A general decision theory for Huber's ϵ-contamination model

Mengjie Chen

Department of Biostatistics, University of North Carolina, Chapel Hill, NC 27599
e-mail: mengjie@email.unc.edu

Chao Gao

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

and

Zhao Ren

Department of Statistics, University of Pittsburgh, Pittsburgh, PA 15260
e-mail: zren@pitt.edu

Abstract: Today's data pose unprecedented challenges to statisticians. It may be incomplete, corrupted or exposed to some unknown source of contamination. We need new methods and theories to grapple with these challenges. Robust estimation is one of the revived fields with potential to accommodate such complexity and glean useful information from modern datasets. Following our recent work on high dimensional robust covariance matrix estimation, we establish a general decision theory for robust statistics under Huber's ϵ-contamination model. We propose a solution using Scheffé estimate to a robust two-point testing problem that leads to the construction of robust estimators adaptive to the proportion of contamination. Applying the general theory, we construct robust estimators for nonparametric density estimation, sparse linear regression and low-rank trace regression. We show that these new estimators achieve the minimax rate with optimal dependence on the contamination proportion. This testing procedure, Scheffé estimate, also enjoys an optimal rate in the exponent of the testing error, which may be of independent interest.

MSC 2010 subject classifications: Primary 62C20, 62G35; secondary 62H12.
Keywords and phrases: Robust statistics, robust testing, minimax rate, density estimation, sparse linear regression, trace regression.

1. Introduction

In Huber's pathbreaking papers [10, 11] on robust estimation theory, he proposed the ϵ-contamination model

$$(1 - \epsilon)P_\theta + \epsilon Q. \tag{1}$$

Under this model, data are drawn from (1) with probability of ϵ to be contaminated by some arbitrary distribution Q. Given i.i.d. observations from (1), the objective is to estimate θ robust to the contamination from Q. It has been discussed in [4] that Huber's ϵ-contamination model provides a favored framework which allows a joint study of statistical efficiency and robustness. In other words, the optimality of an estimator under Huber's ϵ-contamination model indicates that it achieves statistical efficiency and robustness simultaneously. However, not much attention has been paid to this framework in nonparametric and high-dimensional statistics. Inspired by Tukey's work on data depth, we proposed a new concept, matrix depth, for robust estimation of covariance matrix in high dimension in our previous work [4]. We established the optimality of the proposed estimator under Huber's ϵ-contamination model for several covariance matrix classes. This work leaves an important problem open: whether there exists a general rule for minimax rate under Huber's ϵ-contamination model?

To address this problem in this paper, we investigate the following quantity

$$\inf_{\hat{\theta}} \sup_{\theta \in \Theta, Q} \mathbb{E}_{(\epsilon, \theta, Q)} L(\hat{\theta}, \theta), \tag{2}$$

the robust minimax risk for a given parameter space Θ and a loss function $L(\cdot, \cdot)$. The expectation $\mathbb{E}_{(\epsilon, \theta, Q)}$ is determined by the probability (1), and the supreme is taken over all $\theta \in \Theta$ and Q in the class of all probability distributions. When the loss function takes the form of squared total variation distance, we can construct a general robust estimator $\hat{\theta}$, such that the robust minimax risk (2) is upper bounded by some universal constant times

$$\min_{\delta > 0} \left\{ \frac{\log \mathcal{M}(\delta, \Theta, \mathsf{TV}(\cdot, \cdot))}{n} + \delta^2 \right\} \vee \epsilon^2, \tag{3}$$

where $\mathcal{M}(\delta, \Theta, \mathsf{TV}(\cdot, \cdot))$ denotes the δ-covering number of Θ using the total variation distance. This rate (3) consists of two parts. The first part is a common bias variance trade-off term in the classical decision theory without taking account of contamination. The second part is a term contributed by unknown contamination of the data. Comparing the rate (3) to the general lower bound for the ϵ-contamination model derived in our previous work [4], we immediately find that (3) is the minimax rate for the risk in (2). This is the main contribution of our paper.

The construction of rate-optimal robust estimators is enabled in this paper by a novel analysis of the robust testing procedure called Scheffé estimate that was first proposed in [6]. For the robust two-point testing problem, we propose a solution using Scheffé estimate, the testing error of which has a desired exponent leading to a rate-optimal estimation procedure. Our new testing theory has advantages over some classical ones. Under the contamination model, the classical Neyman-Pearson approach lacks robust property. The statistical performance of the likelihood ratio test can be compromised even when one contaminated point is included in the data. The robust testing theory established by Le Cam [14] and Birgé [1] is based on Hellinger distance, which gives a sub-optimal rate

for Huber's ϵ-contamination model. The existing optimal testing function for the robust two-point testing problem was constructed by Huber himself [11]. However, his procedure depends on the knowledge of the contamination proportion ϵ in (1). As shown in our previous work, it is impossible to estimate ϵ when Q is not specified. In comparison, our proposed testing function overcomes this and does not depend on ϵ. This feature, together with its robustness and rate-optimal error exponent, makes our method superior to the previous ones.

The rest of the paper is organized as follows. We first introduce the robust testing problem in Section 2 and propose a solution using Scheffé estimate with a sharp testing error bound. In Section 3, we use this robust testing procedure to construct a general estimator that achieves the optimal rate for (3). Then in Section 4, we construct robust estimators for density estimation, sparse linear regression and low-rank trace regression as applications of the general theory. We show that for all these problems, our estimators achieve minimax optimal rates. Finally, we investigate a scenario when the loss function is not equivalent to total variation distance in the discussion section, Section 5. We show that the minimax rate for non-intrinsic loss functions may depend on ϵ in different ways. All technical proofs are gathered 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 $\{1, 2, ..., m\}$. 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 $a_n \le C b_n$ 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 and $\mathrm{supp}(v) = \{j \in [p] : v_j \ne 0\}$ is its support. For a matrix $A \in \mathbb{R}^{p_1 \times p_2}$, $\mathsf{rank}(A)$ denotes its rank, $\mathrm{vec}(A)$ is its vectorization and $\|A\|_\mathrm{F} = \|\mathrm{vec}(A)\|$ is the matrix Frobenius norm. When A is an squared matrix, $\mathsf{Tr}(A)$ denotes its trace. For two probability distributions P_1 and P_2, their total variation distance is $\mathsf{TV}(P_1, P_2) = \sup_B |P_1(B) - P_2(B)|$, and their Hellinger distance is $H(P_1, P_2) = \left[\int \left(\sqrt{dP_1} - \sqrt{dP_2} \right)^2 \right]^{1/2}$.

2. Robust testing

Given i.i.d. observations $X_1, ..., X_n \sim P$, we consider the following robust two-point testing problem originally set up by Huber in [11]:

$$
\begin{aligned}
H_0 : \quad & P \in \{(1-\epsilon)P_0 + \epsilon Q : Q\}, \\
H_1 : \quad & P \in \{(1-\epsilon)P_1 + \epsilon Q : Q\}.
\end{aligned}
$$

In particular, P_0 and P_1 are two fixed distributions and Q is in the class of all probability distributions. When $\epsilon = 0$, it reduces to the classical two-point testing problem studied by Neyman and Pearson [15]. They showed that the likelihood ratio test $\mathbb{I}\left\{ \prod_{i=1}^n \frac{dP_1}{dP_0}(X_i) > t \right\}$ achieves the optimal testing error, which laid the foundation for modern hypothesis testing. However, the likelihood

ratio test is not robust to cases when $\epsilon > 0$. For example, when $P_0 = N(\theta_0, I_p)$ and $P_1 = N(\theta_1, I_p)$, Neyman-Pearson testing statistic involves the calculation of sample mean, which can be arbitrarily away from the true mean due to the existence of contamination from Q.

Huber showed in his seminal work [11, 12] that the exact optimal solution to the robust two-point testing problem is the following testing function:

$$\phi_{\text{Huber}} = \mathbb{I}\left\{\prod_{i=1}^{n}\left[\left(\frac{dP_1}{dP_0}(X_i) \vee c\right) \wedge C\right] > t\right\},$$

for some $0 < c < C < \infty$. It can be seen as a clipped likelihood ratio test. By clipping the likelihood ratio functions that have enormous or infinitesimal values, the influence from outliers can be diminished. When $\epsilon = 0$, the clipping cut-offs become $c = 0$, $C = \infty$, and ϕ_{Huber} naturally reduces to the likelihood ratio test. Though ϕ_{Huber} exactly minimizes the testing error, the clipping cut-offs c and C depend on the knowledge of ϵ, a quantity that characterizes the contamination proportion. Since it is impossible to estimate ϵ when Q is not specified [4], Huber's approach is not adaptive to the contamination proportion ϵ and thus not applicable.

Another work related to the robust testing problem is by Le Cam [14] and Birgé [1]. Instead of testing between two ϵ-contamination neighborhoods, they considered two Hellinger balls:

$$\begin{aligned} H_0: &\quad P \in \{P : H(P, P_0) \leq \tau\}, \\ H_1: &\quad P \in \{P : H(P, P_1) \leq \tau\}. \end{aligned}$$

They constructed a testing function and established the following testing error

$$\begin{aligned} &\sup_{P \in \{P : H(P, P_0) \leq \tau\}} P\phi + \sup_{P \in \{P : H(P, P_1) \leq \tau\}} P(1 - \phi) \\ \leq &\quad 2\exp\left(-\frac{n}{2}\left(H(P_0, P_1) - 2\tau\right)^2\right), \end{aligned} \tag{4}$$

for any $\tau < \frac{1}{2}H(P_0, P_1)$. However, their procedure cannot give optimal rate under Huber's setting. To put an ϵ-contamination neighborhood into a τ-Hellinger ball, the smallest τ would be $\sqrt{2\epsilon}$. That is,

$$\{(1 - \epsilon)P_0 + \epsilon Q : Q\} \subset \left\{P : H(P, P_0) \leq \sqrt{2\epsilon}\right\}.$$

When it comes to estimation, it will result in a sub-optimal ϵ term instead of the optimal ϵ^2 in (3).

