to be sparse. The nuclear norm and the $\ell_{1,1}$ norm constraints are the tightest convex relaxations of the rank and cardinality constraints, respectively (Recht, Fazel, and Parrilo 2010; Jojic, Saria, and Koller 2011). A similar convex relaxation has also been considered in Chen and Huang (2012) and Richard, Savalle, and Vayatis (2012) under the least squares loss.

We now discuss the close connection between our proposed method and that of She and Chen (2017). As shown by Lemma 2 of She and Chen (2017), (4) is equivalent to

$$\underset{\mathbf{A},\mathbf{C}}{\text{minimize}} \frac{1}{2}||\mathbf{Y}-\mathbf{XA}-\mathbf{C}||_F^2+\lambda(||\mathbf{A}||_*+\gamma||\mathbf{A}||_{1,1})+\tau||\mathbf{C}||_{1,1}, \tag{5}$$

where $\mathbf{C}\in\mathbb{R}^{p\times q}$ is a matrix of augmented parameters that models the outliers. Optimization problem (5) is motivated by the mean shift model $\mathbf{Y}=\mathbf{XA}^*+\mathbf{C}^*+\mathbf{E}$, where $\mathbf{A}^*\in\mathbb{R}^{p\times q}$ is a matrix of regression coefficients, $\mathbf{C}^*$ is a sparse matrix that models the outliers, and $\mathbf{E}$ is a matrix of sub-Gaussian random noise with some abuse of notation.

Let $\widehat{\mathbf{A}}$ be a solution obtained from solving (4) or (5). The primary advantage of studying the estimator $\widehat{\mathbf{A}}$ using the proposed framework in (4) is that it allows us to study the estimator $\widehat{\mathbf{A}}$ under a different perspective than that of the mean shift model (She and Chen 2017). We derive the theoretical results under the model $\mathbf{Y}=\mathbf{XA}^*+\mathbf{E}$, where $\mathbf{E}$ are the random noise with bounded $(1+\delta)$th moment condition. In other words, we consider the case when the outliers are modeled as heavy-tailed random noise rather than a mean shift parameter. Under the bounded moment condition, we show that the rate of convergence for $\widehat{\mathbf{A}}$ undergoes a phase transition as a function of δ in Theorem 1. Our results complement that of the theoretical results derived under the mean shift model in She and Chen (2017) and offer a different perspective to $\widehat{\mathbf{A}}$, illustrating the power of Huber loss under the setting when the random noise is heavy-tailed.

Remark 1. Since optimization problems (4) and (5) are equivalent in that they yield the exact same solution $\widehat{\mathbf{A}}$, results such as those of She and Chen (2017) can be obtained by analyzing the solution $\widehat{\mathbf{A}}$ under (5). We leave this for future work.

2.2. Algorithm

In this section, we provide an efficient algorithm to obtain the proposed estimator $\widehat{\mathbf{A}}$. Due to the equivalence between (4) and (5), we start with deriving a block coordinate descent type algorithm for solving (5). One advantage of solving (5) is that the estimated augmented matrix $\widehat{\mathbf{C}}$ for modeling outliers is outputted as a by-product, which may be of interest to practitioners. We note that optimization problem (4) can also be solved directly using an alternating direction method of multipliers (ADMM) algorithm, which we provide in Section A of the online supplementary materials for completeness.

A block coordinate descent algorithm for solving (5) involves updating the matrices $\mathbf{A}$ and $\mathbf{C}$ iteratively while holding the other fixed until convergence. Given a fixed $\mathbf{A}$, the update for $\mathbf{C}$ can be obtained by solving

$$\underset{\mathbf{C}}{\text{minimize}} \frac{1}{2}||\mathbf{Y}-\mathbf{XA}-\mathbf{C}||_F^2+\tau||\mathbf{C}||_{1,1},$$

which yields a closed-form solution of $\widehat{\mathbf{C}}=S(\mathbf{Y}-\mathbf{XA},\tau)$, where S denote the soft-thresholding operator, applied element-wise to a matrix, that is, $S(A_{ij},b)=\text{sign}(A_{ij})\max(|A_{ij}|-b,0)$. Given $\mathbf{C}$, the update for $\mathbf{A}$ can then be obtained by solving

$$\underset{\mathbf{A}}{\text{minimize}} \frac{1}{2}||\mathbf{Y}-\mathbf{XA}-\mathbf{C}||_F^2+\lambda(||\mathbf{A}||_*+\gamma||\mathbf{A}||_{1,1}). \tag{6}$$

Optimization problem (6) does not admit a closed-form solution. To this end, we derive an ADMM algorithm for solving (6) (Eckstein and Bertsekas 1992; Boyd et al. 2010).

Specifically, we note that (6) is equivalent to

$$\begin{aligned}&\underset{\mathbf{A},\mathbf{Z},\mathbf{W}\in\mathbb{R}^{p\times q}}{\text{minimize}} \left\{\tfrac{1}{2}||\mathbf{Y}-\mathbf{XA}-\mathbf{C}||_F^2+\lambda\left(||\mathbf{W}||_*+\gamma||\mathbf{Z}||_{1,1}\right)\right\},\\ &\text{subject to} \qquad \mathbf{W}=\mathbf{A} \quad \text{and} \quad \mathbf{Z}=\mathbf{A}.\end{aligned} \tag{7}$$

The scaled augmented Lagrangian of (7) takes the form

$$\begin{aligned}\mathcal{L}_\rho(\mathbf{A},\mathbf{Z},\mathbf{W},\mathbf{M},\mathbf{N}) = \frac{1}{2}||\mathbf{Y}-\mathbf{XA}-\mathbf{C}||_F^2+\lambda\left(||\mathbf{W}||_*\right.\\ \left.+\gamma||\mathbf{Z}||_{1,1}\right)+\frac{\rho}{2}||\mathbf{W}-\mathbf{A}+\mathbf{N}||_{\text{F}}^2\\ +\frac{\rho}{2}||\mathbf{Z}-\mathbf{A}+\mathbf{M}||_{\text{F}}^2,\end{aligned}$$

where $\mathbf{A},\mathbf{Z},\mathbf{W}$ are the primal variables, and $\mathbf{N}$ and $\mathbf{M}$ are the dual variables. Note that the ADMM algorithm is an iterative algorithm. At the kth iteration, the ADMM algorithm requires the following updates:

1. $\mathbf{A}_{k+1}\leftarrow\underset{\mathbf{A}}{\text{argmin}}\,\mathcal{L}_\rho(\mathbf{A},\mathbf{Z}_k,\mathbf{W}_k,\mathbf{M}_k,\mathbf{N}_k)$.
2. $\mathbf{Z}_{k+1}\leftarrow\underset{\mathbf{Z}}{\text{argmin}}\,\mathcal{L}_\rho(\mathbf{A}_{k+1},\mathbf{Z},\mathbf{W}_k,\mathbf{M}_k,\mathbf{N}_k)$.
3. $\mathbf{W}_{k+1}\leftarrow\underset{\mathbf{W}}{\text{argmin}}\,\mathcal{L}_\rho(\mathbf{A}_{k+1},\mathbf{Z}_{k+1},\mathbf{W},\mathbf{M}_k,\mathbf{N}_k)$.
4. $\mathbf{N}_{k+1}\leftarrow\mathbf{N}_k+\rho(\mathbf{A}_{k+1}-\mathbf{W}_{k+1})$.
5. $\mathbf{M}_{k+1}\leftarrow\mathbf{M}_k+\rho(\mathbf{A}_{k+1}-\mathbf{Z}_{k+1})$.

The derivation for the closed-form updates are standard for least squares loss and are omitted. The details of the proposed algorithm are summarized in Algorithm 1.

Note that the term $(\mathbf{X}^{\text{T}}\mathbf{X}+2\rho\mathbf{I})^{-1}$ can be computed outside of the loop. Thus, the computational bottleneck in each iteration of Algorithm 1 is the singular value decomposition of a $p\times q$ matrix with computational complexity $\mathcal{O}\{p^2q+q^2p+\min(q^3,p^3)\}$.

Algorithm 1 A Block Coordinate Descent Algorithm for Solving (5).

