Bernoulli **26**(2), 2020, 1139–1170
https://doi.org/10.3150/19-BEJ1144

Robust regression via mutivariate regression depth

CHAO GAO

Department of Statistics, University of Chicago, Chicago, IL 60637, USA.
E-mail: chaogao@galton.uchicago.edu

This paper studies robust regression in the settings of Huber's ϵ-contamination models. We consider estimators that are maximizers of multivariate regression depth functions. These estimators are shown to achieve minimax rates in the settings of ϵ-contamination models for various regression problems including nonparametric regression, sparse linear regression, reduced rank regression, etc. We also discuss a general notion of depth function for linear operators that has potential applications in robust functional linear regression.

Keywords: contamination model; data depth; high-dimensional regression; minimax rate; robust statistics

1. Introduction

Regression is probably one of the most important subjects in statistics. The goal is to learn the conditional mean or median of a response $Y \in \mathbb{R}^m$ given a covariate $X \in \mathbb{R}^p$. Its form ranges from classical low-dimensional linear regression to modern nonparametric and high-dimensional models. In this paper, we study robust regression in the setting of Huber's ϵ-contamination model (Huber [15]). Namely, consider i.i.d. observations

$$(X_1, Y_1), \dots, (X_n, Y_n) \sim (1-\epsilon)P_B + \epsilon Q. \tag{1}$$

The distribution P_B models the relation between X and Y via the regression parameter B, and Q is an unknown contamination distribution. We need to learn the regression parameter B. In this setting, there are approximately ϵn observations sampled from Q that do not carry any information about B. Since we do not know which observation is contaminated or not, a procedure to recover B must be robust. To be specific, this paper covers the following list of robust regression problems:

1. *Nonparametric regression.* The relation between x and y is characterized by $y|x \sim N(f(x), 1)$ with some nonparametric function f. The goal is to estimate f using data sampled from $(1-\epsilon)P_f + \epsilon Q$.
2. *Sparse linear regression.* For a scalar response y and a vector covariate X, a linear model is specified by $y|X \sim N(\beta^T X, \sigma^2)$, with some regression vector β assumed to be sparse. The goal is to estimate β with samples from $(1-\epsilon)P_\beta + \epsilon Q$.
3. *Gaussian graphical model.* In this setting, we observe i.i.d. samples from $(1-\epsilon)N(0, \Omega^{-1}) + \epsilon Q$. The goal is to estimate the sparse precision matrix Ω. The sparsity pattern of Ω characterizes the graphical model of conditional dependence. The Gaussian graphical

1350-7265 © 2020 ISI/BS

model is closely related and can be solved by sparse linear regression (Meinshausen and Bühlmann [25]).

4. *Low-rank trace regression.* For a scalar response y and a matrix covariate X, a linear model is specified by $y|X \sim N(\mathsf{Tr}(B^T X), \sigma^2)$. The regression matrix B is assumed to be low-rank, and the goal is to estimate it with samples from $(1-\epsilon)P_B + \epsilon Q$.
5. *Multivariate linear regression.* In this setting, the response is also multivariate. The linear model is specified by $Y|X \sim N(B^T X, \sigma^2 I_m)$. The problem is also termed as multi-task learning. We will show that even there is no relation between the m univariate linear models, estimation of the m columns of B must be done in a joint fashion once the samples are from $(1-\epsilon)P_B + \epsilon Q$.
6. *Multivariate linear regression with group sparsity.* We consider the same model in the last item, and assume that only a subset of the rows of the regression matrix B are nonzero.
7. *Reduced rank regression.* In the same setting of multivariate linear regression, we further assume the regression matrix B is low-rank.

Though the seven problems listed are very different, and the regression parameter we want to recover can be a vector, a matrix or even a function, we consider a unified robust estimation procedure in this paper. In the setting of multivariate linear regression, we use $\mathbb{P}$ to denote the joint distribution of $X \in \mathbb{R}^p$ and $Y \in \mathbb{R}^m$. The multivariate regression depth of $B \in \mathbb{R}^{p\times m}$ is defined as

$$\mathcal{D}_{\mathcal{U}}(B, \mathbb{P}) = \inf_{U\in\mathcal{U}} \mathbb{P}\left\{\left\langle U^T X, Y - B^T X\right\rangle \geq 0\right\}, \tag{2}$$

for some subset $\mathcal{U} \subset \mathbb{R}^{p\times m}\backslash\{0\}$. The definition of multivariate regression depth in the form of (2) first appeared in Mizera [26]. A very similar but earlier definition was proposed in Bern and Eppstein [3]. When $m = 1$, this is reduced to the univariate regression depth in Rousseeuw and Hubert [30]. When observations are sampled from (1), a robust estimator for B is defined as the maximizer of the empirical depth function. That is, $\widehat{B} = \text{argmax}_{B\in\mathcal{B}} \mathcal{D}_{\mathcal{U}}(B, \mathbb{P}_n)$, where $\mathbb{P}_n$ is the empirical measure of (1). With various choices of $\mathcal{B}$ and $\mathcal{U}$, we are able to estimate the regression parameters of all the seven problems listed above. The error rates are proved to be minimax optimal under the ϵ-contamination model.

The ϵ-contamination model was first proposed by Peter Huber [15]. Its properties have been studied by Huber [16], Huber and Strassen [18], Bickel [4], Donoho and Montanari [13] among others. Most early works studied Q with some assumptions. Some recent papers considered the ϵ-contamination model with Q allowed to be any distribution. To be specific, Chen et al. [7,8] showed that the minimax rate of recovering a parameter under the ϵ-contamination model takes a unified formula $\mathcal{R}(\epsilon) \asymp \mathcal{R}(0) \vee \omega(\epsilon, \Theta, L)$. In other words, the minimax rate is determined by two terms. The first term $\mathcal{R}(0)$ is the minimax rate without contamination, and $\omega(\epsilon, \Theta, L)$ is an extra term caused by contamination, where ϵ is the contamination proportion, Θ is the parameter space, and L is the loss function of the problem. Despite the progress of fundamental limits, efficient algorithms of estimation in ϵ-contamination models are usually very hard to find. Two recent papers Lai et al. [21], Diakonikolas et al. [10] proposed efficient algorithms based on the idea of "higher moment certificate" for robust mean estimation. The idea was later extended by Diakonikolas et al. [11], Balakrishnan et al. [2] for robust regression. These results can achieve near-optimal minimax rates under the contamination model in some special cases. Given the general hardness of computational issues, we will study computationally efficient robust regression algorithms under ϵ-contamination models in a separate paper.

Robust regression is a popular subject in statistics. However, most papers studied robust regression without considering an ϵ fraction of contamination (Huber [17], Siegel [32], Rousseeuw and Yohai [29], Leroy and Rousseeuw [31], Fan et al. [37]). The paper Loh and Tan [22] considered contamination, but in a different form from (1). Thus, the performance of many proposed procedures in the literature have not been tested under (1). An example in Chen et al. [8] shows that even procedures with high breakdown points may not achieve the optimal rate of the ϵ-contamination model. Conversely, Chen et al. [8] also shows that a good performance under the ϵ-contamination model must imply a high breakdown point. This serves as the main motivation to study robust regression using ϵ-contamination models. Though sparse linear regression and low-rank trace regression have already been studied in Chen et al. [7] under the ϵ-contamination model, the proposed procedure of Chen et al. [7] is based on robust testing and thus requires the assumption that the regression vector or matrix must have bounded ℓ_2 or Frobenius norm. In contrast, the estimator obtained by maximizing the regression depth does not require this assumption to achieve rate-optimality.

The rest of the paper is organized as follows. Section 2 reviews the definition and properties of the multivariate regression depth function. The applications in robust regression with one response variable are studied in Section 3. The applications in multivariate robust regression are studied in Section 4. Section 5 discusses some extensions of the results for elliptical distributions. A general notion of regression depth for learning linear operators is also discussed in that section. All technical proofs are given in Section 6.

We close this section by introducing the notation used in the paper. For $a, b \in \mathbb{R}$, let $a \vee b = \max(a, b)$ and $a \wedge b = \min(a, b)$. For an integer m, $[m]$ denotes the set $\{1, 2, \dots, m\}$. Given a set S, $|S|$ denotes its cardinality, and $\mathbb{I}_S$ is the associated indicator function. For two positive sequences $\{a_n\}$ and $\{b_n\}$, the relation $a_n \lesssim b_n$ means that $a_n \leq Cb_n$ for some constant $C > 0$, and $a_n \asymp b_n$ if both $a_n \lesssim b_n$ and $b_n \lesssim a_n$ hold. For a vector $v \in \mathbb{R}^p$, $\|v\|$ denotes the ℓ_2 norm, $\|v\|_1$ the ℓ_1 norm and $\text{supp}(v) = \{j \in [p] : v_j \neq 0\}$ is its support. For a matrix $A \in \mathbb{R}^{d_1 \times d_2}$, $\mathsf{rank}(A)$ denotes its rank, $\text{vec}(A)$ is its vectorization, $\|A\|_{\text{F}} = \|\text{vec}(A)\|$ is the matrix Frobenius norm, $\|A\|_{\ell_1} = \max_{1 \leq j \leq d_2} \sum_{i=1}^{d_1} |A_{ij}|$ is the matrix ℓ_1 norm, and the nuclear norm $\|A\|_{\text{N}}$ is its largest singular value. When A is an squared matrix, $\mathsf{Tr}(A)$ denotes its trace. For two matrices $A, B \in \mathbb{R}^{d_1 \times d_2}$, their trace inner product is $\langle A, B\rangle = \mathsf{Tr}(AB^T)$. For two probability distributions P_1 and P_2, their total variation distance is $\mathsf{TV}(P_1, P_2) = \sup_B |P_1(B) - P_2(B)|$. We use $\mathbb{P}$ and $\mathbb{E}$ to denote generic probability and expectation whose distribution is determined from the context.

2. The multivariate regression depth

For a joint probability distribution $\mathbb{P}$ of $X \in \mathbb{R}^p$ and $Y \in \mathbb{R}^m$, the multivariate regression depth of $B \in \mathbb{R}^{p \times m}$ is defined in (2). Even for m independent univariate regression problems, the multivariate regression depth treats the m regression problems in a joint way. Later, we will see this is essential to achieve optimal rates in Huber's ϵ-contamination models.

The multivariate regression depth function is a special case of tangent depth defined by Mizera [26]. A very closely related definition was considered in Bern and Eppstein [3]. Many important properties of the multivariate regression depth are discussed in Mizera [26]. For example, it is invariant with respect to linear transformations when $\mathcal{U} = \mathbb{R}^{p\times m}\backslash\{0\}$ in the sense that for any invertible $G \in \mathbb{R}^{p\times p}$ and $H \in \mathbb{R}^{m\times m}$,

$$\mathcal{D}_{\mathcal{U}}\big(B, \mathcal{L}(X,Y)\big) = \mathcal{D}_{\mathcal{U}}\big(G^{-1}BH^T, \mathcal{L}(GX, HY)\big),$$

where $\mathcal{L}(\cdot)$ denotes the law. We refer the readers to Mizera [26], Bern and Eppstein [3], Rousseeuw and Hubert [30], Struyf and Rousseeuw [33], Amenta et al. [1] for other important properties.

The general multivariate regression depth function covers some important cases. When $m = 1$, it is Rousseeuw and Hubert's univariate regression depth (Rousseeuw and Hubert [30]),

$$\mathcal{D}_{\mathcal{U}}(\beta, \mathbb{P}) = \inf_{u\in\mathcal{U}} \mathbb{P}\big\{u^T X\big(y - X^T\beta\big) \geq 0\big\}. \tag{3}$$

When $p = 1$ and the covariate is 1, it is Tukey's half-space depth (Tukey [35]) for multivariate location estimation,

$$\mathcal{D}_{\mathcal{U}}(\theta, \mathbb{P}) = \inf_{u\in\mathcal{U}} \mathbb{P}\big\{u^T(Y - \theta) \geq 0\big\}. \tag{4}$$

The error rate of maximizing Tukey's depth under the ϵ-contamination model was studied by Chen et al. [7]. Our main results for multivariate regression not only cover univariate regression depth, but also reproduce the result of Chen et al. [7] for Tukey's depth.

Section 3 and Section 4 study the error rates of the estimator

$$\widehat{B} = \operatorname*{argmax}_{B\in\mathcal{B}} \mathcal{D}_{\mathcal{U}}(B, \mathbb{P}_n) \tag{5}$$

for univariate and multivariate regression, respectively. To benchmark our main results, we need to introduce the general minimax lower bound for ϵ-contamination models obtained by Chen et al. [8].

Theorem 2.1 (Chen et al. [8]). *Given a statistical experiment $\{P_\theta : \theta \in \Theta\}$ and a loss function $L(\cdot,\cdot)$, define*

$$\omega(\epsilon, \Theta, L) = \sup\big\{L(\theta_1, \theta_2) : \mathsf{TV}(P_{\theta_1}, P_{\theta_2}) \leq \epsilon/(1-\epsilon); \theta_1, \theta_2 \in \Theta\big\}.$$

Suppose there is some $\mathcal{R}(0)$ such that

$$\inf_{\hat{\theta}} \sup_{\theta\in\Theta, Q} \mathbb{P}_{(\epsilon,\theta,Q)}\big\{L(\hat{\theta}, \theta) \geq \mathcal{R}(\epsilon)\big\} \geq c \tag{6}$$

holds for $\epsilon = 0$. Then, (6) also holds for any $\epsilon \in (0,1)$ with $\mathcal{R}(\epsilon) \asymp \mathcal{R}(0) \vee \omega(\epsilon, \Theta)$. The notation $\mathbb{P}_{(\epsilon,\theta,Q)}$ stands for $(1-\epsilon)P_\theta + \epsilon Q$.