In this paper, we propose a solution to the robust two-point testing problem as follows:

$$\phi = \mathbb{I}\left\{|P_n(A) - P_0(A)| > |P_n(A) - P_1(A)|\right\}, \tag{5}$$

where $P_n(\cdot)$ denotes the empirical distribution such that

$$P_n(A) = \frac{1}{n}\sum_{i=1}^{n}\mathbb{I}\{X_i \in A\},$$

and A is chosen as a measurable set that maximally distinguishes P_0 and P_1. That is,

$$A = \arg\max_A |P_0(A) - P_1(A)| = \{p_0 > p_1\}, \tag{6}$$

where p_j is the density function defined as $p_j = \frac{dP_j}{d(P_0+P_1)}$ for $j = 0, 1$. The corresponding estimator of the testing function ϕ is called Scheffé estimate by Devroye and Lugosi in their book [6] under the framework of density estimation. This is built on an L_1-based estimator proposed by Yatracos in [25]. The intuition is that with the set A possessing maximal distinguishing power, we check whether the empirical probability of A is closer to $P_0(A)$ or $P_1(A)$. Since we use summation of indicator functions to collect the information offered by each data point separately, compared to the product form taking by the likelihood ratio test, it is robust to outliers. Moreover, the proposed testing procedure does not depend on the contamination proportion ϵ. The testing error of the proposed procedure is characterized by the following theorem.

Theorem 2.1. *Assume* $\mathsf{TV}(P_0, P_1) > 2\epsilon$. *Then we have*

$$\sup_{P \in \{(1-\epsilon)P_0+\epsilon Q : Q\}} P\phi + \sup_{P \in \{(1-\epsilon)P_1+\epsilon Q : Q\}} P(1 - \phi)$$

$$\leq 4\exp\left(-\frac{1}{2}n\left(\mathsf{TV}(P_0, P_1) - 2\epsilon\right)^2\right).$$

We emphasize that Theorem 2.1 says the exponent of the testing error is proportional to $n\left(\mathsf{TV}(P_0, P_1) - 2\epsilon\right)^2$. Although Scheffé estimate was first proposed in [6] for density estimation problems, this important property on exponent of the testing error was not explicitly explored and thus is new. Compared with Le Cam and Birgé's testing error (4), the exponent of ours is characterized by the total variation distance instead of the Hellinger distance. As we will show in Section 3, this exponent leads to minimax optimal estimation for Huber's ϵ-contamination model.

3. Construction of upper bounds

In this section, we present a general principle for the construction of a robust estimator given i.i.d. observations $X_1, ..., X_n \sim (1-\epsilon)P_\theta + \epsilon Q$ with $\theta \in \Theta$ for some parameter space Θ. We assume that the parameter space Θ is totally bounded. Define $m = \mathcal{M}(\delta, \Theta, \mathsf{TV}(\cdot, \cdot))$ to be the smallest number such that there exists $\{\theta_1, ..., \theta_m\} \subset \Theta$ satisfying that for any $\theta \in \Theta$, there is a $j \in [m]$ such that $\mathsf{TV}(P_{\theta_j}, P_\theta) \leq \delta$. We call $\{\theta_1, ..., \theta_m\} \subset \Theta$ a δ-covering set and $\mathcal{M}(\delta, \Theta, \mathsf{TV}(\cdot, \cdot))$ is the corresponding covering number. The estimator of θ is constructed by performing robust testing (5) for each pair in the δ-covering set and then selecting the most favorable one. To be specific, given i.i.d. observations, for any $j \neq k$, define the testing function

$$\phi_{jk} = \mathbb{I}\left\{ \left| \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}\{p_{\theta_j}(X_i) > p_{\theta_k}(X_i)\} - P_{\theta_j}(p_{\theta_j}(X) > p_{\theta_k}(X)) \right| \right.$$

$$> \left| \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}\{p_{\theta_j}(X_i) > p_{\theta_k}(X_i)\} - P_{\theta_k}(p_{\theta_j}(X) > p_{\theta_k}(X)) \right| \right\},$$

where $p_\theta = \frac{dP_\theta}{d\mu}$ is the density function for some common dominating measure μ. When $\phi_{jk} = 1$, θ_k is favored over θ_j. When $\phi_{jk} = 0$, θ_j is favored over θ_k. Finally, the robust estimator is defined as $\hat{\theta} = \theta_{\hat{j}}$ with

$$\hat{j} = \arg \min_{j \in [m]} \sum_{k \neq j} \phi_{jk}. \tag{7}$$

That is to say, the final estimator wins the maximum number of pair-wise competitions. When (7) has multiple minimizers, $\hat{j}$ is understood to be any one of them. This estimator $\theta_{\hat{j}}$ is also called Scheffé tournament winner in [6] within the framework of density estimation. The detailed comparison will be discussed in Remark 3.2 later. Since the testing procedure introduced in Section 2 is adaptive for the contamination proportion ϵ, the estimator (7) is also adaptive for ϵ. The estimation error is upper bounded by the following theorem.

Theorem 3.1. *Assume $\eta > 8(\epsilon + \delta)$. For the estimator $\hat{\theta}$ defined above, we have*

$$\sup_{\theta \in \Theta, Q} \mathbb{P}_{(\epsilon, \theta, Q)} \left\{ \mathsf{TV}(P_{\hat{\theta}}, P_\theta) > \eta + \delta \right\}$$

$$\leq 4\mathcal{M}^2(\delta, \Theta, \mathsf{TV}(\cdot, \cdot)) \exp\left(-\frac{1}{2} n(\eta/4 - 2(\epsilon + \delta))^2 \right),$$

where the probability $\mathbb{P}_{(\epsilon, \theta, Q)}$ is defined in (1).

The theorem immediately implies the convergence rate (3) when we let

$$\eta^2 = C \left[\left\{ \frac{\log \mathcal{M}(\delta, \Theta, \mathsf{TV}(\cdot, \cdot))}{n} + \delta^2 \right\} \vee \epsilon^2 \right]$$

for some large constant C and then minimize the rate over δ. To show the rate (3) implied by Theorem 3.1 is minimax optimal, we first review a general lower bound result in [4].

Theorem 3.2 (Chen, Gao & Ren (2015) [4]). *$L(\cdot, \cdot)$ is a loss function defined on the parameter space Θ. Define*

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

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

$$\inf_{\hat{\theta}} \sup_{\theta \in \Theta, Q} \mathbb{P}_{(\epsilon, \theta, Q)} \left\{ L(\hat{\theta}, \theta) \geq \mathcal{R}(\epsilon) \right\} \geq c \tag{8}$$

holds for $\epsilon = 0$. Then, (8) holds for $\mathcal{R}(\epsilon) \asymp \mathcal{R}(0) \vee \omega(\epsilon, \Theta)$.

Theorem [3.2] provides a lower bound for general loss functions. The quantity $\omega(\epsilon, \Theta)$ is called modulus of continuity defined by Donoho and Liu [8, 7]. For total variation loss, $\omega(\epsilon, \Theta) \asymp \epsilon$. Moreover, a general lower bound result by Yang and Barron [24] implies the formula

$$\mathcal{R}^2(0) \asymp \min_{\delta > 0} \left\{ \frac{\log \mathcal{M}(\delta, \Theta, \mathsf{TV}(\cdot, \cdot))}{n} + \delta^2 \right\},$$

under very mild conditions. Hence, (3) is also the minimax lower bound for the problem.

Remark 3.1. *Both Theorem [3.1] and Theorem [3.2] are stated in probability. To obtain the same conclusion in expectation as defined by (2), observe that the in-probability lower bound directly implies an in-expectation lower bound via Markov inequality. The in-expectation upper bound can be calculated by integrating over the tail probability of Theorem [3.1].*

For some parametric and high-dimensional models, the notion of global covering number may not provide a tight upper bound. We show an improvement of Theorem [3.1] by using the notion of local covering number. Let $\Theta' = \{\theta_1, ..., \theta_m\}$ be a δ-covering set for Θ. For any integer l, define

$$D_l(\delta) = \max_{\theta_0 \in \Theta'} |\{\theta \in \Theta' : l\delta < \mathsf{TV}(P_\theta, P_{\theta_0}) \leq (l+1)\delta\}|.$$

Theorem 3.3. *Let L be any number such that $\frac{L}{4}\delta - 2\epsilon - 2\delta > 0$ and $L/4$ is an integer. For the estimator $\hat{\theta}$ defined by (7), we have*

$$\sup_{\theta \in \Theta, Q} \mathbb{P}_{(\epsilon, \theta, Q)} \left\{ \mathsf{TV}(P_{\hat{\theta}}, P_\theta) > (L+1)\delta \right\}$$

$$\leq \ 2 \sum_{l \geq L/4} D_l(\delta) \exp\left(-\frac{1}{2}n \left(l\delta - 2(\epsilon + \delta) \right)^2 \right)$$

$$+ 2 \left[\sum_{l=0}^{L/4-1} D_l(\delta) \right] \sum_{l \geq L} D_l(\delta) \exp\left(-\frac{1}{2}n \left((l - 3L/4)\delta - 2(\epsilon + \delta) \right)^2 \right),$$

where the probability $\mathbb{P}_{(\epsilon, \theta, Q)}$ is defined in (1).