1. **Initialize** the parameters:
   (a) augmented variable $\mathbf{C}$ to the zero matrix.
   (b) primal variables $\mathbf{A}$, $\mathbf{Z}$, and $\mathbf{W}$ to the zero matrix.
   (c) dual variables $\mathbf{B}_Z$ and $\mathbf{B}_W$ to the zero matrix.
   (d) constants $\rho > 0$, $\tau > 0$, $\lambda > 0$, $\gamma > 0$, and $\epsilon > 0$.
2. **Iterate** until the stopping criterion $||\mathbf{A}^{t+1} - \mathbf{A}^t||_{\mathrm{F}}^2 / ||\mathbf{A}^t||_{\mathrm{F}}^2 \leq \epsilon$ is met, where $\mathbf{A}^t$ is the value of $\mathbf{A}$ obtained at the tth iteration of the block coordinate descent algorithm.
   (a) $\mathbf{C} = S(\mathbf{Y} - \mathbf{XA}, \tau)$, where S denote the soft-thresholding operator, applied element-wise to a matrix: $S(A_{ij}, b) = \mathrm{sign}(A_{ij}) \max(|A_{ij}| - b, 0)$.
   (b) **Iterate** the following until the stopping criterion $||\mathbf{A}_{k+1}^t - \mathbf{A}_k^t||_{\mathrm{F}}^2 / ||\mathbf{A}_k^t||_{\mathrm{F}}^2$ is met, where $\mathbf{A}_k^t$ is the value of $\mathbf{A}$ obtained at the kth iteration of the following ADMM algorithm:
      i. $\mathbf{A} = (\mathbf{X}^{\mathrm{T}}\mathbf{X} + 2\rho\mathbf{I})^{-1}\{\rho(\mathbf{N} + \mathbf{W} + \mathbf{Z} + \mathbf{M}) + \mathbf{X}^{\mathrm{T}}(\mathbf{Y} - \mathbf{C})\}$.
      ii. $\mathbf{Z} = S(\mathbf{A} - \mathbf{M}, \lambda\gamma/\rho)$.
      iii. $\mathbf{W} = \sum_j \max\left(\omega_j - \lambda/\rho, 0\right) \mathbf{a}_j \mathbf{b}_j^{\mathrm{T}}$, where $\sum_j \omega_j \mathbf{a}_j \mathbf{b}_j^{\mathrm{T}}$ is the singular value decomposition of $\mathbf{A} - \mathbf{N}$.
      iv. $\mathbf{M} = \mathbf{M} + \mathbf{Z} - \mathbf{A}$.
      v. $\mathbf{N} = \mathbf{N} + \mathbf{W} - \mathbf{A}$.

3. Statistical Theory

We study the theoretical properties of $\widehat{\mathbf{A}}$ obtained from solving (4). Let $\mathbb{V}_{p,q} = \{\mathbf{U} \in \mathbb{R}^{p\times q} : \mathbf{U}^{\mathrm{T}}\mathbf{U} = \mathbf{I}_q\}$ be the Stiefel manifold of $p \times q$ orthonormal matrices. Throughout the theoretical analysis, we assume that $\mathbf{A}^*$ can be decomposed as

$$\mathbf{A}^* = \mathbf{U}^* \Lambda^* (\mathbf{V}^*)^{\mathrm{T}} = \sum_{k=1}^{r} \lambda_k^* \mathbf{u}_k^* (\mathbf{v}_k^*)^{\mathrm{T}}, \tag{8}$$

where $\mathbf{U}^* \in \mathbb{V}_{p,r}$, $\mathbf{V}^* \in \mathbb{V}_{q,r}$, $\max_k ||\mathbf{u}_k^*||_0 \leq s_u$, and $\max_k ||\mathbf{v}_k^*||_0 \leq s_v$ with $s_u, s_v \ll n$, $r \ll n$, and $rs_u s_v \ll n$. Consequently, $\mathbf{A}^*$ is sparse and low rank. Let $\mathcal{S} = \mathrm{supp}(\mathbf{A}^*)$ be the support set of $\mathbf{A}^*$ with cardinality $|\mathcal{S}| = s$, that, $\mathcal{S}$ contains indices for the nonzero elements in $\mathbf{A}^*$. Note that $s \leq rs_u s_v$.

For simplicity, we consider the case of fixed design matrix $\mathbf{X}$ and assume that the covariates are standardized such that $\max_{i,j} |X_{ij}| = 1$. To characterize the heavy-tailed random noise, we impose a bounded moment condition on the random noise.

Condition 1 (Bounded Moment Condition). Let $\delta > 0$. Assume that each entry of the random error matrix $\mathbf{E}$ in (1) has bounded $(1 + \delta)$th moment, and let

$$v_\delta \equiv \max_{i,j} \mathbb{E}\left(|E_{ij}|^{1+\delta}\right) < \infty.$$

Condition 1 is a relaxation of the commonly used sub-Gaussian assumption to accommodate heavy-tailed random noise. For instance, the t-distribution with degrees of freedom larger than one can be accommodated by the bounded moment condition. This condition has also been used in the context of high-dimensional Huber linear regression (Sun, Zhou, and Fan 2018). Note that Condition 1 allows for heterogeneous random noise as long as the random noise has at least bounded $(1+\delta)$th moment.

Let $\mathbf{H}_\tau(\mathbf{A})$ be the Hessian matrix of the Huber loss function $\ell_\tau(\mathbf{Y} - \mathbf{XA})/n$ in (3). In addition to the random noise, the Hessian matrix is a function of the parameter $\mathbf{A}$, and $\mathbf{H}_\tau(\mathbf{A})$ may equal zero for some $\mathbf{A}$, because the Huber loss is linear at the tails. To avoid singularity of $\mathbf{H}_\tau(\mathbf{A})$, we will study the Hessian matrix in a local neighborhood of $\mathbf{A}^*$. To this end, we define and impose conditions on the localized restricted eigenvalues of $\mathbf{H}_\tau(\mathbf{A})$.

Definition 2 (Localized Restricted Eigenvalues). The minimum and maximum localized restricted eigenvalues for $\mathbf{H}_\tau(\mathbf{A})$ are defined as

$$\kappa_-(\mathbf{H}_\tau(\mathbf{A}), \xi, \eta) = \inf_{\mathbf{U},\mathbf{A}} \left\{ \frac{\mathrm{vec}(\mathbf{U})^{\mathrm{T}} \mathbf{H}_\tau(\mathbf{A}) \mathrm{vec}(\mathbf{U})}{||\mathbf{U}||_{\mathrm{F}}^2} : (\mathbf{A}, \mathbf{U}) \in \mathcal{C}(m, \xi, \eta) \right\},$$

$$\kappa_+(\mathbf{H}_\tau(\mathbf{A}), \xi, \eta) = \sup_{\mathbf{U},\mathbf{A}} \left\{ \frac{\mathrm{vec}(\mathbf{U})^{\mathrm{T}} \mathbf{H}_\tau(\mathbf{A}) \mathrm{vec}(\mathbf{U})}{||\mathbf{U}||_{\mathrm{F}}^2} : (\mathbf{A}, \mathbf{U}) \in \mathcal{C}(m, \xi, \eta) \right\},$$

where

$$\mathcal{C}(m, \xi, \eta) = \{(\mathbf{A}, \mathbf{U}) \in \mathbb{R}^{p\times q} \times \mathbb{R}^{p\times q} : \mathbf{U} \neq \mathbf{0}, \mathcal{S} \subseteq J, |J| \leq m, ||\mathbf{U}_{\mathcal{S}^c}||_{1,1} \leq \xi ||\mathbf{U}_{\mathcal{S}}||_{1,1}, ||\mathbf{A} - \mathbf{A}^*||_{1,1} \leq \eta\}.$$

Condition 2. There exist constants $0 < \kappa_{\mathrm{lower}} \leq \kappa_{\mathrm{upper}} < \infty$ such that the localized restricted eigenvalues of $\mathbf{H}_\tau$ are lower- and upper-bounded by

$$\kappa_{\mathrm{lower}}/2 \leq \kappa_-(\mathbf{H}_\tau(\mathbf{A}), \xi, \eta) \leq \kappa_+(\mathbf{H}_\tau(\mathbf{A}), \xi, \eta) \leq \kappa_{\mathrm{upper}}.$$

A similar type of localized condition was proposed in Fan et al. (2018) for general loss functions and in Sun, Zhou, and Fan (2018) for the analysis of robust linear regression in high dimensions. In what follows, we justify Condition 2 by showing that it is implied by the restricted eigenvalue condition on the empirical Gram matrix $\mathbf{S} = \mathbf{X}^{\mathrm{T}}\mathbf{X}/n$. To this end, we define the restricted eigenvalues of a matrix and then place a condition on the restricted eigenvalues of $\mathbf{S}$.

Definition 3 (Restricted Eigenvalues of a Matrix). Given $\xi > 1$, the minimum and maximum restricted eigenvalues of $\mathbf{S}$ are defined as

$$\rho_-(\mathbf{S}, \xi, m) = \inf_{\mathbf{U}} \left\{ \frac{\mathrm{tr}(\mathbf{U}^{\mathrm{T}}\mathbf{S}\mathbf{U})}{||\mathbf{U}||_{1,2}^2} : \mathbf{U} \in \mathbb{R}^{p\times q}, \mathbf{U} \neq \mathbf{0}, \mathcal{S} \subseteq J, |J| \leq m, ||\mathbf{U}_{J^c}||_{1,1} \leq \xi ||\mathbf{U}_J||_{1,1} \right\},$$

$$\rho_+(\mathbf{S}, \xi, m) = \sup_{\mathbf{U}} \left\{ \frac{\mathrm{tr}(\mathbf{U}^{\mathrm{T}}\mathbf{S}\mathbf{U})}{||\mathbf{U}||_{1,2}^2} : \mathbf{U} \in \mathbb{R}^{p\times q}, \mathbf{U} \neq \mathbf{0}, \mathcal{S} \subseteq J, |J| \leq m, ||\mathbf{U}_{J^c}||_{1,1} \leq \xi ||\mathbf{U}_J||_{1,1} \right\},$$

respectively.