Theorem 2.1 gives a general minimax lower bound for parameter estimation in the settings of ϵ-contamination models. The quantity $\omega(\epsilon, \Theta, L)$ is called modulus of continuity (Donoho and Liu [12]), which characterizes the ability of a loss function $L(\cdot,\cdot)$ to distinguish between two parameters whose corresponding probability distributions are $\epsilon/(1-\epsilon)$ close in total variation distance. The rate $\mathcal{R}(\epsilon) \asymp \mathcal{R}(0) \vee \omega(\epsilon, \Theta, L)$ is the best possible one that can be achieved by any procedure. For many loss functions, $\omega(\epsilon, \Theta, L)$ is at the order of ϵ^2. We will show that the estimator induced by the multivariate depth function is able to achieve the rate $\mathcal{R}(\epsilon) \asymp \mathcal{R}(0) \vee \omega(\epsilon, \Theta, L)$ for all the seven regression problems considered in the paper.

3. Applications of regression depth

3.1. Nonparametric regression

Consider the nonparametric regression model $y = f(x) + z$. To be specific, we use the distribution P_f to denote the sampling process that first sample $x \sim \text{Unif}[0, 1]$ and then sample $y|x \sim N(f(x), 1)$. The regression function admits the expansion $f(x) = \sum_{j=1}^{\infty} \beta_j \phi_j(x)$, where $\{\phi_j\}_{j=1}^{\infty}$ is the Fourier basis on $L^2[0, 1]$. We assume the true regression function belongs to the following Sobolev ball:

$$S_\alpha(M) = \left\{ f = \sum_{j=1}^{\infty} \beta_j \phi_j : \sum_{j=1}^{\infty} j^{2\alpha} \beta_j^2 \leq M^2 \right\}.$$

The smoothness parameter $\alpha > 0$ and radius $M > 0$ are assumed as constants throughout the section.

Define the vector of infinite size $X = \{\phi_j(x)\}_{j\in[\infty]} \in \mathbb{R}^{\infty}$. Then, the model becomes $y = \beta^T X + z$. Recovery of f is equivalent to recovery of $\beta \in \mathbb{R}^{\infty}$. Define

$$\mathcal{U}_k = \left\{ u \in \mathbb{R}^{\infty} \backslash \{0\} : u_j = 0 \text{ for all } j > k \right\}.$$

We use the univariate regression depth (3) to estimate the Fourier coefficients β by

$$\hat{\beta} = \underset{\beta \in \mathcal{U}_k}{\operatorname{argmax}}\, \mathcal{D}_{\mathcal{U}_k}\left(\beta, \left\{(X_i, y_i)\right\}_{i=1}^n\right). \tag{7}$$

To be specific, the empirical regression depth for this problem is

$$\mathcal{D}_{\mathcal{U}_k}\left(\beta, \left\{(X_i, y_i)\right\}_{i=1}^n\right) = \inf_{u\in\mathcal{U}_k} \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}\left\{ \left(\sum_{j=1}^{\infty} u_j \phi_j(x_i) \right) \left(y_i - \sum_{j=1}^{\infty} \beta_j \phi_j(x_i) \right) \geq 0 \right\}.$$

Since the regression function is in the space $S_\alpha(M)$, we expect that β_j's are negligible for high frequencies, and thus the regression depth does not need to involve frequencies after some level k.

We first give a result for the uniform convergence of the empirical regression depth.

Proposition 3.1. *For any probability measure $\mathbb{P}$ and its associated empirical measure $\mathbb{P}_n$, we have for any $\delta > 0$,*

$$\sup_{\beta\in\mathcal{U}_k}\left|\mathcal{D}_{\mathcal{U}_k}(\beta,\mathbb{P}_n)-\mathcal{D}_{\mathcal{U}_k}(\beta,\mathbb{P})\right|\leq C\sqrt{\frac{k}{n}}+\sqrt{\frac{\log(1/\delta)}{2n}},$$

with probability at least $1-2\delta$, where $C > 0$ is some absolute constant.

Using this result, we can study the convergence rate of the estimator (7) in the setting of the ϵ-contamination model. Namely, consider i.i.d. observations from $\mathbb{P}_{(\epsilon,f,Q)} = (1-\epsilon)P_f + \epsilon Q$.

Theorem 3.1. *Consider the estimator $\hat{f} = \sum_j \hat{\beta}_j\phi_j$ with $k = \lceil n^{\frac{1}{2\alpha+1}}\rceil$. Assume that $\epsilon^2 + n^{-\frac{2\alpha}{2\alpha+1}}$ is sufficiently small. Then, we have*

$$\|\hat{f}-f\|^2 = \int_0^1\left(\hat{f}(x)-f(x)\right)^2 dx \leq C\left(n^{-\frac{2\alpha}{2\alpha+1}}\vee\epsilon^2\right),$$

with $\mathbb{P}_{(\epsilon,f,Q)}$-probability at least $1-\exp(-C'(n^{\frac{1}{2\alpha+1}}+n\epsilon^2))$ uniformly over all Q and $f \in S_\alpha(M)$, where C, C' are some absolute constants.

The rate consists of two terms. The first term $n^{-\frac{2\alpha}{2\alpha+1}}$ is the classical minimax rate for nonparametric function estimation in the space $S_\alpha(M)$. See Tsybakov [34], Johnstone [19] for details. The second term ϵ^2 characterizes the influence of contamination. It is not hard to check that the modulus of continuity for the loss $\|\cdot\|^2$ is of order ϵ^2. Thus, the rate $n^{-\frac{2\alpha}{2\alpha+1}}\vee\epsilon^2$ is minimax optimal by Theorem 2.1.

Given that the minimax rate is $n^{-\frac{2\alpha}{2\alpha+1}}\vee\epsilon^2$, a necessary and sufficient condition to achieve the rate $n^{-\frac{2\alpha}{2\alpha+1}}$ as if there is no contamination is $\epsilon \lesssim n^{-\frac{\alpha}{2\alpha+1}}$. Hence, in order to achieve the minimax rate for $\epsilon = 0$, a rate-optimal robust estimator can tolerate at most $n\epsilon \lesssim n^{\frac{\alpha+1}{2\alpha+1}}$ contaminated observations. The number $n^{-\frac{\alpha}{2\alpha+1}}$ can be interpreted as the order of the minimax-rate breakdown point, because the minimax rate will change from $n^{-\frac{2\alpha}{2\alpha+1}}$ to ϵ^2 as soon as $\epsilon \gtrsim n^{-\frac{\alpha}{2\alpha+1}}$. It is interesting to note that a larger α implies a smaller order of $n^{-\frac{\alpha}{2\alpha+1}}$.

3.2. Sparse linear regression

Consider the sparse linear regression model, where the response and covariate are linked by the equation $y = \beta^T X + \sigma z$. The regression vector β is assumed to belong to the sparse set:

$$\Theta_s = \left\{\beta\in\mathbb{R}^p\backslash\{0\} : \sum_{j=1}^p \mathbb{I}\{\beta_j\neq 0\}\leq s\right\}. \tag{8}$$

The joint distribution $(X, y) \sim P_\beta$ is specified by the sampling process $X \sim N(0,\Sigma)$ and $y|X \sim N(\beta^T X, \sigma^2)$. For simplicity of notation, we suppress the dependence on Σ and σ^2 for P_β.

Using the univariate regression depth function (3), we define a sparse estimator by

$$\hat{\beta} = \operatorname*{argmax}_{\beta \in \Theta_s} \mathcal{D}_{\Theta_{2s}}\big(\beta, \{(X_i, y_i)\}_{i=1}^n\big). \tag{9}$$

We take advantage of the sparsity of the problem by setting $\mathcal{U} = \Theta_{2s}$ and $\mathcal{B} = \Theta_s$ in (5). For this sparse regression depth, its uniform convergence property is given by the following proposition.

Proposition 3.2. *For any probability measure $\mathbb{P}$ and its associated empirical measure $\mathbb{P}_n$, we have for any $\delta > 0$,*

$$\sup_{\beta \in \Theta_s} \big|\mathcal{D}_{\Theta_{2s}}(\beta, \mathbb{P}_n) - \mathcal{D}_{\Theta_{2s}}(\beta, \mathbb{P})\big| \leq C\sqrt{\frac{s \log(\frac{ep}{s})}{n}} + \sqrt{\frac{\log(1/\delta)}{2n}},$$

with probability at least $1 - 2\delta$, where $C > 0$ is some absolute constant.

Before giving the convergence rate of (9), we need to define the following quantity:

$$\kappa = \inf_{|\operatorname{supp}(v)|=2s} \frac{\|\Sigma^{1/2} v\|}{\|v\|}.$$

This is called restricted eigenvalue in sparse linear regression literature. Now we are ready for the main results. Consider i.i.d. observations from $\mathbb{P}_{(\epsilon,\beta,Q)} = (1-\epsilon)P_\beta + \epsilon Q$.

Theorem 3.2. *Consider the estimator $\hat{\beta}$. Assume that $\epsilon^2 + \frac{s\log(\frac{ep}{s})}{n}$ is sufficiently small. Then, we have*

$$\|\hat{\beta} - \beta\|_\Sigma^2 = \big\|\Sigma^{1/2}(\hat{\beta} - \beta)\big\|^2 \leq C\sigma^2\bigg(\frac{s \log(\frac{ep}{s})}{n} \vee \epsilon^2\bigg), \tag{10}$$

$$\|\hat{\beta} - \beta\|^2 \leq C\frac{\sigma^2}{\kappa^2}\bigg(\frac{s \log(\frac{ep}{s})}{n} \vee \epsilon^2\bigg), \tag{11}$$

$$\|\hat{\beta} - \beta\|_1^2 \leq C\frac{\sigma^2}{\kappa^2}\bigg(\frac{s^2 \log(\frac{ep}{s})}{n} \vee s\epsilon^2\bigg), \tag{12}$$

with $\mathbb{P}_{(\epsilon,\beta,Q)}$-probability at least $1 - \exp(-C'(s\log(\frac{ep}{s}) + n\epsilon^2))$ uniformly over all Q and $\beta \in \Theta_s$, where C, C' are some absolute constants.

The rates are given in prediction loss, squared ℓ_2 loss and squared ℓ_1 loss, respectively. The rate for the prediction loss does not depend on the covariance of the covariates Σ. On the other hand, the rates for the squared ℓ_2 loss and the squared ℓ_1 loss depend on Σ through a κ^{-2} factor.

These rates were also obtained by Chen et al. [7] under the ϵ-contamination model with a testing-based estimator. However, their results only hold for a subset of Θ_s. In particular, they need to further impose two extra assumptions that $\|\beta\|$ is bounded by the order of σ/κ and the largest $2s$-sparse eigenvalue of Σ is at the order of κ. In contrast, Theorem 3.2 removes these two assumptions and the convergence rates hold uniformly for all $\beta \in \Theta_s$.

When $\epsilon = 0$, the rates obtained in Theorem 3.2 are all minimax optimal by Ye and Zhang [40], Raskutti et al. [27], Verzelen [36]. Though most lower bound results in the literature for sparse linear regression are for fixed design. They can be easily modified into the random design setting considered here. The details are referred to the related discussion in Raskutti et al. [27], Chen et al. [7].

For a general $\epsilon > 0$, it is direct to check that

$$\omega\big(\epsilon, \Theta_s, \|\cdot\|_\Sigma^2\big) \asymp \sigma^2\epsilon^2,$$

$$\omega\big(\epsilon, \Theta_s, \|\cdot\|^2\big) \asymp \frac{\sigma^2\epsilon^2}{\kappa^2},$$

$$\omega\big(\epsilon, \Theta_s, \|\cdot\|_1^2\big) \asymp \frac{s\sigma^2\epsilon^2}{\kappa^2}.$$

Thus, by Theorem 2.1, the rates are also minimax optimal for $\epsilon > 0$.

Theorem 3.2 and the minimax lower bound of the problem shows that the minimax rates are determined by the trade-off between $\frac{s\log(\frac{ep}{s})}{n}$ and ϵ^2. When $\epsilon^2 \lesssim \frac{s\log(\frac{ep}{s})}{n}$, the term $\frac{s\log(\frac{ep}{s})}{n}$ dominates, and the minimax rates are the same as those for $\epsilon = 0$. In this regime, the contamination has no effect on the minimax rates. Note that $\epsilon \lesssim \frac{s\log(\frac{ep}{s})}{n}$ means that a rate-optimal estimator is able to tolerate at most $n\epsilon \lesssim \sqrt{ns\log(\frac{ep}{s})}$ contaminated observations before the minimax rate is changed. It is interesting that $\sqrt{ns\log(\frac{ep}{s})}$ is an increasing function of the sparsity level s. Similar remarks also apply to the other regression problems considered in the paper.

3.3. Gaussian graphical model

In this section, we consider the Gaussian graphical model $P_\Omega = N(0, \Omega^{-1})$. The precision matrix Ω belongs to the following sparse class:

$$\mathcal{F}_s(M) = \left\{\Omega = \Omega^T \in \mathbb{R}^{p\times p} : M^{-1} \leq \lambda_{\min}(\Omega) \leq \lambda_{\max}(\Omega) \leq M, \max_{1\leq i\leq p}\sum_{j=1}^p \mathbb{I}\{\Omega_{ij} \neq 0\} \leq s\right\}.$$

The notation $\lambda_{\min}(\cdot)$ and $\lambda_{\max}(\cdot)$ stand for the smallest and the largest eigenvalues. This class was previously considered in Ren et al. [28]. We assume the number M is a constant throughout this section.