Remark 3.2. *A closely related estimator called minimum distance estimator was first proposed by Yatracos in [25]. Later, built on Yatracos' method, Devroye and Lugosi in their book [6] further proposed a similar estimator called Scheffé tournament winner as the one in (7) within the framework of density estimation. While the theoretical decision framework under the Huber's ϵ-contamination model was not considered in either [25] or [6], it is worthwhile to point out another subtle while essential difference between the results in those early works and the current paper. Since the analysis of main results in [25] and [6] are similar, we only focus on the analysis for the minimum distance estimator in this paper. The minimum distance estimator $\hat{\theta}^Y = \theta_{\hat{j}}$ is defined as*

$$\hat{j} = \arg\min_{j \in [m]} \sup_{A \in \mathcal{A}} \left| P_n(A) - P_{\theta_j}(A) \right|,$$

where the Yatracos class $\mathcal{A} = \{\{p_{\theta_i} > p_{\theta_j}\} : i \neq j \in [m]\}$ is the collection of the sets A in (5) applied on each pair of distributions indexed by the δ-covering set $\{\theta_1, \ldots, \theta_m\}$. This estimator, instead of being explicitly built on pair-wise competitions, minimizes the distance to the empirical measure uniformly over the Yatracos class $\mathcal{A}$. Consequently, it is unlikely for $\hat{\theta}^Y$ to take advantage of the more delicate local covering set in a layer-by-layer fashion as stated in Theorem 3.3 for our analysis. In this sense, the optimality cannot be achieved for several high-dimensional models. See Sections 4.2 and 4.3 for two such examples.

4. Applications

To illustrate the theorems in Section 3, here we present their applications on three problems: density estimation with Hölder smoothness, sparse linear regression and low-rank trace regression.

4.1. Density estimation

Consider i.i.d. observation $X_1, ..., X_n \sim \mathbb{P}_{(\epsilon, f, Q)} = (1 - \epsilon)P_f + \epsilon Q$, where $f = \frac{dP}{d\lambda}$ is the density function of P_f supported on $[0, 1]$ with respect to the Lebesgue measure. We consider the Hölder class for the density function on $[0, 1]$. Let $\{\phi_{lk}\}_{l \geq 0, 0 \leq k \leq 2^l - 1}$ be an orthogonal wavelet basis on the interval $[0, 1]$. The precise construction of the wavelet basis is referred to [5]. Define the following density class:

$$
\begin{aligned}
&\mathcal{H}_{den}(\beta, M) \hspace{8cm} (9) \\
&= \left\{ f = \sum_{\substack{l \geq 0, \\ 0 \leq k \leq 2^l - 1}} f_{lk} \psi_{lk} : f \geq 0, \int_0^1 f = 1, \sup_{\substack{l \geq 0, \\ 0 \leq k \leq 2^l - 1}} 2^{l(1/2 + \beta)} |f_{lk}| \leq M \right\},
\end{aligned}
$$

where $\beta > 0$ is the smoothness index of the function class. The constant $M > 0$ is the radius of the class. By [24],

$$
\log \mathcal{M}\left(\delta, \mathcal{H}_{den}(\beta, M), \mathsf{TV}(\cdot, \cdot)\right) \asymp \delta^{-1/\beta}.
$$

Therefore, using the estimator (7) with $\delta \asymp n^{-\frac{\beta}{2\beta + 1}}$, Theorem 3.1 implies the following convergence rate.

Corollary 4.1. *For the Hölder class $\mathcal{H}_{den}(\beta, M)$, there are some constants C, C', such that*

$$
\|\hat{f} - f\|_1^2 \leq C \left(n^{-\frac{2\beta}{2\beta + 1}} \vee \epsilon^2 \right),
$$

with $\mathbb{P}_{(\epsilon, f, Q)}$-probability at least $1 - \exp\left(-C' \left(n^{\frac{1}{2\beta + 1}} + n\epsilon^2 \right) \right)$ uniformly over all Q and $f \in \mathcal{H}_{den}(\beta, M)$.

Given the equation $\mathsf{TV}(P_{f_1}, P_{f_2}) = \frac{1}{2}\|f_1 - f_2\|_1$, Corollary 4.1 states the convergence result in the squared ℓ_1 distance. Combining with Theorem 3.2 and the discussion thereafter, which implies $n^{-\frac{2\beta}{2\beta+1}} \vee \epsilon^2$ is also the minimax lower bound, we conclude it is the minimax rate for this problem. When $\epsilon^2 \lesssim n^{-\frac{2\beta}{2\beta+1}}$, the rate is dominated by $n^{-\frac{2\beta}{2\beta+1}}$. This is the minimax rate for density estimation when there is no contamination. When $n^{-\frac{2\beta}{2\beta+1}} \lesssim \epsilon^2$, the rate is dominated by ϵ^2. Therefore, the maximum expected number of outliers that can be tolerated without breaking down the usual minimax rate is $n\epsilon \asymp n^{\frac{\beta+1}{2\beta+1}}$.

4.2. Sparse linear regression

For the linear regression model, we consider a random design setting

$$y_i = X_i^T \theta + z_i,$$

where without contamination $X_i \sim N(0, \Sigma)$ and $z_i \sim N(0, \sigma^2)$ are independent. Under the current setting, both the design and the response in the model can be contaminated. That is, we have i.i.d. observations $(X_1, y_1), ..., (X_n, y_n) \sim \mathbb{P}_{(\epsilon, \theta, Q)} = (1 - \epsilon)P_\theta + \epsilon Q$, where the P_θ denotes the probability distribution of

$$p(X, y) = p(X)p(y|X),$$

with $p(X) = N(0, \Sigma)$ and $p(y|X) = N(X^T\theta, \sigma^2)$.

Given the covariance matrix Σ of X, we further impose the *sparse eigenvalue conditions* as follows,

$$\inf_{|\mathrm{supp}(v)|=2s} \|\Sigma^{1/2}v\|/\|v\| \geq \kappa, \tag{10}$$

$$\sup_{|\mathrm{supp}(v)|=2s} \|\Sigma^{1/2}v\|/\|v\| \leq \kappa_u. \tag{11}$$

In addition, we assume $\kappa_u \asymp \kappa$. In other words, the upper and lower sparse eigenvalues are at the same order, which is satisfied, for example, if all eigenvalues of Σ are at the same order. Given noise level σ and sparse eigenvalue level κ, we consider the following sparse set as the parameter space for θ:

$$\Theta(s, M, \sigma, \kappa) = \left\{ \theta \in \mathbb{R}^p : |\mathrm{supp}(\theta)| \leq s, \|\theta\| \leq M\sigma/\kappa \right\},$$

where $s > 0$ is the sparsity of the regression coefficients and $M > 0$ is assumed to be a constant.

Remark 4.1. *The total variation distance $\mathsf{TV}(P_\theta, P_{\theta'})$ is upper bounded by $C\|\Sigma^{1/2}(\theta - \theta')\|/\sigma$ with some constant $C > 0$. Therefore, we impose an upper bound $\|\theta\| \leq M\sigma/\kappa$ for the parameter θ when defining parameter space $\Theta(s, M, \sigma, \kappa)$ to guarantee that the parameter space is totally bounded under the loss $\mathsf{TV}(\cdot, \cdot)$. This is a natural condition and is assumed at the beginning of Section 3. For this totally bounded parameter space $\Theta(s, M, \sigma, \kappa)$, the equivalence of $\mathsf{TV}(P_\theta, P_{\theta'})$ and $\|\Sigma^{1/2}(\theta - \theta')\|/\sigma$ can be further established. See Lemma 6.2 for details.*

For this set, we will show that

$$\log D_l(\delta) \lesssim s \log \frac{ep}{s} + s \log(l+1).$$

Then, using the estimator (7) with $\delta \asymp \sqrt{\frac{s \log \frac{ep}{s}}{n}}$, Theorem 3.3 implies the following convergence rate.

Corollary 4.2. *We assume $s \log \frac{ep}{s} \leq cn$ with some sufficiently small $c > 0$. Then, there are some constants C, C', such that*

$$\|\Sigma^{1/2}(\hat{\theta} - \theta)\|^2 \leq C\sigma^2 \left(\frac{s \log \frac{ep}{s}}{n} \vee \epsilon^2 \right)$$

$$\|\hat{\theta} - \theta\|^2 \leq C\frac{\sigma^2}{\kappa^2} \left(\frac{s \log \frac{ep}{s}}{n} \vee \epsilon^2 \right),$$

with $\mathbb{P}_{(\epsilon,\theta,Q)}$-probability at least $1 - \exp\left(-C'\left(s \log \frac{ep}{s} + n\epsilon^2\right)\right)$ uniformly over $\theta \in \Theta(s, M, \sigma, \kappa)$ and all Q.

We use Theorem 3.3 instead of Theorem 3.1 to derive Corollary 4.2, because Theorem 3.1 uses global metric entropy and will cause an extra logarithmic factor in the convergence rate. For the prediction error loss $\|\Sigma^{1/2}(\hat{\theta} - \theta)\|^2$, the rate does not depend on the covariance Σ of the design matrix. On the other hand, the rate for the estimation error loss $\|\hat{\theta} - \theta\|^2$ depends on the sparse eigenvalue κ of Σ.