Condition 3. There exist constants $0 < \kappa_{\mathrm{lower}} \leq \kappa_{\mathrm{upper}} < \infty$ such that the restricted eigenvalues of $\mathbf{S}$ are lower- and upper-bounded by

$$\kappa_{\mathrm{lower}} \leq \rho_-(\mathbf{S}, \xi, m) \leq \rho_+(\mathbf{S}, \xi, m) \leq \kappa_{\mathrm{upper}}.$$

Condition 3 is a variant of the restricted eigenvalue condition that is commonly used in high-dimensional nonasymptotic analysis. It can be shown that Condition 3 holds with high probability if each row of $\mathbf{X}$ is a sub-Gaussian random vector.

Under Condition 3, we now show that the localized restricted eigenvalues for the Hessian matrix are bounded with high probability under conditions on the robustification parameter τ and the sample size n. That is, we prove that the localized restricted eigenvalues condition in Condition 2 holds with high probability under Condition 3. The result is summarized in the following lemma.

Lemma 1. Consider $\mathbf{A} \in \mathcal{C}(m, \xi, \eta)$ where $\mathcal{C}(m, \xi, \eta)$ is as defined in Definition 2. Let $\tau \geq \max(8\eta, C \cdot (m v_\delta)^{1/(1+\delta)})$ and let $n > C' \cdot m^2 \log(pq)$ for sufficiently large constants $C, C' > 0$. Under Conditions 1 and 3, there exists constants κ_{lower} and κ_{upper} such that the localized restricted eigenvalues of $\mathbf{H}_\tau(\mathbf{A})$ satisfy

$$0 < \kappa_{\text{lower}}/2 \leq \kappa_-(\mathbf{H}_\tau(\mathbf{A}), \xi, \eta) \leq \kappa_+(\mathbf{H}_\tau(\mathbf{A}), \xi, \eta) \leq \kappa_{\text{upper}} < \infty$$

with probability at least $1 - (pq)^{-1}$.

Lemma 1 shows that Condition 2 holds with high probability, as long as Condition 3 on the empirical Gram matrix $\mathbf{S}$ holds. Note that the constants κ_{lower} and κ_{upper} also appear in Condition 3.

We now present our main results on the estimation error of $\widehat{\mathbf{A}}$ under the Frobenius norm and nuclear norm in the following theorem. For simplicity, we will present our main results conditioned on the event that Conditions 1–2 hold.

Theorem 1. Let $\widehat{\mathbf{A}}$ be a solution to (4) with tuning parameters

$$\tau \gtrsim \left(\frac{n v_\delta}{\log(pq)}\right)^{1/\min\{(1+\delta),2\}},$$

$$\lambda \gtrsim v_\delta^{1/\min(1+\delta,2)} \left(\frac{\log(pq)}{n}\right)^{\min\{\delta/(1+\delta),1/2\}}$$

and $\gamma > 2.5$. Suppose that Conditions 1–2 hold with $\xi = (2\gamma + 5)/(2\gamma - 5)$, $\kappa_{\text{lower}} > 0$ and $\eta \gtrsim \kappa_{\text{lower}}^{-1} \lambda r(s_u + s_v)$. Assume that $n > C(r s_u s_v)^2 \log(pq)$ for some sufficiently large universal constant $C > 0$. Then, with probability at least $1 - (pq)^{-1}$, we have

$$\|\widehat{\mathbf{A}} - \mathbf{A}^*\|_{\text{F}} \lesssim \kappa_{\text{lower}}^{-1} v_\delta^{1/\min\{1+\delta,2\}} \sqrt{r s_u s_v} \left\{\frac{\log(pq)}{n}\right\}^{\min\{\delta/(1+\delta),1/2\}},$$

$$\|\widehat{\mathbf{A}} - \mathbf{A}^*\|_{*} \lesssim \kappa_{\text{lower}}^{-1} v_\delta^{1/\min\{1+\delta,2\}} r s_u s_v \left\{\frac{\log(pq)}{n}\right\}^{\min\{\delta/(1+\delta),1/2\}}.$$

Theorem 1 establishes the nonasymptotic convergence rates of our proposed estimator under both Frobenius and nuclear norms in the high-dimensional setting. To the best of our knowledge, we are the first to establish such results on the estimation error for robust sparse reduced rank regression under heavy-tailed random noise. By contrast, most of the existing work on reduced rank regression focuses on rank selection consistency and prediction consistency (Bunea, She, and Wegkamp 2011, 2012). Moreover, in order to achieve consistency, the number of covariates and the number of responses need to be much smaller than the sample size (Candes et al. 2011; She 2017). This motivates the use of the sparsity inducing penalty to reduced rank regression to accommodate high-dimensional covariates and responses. When the random noise has second or higher moments, that is, $\delta \geq 1$, our proposed estimator achieves a parameteric rate of convergence as if sub-Gaussian random noise were assumed. It achieves a slower rate of convergence only when the random noise is extremely heavy-tailed, that is, $0 < \delta < 1$.

Remark 2. As pointed out by the referees and the associate editor, several authors have considered sparse reduced rank regression using row-wise sparsity inducing penalty with low rank constraint or the nuclear norm penalty (Chen and Huang 2012; Bunea, She, and Wegkamp 2012; She 2017). We want to emphasize that our proposed method considers response selection while row-wise sparsity penalty does not. Consequently, we can relax the dependence of our result on q, by only involving the logarithmic of q. Because of this, it is difficult to compare the convergence rate between the two approaches directly.

Remark 3. The restricted eigenvalue type conditions are needed for establishing the estimation error of $\widehat{\mathbf{A}}$. From Theorem 1, a prediction error bound for $||\mathbf{X}(\widehat{\mathbf{A}} - \mathbf{A}^*)||_{\text{F}}^2$ can be obtained directly. On the other hand, if the risk excess error is of interest, then the restricted eigenvalue type conditions can be removed.

Intuitively, one might expect the optimal rate of convergence under the Frobenius norm to have the form $\|\widehat{\mathbf{A}} - \mathbf{A}^*\|_{\text{F}} \lesssim \sqrt{r(s_u+s_v)}\,\{\log(pq)/n\}^{\min\{\delta/(1+\delta),1/2\}}$, since there are a total of roughly $r(s_u + s_v)$ nonzero parameters to be estimated in $\mathbf{A}^*$ as defined in (8). To validate the aforementioned intuition, we provide the minimax lower bound for sparse and low rank estimation under the Gaussian random noise in Theorem 2. The minimax lower bound under random noise with $(1 + \delta)$th bounded moment remains an open problem and we leave it for future work. We consider the following family of all rank r sparse matrices:

$$\begin{aligned}\mathcal{F} = \Big\{ & \mathbf{A} = \mathbf{U}\Lambda\mathbf{V}^T \in \mathbb{R}^{p\times q} : \operatorname{rank}(\mathbf{A}) \leq r, \lambda_1(\mathbf{A}) \\ & \geq \cdots \geq \lambda_r(\mathbf{A}) \geq 0, \lambda_k(\mathbf{A}) = 0, r < k \leq \max(p,q), \\ & \max_{1\leq k\leq r} ||\mathbf{u}_k||_0 \leq s_u, \max_{1\leq k\leq r} ||\mathbf{v}_k||_0 \leq s_v \Big\},\end{aligned} \tag{9}$$

where $\lambda_k(\mathbf{A})$ be the kth largest singular value of $\mathbf{A}$.

Theorem 2. Assume that each entry of the random noise E_{ij} are independent and identically distributed from a standard normal distribution. Suppose that $2(r-1)\max(s_u, s_v) \leq \min(p, q)$. Then, we have

$$\inf_{\widehat{\mathbf{A}}} \sup_{\mathbf{A}^*\in\mathcal{F}} \mathbb{E}||\widehat{\mathbf{A}} - \mathbf{A}^*||_{\text{F}}^2 \geq C_1 \left(\frac{r(s_u+s_v)}{n} + \frac{r s_u}{n}\log\frac{ep}{s_u} + \frac{r s_v}{n}\log\frac{eq}{s_v}\right),$$

$$\inf_{\widehat{\mathbf{A}}} \sup_{\mathbf{A}^*\in\mathcal{F}} \mathbb{E}||\widehat{\mathbf{A}} - \mathbf{A}^*||_{*}^2 \geq C_1 \left(\frac{r^2(s_u+s_v)}{n\log r} + \frac{r^2 s_u}{n}\log\frac{ep}{s_u} + \frac{r^2 s_v}{n}\log\frac{eq}{s_v}\right),$$

where C_1 is a constant depending only on κ_{upper}. Moreover, under the scaling condition $\{\max(s_u, s_v)\}^3 pq \lesssim \min(p^3, q^3)$, we

have

$$\inf_{\widehat{\mathbf{A}}} \sup_{\mathbf{A}^* \in \mathcal{F}} \mathbb{E}||\widehat{\mathbf{A}} - \mathbf{A}^*||_{\mathrm{F}}^2 \geq C_2 \left(\tfrac{r(s_u+s_v)}{n} + \tfrac{r(s_u+s_v)}{n} \log(pq) \right),$$
$$\inf_{\widehat{\mathbf{A}}} \sup_{\mathbf{A}^* \in \mathcal{F}} \mathbb{E}||\widehat{\mathbf{A}} - \mathbf{A}^*||_{*}^2 \geq C_2 \left(\tfrac{r^2(s_u+s_v)}{n \log r} + \tfrac{r^2(s_u+s_v)}{n} \log(pq) \right), \tag{10}$$

where C_2 is a constant depending only on κ_{upper}.

