The Annals of Statistics
2018, Vol. 46, No. 5, 1932–1960
https://doi.org/10.1214/17-AOS1607

ROBUST COVARIANCE AND SCATTER MATRIX ESTIMATION UNDER HUBER'S CONTAMINATION MODEL

BY MENGJIE CHEN*, CHAO GAO*,[1] AND ZHAO REN†

University of Chicago and University of Pittsburgh†*

Covariance matrix estimation is one of the most important problems in statistics. To accommodate the complexity of modern datasets, it is desired to have estimation procedures that not only can incorporate the structural assumptions of covariance matrices, but are also robust to outliers from arbitrary sources. In this paper, we define a new concept called matrix depth and then propose a robust covariance matrix estimator by maximizing the empirical depth function. The proposed estimator is shown to achieve minimax optimal rate under Huber's ε-contamination model for estimating covariance/scatter matrices with various structures including bandedness and sparsity.

1. Introduction. Covariance matrix estimation is one of the most important problems in statistics. The last decade has witnessed the rapid development of statistical theory for covariance matrix estimation under high dimensional settings. Starting from the seminal works of Bickel and Levina [1, 2], covariance matrices with a list of different structures can be estimated with optimal theoretical guarantees. Examples include the bandable matrix [9], sparse matrix [10, 34], Toeplitz matrix [7] and spiked matrix [3, 6]. For a recent comprehensive review on this topic, see [8]. However, these works do not take into account the heavy-tailedness of data and the possible presence of outliers. All these methods are based on sample covariance matrix, which is shown to have a $1/(n+1)$ breakdown point [25]. This means that even if there exists only one arbitrary outlier in the whole dataset, the statistical performance of the estimator can be totally compromised. In this paper, we attempt to tackle the problems of robust covariance matrix estimation under high-dimensional settings.

To be more specific, we consider the distribution $(1-\varepsilon)N(0,\Sigma)+\varepsilon Q$, where Q is an arbitrary distribution that models the outliers and ε is the proportion of contamination. Given i.i.d. observations $X_1,\dots,X_n$ from this distribution, there are approximately $n\varepsilon$ of them distributed according to Q, which can influence the performance of an estimator without robustness property. This setting is called the ε-contamination model, first proposed in a path-breaking paper by Huber [30].

Received March 2016; revised June 2017.
[1]Supported in part by NSF Grant DMS-1712957.
MSC2010 subject classifications. Primary 62H12; secondary 62C20.
Key words and phrases. Data depth, Minimax rate, high-dimensional statistics, outliers, contamination model, breakdown point.

In this paper, Huber proposed a robust location estimator and proved its minimax optimality under the ε-contamination model. His work suggests an estimator that is optimal under the ε-contamination model must achieve statistical efficiency and resistance to outliers simultaneously. Therefore, we view the ε-contamination model as a natural framework to develop theories of robust estimation of covariance matrices. The goal of this paper is to propose an estimator of Σ that achieves the minimax rate under Huber's ε-contamination model.

To obtain a robust covariance matrix estimator, we propose a new concept called the matrix depth. For a p-variate distribution $X \sim \mathbb{P}$, the matrix depth of a positive semidefinite $\Gamma \in \mathbb{R}^{p\times p}$ with respect to $\mathbb{P}$ is defined as

$$\mathcal{D}(\Gamma, \mathbb{P}) = \inf_{\|u\|=1} \min\bigl\{\mathbb{P}\bigl\{|u^T X|^2 \le u^T \Gamma u\bigr\}, \mathbb{P}\bigl\{|u^T X|^2 \ge u^T \Gamma u\bigr\}\bigr\}. \tag{1}$$

We will show that for $\mathbb{P} = N(0, \Sigma)$, the deepest matrix is $\beta\Sigma$ for some constant multiplier $\beta > 0$. Thus, a natural estimator for Σ is $\hat{\Gamma}/\beta$ with $\hat{\Gamma} = \arg\max_{\Gamma\ge 0} \mathcal{D}(\Gamma, \mathbb{P}_n)$. Here, we use the notation $\mathbb{P}_n$ to denote the empirical distribution.