When $\epsilon = 0$, both rates in Corollary 4.2 are known to be minimax optimal. Indeed, the lower bounds with explicit dependence on κ and σ can be found in Theorem 1(b) and Theorem 3(b) of [19], by observing $\kappa_u \asymp \kappa$. In particular, one can easily check that the least favorable subset in the lower bound construction (Proof of Theorem 1(b) in [19]) is contained in our parameter space $\Theta(s, M, \sigma, \kappa)$ for any fixed M. We emphasize that although an empirical version of sparse eigenvalue conditions of (10)-(11) is used in Assumption 3 of [19], it is well known that (see, for example, [22]) under the assumption $s \log \frac{ep}{s} \leq cn$ with some sufficiently small $c > 0$ and $\kappa_u \asymp \kappa$, our population sparse eigenvalue conditions (10)-(11) imply the empirical version with values at the same order of κ and κ_u with probability at least $1 - \exp(-Cn)$. For $\epsilon > 0$, due to the equivalence of total variation distance $\mathsf{TV}(P_\theta, P_{\theta'})$ and $\|\Sigma^{1/2}(\theta - \theta')\|/\sigma$ as well as $\kappa_u \asymp \kappa$, the modulus of continuity for the prediction error loss scales as $\omega(\epsilon, \Theta) \asymp \sigma\epsilon$, while for the estimation error loss, it scales as $\omega(\epsilon, \Theta) \asymp \sigma\epsilon/\kappa$. Hence, by Theorem 3.2, both rates in Corollary 4.2 are minimax optimal.

4.3. Low-rank trace regression

Consider the observation pair (X_i, y_i) satisfying the model

$$y_i = \mathsf{Tr}(X_i^T A) + z_i,$$

where $X_i \in \mathbb{R}^{p_1 \times p_2}$ is an observed design matrix and $A \in \mathbb{R}^{p_1 \times p_2}$ is an unknown low-rank signal matrix. The problem of recovering a high-dimensional low-rank matrix has been considered in [20, 3, 21, 13]. However, these results all assume the data are generated without contamination. In many practical situations, both the design and the response can be contaminated. For some covariance matrix $\Sigma \in \mathbb{R}^{p_1 p_2 \times p_1 p_2}$ and some number $\sigma > 0$, we assume i.i.d. observations $(X_1, y_1), ..., (X_n, y_n) \sim \mathbb{P}_{(\epsilon, A, Q)} = (1 - \epsilon)P_A + \epsilon Q$, where P_A denotes the probability distribution

$$p(X, y) = P(X)P(y|X),$$

with $p(X)$ referring to $\mathrm{vec}(X) \sim N(0, \Sigma)$ and $p(y|X) = N(\mathsf{Tr}(X^T A), \sigma^2)$.

Given the covariance matrix Σ of $\mathrm{vec}(X)$, we further impose the *restricted isometry condition* as follows,

$$\kappa \leq \inf_{\mathsf{rank}(A) \leq 2r} \frac{\|\Sigma^{1/2}\mathrm{vec}(A)\|}{\|A\|_{\mathrm{F}}} \leq \sup_{\mathsf{rank}(A) \leq 2r} \frac{\|\Sigma^{1/2}\mathrm{vec}(A)\|}{\|A\|_{\mathrm{F}}} \leq \kappa_u, \qquad (12)$$

In addition, we assume $\kappa_u \asymp \kappa$. A special case would be that all eigenvalues of Σ are at the same order. Given noise level σ and κ from the restricted isometry condition, we assume the coefficient matrix A is in a low-rank matrix class defined as

$$\mathcal{A}(r, M, \sigma, \kappa) = \left\{ A \in \mathbb{R}^{p_1 \times p_2} : \mathsf{rank}\,(A) \leq r, \|A\|_{\mathrm{F}} \leq M\sigma/\kappa \right\}.$$

The number $r > 0$ upper bounds the rank. We assume M is a constant throughout this section.

Remark 4.2. *The total variation distance* $\mathsf{TV}(P_A, P_{A'})$ *is upper bounded by* $C\|\Sigma^{1/2}(\mathrm{vec}(A) - \mathrm{vec}(A'))\|/\sigma$ *with some constant* $C > 0$. *Consequently, we impose an upper bound* $\|A\|_{\mathrm{F}} \leq M\sigma/\kappa$ *in parameter space* $\mathcal{A}(r, M, \sigma, \kappa)$ *to guarantee the parameter space is totally bounded under the loss* $\mathsf{TV}(P_A, P_{A'})$. *Similar to the setting of sparse linear regression, the equivalence of* $\mathsf{TV}(P_A, P_{A'})$ *and* $\|\Sigma^{1/2}(\mathrm{vec}(A) - \mathrm{vec}(A'))\|/\sigma$ *can be further established for* $\mathcal{A}(r, M, \sigma, \kappa)$ *in the low-rank trace regression setting. See Lemma 6.2 for details.*

For this low-rank matrix class, we will show that

$$\log D_l(\delta) \lesssim r(p_1 + p_2)\log(l + 1).$$

Then, for the estimator (7) with $\delta \asymp \sqrt{\frac{r(p_1 + p_2)}{n}}$, Theorem 3.3 implies the following convergence rate.

Corollary 4.3. *Assume* $r(p_1 + p_2) \leq cn$ *with some sufficiently small* $c > 0$. *Then, there are constants* C, C', *such that*

$$\|\hat{A} - A\|_{\mathrm{F}}^2 \leq C\frac{\sigma^2}{\kappa^2}\left(\frac{r(p_1 + p_2)}{n} \vee \epsilon^2\right),$$

with $\mathbb{P}_{(\epsilon, A, Q)}$*-probability at least* $1 - \exp\left(-C'\left(r(p_1 + p_2) + n\epsilon^2\right)\right)$ *uniformly over* $A \in \mathcal{A}(r, M, \sigma, \kappa)$ *and all* Q.

The rate consists of two parts. The first part is the usual low-rank matrix estimation rate $\frac{\sigma^2 r(p_1+p_2)}{\kappa^2 n}$, which is known to be minimax optimal with explicit dependence on both σ and κ when $\epsilon = 0$. See, for example, Theorem 5 of [13]. To interpret this lower bound in [13], we emphasize that a similar restricted isometry condition as in (12) is imposed in Assumption 2 of [13] with $\mu \asymp \kappa^{-1}$ and $\|A\|_{L_2(\Pi)} \asymp \|\Sigma^{1/2}\mathrm{vec}(A)\|$ in our setting respectively. In addition, it is easy to calculate that the least favorable subset $\mathcal{B}(\mathcal{C})$ in the construction of lower bound in [13] is indeed contained in our parameter space with any fixed M, due to the condition that $r(p_1 + p_2) \leq cn$ with some sufficiently small $c > 0$. The second part is $\frac{\sigma^2 \epsilon^2}{\kappa^2}$, which is contributed by the modulus of continuity $\omega^2(\epsilon, \mathcal{A})$ for this problem, noting that $\kappa_u \asymp \kappa$ in (12) and the equivalence of $\mathsf{TV}(P_A, P_{A'})$ and $\|\Sigma^{1/2}(\mathrm{vec}(A) - \mathrm{vec}(A'))\|/\sigma$ in Remark 4.2. Therefore, by Theorem 3.2, the upper bound in Corollary 4.3 is minimax optimal.

5. Discussion with an example under supreme norm

This paper gives a general framework to construct robust estimators under Huber's ϵ-contamination model. The key idea of the construction lies in the robust testing procedure Scheffé estimate. We emphasize that this robust testing procedure enjoys a desired error exponent that depends on the total variation distance, which is intrinsic to Huber's robust setting. This new result is stated precisely in Theorem 2.1. Consequently, the rate-optimal estimators that we present in Section 4 all depend on the general theorems in Section 3 under loss functions that are equivalent to the total variation distance. However, it is unknown whether the theory can be extended to some important loss functions that are not equivalent to the total variation distance. In this section, we give an example for a supreme norm loss function in the context of a nonparametric white noise model. We show that the minimax rate of the problem depends on the contamination proportion in a different way. The general treatment for non-intrinsic loss functions will be considered as future projects.

The white noise model [17] is considered to be a standard nonparametric model for function estimation [2, 16]. By observing the stochastic process

$$dY_t = f(t)dt + \frac{1}{\sqrt{n}}dW_t, \quad t \in [0,1], \tag{13}$$

with a standard Wiener process $\{W_t\}_{t \in [0,1]}$, the goal is to estimate the function f. Equivalently, (13) can be written as an i.i.d. model. That is, we observe i.i.d. stochastic processes $\{Y_{t,1}\}_{t \in [0,1]}, ..., \{Y_{t,n}\}_{t \in [0,1]} \sim P_f$, where P_f denotes the probability distribution

$$dY_{t,i} = f(t)dt + dW_{t,i}, \tag{14}$$

Under Huber's framework, there is an ϵ probability of contamination, and we observe i.i.d. stochastic processes $\{Y_{t,1}\}_{t \in [0,1]}, ..., \{Y_{t,n}\}_{t \in [0,1]} \sim \mathbb{P}_{(\epsilon, f, Q)} = (1 -$

$\epsilon)P_f + \epsilon Q$. We use a slightly modified version of Hölder class defined in (9):

$$\mathcal{H}(\beta, M) = \left\{ f = \sum_{l \geq 0, 0 \leq k \leq 2^l - 1} f_{lk}\psi_{lk} : \sup_{l \geq 0, 0 \leq k \leq 2^l - 1} 2^{l(1/2+\beta)}|f_{lk}| \leq M \right\},$$

where $\{\phi_{lk}\}_{l \geq 0, 0 \leq k \leq 2^l - 1}$ is an orthogonal wavelet basis on the interval $[0, 1]$, see [5] for the detailed construction.