If we assume that the random noise has at least a bounded second moment, that is, $\delta \geq 1$, then the rate of convergence in Theorem 1 reduces to the following:

$$\begin{aligned} \left\|\widehat{\mathbf{A}} - \mathbf{A}^*\right\|_{\mathrm{F}}^2 &\lesssim \kappa_{\text{lower}}^{-2} v_\delta r s_u s_v \tfrac{\log(pq)}{n}, \\ \left\|\widehat{\mathbf{A}} - \mathbf{A}^*\right\|_{*}^2 &\lesssim \kappa_{\text{lower}}^{-2} v_\delta r^2 s_u^2 s_v^2 \tfrac{\log(pq)}{n}. \end{aligned} \tag{11}$$

Comparing (11) and (10), we see that under the Frobenius norm, the rate of convergence for our proposed method is slower than that of the minimax optimal rate by a scaling factor of $\sqrt{s_u s_v/(s_u+s_v)}$. The lost of the scaling factor is due to the convex relaxation (4) where we estimate $\mathbf{A}$ directly rather than estimating the sparse singular vectors $\mathbf{u}_k$ and $\mathbf{v}_k$ separately. Finally, we note that the above results also illustrate the power of Huber loss: the rate of convergence of the proposed method under bounded second moment condition on the random noise matches that of the optimal rate derived under Gaussian error up to a scaling factor.

Remark 4. The class $\mathcal{F}$ defined in (9) is a sub-class of rs_us_v-sparse and r-rank matrices. Thus, the derived lower bound in Theorem 2 is also a lower bound for element-wise sparse and low-rank matrices. Here we use $\mathcal{F}$ instead of a class of sparse and low rank matrices because it is needed to obtain the nuclear norm convergence rate. Moreover, as the sparse singular value decomposition structure in $\mathcal{F}$ naturally gives a low rank and sparse $\mathbf{A}$ with at most rs_us_v nonzeros, we are able to obtain the convergence rate under the Frobenius norm. We emphasize that rs_us_v is a tight upper bound for the number of nonzeros in $\mathbf{A}$. To see this, considering the case where $r = s_v = 1$, then $rs_us_v = s_u$ and the coefficient matrix $\mathbf{A}$ has exactly s_u nonzeros.

4. Numerical Studies

We perform extensive numerical studies to evaluate the performance of our proposal for robust sparse reduced rank regression. Seven approaches are compared in our numerical studies: classical reduced rank regression, `classical`; our proposal with Huber loss, `hubersrrr`; our proposal with squared error loss (with $\tau \to \infty$), `srrr`; signal extraction approach for sparse multivariate response (Luo and Qi 2017), `SiER`; robust reduced rank regression with an additional mean parameter that models element-wise outliers (She and Chen 2017), `r4`; penalized reduced rank regression via an adaptive nuclear norm (Chen, Dong, and Chan 2013), `rrr`; and the penalized reduced rank regression via a ridge penalty (Mukherjee and Zhu 2011), `rrridge`. The proposals `classical`, `rrridge`, `rrr`, and `r4` do not assume sparsity on the regression coefficients. Among the seven proposals, only `hubersrrr` and `r4` are robust against outliers.

For all of our numerical studies, we generate each row of $\mathbf{X}$ from a multivariate normal distribution with mean zero and covariance matrix Σ, where $\Sigma_{ij} = 0.5^{|i-j|}$ for $1 \leq i,j \leq p$. Then, all elements of $\mathbf{X}$ are divided by the maximum absolute value of $\mathbf{X}$ such that $\max_{i,j} |X_{ij}| = 1$. The response matrix $\mathbf{Y}$ is then generated according to $\mathbf{Y} = \mathbf{X}\mathbf{A}^* + \mathbf{E}$. We consider two different types of outliers: (a) heavy-tailed random noise $\mathbf{E}$, and (b) contamination of some percentage of the elements of $\mathbf{Y}$. We simulate data with sparse and nonsparse low rank matrix $\mathbf{A}^*$. The details for the different scenarios will be specified in Section 4.1.

Our proposal `hubersrrr` involves three tuning parameters. We select the tuning parameters using 5-fold cross-validation with the absolute median loss as the criterion: we vary λ across a fine grid of values, consider four values of $\gamma = \{2.5, 3, 3.5, 4\}$ as suggested by Theorem 1, and considered a range of the robustification parameter $\tau = c\{n/\log(pq)\}^{1/2}$, where $c = \{0.4, 0.45, \ldots, 1.45, 1.5\}$. The tuning parameters for `srrr` are selected in a similar fashion with $\tau \to \infty$. We note that the tuning parameters can also be selected using a calibrated structured cross-validation proposed in She and Tran (2019). For scenarios with nonsparse regression coefficients, we simply set $\gamma = 0$ for `hubersrrr` and `srrr` for fair comparison against other approaches that do not assume sparsity. For `r3`, we select the tuning parameter using five different information criteria implemented in the `R` package `rrpack` (Chen, Dong, and Chan 2013), and report the best result. For `rrridge`, we specify the correct rank for $\mathbf{A}^*$ and consider a fine grid of tuning parameters for the ridge penalty and report the best result. For `classical`, we specify the correct rank for $\mathbf{A}^*$. We implement `SiER` using the default option in the `R` package `SiER`. The method `r4` has two tuning parameters that control the sparsity of the mean shift parameter for modeling outliers and the rank of $\mathbf{A}^*$. We implement `r4` by specifying the correct rank of $\mathbf{A}^*$, and choose the sparsity tuning parameter according to 5-fold cross-validation. Since `r4` is nonconvex, the final solution may depend on the initialization of the parameter of interest. We input the true regression coefficients $\mathbf{A}^*$ as an initial estimator for `r4`. In other words, we give a major advantage to `rrridge`, `r4`, and `classical` in that we provide the rank of $\mathbf{A}^*$ as an input as well as $\mathbf{A}^*$ as an initializer.

To evaluate the performance across different methods, we calculate the difference between the estimated regression coefficients $\widehat{\mathbf{A}}$ and the true coefficients $\mathbf{A}^*$ under the Frobenius norm. In addition, for scenarios with in which $\mathbf{A}^*$ is sparse, we calculate the true and false positive rates (TPR and FPR), defined as the proportion of correctly estimated nonzeros in the true parameter, and the proportion of zeros that are incorrectly estimated to be nonzero in the true parameter, respectively. Since some existing approaches are not applicable in the high-dimensional setting, we perform numerical studies under the low-dimensional setting in which $n \geq p$ in Section 4.1. We then illustrate the performance of our proposed methods, `hubersrrr` and `srrr`, compared to `SiER` and `r4`, in the high-dimensional setting in Section 4.2. In Section 4.3, we illustrate the phase transition phenomenon in Theorem 1 via numerical studies. In Section 4.4, we assess whether the proposed estimator estimates the rank accurately by plotting the top ten singular values of $\widehat{\mathbf{A}}$.

Table 1. The mean (and standrard error) of the difference between the estimated regression coefficients and the true regression coefficients under the Frobenius norm, averaged over 100 datasets, in the setting where $\mathbf{A}^*$ is not sparse, with $n = 200$, $p = 50$, and $q = 10$.

Rank of $\mathbf{A}^*$	Methods	Random noise			Data contamination		
		Normal	t	Log-normal	0%	5%	10%
1	`classical`	6.36(0.08)	69.97(12.90)	16.44(0.71)	6.36(0.08)	13.55(0.24)	21.30(0.40)
	`rrr`	5.80(0.07)	21.64(4.84)	10.71(0.19)	5.80(0.07)	10.25(0.14)	12.25(0.12)
	`rrridge`	5.42(0.06)	14.64(0.99)	9.22(0.17)	5.42(0.06)	9.04(0.10)	10.87(0.12)
	`SiER`	6.48(0.09)	35.20(6.18)	12.70(0.26)	6.48(0.09)	11.98(0.13)	14.13(0.16)
	`r4`	6.65(0.08)	4.65(0.07)	8.38(0.15)	6.65(0.08)	6.93(0.09)	7.35(0.08)
	`srrr`	7.13(0.08)	37.32(12.00)	10.37(0.13)	7.13(0.08)	10.44(0.09)	11.66(0.09)
	`hubersrrr`	7.17(0.08)	6.57(0.09)	6.70(0.08)	7.17(0.08)	7.72(0.08)	8.16(0.08)
2	`classical`	10.34(0.12)	61.59(5.70)	21.83(0.56)	10.34(0.12)	20.90(0.24)	28.15(0.28)
	`rrr`	6.09(0.09)	25.89(1.88)	12.08(0.32)	6.09(0.09)	12.20(0.17)	16.81(0.23)
	`rrridge`	9.16(0.09)	21.71(0.53)	15.16(0.20)	9.16(0.09)	15.25(0.12)	18.27(0.13)
	`SiER`	6.29(0.10)	32.89(2.66)	12.86(0.40)	7.19(0.08)	12.71(0.20)	17.20(0.25)
	`r4`	10.59(0.11)	7.36(0.09)	12.66(0.16)	11.63(0.13)	11.23(0.12)	11.94(0.12)
	`srrr`	8.65(0.11)	31.32(4.57)	14.11(0.23)	8.69(0.11)	14.70(0.16)	17.84(0.17)
	`hubersrrr`	8.67(0.11)	7.61(0.11)	7.81(0.11)	8.70(0.11)	9.57(0.12)	10.33(0.13)

NOTE: Three distributions of random noise are considered: normal, t, and log-normal. We also considered contaminating 5% or 10% of the elements of $\mathbf{Y}$.