For a random vector $X \sim N(0, \Omega^{-1})$, the sparsity pattern of Ω characterizes the graphical model of conditional dependence. In particular, $\Omega_{ij} = 0$ if and only if X_i is independent of X_j given all remaining variables.

Moreover, there is simple linear model that links X_j and X_{-j}, where we use X_{-j} to denote the $(p-1)$-dimensional subvector of X excluding the jth variable. Define $\beta_{(j)} = -\Omega_{jj}^{-1}\Omega_{-j,j}$, then

$$X_j = \beta_{(j)}^T X_{-j} + \xi_j, \tag{13}$$

where the noise has distribution $\xi_j \sim N(0, \Omega_{jj}^{-1})$ and is independent of X_{-j}. Methods based on (13) are proposed in the literature to estimate Ω. See Meinshausen and Bühlmann [25], Yuan [41] for some examples.

With i.i.d. observations from $\mathbb{P}_{(\epsilon,\Omega,Q)}$, we discuss how to explore the linear model (13) to estimate the precision matrix Ω in a robust way. For each $j \in [n]$, we need to estimate $\beta^{(j)}$ and the variance of ξ_j, which is Ω_{jj}^{-1}, respectively. Without loss of generality, assume the sample size n is even. We split the data into two halves. We use the first half to estimate $\beta_{(j)}$ by

$$\hat{\beta}_{(j)} = \operatorname*{argmax}_{\beta \in \Theta_s} \mathcal{D}_{\Theta_{2s}}\big(\beta, \big\{(X_{-j,i}, X_{ji})\big\}_{i=1}^{n/2}\big).$$

The set Θ_s is defined in (8) with the dimension p replaced by $p-1$. The convergence rate of $\hat{\beta}_{(j)}$ is given by Theorem 3.2. We then use the second half of the data together with $\hat{\beta}_{(j)}$ to estimate the variance of ξ_j. For each $i = n/2+1, \dots, n$, define the residue

$$w_{ji} = \big(X_{ji} - \hat{\beta}_{(j)}^T X_{-j,i}\big)^2.$$

Then, we estimate Ω_{jj}^{-1} by median absolute deviation,

$$\widehat{\Omega}_{jj}^{-1} = \frac{\operatorname{Median}(\{w_{ji}\}_{i=n/2+1}^{n})}{[\Phi^{-1}(3/4)]^2},$$

where $\Phi(\cdot)$ is the cumulative distribution function of $N(0,1)$. The convergence rate of $\widehat{\Omega}_{jj}^{-1}$ is given by Chen et al. [8].

Now we are ready to define the estimator of the precision matrix Ω by assembling all pieces. For the jth column, its jth entry is estimated by $\widehat{\Omega}_{jj}$. The remaining entries are estimated by $\widehat{\Omega}_{-j,j} = -\widehat{\Omega}_{jj}\hat{\beta}_{(j)}$. The convergence rate of the estimator $\widehat{\Omega}$ is given by the following theorem.

Theorem 3.3. *Consider the estimator $\widehat{\Omega}$. Assume that $\epsilon^2 + \frac{s\log(\frac{ep}{s})}{n}$ is sufficiently small. Then, we have*

$$\|\widehat{\Omega} - \Omega\|_{\ell_1}^2 \le C\left(\frac{s^2\log(\frac{ep}{s})}{n} \vee s\epsilon^2\right),$$

with $\mathbb{P}_{(\epsilon,\Omega,Q)}$-probability at least $1 - \exp(-C'(s\log(\frac{ep}{s}) + n\epsilon^2))$ uniformly over all Q and $\Omega \in \mathcal{F}_s(M)$, where C, C' are some absolute constants.

Theorem 3.3 gives the error rate of $\widehat{\Omega}$ in terms of squared matrix ℓ_1 norm. Note that the estimator $\widehat{\Omega}$ may not be positive semidefinite. A simple projection step discussed in Yuan [41] leads to a positive semidefinite estimator with the same error rate.

The minimax lower bound of the problem is given by the following theorem.

Theorem 3.4. *Assume $p > c_1 n^{\beta}$ for some constants $\beta > 1$ and $c_1 > 0$, and $\frac{s^2(\log p)^3}{n}$ is sufficiently small. Then,*

$$\inf_{\widehat{\Omega}} \sup_{\Omega\in\mathcal{F}_s(M),Q} \mathbb{P}_{(\epsilon,\Omega,Q)}\left\{\|\widehat{\Omega}-\Omega\|_{\ell_1}^2 > C\left(\frac{s^2\log p}{n} \vee s\epsilon^2\right)\right\} \geq c,$$

for some constants $C, c > 0$.

Proof. By Theorem 2.1, the minimax lower is in the form of $\mathcal{R}(0) \vee \omega(\epsilon, \mathcal{F}_s(M), \|\cdot\|_{\ell_1}^2)$. The first term $\mathcal{R}(0)$ has order $\frac{s^2\log p}{n}$, which was proved in Cai et al. [6]. Direct calculation gives the order of the second term $\omega(\epsilon, \mathcal{F}_s(M), \|\cdot\|_{\ell_1}^2) \asymp s\epsilon^2$. □

Combining the conclusions of Theorem 3.3 and Theorem 3.4, we conclude that the minimax rate for estimating Ω under the squared matrix ℓ_1 norm in the setting of ϵ-contamination model is $\frac{s^2\log p}{n} \vee s\epsilon^2$. Moreover, the estimator $\widehat{\Omega}$ based on regression depth is able to achieve the minimax rate.

3.4. Low-rank trace regression

The goal of trace regression is to recover a low-rank matrix $B \in \mathbb{R}^{p_1\times p_2}$ from noisy linear observations specified by the model $y = \mathsf{Tr}(X^T B) + \sigma z$. We denote by P_B the joint distribution of $X \in \mathbb{R}^{p_1\times p_2}$ and $y \in \mathbb{R}$ that follows $\mathrm{vec}(X) \sim N(0,\Sigma)$ and $y|X \sim N(\mathsf{Tr}(X^T B), \sigma^2)$. Again, the dependence on Σ and σ^2 is suppressed for the notation P_B. The matrix B is assumed to belong to the following set:

$$\mathcal{A}_r = \left\{B \in \mathbb{R}^{p_1\times p_2}\backslash\{0\} : \mathrm{rank}(B) \leq r\right\}. \tag{14}$$

The univariate regression depth (3) can be easily adapted to the trace regression problem. That is,

$$\mathcal{D}_{\mathcal{U}}(B,\mathbb{P}) = \inf_{U\in\mathcal{U}} \mathbb{P}\left\{\langle U, X\rangle\left(y - \langle B, X\rangle\right) \geq 0\right\},$$

where $\mathcal{U} \subset \mathbb{R}^{p_1\times p_2}$. We take advantage of the low-rank assumption, and define the estimator by

$$\widehat{B} = \operatorname*{argmax}_{B\in\mathcal{A}_r} \mathcal{D}_{\mathcal{A}_{2r}}\left(B, \{X_i, y_i\}_{i=1}^n\right). \tag{15}$$

We first present a uniform convergence result for regression depth with a low-rank structure.

Proposition 3.3. *For any probability measure $\mathbb{P}$ and its associated empirical measure $\mathbb{P}_n$, we have for any $\delta > 0$,*

$$\sup_{B\in\mathcal{A}_r} \left|\mathcal{D}_{\mathcal{A}_{2r}}(B,\mathbb{P}_n) - \mathcal{D}_{\mathcal{A}_{2r}}(B,\mathbb{P})\right| \leq C\sqrt{\frac{r(p_1+p_2)}{n}} + \sqrt{\frac{\log(1/\delta)}{2n}},$$

with probability at least $1-2\delta$, where $C > 0$ is some absolute constant.

With the uniform convergence of empirical depth, we can determine the convergence rate of the estimator (15). To facilitate the presentation, we define the following quantity:

$$\kappa = \inf_{\operatorname{rank}(A)=2r} \frac{\|\Sigma^{1/2}\operatorname{vec}(A)\|}{\|A\|_{\mathrm{F}}}.$$

Now, consider i.i.d. observations from $\mathbb{P}_{(\epsilon,B,Q)} = (1-\epsilon)P_B + \epsilon Q$.

Theorem 3.5. *Consider the estimator $\widehat{B}$. Assume that $\epsilon^2 + \frac{r(p_1+p_2)}{n}$ is sufficiently small. Then, we have*

$$\big\|\Sigma^{1/2}\big(\operatorname{vec}(\widehat{B}-B)\big)\big\|^2 \le C\sigma^2\bigg(\frac{r(p_1+p_2)}{n} \vee \epsilon^2\bigg),$$

$$\|\widehat{B}-B\|_{\mathrm{F}}^2 \le C\frac{\sigma^2}{\kappa^2}\bigg(\frac{r(p_1+p_2)}{n} \vee \epsilon^2\bigg),$$

$$\|\widehat{B}-B\|_{\mathrm{N}}^2 \le C\frac{\sigma^2}{\kappa^2}\bigg(\frac{r^2(p_1+p_2)}{n} \vee r\epsilon^2\bigg),$$

with $\mathbb{P}_{(\epsilon,B,Q)}$-probability at least $1-\exp(-C'(r(p_1+p_2)+n\epsilon^2))$ uniformly over all Q and $B \in \mathcal{A}_r$, where C, C' are some absolute constants.

Similar to Theorem 3.2, Theorem 3.5 gives rates for prediction loss, squared Frobenius loss and squared nuclear loss, respectively. When $\epsilon = 0$, the three rates are all minimax optimal by Koltchinskii et al. [20]. To see the optimality for $\epsilon > 0$, note that

$$\omega\big(\epsilon, \mathcal{A}_r, \|\cdot\|_\Sigma^2\big) \asymp \sigma^2\epsilon^2,$$

$$\omega\big(\epsilon, \mathcal{A}_r, \|\cdot\|_{\mathrm{F}}^2\big) \asymp \frac{\sigma^2\epsilon^2}{\kappa^2},$$

$$\omega\big(\epsilon, \mathcal{A}_r, \|\cdot\|_{\mathrm{N}}^2\big) \asymp \frac{r\sigma^2\epsilon^2}{\kappa^2}.$$

Thus, by Theorem 2.1, the rates are all minimax optimal.

Results in Chen et al. [7] gave the same rate for trace regression in the setting of ϵ-contamination model. However, they required extra assumptions such as the boundedness of $\|B\|_{\mathrm{F}}$ and of $\|\Sigma\|_{\mathrm{op}}$. In contrast, Theorem 3.5 achieves the minimax rate of the problem without these extra assumptions.

4. Applications of multivariate regression depth

4.1. Multivariate linear regression

Starting from this section, we consider regression problems with multiple responses in the setting of ϵ-contamination model. Consider the model $Y = B^T X + \sigma Z$, where $B \in \mathbb{R}^{p\times m}$. We use P_B

to denote the joint distribution of $X \in \mathbb{R}^p$ and $Y \in \mathbb{R}^m$ specified by $X \sim N(0, \Sigma)$ and $Y|X \sim N(B^T X, \sigma^2 I_m)$. Again, the dependence on Σ and σ^2 is suppressed for the notation P_B. We use the multivariate regression depth (2) for estimating B. The estimator is defined as

$$\widehat{B} = \operatorname*{argmax}_{B \in \mathbb{R}^{p\times m}} \mathcal{D}_{\mathbb{R}^{p\times m}\setminus\{0\}}\big(B, \{X_i, Y_i\}_{i=1}^n\big). \tag{16}$$

Intuitively, the m univariate regression problems are independent, and one can estimate every column of B separately. The rates are optimal when there is no contamination. However, we will show that this strategy does not lead to rate optimality in the setting of ϵ-contamination model.

The uniform convergence of the multivariate regression depth is given by the following proposition.

Proposition 4.1. *For any probability measure $\mathbb{P}$ and its associated empirical measure $\mathbb{P}_n$, we have for any $\delta > 0$,*

$$\sup_{B \in \mathbb{R}^{p\times m}} \left|\mathcal{D}_{\mathbb{R}^{p\times m}\setminus\{0\}}(B, \mathbb{P}_n) - \mathcal{D}_{\mathbb{R}^{p\times m}\setminus\{0\}}(B, \mathbb{P})\right| \leq C\sqrt{\frac{pm}{n}} + \sqrt{\frac{\log(1/\delta)}{2n}},$$

with probability at least $1 - 2\delta$, where $C > 0$ is some absolute constant.

Then, define the quantity

$$\kappa = \inf_{v \neq 0} \frac{\|\Sigma^{1/2} v\|}{\|v\|}. \tag{17}$$

With Proposition 4.1 and the definition of κ, we are ready to present the main result. Consider i.i.d. observations from $\mathbb{P}_{(\epsilon, B, Q)} = (1 - \epsilon) P_B + \epsilon Q$.

Theorem 4.1. *Consider the estimator $\widehat{B}$. Assume that $\epsilon^2 + \frac{pm}{n}$ is sufficiently small. Then, we have*

$$\mathsf{Tr}\big((\widehat{B} - B)^T \Sigma (\widehat{B} - B)\big) \leq C\sigma^2\left(\frac{pm}{n} \vee \epsilon^2\right), \tag{18}$$

$$\|\widehat{B} - B\|_{\mathrm{F}}^2 \leq C\frac{\sigma^2}{\kappa^2}\left(\frac{pm}{n} \vee \epsilon^2\right), \tag{19}$$

with $\mathbb{P}_{(\epsilon, B, Q)}$-probability at least $1 - \exp(-C'(pm + n\epsilon^2))$ uniformly over all Q and $B \in \mathbb{R}^{p\times m}$, where C, C' are some absolute constants.