Our definition of matrix depth is parallel to Tukey's depth function [46] for a location parameter. The deepest vector according to Tukey's depth is a natural extension of median in the multivariate setting, and thus can be used as a robust location estimator. Zuo and Serfling [59] advocated the notion of statistical depth function that satisfies the four properties in [35] and verified that Tukey's depth indeed satisfies all these properties while many other depth functions [35, 42, 43, 51] do not. The multivariate median defined by Tukey's depth was shown to have a high breakdown point [14, 15, 17]. The original proposal of the depth function in [46] not only provides a way for robust location estimation, but also gives a general way to summarize multivariate data. For example, the depth function can be used to define an index of scatteredness of data [60]. Based on the concept of data depth, a data peeling procedure has been proposed to estimate the covariance matrix. Specifically, one may trim the data points according to their depths and use the remaining ones to estimate the covariance [15, 36]. One may also estimate the covariance through a weighted average with weights that are functions of depths [58]. Though the notion of Tukey's depth is closely related to covariance matrix estimation, depth functions that are directly defined on positive semipositive matrices are not well explored in the literature. The need for such a concept has been mentioned in [44] based on a general framework of location depth functions by [40, 41]. A proposal that is close in spirit to ours is [57], which also uses the projection idea in Tukey's location depth. The matrix depth defined in (1) offers another option. Later, we will also define several variants of the matrix depth that take into account the high-dimensional structures such as bandedness and sparsity. Those matrix depth functions are powerful tools for robust estimation of structured covariance matrices.

We apply the proposed robust matrix estimator to the problems of estimating banded covariance matrices, bandable covariance matrices, sparse covariance matrices and sparse principal components. We show that in all of these cases, the estimators defined by the matrix depth functions achieve the minimax rates of the corresponding ε-contamination models under the operator norm. Therefore, the new estimators enjoy both rate optimality and property of resistance to outliers. Interestingly, the minimax rates have a unified expression. That is, $\mathcal{M}(\varepsilon) \asymp \max\{\mathcal{M}(0), \omega(\varepsilon, \mathcal{F})\}$, where $\mathcal{M}(\varepsilon)$ is the minimax rate for the probability class of distributions $(1-\varepsilon)N(0, \Sigma) + \varepsilon Q$ ranging over $\Sigma \in \mathcal{F}$ for some covariance matrix class $\mathcal{F}$ and all probability distributions Q. The first part $\mathcal{M}(0)$ is the classical minimax rate without contamination. The second part is determined by the quantity $\omega(\varepsilon, \mathcal{F})$ called the modulus of continuity. Its definition goes back to the fundamental works of Donoho and Liu [18] and Donoho [16]. A high level interpretation is that the least favorable contamination distribution Q can be chosen in a way that the parameters within $\omega(\varepsilon, \mathcal{F})$ under a given loss cannot be distinguished from each other. We establish this phenomenon rigorously through a general lower bound argument for all ε-contamination models.

Besides ε-contamination models with Gaussian distributions, we show that our proposed estimators also work for general elliptical distributions. To be specific, the setting $(1-\varepsilon)P_\Gamma + \varepsilon Q$ is also considered, where Γ is the scatter matrix under the canonical representation of an elliptical distribution. In fact, a characteristic property of the scatter matrix Γ of an elliptical distribution is $\mathcal{D}(\Gamma, P_\Gamma) = 1/2$. This property allows the depth function to combine naturally with the elliptical family. The resulting estimators are also shown to achieve the optimal convergence rates. To this end, we can claim that the proposed estimator by matrix depth have two extra robustness properties besides its rate optimality: resistance to outliers and insensitivity to heavy-tailedness. In fact, there are many works in the literature on scatter matrix estimation for elliptical distributions, including [38, 48] in classical settings and [22, 26–29, 39, 54–56] in high-dimensional settings. However, it still remains open whether these estimators can achieve the minimax rates of the ε-contamination models.

The ε-contamination model is a setting where a successful estimator should achieve a good convergence rate and robustness simultaneously. By considering a population variation of the breakdown point which we term as δ-breakdown point, we show in Section 6.3 that for a given estimator that has convergence rate δ under the ε-contamination model, its δ-breakdown point is at least ε. This suggests convergence under Huber's ε-contamination model is a more general notion of robustness than the breakdown point and it provides a unified way to study the statistical convergence rate and robustness jointly.