4.1. Low-Dimensional Setting with $n \geq p$

In this section, we perform numerical studies with $n = 200$, $p = 50$, and $q = 10$. We first consider two cases in which $\mathbf{A}^*$ has low rank but is not sparse:

1. Rank one matrix: $\mathbf{A}^* = \mathbf{u}_1\mathbf{v}_1^{\mathrm{T}}$, where each element of $\mathbf{u}_1 \in \mathbb{R}^p$ and $\mathbf{v}_1 \in \mathbb{R}^q$ is generated from a uniform distribution on the interval $[-1, -0.5] \cup [0.5, 1]$.
2. Rank two matrix: $\mathbf{A}^* = \mathbf{u}_1\mathbf{v}_1^{\mathrm{T}} + \mathbf{u}_2\mathbf{v}_2^{\mathrm{T}}$, where each element of $\mathbf{u}_1, \mathbf{u}_2 \in \mathbb{R}^p$ and $\mathbf{v}_1, \mathbf{v}_2 \in \mathbb{R}^q$ is generated from a uniform distribution on the interval $[-1, -0.5] \cup [0.5, 1]$.

We then generate random noise $\mathbf{E} \in \mathbb{R}^{n\times q}$ from three different distributions: (a) the normal distribution $N(0, 4)$, (b) the t-distribution with degrees of freedom 1.5, and (c) the log-normal distribution $\log N(0, 1.2^2)$. Moreover, we consider a contamination scenario in which we generate each element of $\mathbf{E}$ from the $N(0, 4)$ distribution, and then randomly contaminate 5% and 10% of the elements in $\mathbf{Y}$ by replacing them with random values generated from a uniform distribution on the interval $[10, 20]$. The estimation error for each method under the Frobenius norm, averaged over 100 datasets, is reported in Table 1.

From Table 1, we see that `rrr` and `rrridge` outperform all other methods when $\mathbf{A}^*$ is rank one under Gaussian noise. This is not surprising, since `rrr` and `rrridge` are tailored for reduced rank regression without outliers. We see that `hubersrrr` has similar performance to `srrr`, suggesting that there is no loss of efficiency for `hubersrrr` even when there are no outliers. When the random noise is generated from the t-distribution, `r4` has the best performance, followed by `hubersrrr`. Note that `r4` is nonconvex and we provide the true regression coefficients $\mathbf{A}^*$ as an initializer. The estimation errors for methods that do not model the outliers are substantially higher. For log-normal random noise, `hubersrrr` outperforms `r4`. Under the data contamination model, `r4` and `hubersrrr` perform similarly, and both outperform all of the other methods. These results corroborate the observation in She and Chen (2017) that the estimation of low rank matrices is extremely sensitive to outliers. As we increase the contamination percentage of the observed outcomes, we see that the performance of the nonrobust methods deteriorates. Similar results are observed for the case when $\mathbf{A}^*$ has rank two.

Next, we consider two cases in which $\mathbf{A}^*$ is both sparse and low rank:

1. Sparse rank one matrix: $\mathbf{A}^* = \mathbf{u}_1\mathbf{v}_1^{\mathrm{T}}$ with $\mathbf{u}_1 = (\mathbf{1}_4^{\mathrm{T}}, \mathbf{0}_{p-4}^{\mathrm{T}})^{\mathrm{T}}$ and $\mathbf{v}_1 = (\mathbf{1}_4^{\mathrm{T}}, \mathbf{0}_{q-4}^{\mathrm{T}})^{\mathrm{T}}$;
2. Sparse rank two matrix: $\mathbf{A}^* = \mathbf{u}_1\mathbf{v}_1^{\mathrm{T}} + \mathbf{u}_2\mathbf{v}_2^{\mathrm{T}}$ with $\mathbf{u}_1 = (\mathbf{1}_4^{\mathrm{T}}, \mathbf{0}_{p-4}^{\mathrm{T}})^{\mathrm{T}}$, $\mathbf{v}_1 = (\mathbf{1}_4^{\mathrm{T}}, \mathbf{0}_{q-4}^{\mathrm{T}})^{\mathrm{T}}$, $\mathbf{u}_2 = (\mathbf{0}_2^{\mathrm{T}}, \mathbf{1}_4^{\mathrm{T}}, \mathbf{0}_{p-6}^{\mathrm{T}})^{\mathrm{T}}$, and $\mathbf{v}_2 = (\mathbf{0}_2^{\mathrm{T}}, \mathbf{1}_4^{\mathrm{T}}, \mathbf{0}_{q-6}^{\mathrm{T}})^{\mathrm{T}}$.

The heavy-tailed random noise and data contamination scenarios are as described earlier. The results, averaged over 100 datasets, are reported in Table 2.

When $\mathbf{A}^*$ is sparse, `SiER`, `hubersrrr` and `srrr` outperform all of the methods that do not assume sparsity. In particular, we see that `r4` has the worst performance when the random noise is normal or log-normal, or when the data are contaminated. The method `rrr` has an MSE of 5.00 when the data are contaminated, due to the fact that the information criteria always select models with the regression coefficients estimated to be zero. Under the Gaussian error without any outliers, `SiER` has a lower MSE than our proposed method when the rank of $\mathbf{A}^*$ is one. In summary, our proposal `hubersrrr` has the best performance across most scenarios and is robust against different types of outliers.

4.2. High-Dimensional Setting with $p > n$

In this section, we assess the performance of our proposed method in the high-dimensional setting, when the matrix $\mathbf{A}^*$ is sparse. To this end, we perform numerical studies with $q = 10$, $p = 200$, and $n = 150$. The methods `classical`, `rrr`, and `rrridge` do not assume sparsity and do not model outliers, therefore, their results are omitted. We consider low rank and sparse matrices $\mathbf{A}^*$ described in Section 4.1. Similarly, two types of outliers are considered: heavy-tailed random noise and data contamination. The TPR, FPR, and estimation error under Frobenius norm for both types of scenarios, averaged over 100 datasets, are summarized in Tables 3 and 4, respectively.

Table 2. Results for the case where $\mathbf{A}^*$ is sparse, with $n = 200$, $p = 50$, and $q = 10$.

Rank of $\mathbf{A}^*$	Methods	Random noise			Data contamination		
		Normal	*t*-dist	Log-normal	0%	5%	10%
1	classical	6.27(0.09)	52.85(4.06)	17.95(0.60)	6.27(0.09)	15.23(0.23)	22.34(0.29)
	rrr	4.65(0.04)	6.97(0.89)	4.98(0.01)	4.65(0.04)	4.99(0.01)	5.00(0.01)
	rrridge	2.73(0.03)	7.35(0.35)	4.17(0.08)	2.73(0.03)	4.00(0.05)	4.60(0.05)
	SiER	2.16(0.06)	24.42(1.81)	6.46(0.36)	2.16(0.06)	5.62(0.17)	8.67(0.23)
	r4	4.86(0.07)	3.73(0.05)	8.76(0.17)	4.86(0.07)	5.03(0.07)	5.36(0.09)
	srrr	2.57(0.05)	4.95(0.24)	4.43(0.05)	2.57(0.05)	4.30(0.05)	4.78(0.03)
	hubersrrr	2.57(0.04)	2.19(0.04)	2.44(0.05)	2.57(0.04)	2.80(0.05)	3.02(0.05)
2	classical	10.04(0.12)	59.51(4.10)	22.20(0.57)	10.04(0.12)	20.89(0.24)	29.08(0.31)
	rrr	5.25(0.04)	10.07(0.83)	8.18(0.03)	5.25(0.04)	8.20(0.02)	8.24(0.01)
	rrridge	4.36(0.03)	8.92(0.35)	6.00(0.05)	4.36(0.03)	6.08(0.04)	6.82(0.04)
	SiER	3.39(0.04)	22.68(1.94)	5.45(0.28)	3.39(0.04)	5.14(0.10)	6.83(0.14)
	r4	8.71(0.10)	6.42(0.07)	11.31(0.15)	8.71(0.10)	9.07(0.11)	9.61(0.11)
	srrr	3.27(0.04)	7.74(0.09)	5.72(0.10)	3.27(0.04)	5.62(0.06)	6.73(0.07)
	hubersrrr	3.27(0.04)	2.88(0.04)	3.13(0.05)	3.27(0.04)	3.59(0.04)	3.83(0.05)

NOTE: Other details are as in Table 1.