We first remark that the rates for both prediction loss and squared Frobenius loss are minimax optimal. This can be easily seen from Theorem 2.1 and classical multivariate regression results in the literature.

One can also use univariate regression depth to estimate each column of B separately. This leads to the rates $\sigma^2(\frac{pm}{n} \vee (m\epsilon^2))$ and $\frac{\sigma^2}{\kappa^2}(\frac{pm}{n} \vee (m\epsilon^2))$ for the two loss functions, respectively.

Both rates are clearly sub-optimal because of the extra factor of m before ϵ^2. Therefore, in the setting of ϵ-contamination model, even when there is no structural dependence between the columns of B, the matrix B needs to be estimated jointly.

When $p = 1$, the multivariate regression depth is closely related to Tukey's halfspace depth (4). The rate of convergence of Tukey's median was studied by Chen et al. [8] in the setting of ϵ-contamination model. Theorem 4.1 can be viewed as an extension of their result for $p > 1$.

4.2. Multivariate linear regression with group sparsity

We extend the multivariate regression problem $Y = B^T X + \sigma Z$ to a group sparse setting. The regression matrix B is assumed to belong to the following space

$$\Xi_s = \left\{ B \in \mathbb{R}^{p\times m} \backslash \{0\} : \sum_{j=1}^{p} \mathbb{I}\{B_{j*} \neq 0\} \leq s \right\}.$$

The notation B_{j*} stands for the jth row of the matrix B. We take advantage of the group sparse structure and define the estimator by

$$\widehat{B} = \operatorname*{argmax}_{B\in\Xi_s} \mathcal{D}_{\Xi_{2s}}\left(B, \left\{(X_i, Y_i)\right\}_{i=1}^{n}\right).$$

The uniform convergence of the multivariate regression depth with group sparse structure is given by the following proposition.

Proposition 4.2. *For any probability measure $\mathbb{P}$ and its associated empirical measure $\mathbb{P}_n$, we have for any $\delta > 0$,*

$$\sup_{B\in\Xi_s} \left|\mathcal{D}_{\Xi_{2s}}(B, \mathbb{P}_n) - \mathcal{D}_{\Xi_{2s}}(B, \mathbb{P})\right| \leq C\sqrt{\frac{ms + s\log(\frac{ep}{s})}{n}} + \sqrt{\frac{\log(1/\delta)}{2n}},$$

with probability at least $1 - 2\delta$, where $C > 0$ is some absolute constant.

Proposition 4.2 is an extension of both Proposition 3.2 and Proposition 4.1. The rate consists of two parts. The first part $\frac{ms}{n}$ is determined by the number of parameters. Since there are only s nonzero rows of B, the number of parameters is ms. The second part $\frac{s\log(\frac{ep}{s})}{n}$ is determined by the model selection complexity. Given the sparsity level s, there are $\binom{p}{s}$ possible models with different row supports. This contributes to the rate $n^{-1}\log\binom{p}{s} \asymp \frac{s\log(\frac{ep}{s})}{n}$.

Define the quantity

$$\kappa = \inf_{|\operatorname{supp}(v)|=2s} \frac{\|\Sigma^{1/2}v\|}{\|v\|}. \tag{20}$$

We are now ready to give the main result. Consider i.i.d. observations from $\mathbb{P}_{(\epsilon,B,Q)} = (1 - \epsilon)P_B + \epsilon Q$.

Theorem 4.2. *Consider the estimator* $\widehat{B}$. *Assume that* $\epsilon^2 + \frac{ms+s\log(\frac{ep}{s})}{n}$ *is sufficiently small. Then, we have*

$$\mathsf{Tr}\big((\widehat{B}-B)^T\Sigma(\widehat{B}-B)\big) \leq C\sigma^2\left(\frac{ms+s\log(\frac{ep}{s})}{n} \vee \epsilon^2\right), \tag{21}$$

$$\|\widehat{B}-B\|_{\mathrm{F}}^2 \leq C\frac{\sigma^2}{\kappa^2}\left(\frac{ms+s\log(\frac{ep}{s})}{n} \vee \epsilon^2\right), \tag{22}$$

with $\mathbb{P}_{(\epsilon,B,Q)}$*-probability at least* $1-\exp(-C'(ms+s\log(\frac{ep}{s})+n\epsilon^2))$ *uniformly over all* Q *and* $B \in \Xi_s$, *where* C, C' *are some absolute constants.*

Theorem 4.2 is an extension of both Theorem 3.2 and Theorem 4.1. When $m = 1$, the problem is reduced to sparse linear regression, and $\widehat{B}$ in (20) is the same as $\hat{\beta}$ in (9). Thus, the rates given by Theorem 4.2 recovers those of Theorem 3.2. When $s = 1$, this is the setting of multivariate linear regression without the group sparse structure, and the rates of Theorem 4.2 recover those of Theorem 4.1.

The rates given by Theorem 4.2 are minimax optimal by Theorems 2.1 and Lounici et al. [23].

4.3. Reduced rank regression

The final application is for reduced rank regression. In the multivariate linear regression setting $Y = B^TX + \sigma Z$, the regression matrix B is assumed to be low-rank. In particular, $B \in \mathcal{A}_r$, where $\mathcal{A}_r$ is defined in (14), except that the dimension $p_1 \times p_2$ is replaced by $p \times m$. Some recent progresses on this topic were made by Bunea et al. [5], Ma and Sun [24] and references therein.

Define the estimator by

$$\widehat{B} = \operatorname*{argmax}_{B\in\mathcal{A}_r} \mathcal{D}_{\mathcal{A}_{2r}}\big(B, \{(X_i, Y_i)\}_{i=1}^n\big). \tag{23}$$

We give the uniform convergence of the empirical depth function.

Proposition 4.3. *For any probability measure* $\mathbb{P}$ *and its associated empirical measure* $\mathbb{P}_n$, *we have for any* $\delta > 0$,

$$\sup_{B\in\mathcal{A}_r} \big|\mathcal{D}_{\mathcal{A}_{2r}}(B,\mathbb{P}_n) - \mathcal{D}_{\mathcal{A}_{2r}}(B,\mathbb{P})\big| \leq C\sqrt{\frac{r(p+m)}{n}} + \sqrt{\frac{\log(1/\delta)}{2n}},$$

with probability at least $1-2\delta$, *where* $C > 0$ *is some absolute constant.*

Note that Proposition 4.3 is an extension of Proposition 4.1. For a full rank matrix, $r = p \wedge m$, and therefore $r(p+m) \asymp pm$.

To present the error rate of (23), define the quantity

$$\kappa = \inf_{v \neq 0} \frac{\|\Sigma^{1/2} v\|}{\|v\|}.$$

Consider i.i.d. observations from $\mathbb{P}_{(\epsilon, B, Q)} = (1-\epsilon) P_B + \epsilon Q$.

Theorem 4.3. *Consider the estimator $\widehat{B}$. Assume that $\epsilon^2 + \frac{r(p+m)}{n}$ is sufficiently small. Then, we have*

$$\mathsf{Tr}\big((\widehat{B} - B)^T \Sigma (\widehat{B} - B)\big) \le C\sigma^2 \left(\frac{r(p+m)}{n} \vee \epsilon^2\right),$$

$$\|\widehat{B} - B\|_{\mathrm{F}}^2 \le C \frac{\sigma^2}{\kappa^2} \left(\frac{r(p+m)}{n} \vee \epsilon^2\right),$$

$$\|\widehat{B} - B\|_{\mathrm{N}}^2 \le C \frac{\sigma^2}{\kappa^2} \left(\frac{r^2(p+m)}{n} \vee r\epsilon^2\right),$$

with $\mathbb{P}_{(\epsilon, B, Q)}$-probability at least $1 - \exp(-C'(r(p+m) + n\epsilon^2))$ uniformly over all Q and $B \in \mathcal{A}_r$, where C, C' are some absolute constants.

Theorem 4.3 gives rates in terms of prediction loss, squared Frobenius loss and squared nuclear loss. The rates are identical to those of Theorem 3.5 for low-rank trace regression, with $p+m$ corresponding to $p_1 + p_2$ in Theorem 3.5. This is due to the similarity of the two problems. In both problems, the regression matrix B is assumed to belong to the low-rank set $\mathcal{A}_r$. The only difference is that for trace regression, the response is univariate and the covariate is a matrix, and for reduced rank regression, the response is multivariate and the covariate is a vector.

Applying the minimax lower bounds in Ma and Sun [24], we find that the rates given by Theorem 4.3 are optimal when $\epsilon = 0$. Though the lower bounds in Ma and Sun [24] are for a fixed design setting and they did not give explicit dependence on κ, the results can be easily modified to the random design setting considered here. The dependence on κ can be made explicit as well. We refer the readers to the discussion in Raskutti et al. [27], Chen et al. [7] for details. In addition, Theorem 2.1 and similar calculations of modulus of continuity as in Section 3.4 imply that the rates given by Theorem 4.3 are also optimal when $\epsilon > 0$.

5. Discussion

5.1. Extension to general error distributions

The error distributions we consider in Section 3 and Section 4 are all Gaussian. This assumption can be greatly relaxed. In this section, we consider error distributions that have elliptical shapes.

Definition 5.1. A random vector $W \sim EC(0, \Gamma, F)$ is distributed according to a centered continuous elliptical distribution with a scatter matrix $\Gamma \in \mathbb{R}^{d \times d}$ and a marginal cumulative distribution function F if and only if $W = \Gamma^{1/2}E$, and $F(t) = \mathbb{P}(u^T E \leq t\|u\|)$ does not depend on $u \in \mathbb{R}^d$. Moreover, there is a density function f, such that $f(0) = 1$ and $F(t) = \int_{-\infty}^{t} f(s)\, ds$.

A more general definition of elliptical distributions is referred to Fang et al. [14]. Here, we only consider those that have marginal densities. Without loss of generality, we impose the constraint that $f(0) = 1$. Otherwise, the scatter matrix Γ would only be defined up to a multiplicative factor. When the dimension is 1, the definition covers all random variables with symmetric density functions centered at zero.

Consider the setting of multivariate linear regression in Section 4.1. The regression model P_B for $X \in \mathbb{R}^p$ and $Y \in \mathbb{R}^m$ is specified by the sampling process $X \sim N(0, \Sigma)$ and $(Y - B^T X)|X \sim EC(0, \Gamma, F)$. The dependence on Σ, Γ, F are suppressed in the notation of P_B. For i.i.d. observations generated by $\mathbb{P}_{(\epsilon,B,Q)} = (1-\epsilon)P_B + \epsilon Q$, the results of Theorem 4.1 are extended to the following theorem.

Theorem 5.1. *Consider the estimator $\widehat{B}$ defined in (16). Assume that $\epsilon^2 + \frac{pm}{n}$ is sufficiently small. Moreover, there are some positive constants c_1 and c_2 such that $\min_{|t|\leq c_1} f(t) \geq c_2$. Then, we have*

$$\mathsf{Tr}\big((\widehat{B} - B)^T \Sigma(\widehat{B} - B)\big) \leq C\sigma^2\left(\frac{pm}{n} \vee \epsilon^2\right), \tag{24}$$

$$\|\widehat{B} - B\|_{\mathrm{F}}^2 \leq C\frac{\sigma^2}{\kappa^2}\left(\frac{pm}{n} \vee \epsilon^2\right), \tag{25}$$

with $\mathbb{P}_{(\epsilon,B,Q)}$-probability at least $1 - \exp(-C'(pm + n\epsilon^2))$ uniformly over all Q and $B \in \mathbb{R}^{p\times m}$, where κ is defined in (17), $\sigma^2 = \|\Gamma\|_{\mathrm{op}}$ and C, C' are some absolute constants.

In addition to Theorem 4.1, all the other results (except those of Gaussian graphical model) in Section 3 and Section 4 can be extended to the setting of general elliptical error distributions. The results are the same and thus the details are omitted. Theorem 5.1 implies that the regression depth maximizer is not only robust to contamination, but is also robust to general error distributions.

The results can also be extended to error distributions that are not symmetric. For a univariate error distribution, this corresponds to robust median regression under the contamination model. However, the results would be less interpretable for a multivariate asymmetric error distribution.

Besides the error distribution, the Gaussian assumption for the covariates can also be extended similarly. This requires significantly more technical details and there are more than one way to do it. We therefore do not explore all the possibilities here.

5.2. A general notion of depth for linear operators

Consider a general covariate space $\mathcal{X}$ and a general response space $\mathcal{Y}$. We assume the response space $\mathcal{Y}$ is a Hilbert space equipped with an inner product $\langle\cdot,\cdot\rangle$. Let $\ell(\mathcal{X},\mathcal{Y})$ be a class of linear operators $f : \mathcal{X} \to \mathcal{Y}$. The inner product structure on the response space allows us to define a gen-

Table 1. Examples of Depth Functions

	$\mathcal{X}$	$\mathcal{Y}$
Tukey's depth (Tukey [35])	$\{1\}$	$\mathbb{R}^m$
regression depth (Rousseeuw and Hubert [30])	$\mathbb{R}^p$	$\mathbb{R}$
multivariate regression depth (Bern and Eppstein [3], Mizera [26])	$\mathbb{R}^p$	$\mathbb{R}^m$
depth for functional linear regression	$\mathcal{C}[0,1]$	$\mathbb{R}$
depth for multivariate functional linear regression	$\mathcal{C}[0,1]$	$\mathbb{R}^m$

eral depth function for a linear operator $f \in \ell(\mathcal{X}, \mathcal{Y})$. Given a probability distribution $(X, Y) \sim \mathbb{P}$ on $\mathcal{X} \times \mathcal{Y}$, the depth of an $f \in \ell(\mathcal{X}, \mathcal{Y})$ is defined as

$$\mathcal{D}_{\mathcal{G}}(f, \mathbb{P}) = \inf_{g \in \mathcal{G}} \mathbb{P}\big\{\big\langle g(X), Y - f(X)\big\rangle \geq 0\big\},$$

where $\mathcal{G}$ is a subset of $\ell(\mathcal{X}, \mathcal{Y})$.