The main contribution of the paper is the derivation of the minimax rates for robust covariance matrix estimation under Huber's ε-contamination model, which can be achieved by optimizing over the proposed matrix depth function. We would like to clarify that, in high dimensional settings, the proposed estimators based on

matrix depth are challenging to compute, hence are mainly of theoretical interest. It is interesting and urgent to investigate in the future whether the minimax rates of covariance matrix estimation under Huber's ε-contamination model can be achieved by a provable polynomial-time algorithm. For unstructured covariance matrices under low or moderate dimensions (up to $p = 10$), the proposed depth-based estimators can be used in practice. We provide an algorithm and perform some numerical studies in the Supplementary Material [11]. An R package is available on the Github at https://github.com/ChenMengjie/DepthDescent.

The paper is organized as follows. First, we revisit Tukey's location depth in Section 2 and discuss the convergence rate of the associated multivariate median. The matrix depth is introduced in Section 3 and we use it as a tool to solve various robust structured covariance matrix estimation problems. In Section 4, we discuss the relationship between matrix depth and elliptical distributions. Results of covariance matrix estimation are extended to scatter matrix estimation for elliptical distributions. Section 5 presents a general result on minimax lower bound for the ε-contamination model. In Section 6, we discuss some related topics on robust statistics including the connection between breakdown point and the ε-contamination model as well as an extension of our notion of matrix depth function to the setting with noncentered observations. All proofs of the theoretical results are given in Section 7 and the Supplementary Material [11]. The Supplementary Material [11] also include some numerical studies of the proposed estimators for unstructured covariance matrices when the dimension is low or moderate.

We close this section by introducing some notation. Throughout the paper, we assume the covariance or scatter matrix of interest is not a zero matrix. Given an integer d, we use $[d]$ to denote the set $\{1, 2, \ldots, d\}$. For a vector $u = (u_i)$, $\|u\| = \sqrt{\sum_i u_i^2}$ denotes the ℓ_2 norm. For a matrix $A = (A_{ij})$, we use $s_k(A)$ to denote its kth singular value. The largest and the smallest singular values are denoted as $s_{\max}(A)$ and $s_{\min}(A)$, respectively. The operator norm of A is denoted by $\|A\|_{\mathrm{op}} = s_{\max}(A)$ and the Frobenius norm by $\|A\|_{\mathrm{F}} = \sqrt{\sum_{ij} A_{ij}^2}$. When $A = A^T \in \mathbb{R}^{p\times p}$ is symmetric, $\mathrm{diag}(A)$ means the diagonal matrix whose (i, i)th entry is A_{ii}. Given a subset $J \subset [p]$, A_{JJ} is an $|J| \times |J|$ submatrix, where $|J|$ means the cardinality of J. The set $S^{p-1} = \{u \in \mathbb{R}^p : \|u\| = 1\}$ is the unit sphere in $\mathbb{R}^p$. Given two numbers $a, b \in \mathbb{R}$, we use $a \vee b = \max(a, b)$ and $a \wedge b = \min(a, b)$. For two positive sequences $\{a_n\}, \{b_n\}$, $a_n \lesssim b_n$ means $a_n \le Cb_n$ for some constant $C > 0$ independent of n, and $a_n \asymp b_n$ means $a_n \lesssim b_n$ and $b_n \lesssim a_n$. Given two probability distributions $\mathbb{P}, \mathbb{Q}$, the total variation distance is defined by $\sup_B |\mathbb{P}(B) - \mathbb{Q}(B)|$, and the Kullback–Leibler divergence is defined by $D(\mathbb{P} \,\|\, \mathbb{Q}) = \int \log \frac{d\mathbb{P}}{d\mathbb{Q}}\, d\mathbb{P}$. Throughout the paper, C, c and their variants denote generic constants that do not depend on n. Their values may change from line to line.