Table 3. Results for the case when $\mathbf{A}^*$ is sparse and rank one in the high-dimensional setting with $n = 150$, $p = 200$, and $q = 10$.

Rank of $\mathbf{A}^*$	Methods	Random noise			Data contamination		
		Normal	*t*-dist	Log-normal	0%	5%	10%
r4	Frobenius	19.74(0.28)	15.03(0.77)	27.07(0.42)	19.74(0.28)	21.41(0.33)	25.69(0.67)
	Frobenius	2.95(0.09)	47.02(9.48)	12.84(0.62)	2.95(0.09)	9.64(0.22)	13.08(0.20)
SiER	TPR	0.99(0.01)	0.20(0.02)	0.42(0.04)	0.99(0.01)	0.50(0.04)	0.30(0.03)
	FPR	0.05(0.01)	0.10(0.01)	0.09(0.01)	0.05(0.01)	0.08(0.01)	0.08(0.01)
	Frobenius	3.77(0.05)	6.23(1.23)	4.98(0.01)	3.77(0.05)	4.98(0.16)	5.00(0.01)
srrr	TPR	0.95(0.01)	0.01(0.01)	0.10(0.02)	0.95(0.01)	0.14(0.02)	0.04(0.01)
	FPR	0.12(0.01)	0.01(0.01)	0.01(0.01)	0.12(0.01)	0.02(0.01)	0.01(0.01)
	Frobenius	3.78(0.05)	3.17(0.06)	3.58(0.06)	3.78(0.05)	4.05(0.05)	4.23(0.05)
hubersrrr	TPR	0.95(0.01)	0.99(0.01)	0.97(0.01)	0.95(0.01)	0.90(0.02)	0.82(0.03)
	FPR	0.13(0.01)	0.11(0.01)	0.15(0.01)	0.13(0.01)	0.13(0.01)	0.12(0.01)

NOTE: Three distributions of random noise are considered: normal, *t*, and log-normal. We also considered contaminating 5% or 10% of the elements of $\mathbf{Y}$. We report the mean (and standard error) of the true and false positive rates, and the difference between $\widehat{\mathbf{A}}$ and $\mathbf{A}^*$ under Frobenius norm, averaged over 100 datasets.

Table 4. Results for the case when $\mathbf{A}^*$ is sparse and rank two in the high-dimensional setting with $n = 150$, $p = 200$, and $q = 10$.

Rank of $\mathbf{A}^*$	Methods	Random noise			Data contamination		
		Normal	*t*-dist	Log-normal	0%	5%	10%
r4	Frobenius	28.83(0.28)	24.13(0.79)	31.32(0.42)	28.83(0.28)	32.58(0.46)	40.96(0.88)
	Frobenius	3.98(0.06)	45.53(9.54)	10.00(0.71)	3.98(0.06)	6.79(0.19)	10.29(0.34)
SiER	TPR	0.93(0.01)	0.35(0.09)	0.69(0.03)	0.93(0.01)	0.79(0.02)	0.64(0.03)
	FPR	0.05(0.01)	0.09(0.01)	0.06(0.01)	0.05(0.01)	0.05(0.01)	0.07(0.01)
	Frobenius	4.66(0.05)	9.37(1.21)	7.47(0.09)	4.66(0.05)	7.32(0.06)	7.92(0.05)
srrr	TPR	0.96(0.01)	0.06(0.02)	0.42(0.03)	0.96(0.01)	0.54(0.02)	0.25(0.02)
	FPR	0.16(0.01)	0.01(0.01)	0.04(0.01)	0.16(0.01)	0.07(0.01)	0.03(0.01)
	Frobenius	4.67(0.05)	4.02(0.06)	4.43(0.07)	4.67(0.05)	5.02(0.06)	5.32(0.06)
hubersrrr	TPR	0.96(0.01)	0.99(0.01)	0.97(0.01)	0.96(0.01)	0.94(0.01)	0.93(0.01)
	FPR	0.16(0.01)	0.16(0.01)	0.18(0.01)	0.16(0.01)	0.16(0.01)	0.16(0.01)

NOTE: Other details are as in Table 3.

We see that for Gaussian random noise, SiER has the lowest estimation error, followed by hubersrrr and srrr. We note that hubersrrr and srrr have similar results, indicating that there is little loss of efficiency when there are no outliers. However, in scenarios in which the random noise is heavy-tailed, hubersrrr has high TPR, low FPR, and low Frobenius norm compared to all of the other methods. In fact, we see that when the random noise is heavy-tailed, the TPR and FPR of srrr and SiER are very low. We see similar performance for the case when the data are contaminated in Table 4. These results suggest that hubersrrr should be preferred in all scenarios since it allows accurate estimation of $\mathbf{A}^*$ when the random noise are heavy-tailed, or under data contamination. Moreover, there is little loss of efficiency compared to srrr and SiER when there are no outliers.

4.3. Phase Transition Phenomenon

Similar to Section 4.2, we generate the response matrix $\mathbf{Y} = \mathbf{X}\mathbf{A}^* + \mathbf{E}$ with a sparse rank two matrix $\mathbf{A}^* = \mathbf{u}_1\mathbf{v}_1^{\mathrm{T}} + \mathbf{u}_2\mathbf{v}_2^{\mathrm{T}}$, where $\mathbf{u}_1 = (\mathbf{1.5}_4^{\mathrm{T}}, \mathbf{0}_{p-4}^{\mathrm{T}})^{\mathrm{T}}$, $\mathbf{v}_1 = (\mathbf{1.5}_4^{\mathrm{T}}, \mathbf{0}_{q-4}^{\mathrm{T}})^{\mathrm{T}}$, $\mathbf{u}_2 = (\mathbf{0}_2^{\mathrm{T}}, \mathbf{1.5}_4^{\mathrm{T}}, \mathbf{0}_{p-6}^{\mathrm{T}})^{\mathrm{T}}$, and $\mathbf{v}_2 = (\mathbf{0}_2^{\mathrm{T}}, \mathbf{1.5}_4^{\mathrm{T}}, \mathbf{0}_{q-6}^{\mathrm{T}})^{\mathrm{T}}$. To validate the phase transition behavior, we generate each element of the random noise $\mathbf{E}$ from a t_{df} distribution with degrees of freedom df. The t_{df} distribution has a finite $(1 + \delta)$th moment provided that the degrees of freedom is larger than $1+\delta$. We take $\mathrm{df} = \{1.5, 1.6, \ldots, 3.5\}$. The

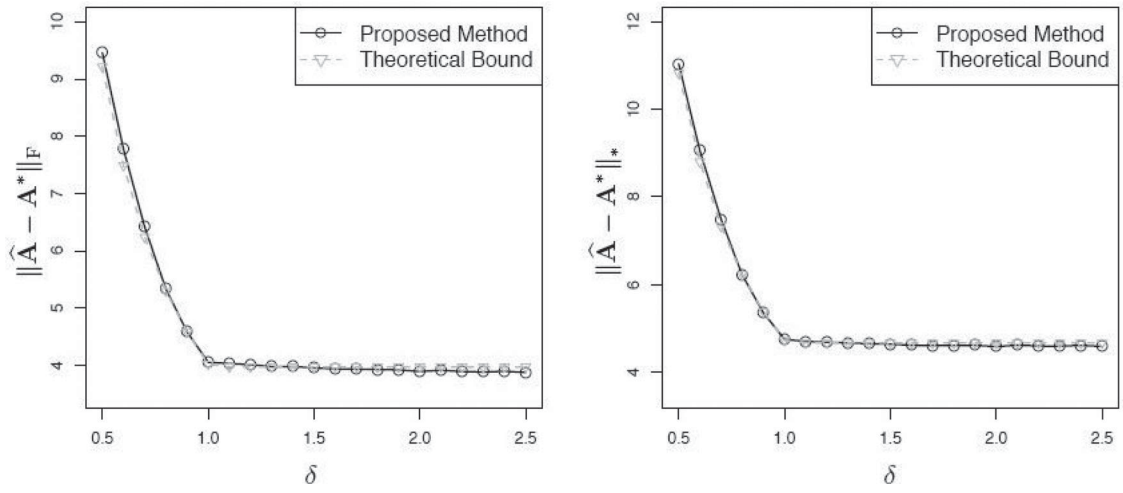

Figure 1. Estimation error for $\widehat{\mathbf{A}}$ under the Frobenius and nuclear norm across a range of δ, averaged over 200 datasets. The black solid line is the estimation error for $\widehat{\mathbf{A}}$ and the gray dash line is the theoretical bound.