We are going to construct an estimator that achieves the optimal rate under the supreme loss $\|\hat{f} - f\|_\infty$. Let L be the largest integer such that $2^L \leq \left(\frac{\log n}{n} \vee \epsilon^2 \right)^{-\frac{1}{2\beta+1}}$. The estimator is $\hat{f} = \sum_{0 \leq l \leq L, 0 \leq k \leq 2^l - 1} \hat{f}_{lk}\psi_{lk}$ for

$$\hat{f}_{lk} = \text{Median}\left(\{y_{lk,i}\}_{i=1}^n \right),$$

where $y_{lk,i} = \int_0^1 \psi_{lk}(t)dY_{t,i}$ are empirical wavelet coefficients.

Theorem 5.1. *Assume $\epsilon < 1/4$. For the Hölder class $\mathcal{H}(\beta, M)$, there are constants C, C', such that*

$$\|\hat{f} - f\|_\infty^2 \leq C \left[\left(\frac{n}{\log n} \right)^{-\frac{2\beta}{2\beta+1}} \vee \epsilon^{\frac{4\beta}{2\beta+1}} \right],$$

with $\mathbb{P}_{(\epsilon,f,Q)}$-probability at least $1 - \exp\left(-C'\left(\log n + n\epsilon^2 \right) \right)$ uniformly over $f \in \mathcal{H}(\beta, M)$ and all Q.

This theorem characterizes the upper bound of this problem. By applying Theorem 3.2, we show it is also the minimax lower bound.

Corollary 5.1. *There are some constants $C, c > 0$ such that*

$$\inf_{\hat{f}} \sup_{f \in \mathcal{H}(\beta,M),Q} \mathbb{P}_{(\epsilon,f,Q)} \left\{ \|\hat{f} - f\|_\infty^2 > C \left[\left(\frac{n}{\log n} \right)^{-\frac{2\beta}{2\beta+1}} \vee \epsilon^{\frac{4\beta}{2\beta+1}} \right] \right\} > c.$$

Combining Theorem 5.1 and Corollary 5.1, we conclude that $\left(\frac{n}{\log n} \right)^{-\frac{2\beta}{2\beta+1}} \vee \epsilon^{\frac{4\beta}{2\beta+1}}$ is the minimax rate for estimating a nonparametric drift function f under the supreme loss in Huber's framework. Compared with Corollary 4.1, the dependence on the contamination proportion is through $\epsilon^{\frac{4\beta}{2\beta+1}}$ instead of the usual ϵ^2 for the total variation loss. This is because for the supreme loss, $\epsilon^{\frac{2\beta}{2\beta+1}}$ is the modulus of continuity defined in Theorem 3.2. When $\epsilon = 0$, the rate reduces to the usual nonparametric rate for supreme loss [23].

Remark 5.1. *Note that the estimator $\hat{f}$ does not use the general construction in Section 3. As a consequence, it requires the knowledge of the contamination proportion ϵ. However, it reveals a minimax rate with an interesting dependence on ϵ, which is different from the rates of the estimators in Section 3 and Section 4. It is of great interest to us how to construct an estimator that is adaptive to ϵ for the supreme loss. A more general open question is to seek ways of construction of estimators for other non-intrinsic loss functions.*

6. Proofs

This section collects the proofs of all technical results in the paper. The proofs of the results in Section 2 and Section 3 are given in Section 6.1. The proofs of the results in Section 4 and Section 5 are given in Section 6.2 and Section 6.3, respectively.

6.1. Proofs in Section 2 and Section 3

Before stating the proofs of the main theorems, we need the following lemma to upper bound the testing error with respect to distributions in a total variation neighborhood.

Lemma 6.1. *Consider the testing function ϕ in the form of (5). Assume $\mathsf{TV}(P_0, P_1) > 2\xi$, and then*

$$\sup_{\{P:\mathsf{TV}(P,P_0)\leq\xi\}} P\phi \quad \leq \quad 2\exp\left(-\frac{1}{2}n\left(\mathsf{TV}(P_0,P_1)-2\xi\right)^2\right),$$

$$\sup_{\{P:\mathsf{TV}(P,P_1)\leq\xi\}} P(1-\phi) \quad \leq \quad 2\exp\left(-\frac{1}{2}n\left(\mathsf{TV}(P_0,P_1)-2\xi\right)^2\right).$$

Proof. Since the proofs of the two inequalities are the same, we only give details for the first one. For any P such that $\mathsf{TV}(P, P_0) \leq \xi$, we have

$$
\begin{aligned}
P\phi &= P\left\{|P_n(A) - P_0(A)| > |P_n(A) - P_1(A)|\right\} \\
&\leq P\left\{|P_n(A) - P_0(A)| > |P_0(A) - P_1(A)| - |P_n(A) - P_0(A)|\right\} \quad (15) \\
&= P\left\{2|P_n(A) - P_0(A)| > \mathsf{TV}(P_0, P_1)\right\} \quad (16) \\
&\leq P\left\{2|P_n(A) - P(A)| > \mathsf{TV}(P_0, P_1) - 2\xi\right\} \quad (17) \\
&\leq 2\exp\left(-\frac{1}{2}n\left(\mathsf{TV}(P_0,P_1)-2\xi\right)^2\right). \quad (18)
\end{aligned}
$$

The inequality (15) is due to triangle inequality. By rearrangement and the definition of total variation distance, we get (16). Then, (17) is obtained through triangle inequality and the fact that $|P(A) - P_0(A)| \leq \mathsf{TV}(P, P_0) \leq \xi$. Finally, (18) is by Hoeffding's inequality. Taking supreme over the set $\{P : \mathsf{TV}(P, P_0) \leq \xi\}$, the proof is complete. □

Now we are ready to give the proofs of the main theorems.

Proof of Theorem 2.1. Note that

$$\{(1-\epsilon)P_0 + \epsilon Q : Q\} \subset \{P : \mathsf{TV}(P, P_0) \leq \epsilon\},$$

and

$$\{(1-\epsilon)P_1 + \epsilon Q : Q\} \subset \{P : \mathsf{TV}(P, P_1) \leq \epsilon\}.$$

Thus, the proof is complete. □

Proof of Theorem 3.1. Let us use the notation $\phi_j = \sum_{k \neq j} \phi_{jk}$ and $\Theta_j = \{\theta \in \Theta : \mathsf{TV}(P_\theta, P_{\theta_j}) \leq \delta\}$. For some $c \in (0,1)$, let $\mathcal{N}_j = \{k \neq j : \mathsf{TV}(P_{\theta_k}, P_{\theta_j}) \leq c\eta\}$. Then, for $P = (1-\epsilon)P_\theta + \epsilon Q$ with any $\theta \in \Theta_j$ and any Q, we have

$$P\left\{\mathsf{TV}(P_{\hat{\theta}}, P_\theta) > \eta + \delta\right\}$$

$$\leq P\left\{\mathsf{TV}(P_{\theta_{\hat{j}}}, P_{\theta_j}) > \eta\right\} \tag{19}$$

$$\leq P\left\{\phi_j \geq \min_{\{k:\mathsf{TV}(P_{\theta_k}, P_{\theta_j}) > \eta\}} \phi_k\right\} \tag{20}$$

$$\leq P\{\phi_j > |\mathcal{N}_j|\} + P\left\{\min_{\{k:\mathsf{TV}(P_{\theta_k}, P_{\theta_j}) > \eta\}} \phi_k < |\mathcal{N}_j| + 1\right\} \tag{21}$$

$$\leq P\{\phi_{jk} = 1 \text{ for some } k \notin \mathcal{N}_j\}$$
$$+ \sum_{\{k:\mathsf{TV}(P_{\theta_k}, P_{\theta_j}) > \eta\}} P\{\phi_{kl} = 0 \text{ for some } l \in \mathcal{N}_j \cup \{j\}\}$$

$$\leq \sum_{k \notin \mathcal{N}_j} P\phi_{jk} + \sum_{\{k:\mathsf{TV}(P_{\theta_k}, P_{\theta_j}) > \eta\}} \sum_{l \in \mathcal{N}_j \cup \{j\}} P(1 - \phi_{kl}) \tag{22}$$

$$\leq 2\mathcal{M}(\delta, \Theta, \mathsf{TV}(\cdot, \cdot)) \exp\left(-\frac{1}{2}n\left(c\eta - 2(\epsilon + \delta)\right)^2\right) \tag{23}$$
$$+ 2\mathcal{M}^2(\delta, \Theta, \mathsf{TV}(\cdot, \cdot)) \exp\left(-\frac{1}{2}n\left((1-c)\eta - 2(\epsilon + \delta + c\eta)\right)^2\right).$$

The inequality (19) is by $\theta \in \Theta_j$. Suppose $\phi_j < \min_{\{k:\mathsf{TV}(P_{\theta_k}, P_{\theta_j}) > \eta\}} \phi_k$, we must have $\mathsf{TV}(P_{\theta_{\hat{j}}}, P_{\theta_j}) \leq \eta$ by the definition of $\hat{j}$ in (7). Therefore,

$$\left\{\mathsf{TV}(P_{\theta_{\hat{j}}}, P_{\theta_j}) > \eta\right\} \subset \left\{\phi_j \geq \min_{\{k:\mathsf{TV}(P_{\theta_k}, P_{\theta_j}) > \eta\}} \phi_k\right\},$$

which implies (20). The inequality (21) uses the fact that $\{x \geq y\} \subset \{x > z\} \cup \{y < z + 1\}$. Finally, (23) is obtained by applying Lemma 6.1 with the relations

$$\{(1-\epsilon)P_\theta + \epsilon Q : \theta \in \Theta_j, Q\} \subset \{P : \mathsf{TV}(P, P_{\theta_j}) \leq \epsilon + \delta\},$$

and

$$\{(1-\epsilon)P_\theta + \epsilon Q : \theta \in \Theta_j, Q\} \subset \{P : \mathsf{TV}(P, P_{\theta_l}) \leq \epsilon + \delta + c\eta\}.$$