2. Prologue: Robust location estimation. We start by the problem of robust location estimation. Consider i.i.d. observations $X_1, \ldots, X_n \sim \mathbb{P}_{(\varepsilon,\theta,Q)} =$

$(1-\varepsilon)P_\theta + \varepsilon Q$, where $P_\theta = N(\theta, I_p)$. The goal is to estimate the location parameter θ from the contaminated data $\{X_i\}_{i=1}^n$. It is known that the sample average is not robust because of its sensitivity to outliers. We consider Tukey's median ([45, 46], see [47] as well) as a robust estimator of the location θ. First, we need to introduce Tukey's depth function. For any $\eta \in \mathbb{R}^p$ and a distribution $\mathbb{P}$ on $\mathbb{R}^p$, the Tukey's depth of η with respect to $\mathbb{P}$ is defined as

$$\mathcal{D}(\eta, \mathbb{P}) = \inf_{u \in S^{p-1}} \mathbb{P}\{u^T X \le u^T \eta\} \qquad \text{where } X \sim \mathbb{P}.$$

Given i.i.d. observations $\{X_i\}_{i=1}^n$, the Tukey's depth of η with respect to the observations $\{X_i\}_{i=1}^n$ is defined as

$$\mathcal{D}(\eta, \{X_i\}_{i=1}^n) = \mathcal{D}(\eta, \mathbb{P}_n) = \min_{u \in S^{p-1}} \frac{1}{n} \sum_{i=1}^n \mathbb{I}\{u^T X_i \le u^T \eta\},$$

where $\mathbb{P}_n = \frac{1}{n}\sum_{i=1}^n \delta_{X_i}$ is the empirical distribution. Then Tukey's median is defined to be the deepest point with respect to the observations, that is,

$$\hat{\theta} = \arg\max_{\eta \in \mathbb{R}^p} \mathcal{D}(\eta, \{X_i\}_{i=1}^n). \tag{2}$$

When (2) has multiple maxima, $\hat{\theta}$ is understood as any vector that attains the deepest level. The convergence rate of $\hat{\theta}$ is stated in the following theorem.

THEOREM 2.1. *Consider Tukey's median $\hat{\theta}$. Assume that $\varepsilon < 1/5$. Then there exist absolute constants $C, C_1 > 0$, such that for any $\delta \in (0, 1/2)$ satisfying $C_1(\frac{p}{n} + \frac{\log(1/\delta)}{n}) < 1$, we have*

$$\|\hat{\theta} - \theta\|^2 \le C\left(\left(\frac{p}{n} \vee \varepsilon^2\right) + \frac{\log(1/\delta)}{n}\right),$$

with $\mathbb{P}_{(\varepsilon,\theta,Q)}$-probability at least $1 - 2\delta$ uniformly over all θ and Q.

REMARK 2.1. By scrutinizing the proof of Theorem 2.1, the result can hold for any $\varepsilon < 1/3 - c'$ for an arbitrarily small constant c'. The critical threshold $1/3$ has a meaning of the highest breakdown point for Tukey's median [15, 17]. Further discussion on the connection between the breakdown point and the ε-contamination model is given in Section 6.

REMARK 2.2. Theorem 2.1 is valid for identity covariance matrix. For a more general case $P_\theta = N(\theta, \Sigma)$, as long as $s_{\max}(\Sigma) \le M$ with some absolute constant $M > 0$, the result remains valid. In addition, the result can also be extended to the class of elliptical distributions considered in Section 4.

To the best of our knowledge, Theorem 2.1 is the first result in the literature that gives an error bound for Tukey's median under Huber's ε-contamination model. It says that the convergence rate of Tukey's median is p/n in terms of the squared ℓ_2 loss when $\varepsilon^2 \lesssim p/n$. Otherwise, the rate is ε^2. Therefore, as long as the number of outliers from Q is at the order of $n\varepsilon = O(\sqrt{np})$, the convergence rate of Tukey's median is identical to the case when $\varepsilon = 0$. The next theorem shows that Tukey's median is optimal under the ε-contamination model in a minimax sense.

THEOREM 2.2. *There exist some absolute constants $C, c > 0$ such that*

$$\inf_{\hat{\theta}} \sup_{\theta, Q} \mathbb{P}_{(\varepsilon,\theta,Q)} \left\{ \|\hat{\theta} - \theta\|^2 \geq C\left(\frac{p}{n} \vee \varepsilon^2\right) \right\} \geq c,$$

for any $\varepsilon \in [0, 1]$.

Theorem 2.2 provides a minimax lower bound for the ε-contamination model. It implies that as long as $\varepsilon^2 \gtrsim p/n$, the usual minimax rate p/n for estimating θ is no longer achievable. It also justifies the optimality of Tukey's median from a minimax perspective. To summarize, Theorems 2.1 and 2.2 jointly provide a framework for robust statistics that characterize both rate optimality and resistance to outliers simultaneously.