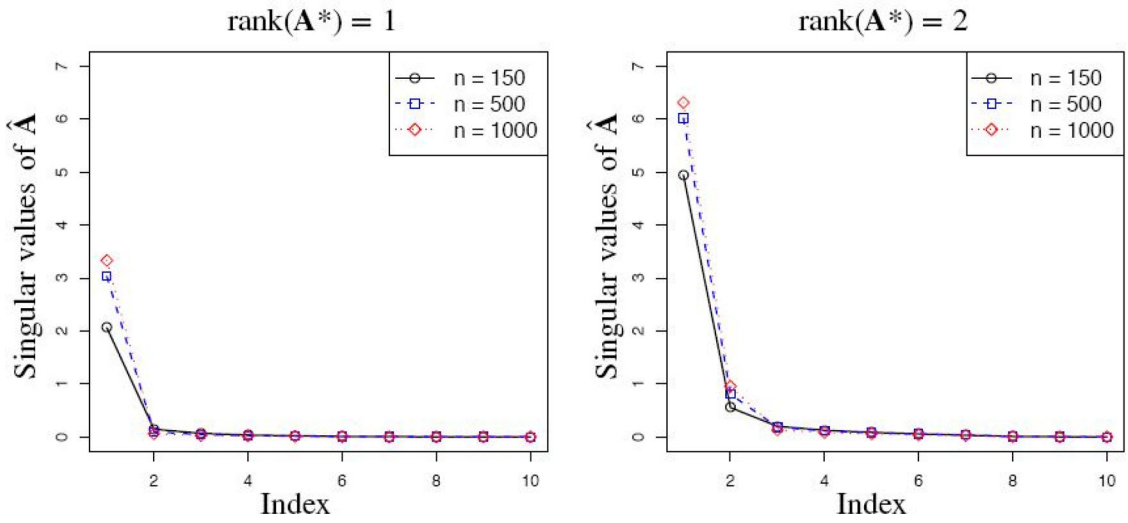

Figure 2. Singular values for $\widehat{\mathbf{A}}$ averaged over 100 datasets. The left and right panels present results for simulation settings with rank($\mathbf{A}^*$) = 1 and rank($\mathbf{A}^*$) = 2, respectively. The black, blue, and red lines correspond to the results for the case when $p = 200$ and $n = \{150, 500, 1000\}$, respectively.

differences between $\widehat{\mathbf{A}}$ and $\mathbf{A}^*$ under the Frobenius and nuclear norm across different values of δ, averaged over 200 replications, are reported in Figure 1 for the case when $n = 400, p = 500$, and $q = 10$. We also calculate the theoretical bound as a function of δ, that is, $\nu_\delta^{1/\min(1+\delta,2)} \{\log(pq)/n\}^{\min\{\delta/(1+\delta),1/2\}}$, where we set $\nu_\delta = 20$ to obtain the curves in Figure 1.

From Figure 1, we see that the curve obtained theoretically matches with the estimation error of $\widehat{\mathbf{A}}$ under both the Frobenius norm and nuclear norm, across a range of δ. In particular, we observe that there is a phase transition phenomenon: when $\delta < 1$, the estimation error decreases as δ increases; when $\delta \geq 1$, the estimation error under both Frobenius and nuclear norms are flat even when we increase δ.

4.4. Rank Selection Consistency

We consider the simulation setting in Section 4.2 with random noise simulated from the t-distribution with degrees of freedom 1.5. The top 10 singular values for $\widehat{\mathbf{A}}$ obtained from our proposed method with $n = \{150, 500, 1000\}$ and $p = 200$ are plotted in Figure 2. We see from the left panel of Figure 2 that the largest singular value of $\widehat{\mathbf{A}}$ is estimated to be large and the rest of the singular values are approximately zero when rank($\mathbf{A}^*$) = 1. On the other hand, when rank($\mathbf{A}^*$) = 2, our proposed method estimates both the first and second largest singular value to be significantly different from zero, while the rest of the singular values are estimated to be approximately zero. These results suggest that the proposed estimator can estimate the rank of $\mathbf{A}^*$ quite accurately.

5. Data Application

We apply the proposed robust sparse reduced rank regression to the *Arabidopsis thaliana* dataset, which consists of gene expression measurements for $n = 118$ samples (Rodrígues-Concepción and Boronat 2002; Wille et al. 2004; Ma, Gong, and Bohnert 2007; Tan, Witten, and Shojaie 2015; She and Chen 2017). It is known that isoprenoids play many important roles in biochemical functions such as respiration, photosynthesis, and regulation of growth in plants. Here, we explore the

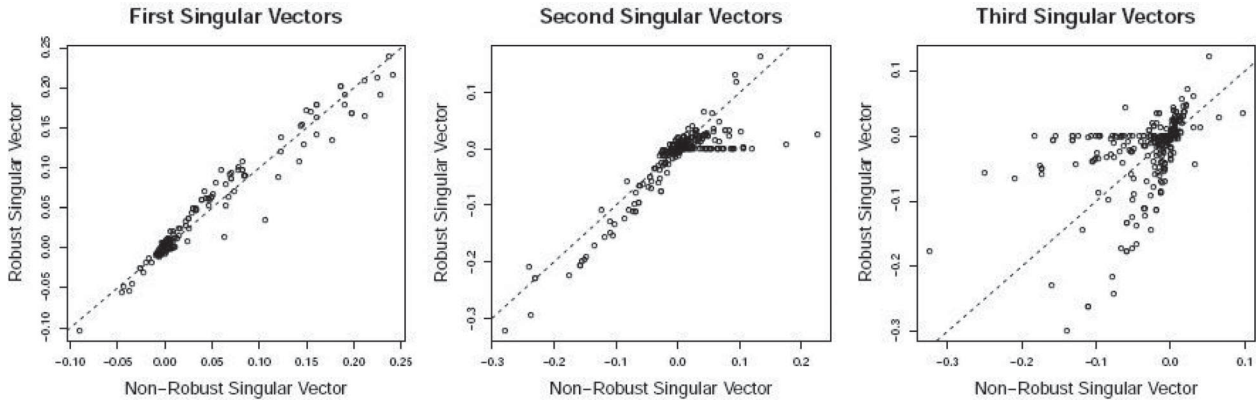

Figure 3. Scatterplots of the leading right singular vectors of $\widehat{\mathbf{XA}}_{\texttt{hubersrrr}}$ and $\widehat{\mathbf{XA}}_{\texttt{srrr}}$.

connection between two isoprenoid biosynthesis pathways and some downstream pathways.

Similar to She and Chen (2017), we treat the $p = 39$ genes from two isoprenoid biosynthesis pathways as the predictors, and treat the $q = 795$ genes from 56 downstream pathways as the response. Thus, $\mathbf{X} \in \mathbb{R}^{118\times 39}$ and $\mathbf{Y} \in \mathbb{R}^{118\times 795}$, and we are interested in fitting the model $\mathbf{Y} = \mathbf{XA} + \mathbf{E}$. We scale each element of $\mathbf{X}$ such that $\max_{i,j} |X_{ij}| = 1$, and standardize each column of $\mathbf{Y}$ to have mean zero and standard deviation one.

In Section 4.2, we illustrated with numerical studies that if the response variables are heavy-tailed, sparse reduced rank regression with squared error loss will lead to incorrect estimates. We now illustrate the difference between solving (4) with Huber loss and squared error loss. We set $\gamma = 3$, and pick λ such that there are 1000 nonzeros in the estimated coefficient matrix. For the robust method, we set the robustification parameter to equal $\tau = 3$ for simplicity. In principle, this quantity can be chosen using cross-validation.

Let $\widehat{\mathbf{A}}_{\texttt{hubersrrr}}$ and $\widehat{\mathbf{A}}_{\texttt{srrr}}$ be the estimated regression coefficients for the robust and nonrobust methods, respectively. To measure the difference between the two approaches in terms of regression coefficients and prediction, we compute the quantities $||\widehat{\mathbf{A}}_{\texttt{hubersrrrst}} - \widehat{\mathbf{A}}_{\texttt{srrr}}||_{\mathrm{F}}/||\widehat{\mathbf{A}}_{\texttt{hubersrrr}}||_{\mathrm{F}} \approx 37\%$ and $||\mathbf{X}\widehat{\mathbf{A}}_{\texttt{hubersrrr}} - \mathbf{X}\widehat{\mathbf{A}}_{\texttt{srrr}}||_{\mathrm{F}}/||\mathbf{X}\widehat{\mathbf{A}}_{\texttt{hubersrrr}}||_{\mathrm{F}} \approx 35\%$.

Figure 3 displays scatterplots of the right singular vectors of $\mathbf{X}\widehat{\mathbf{A}}_{\texttt{srrr}}$ against the right singular vectors of $\mathbf{X}\widehat{\mathbf{A}}_{\texttt{hubersrrr}}$. We see that while the first singular vectors are similar between the two methods, the second and third singular vectors are very different. These results suggest that the regression coefficients and model predictions can be quite different between robust and nonrobust methods when there are outliers, and that care needs to be taken during model fitting.

Next, we assess the prediction accuracy of our proposed method under both Huber loss and squared error loss. Specifically, we split the data into training set with $n_{\text{train}} = 100$ and test set with $n_{\text{test}} = 18$. Then, we fit the proposed method on the training set with tuning parameters selected using cross-validation similar to that of described in Section 4, and evaluate the prediction accuracy on the test set. We repeat this procedure 1000 times. The prediction error under the Huber and squared error loss are 8836 and 8907, with standard errors 63 and 64, respectively. The improvement for using the Huber loss is mild in this dataset, mainly due to the fact that the outliers themselves are not very large. To further illustrate the advantage of the Huber loss, we repeat the aforementioned analysis with one entry of the response matrix perturbed with the number 50. The prediction error are now 8837 and 9115 for the Huber loss and squared error loss, with a standard error of 63 and 66, respectively. In summary, we see that the prediction error under the Huber loss does not change since it is robust to outliers, whereas the prediction error under squared error loss increases significantly.