The proof is complete by choosing $c = \frac{1}{4}$. $\qquad\square$

Proof of Theorem 3.3. Let us use the notation $\phi_j = \sum_{k \neq j} \phi_{jk}$ and $\Theta_j = \{\theta \in \Theta : \mathsf{TV}(P_\theta, P_{\theta_j}) \leq \delta\}$. For some $c \in (0,1)$, let $\mathcal{N}_j = \{k \neq j : \mathsf{TV}(P_{\theta_k}, P_{\theta_j}) \leq L\delta/4\}$. Then, for $P = (1-\epsilon)P_\theta + \epsilon Q$ with any $\theta \in \Theta_j$ and any Q, we have

$$P\left\{\mathsf{TV}(P_{\hat{\theta}}, P_\theta) > (L+1)\delta\right\}$$

$$\leq \sum_{\{k:\mathsf{TV}(P_{\theta_k},P_{\theta_j})> \frac{L\delta}{4}\}} P\phi_{jk} + \sum_{\{k:\mathsf{TV}(P_{\theta_k},P_{\theta_j})>L\delta\}} \sum_{\{t:\mathsf{TV}(P_{\theta_t},P_{\theta_j})\leq \frac{L\delta}{4}\}} P(1-\phi_{kt}).$$

This is by the same argument for deriving (22) in the proof of Theorem 3.1. Then, we have

$$\sum_{\{k:\mathsf{TV}(P_{\theta_k},P_{\theta_j})>L\delta/4\}} P\phi_{jk}$$

$$\leq \sum_{l\geq L/4} \sum_{\{k:l\delta<\mathsf{TV}(P_{\theta_k},P_{\theta_j})\leq(l+1)\delta\}} P\phi_{jk}$$

$$\leq 2\sum_{l\geq L/4} D_l(\delta)\exp\left(-\frac{1}{2}n(l\delta-2(\epsilon+\delta)^2)\right),$$

where the last inequality is by

$$|\{k:l\delta<\mathsf{TV}(P_{\theta_k},P_{\theta_j})\leq(l+1)\delta\}|\leq D_l(\delta), \tag{24}$$

and Lemma 6.1 with the relation

$$\{(1-\epsilon)P_\theta+\epsilon Q:\theta\in\Theta_j,Q\}\subset\{P:\mathsf{TV}(P,P_{\theta_j})\leq\epsilon+\delta\}.$$

We also have

$$\sum_{\{k:\mathsf{TV}(P_{\theta_k},P_{\theta_j})>L\delta\}} \sum_{\{t:\mathsf{TV}(P_{\theta_t},P_{\theta_j})\leq L\delta/4\}} P(1-\phi_{kt})$$

$$\leq \sum_{l\geq L} \sum_{\{k:l\delta<\mathsf{TV}(P_{\theta_k},P_{\theta_j})\leq(l+1)\delta\}} \sum_{\{t:\mathsf{TV}(P_{\theta_t},P_{\theta_j})\leq L\delta/4\}} P(1-\phi_{kt})$$

$$\leq 2\left[\sum_{l=0}^{L/4-1} D_l(\delta)\right]\sum_{l\geq L} D_l(\delta)\exp\left(-\frac{1}{2}n\left(l\delta-L\delta/4-2(\epsilon+\delta+L\delta/4)\right)^2\right),$$

where the last inequality follows from (24),

$$\left|\{t\neq j:\mathsf{TV}(P_{\theta_t},P_{\theta_j})\leq L\delta/4\}\right|\leq\sum_{l=0}^{L/4-1} D_l(\delta),$$

and Lemma 6.1 with the relations

$$\{(1-\epsilon)P_\theta+\epsilon Q:\theta\in\Theta_j,Q)\}\subset\{P:\mathsf{TV}(P,P_{\theta_t})\leq\epsilon+\delta+L\delta/4\}$$

for any θ_t such that $\mathsf{TV}(P_{\theta_t},P_{\theta_j})\leq L\delta/4$. Combining the bounds above, the proof is complete. □

6.2. Proofs in Section 4

First, we give a lemma that establishes the equivalence between total variation distance and ℓ_2 norm for linear regression and trace regression.

Lemma 6.2. *Assume the upper sparse eigenvalue condition in (11) holds. For P_θ specified in Section 4.2, there are constants C_1, C_2, such that*

$$C_1 \frac{\|\Sigma^{1/2}(\theta - \theta')\|}{\sigma} \le \mathsf{TV}(P_\theta, P_{\theta'}) \le C_2 \frac{\|\Sigma^{1/2}(\theta - \theta')\|}{\sigma},$$

for any $\theta, \theta' \in \Theta(s, M, \sigma, \kappa)$. Similarly, assume the restricted isometry condition in (12) holds. For P_A specified in Section 4.3, there are constants C_1, C_2, such that

$$C_1 \frac{\|\Sigma^{1/2}(\mathrm{vec}(A) - \mathrm{vec}(A'))\|}{\sigma} \le \mathsf{TV}(P_A, P_{A'}) \le C_2 \frac{\|\Sigma^{1/2}(\mathrm{vec}(A) - \mathrm{vec}(A'))\|}{\sigma},$$

for any $A, A' \in \mathcal{A}(r, M, \sigma, \kappa)$.

Proof. Since the proofs of the two inequalities are nearly identical, we only give details for the first one. The density function of P_θ is

$$(2\pi)^{-p/2} |\Sigma|^{-1/2} e^{-\frac{1}{2} X^T \Sigma^{-1} X} \times \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(y - X^T \theta)^2},$$

where $|\Sigma|$ is the determinant of Σ. Therefore, by the definition of total variation distance, we have

$$
\begin{aligned}
&\mathsf{TV}(P_\theta, P_{\theta'}) \\
= \ & P_\theta \left\{ (y - X^T \theta)^2 < (y - X^T \theta')^2 \right\} - P_{\theta'} \left\{ (y - X^T \theta)^2 < (y - X^T \theta')^2 \right\}.
\end{aligned}
$$

Note that

$$
\begin{aligned}
&P_\theta \left\{ (y - X^T \theta)^2 < (y - X^T \theta')^2 \right\} \\
= \ & P_\theta \left\{ \frac{(y - X^T \theta)}{\sigma} \frac{(X^T(\theta - \theta'))}{|(X^T(\theta - \theta'))|} > -\frac{|X^T(\theta - \theta')|}{2\sigma} \right\} \\
= \ & \mathbb{E}\Phi \left(\frac{|X^T(\theta - \theta')|}{2\sigma} \right),
\end{aligned}
$$

where Φ is the cumulative distribution function of $N(0,1)$ and the last equality is because $\frac{(y - X^T\theta)}{\sigma} \frac{(X^T(\theta - \theta'))}{|(X^T(\theta - \theta'))|}$ is distributed by $N(0,1)$ conditioning on X. Hence,

$$\mathsf{TV}(P_\theta, P_{\theta'}) = 2\mathbb{E}\Phi \left(\frac{|X^T(\theta - \theta')|}{2\sigma} \right) - 1 = \mathbb{E} \int_{-\frac{|X^T(\theta - \theta')|}{2\sigma}}^{\frac{|X^T(\theta - \theta')|}{2\sigma}} \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}} dt. \quad (25)$$

An upper bound for (25) is

$$\frac{1}{\sqrt{2\pi}} \mathbb{E} \frac{|X^T(\theta - \theta')|}{\sigma} = \frac{\|\Sigma^{1/2}(\theta - \theta')\|}{\sigma\sqrt{2\pi}} \mathbb{E}|Z|,$$

for $Z \sim N(0,1)$. A lower bound for (25) is

$$\mathbb{E} \frac{1}{\sqrt{2\pi}} e^{-\frac{|X^T(\theta - \theta')|^2}{8\sigma^2}} \frac{|X^T(\theta - \theta')|}{\sigma}$$

$$
\begin{aligned}
&= \frac{\|\Sigma^{1/2}(\theta - \theta')\|}{\sigma\sqrt{2\pi}} \mathbb{E} e^{-\frac{\|\Sigma^{1/2}(\theta-\theta')\|^2}{8\sigma^2}|Z|^2} |Z| \\
&\geq \frac{\|\Sigma^{1/2}(\theta - \theta')\|}{\sigma\sqrt{2\pi}} \mathbb{E} e^{-C^2 M^2 |Z|^2/2} |Z|,
\end{aligned}
$$

where $Z \sim N(0,1)$ and the last inequality follows from the upper sparse eigenvalue condition in (11) and the fact $\|\theta\| \leq M\sigma/\kappa$. Hence, we have proved that

$$
C_1 \frac{\|\Sigma^{1/2}(\theta - \theta')\|}{\sigma} \leq \mathsf{TV}(P_\theta, P_{\theta'}) \leq C_2 \frac{\|\Sigma^{1/2}(\theta - \theta')\|}{\sigma},
$$

with $C_1 = \frac{1}{\sqrt{2\pi}} \mathbb{E} e^{-C^2 M^2 |Z|^2/2} |Z|$ and $C_2 = \frac{1}{\sqrt{2\pi}} \mathbb{E} |Z|$. $\qquad\square$

With the help of the above lemma, we are ready to give the proofs of the results in Section 4.