Another natural robust estimator for location is the componentwise median, defined as $\hat{\theta} = (\hat{\theta}_1, \ldots, \hat{\theta}_p)^T$ with $\hat{\theta}_j = \mathsf{Median}(\{X_{ij}\}_{i=1}^n)$. We show that the componentwise median has an inferior convergence rate via the following proposition.

PROPOSITION 2.1. *Consider the componentwise median $\hat{\theta}$. There exist absolute constants $C, c > 0$ such that*

$$\sup_{\theta, Q} \mathbb{P}_{(\varepsilon,\theta,Q)} \left\{ \|\hat{\theta} - \theta\|^2 \geq Cp\left(\frac{1}{n} \vee \varepsilon^2\right) \right\} \geq c,$$

for any $\varepsilon \in [0, 1]$.

Obviously, $p(n^{-1} \vee \varepsilon^2)$ is also the upper bound for $\hat{\theta}$ by applying Theorem 2.1 to each coordinate. Since $p/n \vee \varepsilon^2 \ll p(n^{-1} \vee \varepsilon^2)$ when $\varepsilon^2 \gtrsim 1/n$, the componentwise median has a slower convergence rate. It achieves the rate p/n only when $n\varepsilon = O(\sqrt{n})$. Therefore, to preserve the rate p/n, the componentwise median can tolerate at most $O(\sqrt{n})$ number of outliers, whereas Tukey's median can tolerate $O(\sqrt{pn})$.

3. Robust covariance matrix estimation. In this section, we consider estimating covariance matrices under the ε-contamination model. The model is represented as $\mathbb{P}_{(\varepsilon,\Sigma,Q)} = (1 - \varepsilon)P_\Sigma + \varepsilon Q$, where $P_\Sigma = N(0, \Sigma)$ and Q is any distribution. Motivated by Tukey's depth function for location parameters, we introduce a new concept called matrix depth. The robust matrix estimator is defined as the deepest covariance matrix with respect to the observations. This estimator achieves minimax optimal rates under the ε-contamination model.

3.1. *Matrix depth.* The main idea of Tukey's median is to project multivariate data onto all one-dimensional subspaces and obtain the deepest point by evaluating depths in those one-dimensional subspaces. Such an idea can be used to estimate covariance matrices. Formally speaking, for $X \sim N(0, \Sigma)$, the population median of $|u^T X|^2$ is $\beta u^T \Sigma u$ for every $u \in S^{p-1}$ with some absolute constant β defined later. Thus, an estimator of Σ can be obtained by estimating variance on every direction.

Inspired by the above idea, we define the matrix depth of a positive semidefinite $\Gamma \in \mathbb{R}^{p\times p}$ with respect to a distribution $\mathbb{P}$ as

$$\mathcal{D}(\Gamma, \mathbb{P}) = \inf_{u\in S^{p-1}} \min\bigl\{\mathbb{P}\bigl\{|u^T X|^2 \le u^T \Gamma u\bigr\}, \mathbb{P}\bigl\{|u^T X|^2 \ge u^T \Gamma u\bigr\}\bigr\},$$

where $X \sim \mathbb{P}$. To adapt to various structure constraints in high-dimensional settings, it is also helpful to define matrix depth by a subset of the directions S^{p-1}. Given a subset $\mathcal{U} \subset S^{p-1}$, the matrix depth of Γ with respect to a distribution $\mathbb{P}$ relative to $\mathcal{U}$ is defined as

$$\mathcal{D}_{\mathcal{U}}(\Gamma, \mathbb{P}) = \inf_{u\in \mathcal{U}} \min\bigl\{\mathbb{P}\bigl\{|u^T X|^2 \le u^T \Gamma u\bigr\}, \mathbb{P}\bigl\{|u^T X|^2 \ge u^T \Gamma u\bigr\}\bigr\},$$

where $X \sim \mathbb{P}$. We adopt the notation $\mathcal{D}_{S^{p-1}}(\Gamma, \mathbb{P}) = \mathcal{D}(\Gamma, \mathbb{P})$, and when $\mathcal{U}$ is a singleton set, we use $\mathcal{D}_u(\Gamma, \mathbb{P})$ instead of $\mathcal{D}_{\{u\}}(\Gamma, \mathbb{P})$. At the population level, the following proposition shows that the true covariance matrix, multiplied by a scalar, is the deepest positive semidefinite matrix.