This general definition not only covers the multivariate regression depth studied in this paper, but also allows the covariate to be a function. Some special cases are listed in Table 1. When $\mathcal{X} \times \mathcal{Y}$ takes $\{1\} \times \mathbb{R}^m$, $\mathbb{R}^p \times \mathbb{R}$ and $\mathbb{R}^p \times \mathbb{R}^m$, respectively, we recover Tukey's depth, regression depth and multivariate regression depth. Moreover, when $\mathcal{X}$ takes the class of all continuous functions on the unit interval $\mathcal{C}[0, 1]$, the depth function can be used for robust functional linear regression. This application will be considered in future projects.

6. Proofs

This section collects the proofs of the results presented in the paper. Section 6.1 proves uniform convergence of all the empirical depth functions used in the paper. This includes the proofs of Propositions 3.1, 3.2, 3.3, 4.1, 4.2 and 4.3. Section 6.2 establishes the curvature of the population depth functions. Finally, in Section 6.3, we prove all the theorems in the paper.

6.1. Uniform convergence of the empirical depth functions

To establish uniform convergence of the empirical depth functions, it is essential to bound $\sup_{A \in \mathcal{A}} |\mathbb{P}_n(A) - \mathbb{P}(A)|$ over a collection $\mathcal{A}$. The first step is to use McDiarmid's bounded difference inequality. The following version can be found in Chapter 3.1 of Devroye and Lugosi [9].

Lemma 6.1. *For any probability measure $\mathbb{P}$ and its associated empirical measure $\mathbb{P}_n$, we have for any $t > 0$,*

$$\sup_{A \in \mathcal{A}} \big|\mathbb{P}_n(A) - \mathbb{P}(A)\big| \leq \mathbb{E}\Big\{\sup_{A \in \mathcal{A}} \big|\mathbb{P}_n(A) - \mathbb{P}(A)\big|\Big\} + t,$$

with probability at least $1 - 2e^{-2nt^2}$.

By Lemma 6.1, it is sufficient to bound the expectation $\mathbb{E}\{\sup_{A\in\mathcal{A}}|\mathbb{P}_n(A)-\mathbb{P}(A)|\}$. This quantity can be controlled by the VC dimension of $\mathcal{A}$. The following lemma can be found in Chapter 4.3 of Devroye and Lugosi [9].

Lemma 6.2. *For any class $\mathcal{A}$ with VC dimension V,*

$$\mathbb{E}\Big\{\sup_{A\in\mathcal{A}}\big|\mathbb{P}_n(A)-\mathbb{P}(A)\big|\Big\}\leq C\sqrt{\frac{V}{n}},$$

where $C>0$ is a universal constant.

Lemma 6.2 suggests that we need to give an upper bound for the VC dimension of the class $\mathcal{A}$. For the depth functions considered in the paper, the relevant class is

$$\mathcal{A}=\big\{\big\{Z\in\mathbb{R}^{d_1\times d_2}:\mathsf{Tr}\big(WZ^T\big)\geq 0\big\}:W\in\mathbb{R}^{d_1\times d_2},\operatorname{rank}(W)\leq r\big\}. \tag{26}$$

Intuitively, the matrix W in $\mathcal{A}$ defined above has at most $r(d_1+d_2)$ degrees of freedom, which suggests a VC dimension bound $r(d_1+d_2)$. It was shown by Wolf et al. [39] that the VC dimension of $\mathcal{A}$ is bounded by $r(d_1+d_2)\log(r(d_1+d_2))$. Using a slightly modified proof, we obtain a bound with the rate $r(d_1+d_2)$ at the cost of a larger constant.

Lemma 6.3. *The VC dimension of* (26) *is bounded by* $8r(d_1+d_2)$.

Proof. For a matrix with rank at most r, it has decomposition $W=\sum_{l=1}^r u_l v_l^T$. Thus, $\mathsf{Tr}(WZ^T)=\sum_{l=1}^r u_l^T Z v_l=\sum_{l=1}^r\sum_{i=1}^{d_1}\sum_{j=1}^{d_2}Z_{ij}u_{li}v_{lj}$ is a polynomial of degree 2 in $r(d_1+d_2)$ variables. According to Warren [38], Wolf et al. [39], if there is some $x\geq r(d_1+d_2)$, such that

$$\left(\frac{8ex}{r(d_1+d_2)}\right)^{r(d_1+d_2)}\leq 2^x \tag{27}$$

holds, then the VC dimension is bounded by x. It is easy to check that $x=8r(d_1+d_2)$ satisfies (27), and thus is an upper bound for the VC dimension. □

Now we are ready to give proofs for all the uniform convergence results of the empirical depth functions.

Proof of Proposition 3.1. For a general multi-task regression depth function, we have

$$\begin{aligned}&\sup_{B\in\mathcal{B}}\big|\mathcal{D}_{\mathcal{U}}(B,\mathbb{P})-\mathcal{D}_{\mathcal{U}}(B,\mathbb{P}_n)\big|\\&\quad\leq\sup_{B\in\mathcal{B}}\sup_{U\in\mathcal{U}}\big|\mathbb{P}_n\big\{\big\langle U^TX,Y-B^TX\big\rangle\geq 0\big\}\\&\qquad-\mathbb{P}\big\{\big\langle U^TX,Y-B^TX\big\rangle\geq 0\big\}\big|.\end{aligned}$$

Since

$$\begin{aligned}&\langle U^T X, Y - B^T X\rangle\\&\quad = \mathsf{Tr}(UYX^T) - \mathsf{Tr}(UB^T XX^T)\\&\quad = \mathsf{Tr}(WZ^T),\end{aligned}$$

where

$$W = W(U, B) = \begin{pmatrix} U & 0 \\ 0 & UB^T \end{pmatrix} \quad \text{and} \quad Z^T = \begin{pmatrix} YX^T & 0 \\ 0 & -XX^T \end{pmatrix}, \tag{28}$$

we have

$$\sup_{B\in\mathcal{B}} \bigl|\mathcal{D}_{\mathcal{U}}(B, \mathbb{P}) - \mathcal{D}_{\mathcal{U}}(B, \mathbb{P}_n)\bigr| \le \sup_{A\in\mathcal{A}} \bigl|\mathbb{P}_n(A) - \mathbb{P}(A)\bigr|. \tag{29}$$

We use $\mathbb{P}$ to denote the distribution of Z with slight abuse of notation. The set $\mathcal{A}$ is defined as

$$\mathcal{A} = \bigl\{\bigl\{Z \in \mathbb{R}^{2p\times(p+m)} : \mathsf{Tr}(WZ^T) \ge 0\bigr\} : W = W(U, B), U \in \mathcal{U}, B \in \mathcal{B}\bigr\}. \tag{30}$$

In the setting of Proposition 3.1, we have $p = m = k$, and

$$W = \begin{pmatrix} u & 0 \\ 0 & u\beta^T \end{pmatrix}, \tag{31}$$

for any $u \in \mathbb{R}^k$ and $\beta \in \mathbb{R}^k$. Hence, W is of rank at most 2. By Lemma 6.1, Lemma 6.2 and Lemma 6.3, we obtain the desired result. □

Proof of Proposition 3.2. The same argument that leads to (29) gives the bound

$$\sup_{\beta\in\Theta_s} \bigl|\mathcal{D}_{\Theta_{2s}}(\beta, \mathbb{P}_n) - \mathcal{D}_{\Theta_{2s}}(\beta, \mathbb{P})\bigr| \le \max_{S_1\in\{S\subset[p]:|S|=s\}, S_2\in\{S\subset[p]:|S|=2s\}} \sup_{A\in\mathcal{A}_{S_1,S_2}} \bigl|\mathbb{P}_n(A) - \mathbb{P}(A)\bigr|,$$

where

$$\mathcal{A}_{S_1,S_2} = \bigl\{\bigl\{Z \in \mathbb{R}^{2p\times(p+1)} : \mathsf{Tr}(WZ^T) \ge 0\bigr\} : W = W(u, \beta), u \in \Theta_{S_2}, \beta \in \Theta_{S_1}\bigr\},$$

and $W(u, \beta)$ is in the form (31). For any subset $S \subset [p]$, Θ_S is defined as

$$\Theta_S = \bigl\{u \in \mathbb{R}^p : u_j = 0 \text{ for all } j \in S^c\bigr\}.$$

By Lemma 6.1 and a union bound argument, we have

$$\begin{aligned}&\sup_{\beta\in\Theta_s} \bigl|\mathcal{D}_{\Theta_{2s}}(\beta, \mathbb{P}_n) - \mathcal{D}_{\Theta_{2s}}(\beta, \mathbb{P})\bigr|\\&\quad \le \max_{S_1\in\{S\subset[p]:|S|=s\}, S_2\in\{S\subset[p]:|S|=2s\}} \mathbb{E}\Bigl\{\sup_{A\in\mathcal{A}_{S_1,S_2}} \bigl|\mathbb{P}_n(A) - \mathbb{P}(A)\bigr|\Bigr\} + t,\end{aligned}$$

with probability at least $1 - 2e^{-2nt^2+4s\log(\frac{ep}{s})}$. Finally, in view of Lemma 6.2, it is sufficient to upper bound the VC dimension of $\mathcal{A}_{S_1,S_2}$. Note that $\mathcal{A}_{S_1,S_2}$ contains matrices of the form (31) with $u \in \Theta_{S_2}$ and $\beta \in \Theta_{S_1}$, the VC dimension is bounded by $16(5s+1)$ according to Lemma 6.3. Hence, we have

$$\sup_{\beta\in\Theta_s} \left|\mathcal{D}_{\Theta_{2s}}(\beta,\mathbb{P}_n) - \mathcal{D}_{\Theta_{2s}}(\beta,\mathbb{P})\right| \le C\sqrt{\frac{s}{n}} + t,$$

with probability at least $1 - 2e^{-2nt^2+4s\log(\frac{ep}{s})}$ for some universal constant $C > 0$. The desired result follows by setting $t^2 = \frac{4s\log(\frac{ep}{s})+\log(1/\delta)}{2n}$. □

Proof of Proposition 3.3. For the trace regression depth function, we have

$$\begin{aligned}
&\sup_{B\in\mathcal{A}_r} \left|\mathcal{D}_{\mathcal{A}_{2r}}(B,\mathbb{P}_n) - \mathcal{D}_{\mathcal{A}_{2r}}(B,\mathbb{P})\right| \\
&\quad \le \sup_{B\in\mathcal{A}_r}\sup_{U\in\mathcal{A}_{2r}} \left|\mathbb{P}_n\left\{\langle U, X\rangle\left(y - \langle B, X\rangle\right) \ge 0\right\} - \mathbb{P}\left\{\langle U, X\rangle\left(y - \langle B, X\rangle\right) \ge 0\right\}\right|.
\end{aligned}$$

Since

$$\begin{aligned}
&\langle U, X\rangle\left(y - \langle B, X\rangle\right) \\
&\quad = U^T X y - \mathsf{Tr}\left(BU^T X X^T\right) \\
&\quad = \mathsf{Tr}\left(W Z^T\right),
\end{aligned}$$

where

$$W = W(U,B) = \begin{pmatrix} U^T & 0 \\ 0 & BU^T \end{pmatrix} \quad \text{and} \quad Z^T = \begin{pmatrix} Xy & 0 \\ 0 & -XX^T \end{pmatrix},$$

and we thus have

$$\sup_{B\in\mathcal{A}_r} \left|\mathcal{D}_{\mathcal{A}_{2r}}(B,\mathbb{P}_n) - \mathcal{D}_{\mathcal{A}_{2r}}(B,\mathbb{P})\right| \le \sup_{A\in\mathcal{A}}\left|\mathbb{P}_n(A) - \mathbb{P}(A)\right|.$$

We use $\mathbb{P}$ to denote the distribution of Z with slight abuse of notation. The set $\mathcal{A}$ is defined as

$$\mathcal{A} = \left\{\left\{Z \in \mathbb{R}^{(p_1+p_2)\times 2p_1} : \mathsf{Tr}\left(WZ^T\right) \ge 0\right\} : W = W(U,B), U \in \mathcal{U}, B \in \mathcal{B}\right\}.$$

By Lemma 6.3, the VC dimension of $\mathcal{A}$ is bounded by $16r(3p_1 + p_2)$. Together with Lemma 6.1 and Lemma 6.2, we obtain the desired result. □

Proof of Proposition 4.1. Using the argument that leads to (29), we have

$$\sup_{B\in\mathbb{R}^{p\times m}} \left|\mathcal{D}_{\mathbb{R}^{p\times m}\setminus\{0\}}(B,\mathbb{P}) - \mathcal{D}_{\mathbb{R}^{p\times m}\setminus\{0\}}(B,\mathbb{P}_n)\right| \le \sup_{A\in\mathcal{A}}\left|\mathbb{P}_n(A) - \mathbb{P}(A)\right|,$$

where $\mathcal{A}$ is defined in (30), which involves matrices W of dimension $2p \times (p+m)$ with rank at most $p \wedge m$. According to Lemma 6.3, its VC dimension is bounded by $8(p \wedge m)(3p+m) \leq 32pm$. Together with Lemma 6.1 and Lemma 6.2, we obtain the desired result. □