Proof of Corollary 4.1. The result directly follows Theorem 3.1 by realizing that $\mathsf{TV}(P_{f_1}, P_{f_2}) = \frac{1}{2}\|f_1 - f_2\|_1$. $\qquad\square$

Proof of Corollary 4.2. We use the estimator (7) with $\delta = \sqrt{\frac{s \log \frac{ep}{s}}{n}}$. Here we work with a δ-packing set $\Theta' = \{\theta_1, ..., \theta_m\}$ of maximum cardinality in the sense that $\min_{i\neq j} \mathsf{TV}(P_{\theta_i}, P_{\theta_j}) \geq \delta$ with the largest possible m. The value $m = \mathcal{N}(\delta, \Theta, \mathsf{TV}(\cdot, \cdot))$ is called δ-packing number. It is easy to see that Θ' is also a δ-covering set and m is equal to δ-covering number up to a constant factor. Indeed, δ-covering and δ-packing numbers are (up to a constant factor) essentially the same, i.e., $\mathcal{M}(\delta, \Theta, \mathsf{TV}(\cdot, \cdot)) \leq \mathcal{N}(\delta, \Theta, \mathsf{TV}(\cdot, \cdot)) \leq \mathcal{M}(\delta/2, \Theta, \mathsf{TV}(\cdot, \cdot))$. See, for example, [18]. According to Lemma 6.2, we have that $\min_{i\neq j} \|\Sigma^{1/2}(\theta_i - \theta_j)\| \geq \sigma\delta/(C_2)$. Hence, for any θ_0, we have

$$
\begin{aligned}
&\quad |\{\theta \in \Theta' : l\delta < \mathsf{TV}(P_\theta, P_{\theta_0}) \leq (l+1)\delta\}| \\
&\leq |\{\theta \in \Theta' : \mathsf{TV}(P_\theta, P_{\theta_0}) \leq (l+1)\delta\}| \\
&\leq \sum_{|S|\leq s} |\{\theta \in \Theta' : \mathrm{supp}(\theta) = S, \mathsf{TV}(P_\theta, P_{\theta_0}) \leq (l+1)\delta\}| \\
&\leq \sum_{|S|\leq s} \left| \{\theta \in \Theta' : \mathrm{supp}(\theta) = S, \|\Sigma^{1/2}(\theta - \theta_0)\| \leq \sigma(l+1)\delta/C_1 \} \right| \\
&\leq \exp\left(s \log \frac{ep}{s}\right)(l+1)^{C_3 s},
\end{aligned}
$$

where the last inequality is through a volume ratio argument [18]. Taking supreme over θ_0, we have

$$
\log D_l(\delta) \leq C_4 \left(s \log \frac{ep}{s} + s \log(l+1)\right).
$$

Using Theorem 3.3 with $L = \left\lfloor C_5 \frac{\delta + \epsilon}{\delta} \right\rfloor$ for some large $C_5 > 0$, direct calculation gives that

$$
\mathsf{TV}(P_{\hat\theta}, P_\theta) \leq C_6(\delta + \epsilon),
$$

for some $C_6 > 0$, with probability at least $1 - \exp\left(-C'n(\delta^2 + \epsilon^2)\right)$, where $\delta = \sqrt{\frac{s \log \frac{ep}{s}}{n}}$. By Lemma 6.2 and the definition of κ, we obtain the convergence rate with the desired loss functions. Thus, the proof is complete. $\qquad\square$

Proof of Corollary 4.3. We use the estimator (7) with $\delta = \sqrt{\frac{r(p_1+p_2)}{n}}$. Similar to the argument in the proof of Corollary 4.2, we work with a δ-packing set $\mathcal{A}' = \{A_1, ..., A_m\}$ of maximum cardinality in the sense that $\min_{i \neq j} \mathsf{TV}(P_{A_i}, P_{A_j}) \geq \delta$ with the largest possible m. It is easy to see $\mathcal{A}'$ is also a δ-covering set. According to Lemma 6.2, we have that $\min_{i \neq j} \|\Sigma^{1/2}(\text{vec}(A_i) - \text{vec}(A_j))\| \geq \sigma\delta/(C_2)$ and consequently $\min_{i \neq j} \|A_i - A_j\|_{\text{F}} \geq C_3\sigma\delta/\kappa$ from the restricted isometry condition in (12). Hence, we have

$$
\begin{aligned}
& |\{A \in \mathcal{A}' : l\delta < \mathsf{TV}(P_A, P_{A_0}) \leq (l+1)\delta\}| \\
\leq \quad & |\{A \in \mathcal{A}' : \mathsf{TV}(P_A, P_{A_0}) \leq (l+1)\delta\}| \\
\leq \quad & \left|\{A \in \mathcal{A}' : \|\Sigma^{1/2}(\text{vec}(A) - \text{vec}(A_0))\| \leq \sigma(l+1)\delta/C_1\}\right| \\
\leq \quad & |\{A \in \mathcal{A}' : \|A - A_0\|_{\text{F}} \leq C_4\sigma(l+1)\delta/\kappa\}| \\
\leq \quad & (l+1)^{C_5 r(p_1+p_2)},
\end{aligned}
$$

where the second inequality follows from Lemma 6.2, the third inequality follows from the restricted isometry condition in (12), and the last inequality is due to Lemma 3.1 of [3] and the fact $\min_{i \neq j} \|A_i - A_j\|_{\text{F}} \geq C_3\sigma\delta/\kappa$. Taking supreme over θ_0, we have

$$\log D_l(\delta) \leq C_5 r(p_1 + p_2) \log(l+1).$$

Using Theorem 3.3 with $L = \left\lfloor C_6 \frac{\delta + \epsilon}{\delta} \right\rfloor$ for some large $C_6 > 0$, direct calculation gives that

$$\mathsf{TV}(P_{\hat{A}}, P_A) \leq C_7(\delta + \epsilon),$$

for some $C_7 > 0$, with probability at least $1 - \exp\left(-C'n(\delta^2 + \epsilon^2)\right)$, where $\delta = \sqrt{\frac{r(p_1+p_2)}{n}}$. By Lemma 6.2 and the definition of κ, we obtain the convergence rate with the desired loss function. Thus, the proof is complete. $\qquad\square$

6.3. Proofs in Section 5

Before stating the proofs of Theorem 5.1 and Corollary 5.1, we present a lemma that establishes equivalence between different loss functions.

Lemma 6.3. *For P_f specified in Section 5, there are constants C_1, C_2, C_3, C_4, such that*

$$C_2^{-1}\|f_1 - f_2\|_\infty \leq \sum_{l \geq 0} 2^{l/2} \max_{0 \leq k \leq 2^l - 1} |f_{1,lk} - f_{2,lk}| \leq C_1^{-1}\|f_1 - f_2\|_\infty,$$

$$C_3\|f_1 - f_2\| \leq \mathsf{TV}(P_{f_1}, P_{f_2}) \leq C_4\|f_1 - f_2\|,$$

for all $f_1, f_2 \in \mathcal{H}(\beta, L)$, where $\{f_{1,lk}\}$ and $f_{2,lk}$ are wavelet coefficients of f_1 and f_2, and $\|\cdot\|$ is understood as both vector and function ℓ_2 norm.

Proof. It is well known that two term $\sum_{l \geq 0} 2^{l/2} \max_{0 \leq k \leq 2^l-1} |f_{1,lk} - f_{2,lk}|$ and $\|f_1 - f_2\|_\infty$ are equivalent in the wavelet literature. See, for example, [9]. The equivalence implies that $\mathcal{H}(\beta, L)$ is a subset of an ℓ_2 ball. Indeed, for any $f \in \mathcal{H}(\beta, L)$, we have that $\|f\| \leq \|f\|_\infty$, which implies

$$\|f\| \leq C_2 \sum_{l \geq 0} 2^{l/2} \max_{0 \leq k \leq 2^l-1} |f_{lk}| \leq C_2 \sum_{l \geq 0} 2^{l/2} M 2^{-l(1/2+\beta)} \leq \frac{1}{1-2^{-\beta}} C_2 M. \tag{26}$$

To study $\mathsf{TV}(P_{f_1}, P_{f_2})$, we use an equivalent model of (14) in terms of wavelet coefficients. That is,

$$y_{lk} = f_{lk} + z_{lk}, \quad l \geq 0, 0 \leq k \leq 2^l - 1, \tag{27}$$

where $\{z_{lk}\}$ are i.i.d. $N(0,1)$. Then, direct calculation gives

$$\mathsf{TV}(P_{f_1}, P_{f_2}) = 2\Phi\left(\frac{\|f_1 - f_2\|}{2}\right) - 1 = \int_{-\frac{\|f_1-f_2\|}{2}}^{\frac{\|f_1-f_2\|}{2}} \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}} dt. \tag{28}$$

An upper bound for (28) is $\frac{1}{\sqrt{2\pi}} \|f_1 - f_2\|$. A lower bound for (28) is

$$\frac{1}{\sqrt{2\pi}} e^{-\frac{\|f_1-f_2\|^2}{8}} \|f_1 - f_2\| \geq \frac{1}{\sqrt{2\pi}} e^{-\frac{C_2^2 M^2}{2(1-2^{-\beta})^2}} \|f_1 - f_2\|,$$

where we have used (26). Thus, the proof is complete. $\square$

The next lemma characterizes the statistical property of a median estimator under Huber's ϵ-contamination model.