PROPOSITION 3.1. *Define β through the equation*

$$\Phi(\sqrt{\beta}) = \frac{3}{4}, \tag{3}$$

where Φ is the cumulative distribution function of $N(0, 1)$. Then, for any $\mathcal{U} \subset S^{p-1}$, we have $\mathcal{D}_{\mathcal{U}}(\beta\Sigma, P_\Sigma) = \frac{1}{2}$.

Given i.i.d. observations $\{X_i\}_{i=1}^n$ from $\mathbb{P}$, the matrix depth of Γ with respect to $\{X_i\}_{i=1}^n$ is defined as

$$\begin{aligned}\mathcal{D}_{\mathcal{U}}\bigl(\Gamma, \{X_i\}_{i=1}^n\bigr) = \min_{u\in\mathcal{U}} \min\Biggl\{ &\frac{1}{n}\sum_{i=1}^n \mathbb{I}\bigl\{|u^T X_i|^2 \le u^T \Gamma u\bigr\}, \\ &\frac{1}{n}\sum_{i=1}^n \mathbb{I}\bigl\{|u^T X_i|^2 \ge u^T \Gamma u\bigr\}\Biggr\}.\end{aligned} \tag{4}$$

Note that there are only $n + 1$ possible values for $\frac{1}{n}\sum_{i=1}^n \mathbb{I}\{|u^T X_i|^2 \le u^T \Gamma u\}$, which allows us to use minimum rather than infimum when defining the empirical matrix depth function in (4). We adopt the notation $\mathcal{D}_{S^{p-1}}(\Gamma, \{X_i\}_{i=1}^n) = \mathcal{D}(\Gamma, \{X_i\}_{i=1}^n)$. A general estimator for $\beta\Sigma$ is given by

$$\hat{\Gamma} = \arg\max_{\Gamma\in\mathcal{F}} \mathcal{D}_{\mathcal{U}}\bigl(\Gamma, \{X_i\}_{i=1}^n\bigr), \tag{5}$$

where $\mathcal{F}$ is some matrix class to be specified later. One can either use $\mathcal{F}$ to impose various structure constraints in high-dimensional settings or use it to promote positive-definiteness of the estimator. The estimator of Σ is

$$\hat{\Sigma} = \hat{\Gamma}/\beta, \tag{6}$$

where β is defined through (3).

3.2. *General covariance matrix.* Consider the following covariance matrix class with bounded spectra:

$$\mathcal{F}(M) = \{\Sigma = \Sigma^T \in \mathbb{R}^{p\times p} : \Sigma \succeq 0, s_{\max}(\Sigma) \le M\},$$

where $\Sigma \succeq 0$ means Σ is positive semidefinite and $M > 0$ is some absolute constant that does not scale with p or n.

To define an estimator, it is natural to pick $\mathcal{U} = S^{p-1}$. Recall we adopt the notation $\mathcal{D}_{S^{p-1}}(\Gamma, \{X_i\}_{i=1}^n) = \mathcal{D}(\Gamma, \{X_i\}_{i=1}^n)$. Define

$$\hat{\Gamma} = \arg\max_{\Gamma\succeq 0} \mathcal{D}(\Gamma, \{X_i\}_{i=1}^n). \tag{7}$$

When (7) has multiple maxima, $\hat{\Gamma}$ is understood as any positive semidefinite matrix that attains the deepest level. A final estimator of Σ is defined by $\hat{\Sigma} = \hat{\Gamma}/\beta$ as in (6). The error bound of $\hat{\Sigma}$ is stated in the following theorem.

THEOREM 3.1. *Assume that $\varepsilon < 1/5$. Then there exist absolute constants $C, C_1 > 0$, such that for any $\delta \in (0, 1/2)$ satisfying $C_1 \frac{p+\log(1/\delta)}{n} < 1$, we have*

$$\|\hat{\Sigma} - \Sigma\|_{\mathrm{op}}^2 \le C\left(\left(\frac{p}{n} \vee \varepsilon^2\right) + \frac{\log(1/\delta)}{n}\right),$$

with $\mathbb{P}_{(\varepsilon,\Sigma,Q)}$-probability at least $1 - 2\delta$ uniformly over all Q and $\Sigma \in \mathcal{F}(M)$.