Proof of Proposition 4.2. The same argument that leads to (29) gives the bound

$$\sup_{B \in \Xi_s} \left|\mathcal{D}_{\Xi_{2s}}(B, \mathbb{P}_n) - \mathcal{D}_{\Xi_{2s}}(B, \mathbb{P})\right| \leq \max_{S_1 \in \{S \subset [p]:|S|=s\}, S_2 \in \{S \subset [p]:|S|=2s\}} \sup_{A \in \mathcal{A}_{S_1,S_2}} \left|\mathbb{P}_n(A) - \mathbb{P}(A)\right|,$$

where

$$\mathcal{A}_{S_1,S_2} = \left\{\left\{Z \in \mathbb{R}^{2p \times (p+m)} : \mathsf{Tr}\left(WZ^T\right) \geq 0\right\} : W = W(U, B), U \in \Xi_{S_2}, B \in \Xi_{S_1}\right\},$$

and $W(U, B)$ is defined in (28). For any subset $S \subset [p]$, Ξ_S is defined as

$$\Xi_S = \left\{U \in \mathbb{R}^{p \times m} : U_{j*} = 0 \text{ for all } j \in S^c\right\}.$$

By Lemma 6.1 and a union bound argument, we have

$$\begin{aligned} &\sup_{B \in \Xi_s} \left|\mathcal{D}_{\Xi_{2s}}(B, \mathbb{P}_n) - \mathcal{D}_{\Xi_{2s}}(B, \mathbb{P})\right| \\ &\quad \leq \max_{S_1 \in \{S \subset [p]:|S|=s\}, S_2 \in \{S \subset [p]:|S|=2s\}} \mathbb{E}\left\{\sup_{A \in \mathcal{A}_{S_1,S_2}} \left|\mathbb{P}_n(A) - \mathbb{P}(A)\right|\right\} + t, \end{aligned}$$

with probability at least $1 - 2e^{-2nt^2 + 4s\log(\frac{ep}{s})}$. Finally, in view of Lemma 6.2, it is sufficient to upper bound the VC dimension of $\mathcal{A}_{S_1,S_2}$. Note that $\mathcal{A}_{S_1,S_2}$ contains matrices of the form (28) with $U \in \Xi_{S_2}$ and $B \in \Xi_{S_1}$, the VC dimension is bounded by $8(2s \wedge m)(5s+m) \leq 64ms$ according to Lemma 6.3. Hence, we have

$$\sup_{B \in \Xi_s} \left|\mathcal{D}_{\Xi_{2s}}(B, \mathbb{P}_n) - \mathcal{D}_{\Xi_{2s}}(B, \mathbb{P})\right| \leq C\sqrt{\frac{ms}{n}} + t,$$

with probability at least $1 - 2e^{-2nt^2 + 4s\log(\frac{ep}{s})}$ for some universal constant $C > 0$. The desired result follows by setting $t^2 = \frac{4s\log(\frac{ep}{s}) + \log(1/\delta)}{2n}$. □

Proof of Proposition 4.3. Using the argument that leads to (29), we have

$$\sup_{B \in \mathcal{A}_r} \left|\mathcal{D}_{\mathcal{A}_{2r}}(B, \mathbb{P}_n) - \mathcal{D}_{\mathcal{A}_{2r}}(B, \mathbb{P})\right| \leq \sup_{A \in \mathcal{A}} \left|\mathbb{P}_n(A) - \mathbb{P}(A)\right|,$$

where $\mathcal{A}$ is defined in (30), which involves matrices W of dimension $2p \times (p+m)$ with rank at most $2r$. According to Lemma 6.3, its VC dimension is bounded by $16r(3p+m)$. Together with Lemma 6.1 and Lemma 6.2, we obtain the desired result. □

6.2. Curvature of the populational depth functions

In addition to the uniform convergence results, another key ingredient we need is the curvature of the population depth function. They are characterized for both univariate regression and multivariate regression by the following two lemmas, respectively.

Lemma 6.4. *Let P_β denote the joint distribution of $(X, y) \in \mathbb{R}^p \times \mathbb{R}$ specified by $X \sim N(0, \Sigma)$ and $y|X \sim N(\beta^T X, \sigma^2)$. For any $\alpha \in \mathbb{R}^p$ such that $\alpha - \beta \in \mathcal{U}$, as long as $\mathcal{D}_{\mathcal{U}}(\alpha, P_\beta) \geq \frac{1}{2} - \eta$ for some $\eta < \frac{5}{12}$, we have*

$$\left\| \Sigma^{1/2}(\alpha - \beta) \right\| \leq C\sigma\eta,$$

where $C > 0$ is some universal constant.

Proof. By the definition of the depth function, we have

$$\mathcal{D}_{\mathcal{U}}(\alpha, P_\beta) = 1 - \sup_{u \in \mathcal{U}} \mathbb{E}\Phi\left(\frac{u^T X X^T (\alpha - \beta)}{\sigma |u^T X|} \right),$$

where $\Phi(\cdot)$ is the cumulative distribution function of $N(0, 1)$. Together with the condition $\mathcal{D}_{\mathcal{U}}(\alpha, P_\beta) \geq \frac{1}{2} - \eta$, we obtain

$$\sup_{u \in \mathcal{U}} \mathbb{E}\Phi\left(\frac{u^T X X^T (\alpha - \beta)}{\sigma |u^T X|} \right) - \Phi(0) \leq \eta.$$

Since $\alpha - \beta \in \mathcal{U}$, we have

$$\mathbb{E}\Phi\left(\frac{|X^T (\alpha - \beta)|}{\sigma} \right) - \Phi(0) \leq \eta.$$

For $Z \sim N(0, 1)$, consider the function $g(t) = \mathbb{E}\Phi(t|Z|)$. It is easy to check that $g(t)$ is increasing in t. Since $g(4) > 11/12$, the fact that $g(t) \leq 1/2 + \eta$ for some $\eta < 5/12$ implies that $t \leq 4$. The definition of $g(t)$ implies that

$$g(t) - \frac{1}{2} = \mathbb{E}\Phi\bigl(t|Z|\bigr) - \Phi(0) \geq \phi(t)\mathbb{E}\min\bigl\{t, t|Z|\bigr\} \geq t\phi(4)\mathbb{E}\min\bigl\{1, |Z|\bigr\},$$

where $\phi(\cdot)$ is the density function of $N(0, 1)$. Therefore,

$$\mathbb{E}\Phi\left(\frac{|X^T (\alpha - \beta)|}{\sigma} \right) - \Phi(0) = g\left(\frac{\|\Sigma^{1/2}(\alpha - \beta)\|}{\sigma} \right) - \frac{1}{2} \geq c\frac{\|\Sigma^{1/2}(\alpha - \beta)\|}{\sigma},$$

where $c = \phi(4)\mathbb{E}\min\{1, |Z|\}$. This leads to the conclusion

$$\left\| \Sigma^{1/2}(\alpha - \beta) \right\| \leq c^{-1}\sigma\eta.$$

Thus, the proof is complete. □

Lemma 6.5. *Let P_B denote the joint distribution of $(X, Y) \in \mathbb{R}^p \times \mathbb{R}^m$ specified by $X \sim N(0, \Sigma)$ and $Y|X \sim N(B^T X, \sigma^2 I_m)$. For any $A \in \mathbb{R}^{p\times m}$ such that $A - B \in \mathcal{U}$, as long as $\mathcal{D}_{\mathcal{U}}(A, P_B) \geq \frac{1}{2} - \eta$ for some $\eta < \frac{3}{20}$, we have*

$$\sqrt{\mathsf{Tr}\big((A-B)^T \Sigma (A-B)\big)} \leq C\sigma\eta,$$

where $C > 0$ is some universal constant.

Proof. By the definition of the depth function, we have

$$\mathcal{D}_{\mathcal{U}}(A, P_B) = 1 - \sup_{U\in\mathcal{U}} \mathbb{E}\Phi\left(\sigma^{-1}\left\langle \frac{U^T X}{\|U^T X\|}, (A-B)^T X\right\rangle\right),$$

where $\Phi(\cdot)$ is the cumulative distribution function of $N(0, 1)$. Together with the condition $\mathcal{D}_{\mathcal{U}}(A, P_B) \geq \frac{1}{2} - \eta$, we obtain

$$\sup_{U\in\mathcal{U}} \mathbb{E}\Phi\left(\sigma^{-1}\left\langle \frac{U^T X}{\|U^T X\|}, (A-B)^T X\right\rangle\right) - \Phi(0) \leq \eta.$$

Sime $A - B \in \mathcal{U}$, we have

$$\mathbb{E}\Phi\left(\frac{\|(A-B)^T X\|}{\sigma}\right) - \Phi(0) \leq \eta. \tag{32}$$

Consider the random variable $Y = \frac{\|(A-B)^T X\|^2}{\mathsf{Tr}((A-B)^T\Sigma(A-B))}$. We need a lower bound for the probability $\mathbb{P}(Y > c)$. By Cauchy–Schwarz inequality, we have

$$\mathbb{E}Y \leq c + \mathbb{E}Y\mathbb{I}\{Y > c\} \leq c + \sqrt{\mathbb{E}Y^2}\sqrt{\mathbb{P}(Y > c)}.$$

This leads to the inequality

$$\sqrt{\mathbb{P}(Y > c)} \geq \frac{\mathbb{E}Y - c}{\sqrt{\mathbb{E}Y^2}}. \tag{33}$$

Thus, we need a lower bound for $\mathbb{E}Y$ and an upper bound for $\mathbb{E}Y^2$. It is easy to see that $\mathbb{E}Y = 1$. To bound $\mathbb{E}Y^2$, we write

$$\big\|(A-B)^T X\big\|^2 = \big\|(A-B)^T \Sigma^{1/2} Z\big\|^2 = \sum_{j=1}^m \big|K_j^T Z\big|^2,$$

where K_j^T is the jth row of $(A-B)^T\Sigma^{1/2}$, and $Z \sim N(0, I_p)$. Thus,

$$\mathsf{Tr}\big((A-B)^T \Sigma (A-B)\big) = \sum_{j=1}^m \|K_j\|^2.$$

Therefore,

$$\begin{aligned}\mathbb{E}\big\|(A-B)^T X\big\|^4 &= \sum_{j=1}^m \sum_{l=1}^m \mathbb{E}\big|K_j^T Z\big|^2 \big|K_l^T Z\big|^2 \\ &= \sum_{j=1}^m \sum_{l=1}^m \|K_j\|^2 \|K_l\|^2 \mathbb{E}\frac{|K_j^T Z|^2}{\|K_j\|^2} \frac{|K_l^T Z|^2}{\|K_l\|^2} \\ &\leq 3 \sum_{j=1}^m \sum_{l=1}^m \|K_j\|^2 \|K_l\|^2 \\ &= 3\left(\sum_{j=1}^m \|K_j\|^2\right)^2.\end{aligned}$$

Hence,

$$\mathbb{E}Y^2 = \frac{\mathbb{E}\|(A-B)^T X\|^4}{(\sum_{j=1}^m \|K_j\|^2)^2} \leq 3.$$

The inequality (33) leads to

$$\mathbb{P}\left(Y > \frac{1}{4}\right) \geq \frac{3}{16}.$$

Now we define the function

$$g(t) = \mathbb{E}\Phi\left(t \frac{\|(A-B)^T X\|}{\sqrt{\mathsf{Tr}((A-B)^T \Sigma (A-B))}}\right) = \mathbb{E}\Phi(t\sqrt{Y}).$$

It is easy to check that $g(t)$ is increasing in t. Moreover,

$$\begin{aligned} g(4) &= \mathbb{E}\Phi(4\sqrt{Y}) \\ &= \mathbb{E}\Phi(4\sqrt{Y})\mathbb{I}\left\{Y > \frac{1}{4}\right\} + \mathbb{E}\Phi(4\sqrt{Y})\mathbb{I}\left\{Y \leq \frac{1}{4}\right\} \\ &\geq \Phi(2)\mathbb{P}\left(Y > \frac{1}{4}\right) + \Phi(0)\mathbb{P}\left(Y \leq \frac{1}{4}\right) \\ &= \big(\Phi(2) - \Phi(0)\big)\mathbb{P}\left(Y > \frac{1}{4}\right) + \frac{1}{2} \\ &\geq \frac{1}{2} + \frac{3}{20}.\end{aligned}$$

Therefore, $g(t) \leq \frac{1}{2} + \eta$ for some $\eta < \frac{3}{20}$ implies that $t \leq 4$. The definition of $g(t)$ implies that

$$\begin{aligned} g(t) - \frac{1}{2} &= \mathbb{E} \int_0^{t\sqrt{Y}} \phi(x)\, dx \\ &\geq \mathbb{E} \int_0^{\min(t\sqrt{Y},t)} \phi(x)\, dx \\ &\geq \mathbb{E}\phi(t) \min(t\sqrt{Y}, t) \\ &\geq t\phi(4)\mathbb{E}\min\{1, \sqrt{Y}\}, \end{aligned}$$

where $\phi(\cdot)$ is the density function of $N(0, 1)$. Finally, we need to lower bound $\mathbb{E}\min\{1, \sqrt{Y}\}$. We have

$$\begin{aligned} \mathbb{E}\min\{1, \sqrt{Y}\} &\geq \frac{1}{2}\mathbb{P}\left(\min\{1, \sqrt{Y}\} > \frac{1}{2}\right) \\ &\geq \frac{1}{2}\mathbb{P}\left(Y > \frac{1}{4}\right) \\ &\geq \frac{3}{32}. \end{aligned}$$

Hence,

$$\begin{aligned} \mathbb{E}\Phi\left(\frac{\|(A-B)^T X\|}{\sigma}\right) - \Phi(0) &= g\left(\frac{\sqrt{\mathsf{Tr}((A-B)^T\Sigma(A-B))}}{\sigma}\right) - \frac{1}{2} \\ &\geq c\frac{\sqrt{\mathsf{Tr}((A-B)^T\Sigma(A-B))}}{\sigma}, \end{aligned}$$

where $c = \frac{3\phi(4)}{32}$. Using (32), we obtain the desired conclusion, and the proof is complete. □

6.3. Proofs of main results

This section gives proofs of Theorems 3.1, 3.2, 3.3, 3.5, 4.1, 4.2 and 4.3 as well as Theorem 5.1. For i.i.d. data $\{(X_i, Y_i)\}_{i=1}^n$ from a contaminated distribution $(1-\epsilon)P + \epsilon Q$, it can be written as $\{(X_i^P, Y_i^P)\}_{i=1}^{n_1} \cup \{(X_i^Q, Y_i^Q)\}_{i=1}^{n_2}$. Marginally, we have $n_2 \sim \text{Binomial}(n, \epsilon)$ and $n_1 = n - n_2$. Conditioning on n_1 and n_2, $\{(X_i^P, Y_i^P)\}_{i=1}^{n_1}$ are i.i.d. from P and $\{(X_i^Q, Y_i^Q)\}_{i=1}^{n_2}$ are i.i.d. from Q. The following lemma (Lemma 7.1 in Chen et al. [8]) controls the ratio n_2/n_1.