Lemma 6.4. *Assume $\epsilon < 1/4$. There exists a constant $C > 0$, such that for each $0 \leq l \leq L$ and $0 \leq k \leq 2^l - 1$, we have*

$$\sup_{f \in \mathcal{H}(\beta, M), Q} \mathbb{P}_{(\epsilon, f, Q)} \left\{ |\hat{f}_{lk} - f_{lk}| > C\left(\sqrt{\frac{\log(1/\delta)}{n}} \vee \epsilon\right) \right\} \leq 2\delta,$$

for any $\delta > 0$ that $\sqrt{\frac{\log(1/\delta)}{n}}$ is sufficiently small.

Proof. Since $y_{lk,i} \sim N(f_{lk}, 1)$, the setting is a special case of Theorem 2.1 in [4]. A careful examination of its proof gives the desired result. $\square$

Now we give the proofs of 5.1 and Corollary 5.1 with the facility of the above two lemmas.

Proof of Theorem 5.1. Note that

$$\sum_{l \geq 0} 2^{l/2} \max_{0 \leq k \leq 2^l-1} |\hat{f}_{lk} - f_{lk}| = \sum_{l \leq L} 2^{l/2} \max_{0 \leq k \leq 2^l-1} |\hat{f}_{lk} - f_{lk}| + \sum_{l > L} 2^{l/2} \max_{0 \leq k \leq 2^l-1} |f_{lk}|.$$

It is sufficient to give upper bounds for the two terms. Since $f \in \mathcal{H}(\beta, M)$,

$$\sum_{l>L} 2^{l/2} \max_{0 \leq k \leq 2^l-1} |f_{lk}| \leq \sum_{l>L} 2^{l/2} M 2^{-l(1/2+\beta)} \leq \frac{2M}{1-2^{-\beta}} \left(\frac{\log n}{n} \vee \epsilon^2 \right)^{\frac{\beta}{2\beta+1}},$$

by the definition of L. Using Lemma 6.4 with $2\delta = \exp\left(-C'(\log n + n\epsilon^2)\right)$ for some constant $C' > 0$ and a union bound argument, we have

$$\max_{l \geq L, 0 \leq k \leq 2^l-1} |\hat{f}_{lk} - f_{lk}| \leq \bar{C} \left(\sqrt{\frac{\log n}{n}} \vee \epsilon \right),$$

with probability at least $1 - \exp\left(-C'(\log n + n\epsilon^2)\right)$. Therefore,

$$\sum_{l \leq L} 2^{l/2} \max_{0 \leq k \leq 2^l-1} |\hat{f}_{lk} - f_{lk}| \leq \bar{C} \left(\sqrt{\frac{\log n}{n}} \vee \epsilon \right) \sum_{l \leq L} 2^{l/2} \leq \tilde{C} \left(\frac{\log n}{n} \vee \epsilon^2 \right)^{\frac{\beta}{2\beta+1}}.$$

Hence,

$$\sum_{l \geq 0} 2^{l/2} \max_{0 \leq k \leq 2^l-1} |\hat{f}_{lk} - f_{lk}| \leq C \left(\frac{\log n}{n} \vee \epsilon^2 \right)^{\frac{\beta}{2\beta+1}},$$

with probability at least $1 - \exp\left(-C'(\log n + n\epsilon^2)\right)$. By Lemma 6.3, the same bound holds for $\|\hat{f} - f\|_\infty$, and the proof is complete. $\square$

Proof of Corollary 5.1. The lower bound $\mathcal{R}(0) \vee \omega(\epsilon, \mathcal{H}(\beta, M))$ immediately follows from Theorem 3.2. In this problem, it is known that $\mathcal{R}(0) \asymp \left(\frac{n}{\log n} \right)^{-\frac{2\beta}{2\beta+1}}$. See, for example, [23]. Therefore, it is sufficient to calculate the modulus of continuity $\omega(\epsilon, \mathcal{H}(\beta, L))$. Define $\bar{l}$ to be the greatest integer such that $2^{\bar{l}(1/2+\beta)}\epsilon \leq M$. Then, let $f_1 = 0$ and $f_2 = f_1 + \epsilon\psi_{\bar{l}1}$. It is easy to see that $f_1, f_2 \in \mathcal{H}(\beta, M)$. By Lemma 6.3, $\mathsf{TV}(P_{f_1}, P_{f_2}) \leq C_4\|f_1 - f_2\| = (2\pi)^{-1/2}\epsilon \leq \epsilon/(1-\epsilon)$, where $C_4 = (2\pi)^{-1/2}$ according to the proof of Lemma 6.3. Moreover, we have that

$$\|f_1 - f_2\|_\infty \geq C_1 \sum_{l \geq 0} 2^{l/2} \max_{0 \leq k \leq 2^l-1} |f_{1,lk} - f_{2,lk}| \geq C_1 2^{\bar{l}/2}\epsilon \gtrsim \epsilon^{\frac{2\beta}{2\beta+1}}.$$

Hence, $\omega(\epsilon, \mathcal{H}(\beta, M)) \gtrsim \epsilon^{\frac{2\beta}{2\beta+1}}$, and the proof is complete. $\square$

Acknowledgements

References

[1] BIRGÉ, L. (1984). Sur un theórème de minimax et son application aux tests. *Probability and Mathematical Statistics* **3** 259–282. MR0764150

[2] BROWN, L. D. and LOW, M. G. (1996). Asymptotic equivalence of nonparametric regression and white noise. *The Annals of Statistics* **24** 2384–2398. MR1425958

[3] CANDES, E. J. and PLAN, Y. (2011). Tight oracle inequalities for low-rank matrix recovery from a minimal number of noisy random measurements. *Information Theory, IEEE Transactions on* **57** 2342–2359. MR2809094

[4] CHEN, M., GAO, C. and REN, Z. (2015). Robust covariance and scatter matrix estimation via matrix depth. *arXiv preprint arXiv:1506.00691.*

[5] COHEN, A., DAUBECHIES, I. and VIAL, P. (1993). Wavelets on the interval and fast wavelet transforms. *Applied and Computational Harmonic Analysis* **1** 54–81. MR1256527

[6] DEVROYE, L. and LUGOSI, G. (2012). *Combinatorial Methods in Density Estimation.* Springer Science & Business Media. MR1843146

[7] DONOHO, D. L. (1994). Statistical estimation and optimal recovery. *The Annals of Statistics* **22** 238–270. MR1272082

[8] DONOHO, D. L. and LIU, R. C. (1991). Geometrizing rates of convergence, III. *The Annals of Statistics* **19** 668–701. MR1105839

[9] HOFFMANN, M., ROUSSEAU, J. and SCHMIDT-HIEBER, J. (2015). On adaptive posterior concentration rates. *The Annals of Statistics* **43** 2259–2295. MR3396985

[10] HUBER, P. J. (1964). Robust estimation of a location parameter. *The Annals of Mathematical Statistics* **35** 73–101. MR0161415

[11] HUBER, P. J. (1965). A robust version of the probability ratio test. *The Annals of Mathematical Statistics* **36** 1753–1758. MR0185747

[12] HUBER, P. J. and STRASSEN, V. (1973). Minimax tests and the Neyman-Pearson lemma for capacities. *The Annals of Statistics* **1** 251–263. MR0356306

[13] KOLTCHINSKII, V., LOUNICI, K. and TSYBAKOV, A. B. (2011). Nuclear-norm penalization and optimal rates for noisy low-rank matrix completion. *The Annals of Statistics* **39** 2302–2329. MR2906869

[14] LE CAM, L. (1973). Convergence of estimates under dimensionality restrictions. *The Annals of Statistics* **1** 38–53. MR0334381

[15] NEYMAN, J. and PEARSON, E. (1933). On the problem of the most efficient tests of statistical hypotheses. *Philosophical Transactions of the Royal Society of London A: Mathematical, Physical and Engineering Sciences* **231** 289–337.

[16] NUSSBAUM, M. (1996). Asymptotic equivalence of density estimation and Gaussian white noise. *The Annals of Statistics* **24** 2399–2430. MR1425959

[17] PINSKER, M. S. (1980). Optimal filtering of square-integrable signals in Gaussian noise. *Problemy Peredachi Informatsii* **16** 52–68. MR0624591

[18] POLLARD, D. (1990). Empirical Processes: Theory and Applications. In *NSF-CBMS Regional Conference Series in Probability and Statistics* i–86. JSTOR. MR1089429

[19] RASKUTTI, G., WAINWRIGHT, M. J. and YU, B. (2011). Minimax rates of estimation for high-dimensional linear regression over ℓ_q-balls. *IEEE Transactions on Information Theory* **57** 6976–6994. MR2882274

[20] RECHT, B., FAZEL, M. and PARRILO, P. A. (2010). Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization. *SIAM review* **52** 471–501. MR2680543

[21] ROHDE, A. and TSYBAKOV, A. B. (2011). Estimation of high-dimensional low-rank matrices. *The Annals of Statistics* **39** 887–930. MR2816342

[22] RUDELSON, M. and ZHOU, S. (2013). Reconstruction from anisotropic random measurements. *IEEE Transactions on Information Theory* **59** 3434–3447. MR3061256

[23] TSYBAKOV, A. B. (2008). *Introduction to Nonparametric Estimation.* Springer Science & Business Media. MR2724359

[24] YANG, Y. and BARRON, A. (1999). Information-theoretic determination of minimax rates of convergence. *The Annals of Statistics* **27** 1564–1599. MR1742500

[25] YATRACOS, Y. G. (1985). Rates of convergence of minimum distance estimators and Kolmogorov's entropy. *The Annals of Statistics* **13** 768–774. MR0790571