The convergence rate for the deepest covariance is $(p/n) \vee \varepsilon^2$ under the squared operator norm. A matching lower bound is given by the following theorem.

THEOREM 3.2. *There exist some absolute constants $C, c > 0$ such that*

$$\inf_{\hat{\Sigma}} \sup_{\Sigma\in\mathcal{F}(M)} \sup_{Q} \mathbb{P}_{(\varepsilon,\Sigma,Q)}\left\{\|\hat{\Sigma} - \Sigma\|_{\mathrm{op}}^2 \ge C\left(\frac{p}{n} \vee \varepsilon^2\right)\right\} \ge c,$$

for any $\varepsilon \in [0, 1]$.

Theorems 3.1 and 3.2 show that the minimax rate for estimating a covariance matrix under Huber's ε-contamination model is $(p/n) \vee \varepsilon^2$. The part p/n is the classical parametric rate [12] for estimating a covariance matrix without contamination under the squared spectral norm.

3.3. *Bandable covariance matrix.* In many high-dimensional applications such as time series data in finance, the covariates of data are collected in an ordered fashion. This leads to a natural banded estimator of the covariance matrix [2, 9]. Define the class of covariance matrices with a banded structure by

$$\mathcal{F}_k = \{\Sigma = (\sigma_{ij}) \succeq 0 : \sigma_{ij} = 0 \text{ if } |i-j| > k\}.$$

Next, we propose a notion of matrix depth function relative to some subset $\mathcal{U}_k \subset S^{p-1}$ defined particularly for the class $\mathcal{F}_k$. For any $l_1, l_2 \in [p]$, define $\mathcal{V}_{[l_1,l_2]} = \{u = (u_i) \in S^{p-1} : u_i = 0 \text{ if } i \notin [l_1, l_2]\}$. Then $\mathcal{V}_{[l_1,l_2]}$ is equivalent to $S^{l_2-l_1}$ on the coordinates $\{l_1, \ldots, l_2\}$. The depth function is defined relatively to the following subset:

$$\mathcal{U}_k = \bigcup_{l=1}^{p+1-2k} \mathcal{V}_{[l,l+2k-1]} \qquad \text{if } 2k \le p \quad \text{and} \quad \mathcal{U}_k = \mathcal{V}_{[1,p]} = S^{p-1} \qquad \text{if } 2k > p.$$

Then a robust covariance matrix estimator with banded structure is defined as

$$\hat{\Gamma} = \arg\max_{\Gamma \in \mathcal{F}_k} \mathcal{D}_{\mathcal{U}_k}(\Gamma, \{X_i\}_{i=1}^n). \tag{8}$$

An estimator for Σ is $\hat{\Sigma} = \hat{\Gamma}/\beta$ as in (6).

To study the convergence rate of $\hat{\Sigma}$, we consider the class $\mathcal{F}_k(M) = \mathcal{F}_k \cap \mathcal{F}(M)$. The convergence rate of $\hat{\Sigma}$ under the ε-contamination model is stated in the following theorem.

THEOREM 3.3. *Assume that $\varepsilon < 1/5$. Then there exist absolute constants $C, C_1 > 0$, such that for any $\delta \in (0, 1/2)$ satisfying $C_1(\frac{k+\log p}{n} + \frac{\log(1/\delta)}{n}) < 1$, we have*

$$\|\hat{\Sigma} - \Sigma\|_{\text{op}}^2 \le C\left(\left(\frac{k+\log p}{n} \vee \varepsilon^2\right) + \frac{\log(1/\delta)}{n}\right),$$

with $\mathbb{P}_{(\varepsilon,\Sigma,Q)}$-probability at least $1-2\delta$ uniformly over all Q and $\Sigma \in \mathcal{F}_k(M)$.

Theorem 3.3 states that the convergence rate for $\hat{\Sigma}$ under the class $\mathcal{F}_k(M)$ is $\frac{k+\log p}{n} \vee \varepsilon^2$. When $\varepsilon^2 \lesssim \frac{k+\log p}{n}$, this is exactly the minimax rate in [9]. Therefore, Theorem 3.3 extends the result of [9] to a robust setting. If the rate $\frac{k+\log p}{n}$ is pursued, then the maximum number of outliers that $\hat{\Sigma}$ can tolerate is $O(\sqrt{n(k+\log p)})$.