Lemma 6.6. *Assume $\epsilon < 1/2$. For any $\delta > 0$ satisfying $n^{-1}\log(1/\delta) < c$ for some sufficiently small constant c, we have*

$$\frac{n_2}{n_1} \leq \frac{\epsilon}{1-\epsilon} + C\sqrt{\frac{\log(1/\delta)}{n}}, \tag{34}$$

with probability at least $1-\delta$, where $C > 0$ is a universal constant.

Now we are ready to prove the main results.

Proof of Theorem 3.1. By Lemma 6.6, we decompose the data $\{(X_i, y_i)\}_{i=1}^n = \{(X_i^P, y_i^P)\}_{i=1}^{n_1} \cup \{(X_i^Q, y_i^Q)\}_{i=1}^{n_2}$. The following analysis is on the intersection of the events of (34) and Proposition 3.1 that holds with probability at least $1-2\delta$. For any $f = \sum_{j=1}^{\infty} \beta_j \phi_j \in S_\alpha(M)$, there exists some $\beta_{[k]} \in \mathcal{U}_k$, such that for the corresponding $f_{[k]}$,

$$\|f_{[k]} - f\|^2 = \|\beta_{[k]} - \beta\|^2 \le C_1 k^{-2\alpha}, \tag{35}$$

for some constant $C_1 > 0$ that only depends on α and M. Recall the notation P_f. By the definition of the depth function and Proposition 3.1, we have

$$\begin{aligned}
\mathcal{D}_{\mathcal{U}_k}(\hat{\beta}, P_f) &\ge \mathcal{D}_{\mathcal{U}_k}\big(\hat{\beta}, \big\{\big(X_i^P, y_i^P\big)\big\}_{i=1}^{n_1}\big) - C\sqrt{\frac{k}{n_1}} - \sqrt{\frac{\log(1/\delta)}{2n_1}} && (36)\\
&\ge \frac{n}{n_1}\mathcal{D}_{\mathcal{U}_k}\big(\hat{\beta}, \big\{(X_i, y_i)\big\}_{i=1}^{n}\big) - \frac{n_2}{n_1} - C\sqrt{\frac{k}{n_1}} - \sqrt{\frac{\log(1/\delta)}{2n_1}} && (37)\\
&\ge \frac{n}{n_1}\mathcal{D}_{\mathcal{U}_k}\big(\beta_{[k]}, \big\{(X_i, y_i)\big\}_{i=1}^{n}\big) - \frac{n_2}{n_1} - C\sqrt{\frac{k}{n_1}} - \sqrt{\frac{\log(1/\delta)}{2n_1}} && (38)\\
&\ge \mathcal{D}_{\mathcal{U}_k}\big(\beta_{[k]}, \big\{\big(X_i^P, y_i^P\big)\big\}_{i=1}^{n_1}\big) - \frac{n_2}{n_1} - C\sqrt{\frac{k}{n_1}} - \sqrt{\frac{\log(1/\delta)}{2n_1}} && (39)\\
&\ge \mathcal{D}_{\mathcal{U}_k}(\beta_{[k]}, P_f) - \frac{n_2}{n_1} - 2C\sqrt{\frac{k}{n_1}} - 2\sqrt{\frac{\log(1/\delta)}{2n_1}}. && (40)
\end{aligned}$$

The inequalities (36) and (40) are by Proposition 3.1. The inequalities (37) and (39) are due to the property of depth function that

$$n_1\mathcal{D}_{\mathcal{U}_k}\big(\beta, \{Y_i\}_{i=1}^{n_1}\big) \ge n\mathcal{D}_{\mathcal{U}_k}\big(\beta, \{X_i\}_{i=1}^{n}\big) - n_2 \ge n_1\mathcal{D}_{\mathcal{U}_k}\big(\beta, \{Y_i\}_{i=1}^{n_1}\big) - n_2,$$

for any $\beta \in \mathcal{U}_k$. The inequality (38) is by the definition of $\hat{\beta}$. Moreover,

$$\begin{aligned}
&\big|\mathcal{D}_{\mathcal{U}_k}(\beta_{[k]}, P_f) - \mathcal{D}_{\mathcal{U}_k}(\beta, P_f)\big| \\
&\quad\le \sup_{u\in\mathcal{U}_k}\big|P_f\big(u^T X\big(y - X^T\beta\big) \ge 0\big) - P_f\big(u^T X\big(y - X^T\beta_{[k]}\big) \ge 0\big)\big| \\
&\quad= \sup_{u\in\mathcal{U}_k}\left|\mathbb{E}\Phi\left(\frac{u^T X X^T(\beta_{[k]} - \beta)}{|u^T X|}\right) - \Phi(0)\right| && (41)\\
&\quad\le \sqrt{\frac{1}{2\pi}}\mathbb{E}\big|X^T(\beta_{[k]} - \beta)\big|
\end{aligned}$$

$$\leq \sqrt{\frac{1}{2\pi}} \sqrt{\mathbb{E}\big(f_{[k]}(x) - f(x)\big)^2} \tag{42}$$

$$= \sqrt{\frac{1}{2\pi}} \| f_{[k]} - f \|$$

$$\leq C_1^{1/2} \sqrt{\frac{1}{2\pi}} k^{-\alpha}, \tag{43}$$

where $\Phi(\cdot)$ is the cumulative distribution function of $N(0, 1)$ in (41) and $x \sim \text{Unif}[0, 1]$ in (42). The inequality (43) is due to (35). Therefore,

$$\mathcal{D}_{\mathcal{U}_k}(\beta_{[k]}, P_f) \geq \frac{1}{2} - C_1^{1/2} \sqrt{\frac{1}{2\pi}} k^{-\alpha}.$$

Together with the inequality (40) and Lemma 6.6, we have

$$\mathcal{D}_{\mathcal{U}_k}(\hat{\beta}, P_f) \geq \frac{1}{2} - \frac{\epsilon}{1-\epsilon} - C_2\left(\sqrt{\frac{k}{n}} + k^{-\alpha} + \sqrt{\frac{\log(1/\delta)}{n}}\right), \tag{44}$$

with probability at least $1 - 2\delta$. At this point, we cannot directly use Lemma 6.4, because $\hat{\beta} - \beta \notin \mathcal{U}_k$. A slightly different argument is needed. Starting from (44), we have

$$\sup_{u \in \mathcal{U}_k} \mathbb{E}\Phi\left(\frac{u^T X X^T (\hat{\beta} - \beta)}{|u^T X|}\right) - \Phi(0) \leq \frac{\epsilon}{1-\epsilon} + C_2\left(\sqrt{\frac{k}{n}} + k^{-\alpha} + \sqrt{\frac{\log(1/\delta)}{n}}\right),$$

where the expectation is only taken over X. The same argument that leads to (43) gives

$$\sup_{u \in \mathcal{U}_k} \mathbb{E}\Phi\left(\frac{u^T X X^T (\hat{\beta} - \beta_{[k]})}{|u^T X|}\right) - \Phi(0) \leq \frac{\epsilon}{1-\epsilon} + C_3\left(\sqrt{\frac{k}{n}} + k^{-\alpha} + \sqrt{\frac{\log(1/\delta)}{n}}\right).$$

Now, since $\hat{\beta} - \beta_{[k]} \in \mathcal{U}_k$, by the same argument in the proof of Lemma 6.4, we have

$$\| \hat{f} - f_{[k]} \| \leq C_4\left(\epsilon + \sqrt{\frac{k}{n}} + k^{-\alpha} + \sqrt{\frac{\log(1/\delta)}{n}}\right).$$

Using (35) again, we have

$$\| \hat{f} - f \| \leq C_5\left(\epsilon + \sqrt{\frac{k}{n}} + k^{-\alpha} + \sqrt{\frac{\log(1/\delta)}{n}}\right).$$

The choice $k = \lceil n^{\frac{1}{2\alpha+1}} \rceil$ completes the proof. □

Proofs of Theorem 3.2 and Theorem 3.5. We first give the proof of Theorem 3.2. By Lemma 6.6, we decompose the data $\{(X_i, y_i)\}_{i=1}^n = \{(X_i^P, y_i^P)\}_{i=1}^{n_1} \cup \{(X_i^Q, y_i^Q)\}_{i=1}^{n_2}$. The following analysis is on the intersection of the events of (34) and Proposition 3.2 that holds with probability at least $1-2\delta$. Recall the notation P_β. Using the same arguments as in (36)–(40), we get

$$\mathcal{D}_{\Theta_{2s}}(\hat{\beta}, P_\beta) \geq \frac{1}{2} - \frac{n_2}{n_1} - 2C\sqrt{\frac{s\log(\frac{ep}{s})}{n_1}} - 2\sqrt{\frac{\log(1/\delta)}{2n_1}}.$$

Lemma 6.6 implies that

$$\mathcal{D}_{\Theta_{2s}}(\hat{\beta}, P_\beta) \geq \frac{1}{2} - \frac{\epsilon}{1-\epsilon} - C_1\left(\sqrt{\frac{s\log(\frac{ep}{s})}{n}} + \sqrt{\frac{\log(1/\delta)}{2n}}\right), \tag{45}$$

with probability at least $1-2\delta$. Since $\hat{\beta}-\beta \in \Theta_{2s}$, we use Lemma 6.4 to deduce (10). The bounds (11) and (12) are direct implications of (10) by the definition of κ. Thus, the proof of Theorem 3.2 is complete. The proof of Theorem 3.5 follows the same argument, and we do not repeat the details. □

Proof of Theorem 3.3. We use $\mathcal{D}_1$ to denote the first half of the data and $\mathcal{D}_2$ to denote the second half. The model $X_j = \beta_{(j)}^T X_{-j} + \xi_j$ is an instance of sparse linear regression in Section 3.2. Thus, the result of Theorem 3.2 implies that

$$\left\|\Sigma_{-j,-j}^{1/2}(\beta_{(j)} - \hat{\beta}_{(j)})\right\|^2 \leq C\left(\frac{s\log(\frac{ep}{s})}{n} \vee \epsilon^2 + \frac{\log(1/\delta)}{n}\right),$$

and

$$\|\hat{\beta}_{(j)} - \beta_{(j)}\|_1^2 \leq C\left(\frac{s^2\log(\frac{ep}{s})}{n} \vee s\epsilon^2 + \frac{s\log(1/\delta)}{n}\right),$$

with probability at least $1-2\delta$. The matrix $\Sigma_{-j,-j}$ is the covariance of X_{-j}. Now we study the error of $\widehat{\Omega}_{jj}^{-1}$. Conditioning on $\mathcal{D}_1$,

$$\begin{aligned} X_j - \hat{\beta}_{(j)}^T X_{-j} &= (\beta_{(j)} - \hat{\beta}_{(j)})^T X_{-j} + \xi_j \\ &\sim (1-\epsilon)N\left(0, \left\|\Sigma_{-j,-j}^{1/2}(\beta_{(j)} - \hat{\beta}_{(j)})\right\|^2 + \Omega_{jj}^{-1}\right) + \epsilon Q. \end{aligned}$$