6. Discussion

We propose robust sparse reduced rank regression for analyzing large, complex, and possibly contaminated data. Our proposal is based on a convex relaxation, and is thus, computationally tractable. We show that our proposal is statistically consistent under both Frobenius and nuclear norms in the high-dimensional setting in which $p > n$. By contrast, most of the existing literature in reduced rank regression focus on prediction and rank selection consistency.

Specifically, we focus on quantifying the tradeoff between heavy-tailness of the random noise and the statistical bias. We show that the proposed robust estimator can achieve exponential-type deviation errors only under bounded low-order moments. Our work offers a different perspective to studying robustness. In particular, our framework is different from the conventional perspective on robust statistics under the Huber's ϵ-contamination model, which focuses on developing robust procedures with a high breakdown point (Huber 1964). The breakdown point of an estimator is defined roughly as the proportion of arbitrary outliers an estimator can tolerate before the estimator produces arbitrarily large estimates, or breaks down (Hampel 1971). Under the conventional perspective, the proposed method, similar to that of the classical Huber regression, has a breakdown point of $1/n$, when both features and responses can be arbitrarily contaminated. We leave the theoretical investigations of robust sparse reduced rank regression with nonconvex truncated losses, that may potentially have resistant estimation, for future work.

Supplementary Materials

The online supplementary materials collect an ADMM algorithm, the proofs for all the theoretical results.

Funding

Tan is supported by NSF DMS 2113356, NSF DMS 1949730, and NIH RF1-MH122833. Sun is supported in part by NSERC grant RGPIN-2018-06484. Witten is supported by NIH R01 GM123993, Simons Investigator for Mathematical Modeling of Living Systems and NSF CAREER award 1252624.

References

Boyd, S., Parikh, N., Chu, E., Peleato, B., and Eckstein, J. (2010), "Distributed Optimization and Statistical Learning via the ADMM," *Foundations and Trends in Machine Learning*, 3, 1–122. [3]

Bunea, F., She, Y., and Wegkamp, M. H. (2011), "Optimal Selection of Reduced Rank Estimators of High-Dimensional Matrices," *The Annals of Statistics*, 39, 1282–1309. [1,5]

Bunea, F., She, Y., and Wegkamp, M. H. (2012), "Joint Variable and Rank Selection for Parsimonious Estimation of High-dimensional Matrices," *The Annals of Statistics*, 40, 2359–2388. [1,5]

Candes, E. J., Li, X., Ma, Y., and Wright, J. (2011), "Robust Principal Component Analysis?" *Journal of ACM*, 58, 1–37. [2,5]

Chen, K., Chan, K.-S., and Stenseth, N. C. (2012), "Reduced Rank Stochastic Regression with a Sparse Singular Value Decomposition," *Journal of the Royal Statistical Society*, Series B, 74, 203–221. [2]

Chen, K., Dong, H., and Chan, K.-S. (2013), "Reduced Rank Regression via Adaptive Nuclear Norm Penalization," *Biometrika*, 100, 901–920. [1,6]

Chen, L., and Huang, J. Z. (2012), "Sparse Reduced-Rank Regression for Simultaneous Dimension Reduction and Variable Selection," *Journal of the American Statistical Association*, 107, 1533–1545. [1,2,3,5]

Cont, R. (2001), "Empirical Properties of Asset Returns: Stylized Facts and Statistical Issues," *Quantitive Finance*, 1, 223–236. [1]

Eckstein, J., and Bertsekas, D. (1992), "On the Douglas–Rachford Splitting Method and the Proximal Point Algorithm for Maximal Monotone Operators," *Mathematical Programming*, 55, 293–318. [3]

Fan, J., Liu, H., Sun, Q., and Zhang, T. (2018), "I-LAMM: For Sparse Learning: Simultaneous Control of Algorithmic Complexity and Statistical Error," *The Annals of Statistics*, 46, 818–841. [4]

Hampel, F. R. (1971), "A General Qualitative Definition of Robustness," *The Annals of Mathematical Statistics*, 42, 1887–1896. [10]

He, X., and Shao, Q.-M. (1996), "A General Bahadur Representation of M-estimators and its Application to Linear Regression with Nonstochastic Designs," *The Annals of Statistics*, 24, 2608–2630. [2]

Huber, P. J. (1964), "Robust Estimation of a Location Parameter," *The Annals of Mathematical Statistics*, 35, 73–101. [2,10]

——— (1973), "Robust Regression: Asymptotics, Conjectures and Monte Carlo," *The Annals of Statistics*, 1, 799–821. [2]

Izenman, A. J. (1975), "Reduced-Rank Regression for the Multivariate Linear Model," *Journal of Multivariate Analysis*, 5, 248–264. [2]

Jojic, V., Saria, S., and Koller, D. (2011), "Convex Envelopes of Complexity Controlling Penalties: The Case Against Premature Envelopment," in *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*. [3]

Liu, G., Lin, Z., Yan, S., Sun, J., Yu, Y., and Ma, Y. (2012), "Robust Recovery of Subspace Structures by Low-Rank Representation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35, 171–184. [1]

Luo, R., and Qi, X. (2017), "Signal Extraction Approach for Sparse Multivariate Response Regression," *Journal of Multivariate Analysis*, 153, 83–97. [1,6]

Ma, S., Gong, Q., and Bohnert, H. (2007), "An Arabidopsis Gene Network Based on the Graphical Gaussian Model," *Genome Research*, 17, 1614–1625. [9]

Mammen, E. (1989), "Asymptotics with Increasing Dimension for Robust Regression with Applications to the Bootstrap," *The Annals of Statistics*, 17, 382–400. [2]

Mukherjee, A., and Zhu, J. (2011), "Reduced Rank Ridge Regression and its Kernel Extensions," *Statistical Analysis and Data Mining: The ASA Data Science Journal*, 4, 612–622. [1,6]

Müller, U. A., Dacorogna, M. M., and Pictet, O. V. (1998), "Heavy Tails in High-Frequency Financial Data," in *A Practical Guide to Heavy Tails: Statistical Techniques and Applications*, eds. R. Adler, R. Feldman, and M. Taqqu, pp. 55–78, Boston: Birkhäuser. [1]

Portnoy, S. (1985), "Asymptotic Behavior of M Estimators of p Regression Parameters When p^2/n is Large; II. Normal Approximation," *The Annals of Statistics*, 13, 1403–1417. [2]

Recht, B., Fazel, M., and Parrilo, P. (2010), "Guaranteed Minimum-Rank Solutions of Linear Matrix Equations via Nuclear Norm Minimization," *SIAM Review*, 52, 471–501. [3]

Richard, E., Savalle, P.-A., and Vayatis, N. (2012), "Estimation of Simultaneously Sparse and Low Rank Matrices." arXiv preprint arXiv:1206.6474. [1,3]

Rodrígues-Concepción, M., and Boronat, A. (2002), "Elucidation of the Methylerythritol Phosphate Pathway for Isoprenoid Biosynthesis in Bacteria and Plastids. A Metabolic Milestone Achieved Through Genomics," *Plant Physiology*, 130, 1079–1089. [9]

She, Y. (2017), "Selective Factor Extraction in High Dimensions," *Biometrika*, 104, 97–110. [1,2,5]

She, Y., and Chen, K. (2017), "Robust Reduced-Rank Regression," *Biometrika*, 104, 633–647. [1,3,6,7,9,10]

She, Y., and Owen, A. B. (2011), "Outlier Detection Using Nonconvex Penalized Regression," *Journal of the American Statistical Association*, 106, 626–639. [1]

She, Y., and Tran, H. (2019), "On Cross-Validation for Sparse Reduced Rank Regression," *Journal of the Royal Statistical Society*, Series B, 81, 145–161. [6]

Sun, Q., Zhou, W., and Fan, J. (2018), "Adaptive Huber Regression," *Journal of the American Statistical Association*, 115, 254–265. [2,4]

Tan, K., Witten, D., and Shojaie, A. (2015), "The Cluster Graphical Lasso for Improved Estimation of Gaussian Graphical Models," *Computational Statistics and Data Analysis*, 85, 23–36. [9]

Wang, Q., Gong, P., Chang, S., Huang, T. S., and Zhou, J. (2016), "Robust Convex Clustering Analysis," in *IEEE 16th International Conference on Data Mining*. [1]

Wille, A., Zimmermann, P., Vranová, E., Fürholz, A., Laule, O., Bleuler, S., Hennig, L., Prelíc, A., Rohr, P., Thiele, L., Zitzler, E., Gruissem, W., and Bühlmann, P. (2004), "Sparse Graphical Gaussian Modeling of the Isoprenoid Gene Network in Arabidopsis thaliana," *Genome Biology*, 5, 1–13. [9]