Theorem 3.1 of Chen et al. [8] implies that

$$\left|\widehat{\Omega}_{jj}^{-1} - \Omega_{jj}^{-1}\right|^2 \leq 2\left\|\Sigma_{-j,-j}^{1/2}(\beta_{(j)} - \hat{\beta}_{(j)})\right\|^4 + C_1\left(\epsilon^2 + \frac{\log(1/\delta)}{n}\right),$$

with probability at least $1-2\delta$. Therefore,

$$\left|\widehat{\Omega}_{jj}^{-1} - \Omega_{jj}^{-1}\right|^2 \leq C_2\left(\epsilon^2 + \left(\frac{s\log(\frac{ep}{s})}{n}\right)^2 + \frac{\log(1/\delta)}{n}\right),$$

with probability at least $1-4\delta$. Combing the bounds above, we have

$$\begin{aligned}\|\widehat{\Omega}_{-j,j}-\Omega_{-j,j}\|_1^2 &= \|\widehat{\Omega}_{jj}\hat{\beta}_{(j)}-\Omega_{jj}\beta_{(j)}\|_1^2 \\ &\leq 2|\widehat{\Omega}_{jj}|^2\|\hat{\beta}_{(j)}-\beta_{(j)}\|_1^2+2\|\beta_{(j)}\|_1^2|\widehat{\Omega}_{jj}-\Omega_{jj}|^2 \\ &\leq C_3\left(\frac{s^2\log(\frac{ep}{s})}{n}\vee s\epsilon^2+\frac{s\log(1/\delta)}{n}\right),\end{aligned}$$

with probability at least $1-4\delta$. Therefore,

$$\|\widehat{\Omega}_{*j}-\Omega_{*j}\|_1^2\leq C_4\left(\frac{s^2\log(\frac{ep}{s})}{n}\vee s\epsilon^2+\frac{s\log(1/\delta)}{n}\right),$$

with probability at least $1-4\delta$. Finally, a union bound argument gives

$$\|\widehat{\Omega}-\Omega\|_{\ell_1}^2=\max_{1\leq j\leq p}\|\widehat{\Omega}_{*j}-\Omega_{*j}\|_1^2\leq C_4\left(\frac{s^2\log(\frac{ep}{s})}{n}\vee s\epsilon^2+\frac{s\log(1/\delta)}{n}\right),$$

with probability at least $1-4p\delta$. Choose $\delta=\exp(-C_5(n\epsilon^2+s\log(ep/s)))$, and the proof is complete. □

Proofs of Theorem 4.1, Theorem 4.2 and Theorem 4.3. We first state the proof of Theorem 4.2. Recall the notation P_B. The same argument that leads to (45) gives

$$\mathcal{D}_{\Xi_{2s}}(\widehat{B},P_B)\geq\frac{1}{2}-\frac{\epsilon}{1-\epsilon}-C_1\left(\sqrt{\frac{ms+s\log(\frac{ep}{s})}{n}}+\sqrt{\frac{\log(1/\delta)}{2n}}\right).$$

Since $\widehat{B}-B\in\Xi_{2s}$, we use Lemma 6.5 to deduce (21). The bound (22) is a direct implication of (21) by the definition of κ. This completes the proof of Theorem 4.2. Setting $s=p$ gives the proof of Theorem 4.1. The proof of Theorem 4.3 follows the same argument, and we omit the details. □

Proof of Theorem 5.1. The proof is the same as that of Theorem 4.1, except that we need to establish a similar curvature result as Lemma 6.5 for the elliptical distribution. The same argument that leads to (32) gives

$$\mathbb{E}F\left(\frac{\|(A-B)^TX\|}{\|\Gamma^{1/2}(A-B)^TX\|}\left\|(A-B)^TX\right\|\right)-F(0)\leq\eta.$$

By the definition $\sigma^2=\|\Gamma\|_{\mathrm{op}}$, we have

$$\mathbb{E}F\left(\frac{1}{\sigma}\left\|(A-B)^TX\right\|\right)-F(0)\leq\eta.$$

Following the proof of Lemma 6.5, it is sufficient to show that $g(t) - 1/2 \geq Ct\mathbb{E}\{1, \sqrt{Y}\}$ for some constant $C > 0$, where $g(t) = \mathbb{E}F(t\sqrt{Y})$. We outline the main step without repeating all the details that have already been used in the proof of Lemma 6.5. The fact that $g(t) \leq \frac{1}{2} + \eta$ for a sufficiently small η implies that $t \leq c_1$. Then,

$$g(t) - \frac{1}{2} \geq t \min_{|t| \leq c_1} f(t)\mathbb{E}\min\{1, \sqrt{Y}\}.$$

Under the assumption that $\min_{|t| \leq c_1} f(t) \geq c_2$, the proof is complete. □

Acknowledgements

The author thanks Zhao Ren and Haoyang Liu for reading the manuscript and offering insightful suggestions. The research of CG is supported in part by NSF grant DMS-1712957 and NSF CAREER Award DMS-1847590.

References

[1] Amenta, N., Bern, M., Eppstein, D. and Teng, S.-H. (2000). Regression depth and center points. *Discrete Comput. Geom.* **23** 305–323. MR1744506 https://doi.org/10.1007/PL00009502

[2] Balakrishnan, S., Du, S.S., Li, J. and Singh, A. (2017). Computationally efficient robust sparse estimation in high dimensions. In *Conference on Learning Theory* 169–212.

[3] Bern, M. and Eppstein, D. (2000). Multivariate regression depth. In *Proceedings of the Sixteenth Annual Symposium on Computational Geometry* (*Hong Kong*, 2000) 315–321. New York: ACM. MR1802280 https://doi.org/10.1145/336154.336218

[4] Bickel, P.J. (1984). Robust regression based on infinitesimal neighbourhoods. *Ann. Statist.* **12** 1349–1368. MR0760693 https://doi.org/10.1214/aos/1176346796

[5] Bunea, F., She, Y. and Wegkamp, M.H. (2011). Optimal selection of reduced rank estimators of high-dimensional matrices. *Ann. Statist.* **39** 1282–1309. MR2816355 https://doi.org/10.1214/11-AOS876

[6] Cai, T.T., Liu, W. and Zhou, H.H. (2016). Estimating sparse precision matrix: Optimal rates of convergence and adaptive estimation. *Ann. Statist.* **44** 455–488. MR3476606 https://doi.org/10.1214/13-AOS1171

[7] Chen, M., Gao, C. and Ren, Z. (2016). A general decision theory for Huber's ϵ-contamination model. *Electron. J. Stat.* **10** 3752–3774. MR3579675 https://doi.org/10.1214/16-EJS1216

[8] Chen, M., Gao, C. and Ren, Z. (2018). Robust covariance and scatter matrix estimation under Huber's contamination model. *Ann. Statist.* **46** 1932–1960. MR3845006 https://doi.org/10.1214/17-AOS1607

[9] Devroye, L. and Lugosi, G. (2001). *Combinatorial Methods in Density Estimation*. *Springer Series in Statistics*. New York: Springer. MR1843146 https://doi.org/10.1007/978-1-4613-0125-7

[10] Diakonikolas, I., Kamath, G., Kane, D.M., Li, J., Moitra, A. and Stewart, A. (2016). Robust estimators in high dimensions without the computational intractability. In 57*th Annual IEEE Symposium on Foundations of Computer Science – FOCS* 2016 655–664. Los Alamitos, CA: IEEE Computer Soc. MR3631028

[11] Diakonikolas, I., Kong, W. and Stewart, A. (2019). Efficient algorithms and lower bounds for robust linear regression. In *Proceedings of the Thirtieth Annual ACM-SIAM Symposium on Discrete Algorithms* 2745–2754. Philadelphia, PA: SIAM. MR3909639 https://doi.org/10.1137/1.9781611975482.170

[12] Donoho, D.L. and Liu, R.C. (1991). Geometrizing rates of convergence. II, III. *Ann. Statist.* **19** 633–667, 668–701. MR1105839 https://doi.org/10.1214/aos/1176348114

[13] Donoho, D.L. and Montanari, A. (2015). Variance breakdown of huber (m)-estimators. $n/p \to m \in (1, \infty)$. Preprint. Available at arXiv:1503.02106.

[14] Fang, K.T., Kotz, S. and Ng, K.W. (1990). *Symmetric Multivariate and Related Distributions*. *Monographs on Statistics and Applied Probability* **36**. London: CRC Press. MR1071174 https://doi.org/10.1007/978-1-4899-2937-2

[15] Huber, P.J. (1964). Robust estimation of a location parameter. *Ann. Math. Stat.* **35** 73–101. MR0161415 https://doi.org/10.1214/aoms/1177703732

[16] Huber, P.J. (1965). A robust version of the probability ratio test. *Ann. Math. Stat.* **36** 1753–1758. MR0185747 https://doi.org/10.1214/aoms/1177699803

[17] Huber, P.J. (1973). Robust regression: Asymptotics, conjectures and Monte Carlo. *Ann. Statist.* **1** 799–821. MR0356373

[18] Huber, P.J. and Strassen, V. (1973). Minimax tests and the Neyman–Pearson lemma for capacities. *Ann. Statist.* **1** 251–263. MR0356306

[19] Johnstone, I.M. (2011). Gaussian estimation: Sequence and wavelet models. Manuscript.

[20] Koltchinskii, V., Lounici, K. and Tsybakov, A.B. (2011). Nuclear-norm penalization and optimal rates for noisy low-rank matrix completion. *Ann. Statist.* **39** 2302–2329. MR2906869 https://doi.org/10.1214/11-AOS894

[21] Lai, K.A., Rao, A.B. and Vempala, S. (2016). Agnostic estimation of mean and covariance. In 57*th Annual IEEE Symposium on Foundations of Computer Science – FOCS* 2016 665–674. Los Alamitos, CA: IEEE Computer Soc. MR3631029

[22] Loh, P.-L. and Tan, X.L. (2018). High-dimensional robust precision matrix estimation: Cellwise corruption under ϵ-contamination. *Electron. J. Stat.* **12** 1429–1467. MR3804842 https://doi.org/10.1214/18-EJS1427

[23] Lounici, K., Pontil, M., van de Geer, S. and Tsybakov, A.B. (2011). Oracle inequalities and optimal inference under group sparsity. *Ann. Statist.* **39** 2164–2204. MR2893865 https://doi.org/10.1214/11-AOS896

[24] Ma, Z. and Sun, T. (2014). Adaptive sparse reduced-rank regression. Preprint. Available at arXiv:1403.1922.

[25] Meinshausen, N. and Bühlmann, P. (2006). High-dimensional graphs and variable selection with the lasso. *Ann. Statist.* **34** 1436–1462. MR2278363 https://doi.org/10.1214/009053606000000281

[26] Mizera, I. (2002). On depth and deep points: A calculus. *Ann. Statist.* **30** 1681–1736. MR1969447 https://doi.org/10.1214/aos/1043351254

[27] Raskutti, G., Wainwright, M.J. and Yu, B. (2011). Minimax rates of estimation for high-dimensional linear regression over ℓ_q-balls. *IEEE Trans. Inform. Theory* **57** 6976–6994. MR2882274 https://doi.org/10.1109/TIT.2011.2165799

[28] Ren, Z., Sun, T., Zhang, C.-H. and Zhou, H.H. (2015). Asymptotic normality and optimalities in estimation of large Gaussian graphical models. *Ann. Statist.* **43** 991–1026. MR3346695 https://doi.org/10.1214/14-AOS1286

[29] Rousseeuw, P. and Yohai, V. (1984). Robust regression by means of S-estimators. In *Robust and Nonlinear Time Series Analysis* (*Heidelberg*, 1983). *Lect. Notes Stat.* **26** 256–272. New York: Springer. MR0786313 https://doi.org/10.1007/978-1-4615-7821-5_15

[30] Rousseeuw, P.J. and Hubert, M. (1999). Regression depth. *J. Amer. Statist. Assoc.* **94** 388–433. MR1702314 https://doi.org/10.2307/2670155

[31] Rousseeuw, P.J. and Leroy, A.M. (1987). *Robust Regression and Outlier Detection*. *Wiley Series in Probability and Mathematical Statistics*: *Applied Probability and Statistics*. New York: Wiley. MR0914792 https://doi.org/10.1002/0471725382

[32] Siegel, A.F. (1982). Robust regression using repeated medians. *Biometrika* **69** 242–244.

[33] Struyf, A. and Rousseeuw, P.J. (1999). Halfspace depth and regression depth characterize the empirical distribution. *J. Multivariate Anal.* **69** 135–153. MR1701410 https://doi.org/10.1006/jmva.1998.1804

[34] Tsybakov, A.B. (2009). *Introduction to Nonparametric Estimation. Springer Series in Statistics*. New York: Springer. MR2724359 https://doi.org/10.1007/b13794

[35] Tukey, J.W. (1975). Mathematics and the picturing of data. In *Proceedings of the International Congress of Mathematicians* (*Vancouver, B. C.*, 1974), *Vol.* 2 523–531. MR0426989

[36] Verzelen, N. (2012). Minimax risks for sparse regressions: Ultra-high dimensional phenomenons. *Electron. J. Stat.* **6** 38–90. MR2879672 https://doi.org/10.1214/12-EJS666

[37] Fan, J. and Li, Q. Wang, Y. (2014). Robust estimation of high-dimensional mean regression. Preprint. Available at arXiv:1410.2150.

[38] Warren, H.E. (1968). Lower bounds for approximation by nonlinear manifolds. *Trans. Amer. Math. Soc.* **133** 167–178. MR0226281 https://doi.org/10.2307/1994937

[39] Wolf, L., Jhuang, H. and Hazan, T. (2007). Modeling appearances with low-rank svm. In 2007 *IEEE Conference on Computer Vision and Pattern Recognition* 1–6. IEEE.

[40] Ye, F. and Zhang, C.-H. (2010). Rate minimaxity of the Lasso and Dantzig selector for the ℓ_q loss in ℓ_r balls. *J. Mach. Learn. Res.* **11** 3519–3540. MR2756192

[41] Yuan, M. (2010). High dimensional inverse covariance matrix estimation via linear programming. *J. Mach. Learn. Res.* **11** 2261–2286. MR2719856

Received March 2017 and revised March 2019