Besides matrices with exact banded structure, we also consider the following class of bandable matrices, in which the variables X_i and X_j become less correlated for larger $|i-j|$. That is,

$$\mathcal{F}_\alpha(M, M_0, M_{\min}) = \Big\{\Sigma = (\sigma_{ij}) \in \mathcal{F}(M) : \max_j \sum_{\{i:|i-j|>k\}} |\sigma_{ij}| \le M_0 k^{-\alpha},$$

$$s_{\min}(\Sigma) > M_{\min}\Big\},$$

where $M_0 > 0$ and $0 < M_{\min} < M$ are some absolute constants that do not scale with p or n. This covariance class is mainly motivated by many scientific applications such as climatology and spectroscopy. See, for example, [24] and [52]. The parameter α specifies how fast the magnitude of σ_{ij} decays to zero along the off-diagonal direction.

THEOREM 3.4. *Consider the robust banded estimator* $\hat{\Sigma}$ *in Theorem 3.3 with* $k = \lceil n^{\frac{1}{2\alpha+1}} \rceil \wedge p$. *In addition, we assume that* $\varepsilon < 1/5$. *Then there exist absolute constants* $C, C_1 > 0$, *such that for any* $\delta \in (0, 1/2)$ *satisfying* $C_1 \frac{\min(n^{\frac{1}{2\alpha+1}} + \log p, p) + \log(1/\delta)}{n} < 1$, *we have*

$$\|\hat{\Sigma} - \Sigma\|_{\mathrm{op}}^2 \le C\left(\left(\min\left\{n^{-\frac{2\alpha}{2\alpha+1}} + \frac{\log p}{n}, \frac{p}{n}\right\} \vee \varepsilon^2\right) + \frac{\log(1/\delta)}{n}\right),$$

with $\mathbb{P}_{(\varepsilon,\Sigma,Q)}$*-probability at least* $1 - 2\delta$ *uniformly over all* $\Sigma \in \mathcal{F}_\alpha(M, M_0, M_{\min})$ *and* Q.

REMARK 3.1. Unlike Theorem 3.3, in Theorems 3.4 we impose a condition $s_{\min}(\Sigma) > M_{\min}$ on the smallest eigenvalue of Σ while the minimax rate-optimal result in Cai, Zhang and Zhou [9] does not require such a condition in the uncontaminated setting. The reason we consider a slightly smaller parameter space is mainly due to our depth-based estimation approach. Indeed, since a bandable matrix Σ is not necessarily banded, the analysis naturally takes a bias-variance tradeoff with the pivotal matrix being $\Sigma_k = (\sigma_{ij}\mathbb{I}\{|i - j| \le k\})$, a banded version of Σ. Our analysis measures the bias via the matrix depth. The condition on $s_{\min}(\Sigma)$ guarantees the proper behavior of the depth of Σ_k, which can be well controlled solely by the bandwidth k.

To close this section, we show in the following theorem that both rates in Theorems 3.3 and 3.4 are minimax optimal under the ε-contamination model.

THEOREM 3.5. *Assume* $p \le \exp(\gamma n)$ *for some* $\gamma > 0$. *There exist some absolute constants* $C, c > 0$ *such that*

$$\inf_{\hat{\Sigma}} \sup_{\Sigma \in \mathcal{F}_k(M)} \sup_{Q} \mathbb{P}_{(\varepsilon,\Sigma,Q)}\left\{\|\hat{\Sigma} - \Sigma\|_{\mathrm{op}}^2 \ge C\left(\frac{k + \log p}{n} \vee \varepsilon^2\right)\right\} \ge c$$

and

$$\inf_{\hat{\Sigma}} \sup_{\Sigma \in \mathcal{F}_\alpha} \sup_{Q} \mathbb{P}_{(\varepsilon,\Sigma,Q)}\left\{\|\hat{\Sigma} - \Sigma\|_{\mathrm{op}}^2 \ge C\left(\min\left\{n^{-\frac{2\alpha}{2\alpha+1}} + \frac{\log p}{n}, \frac{p}{n}\right\} \vee \varepsilon^2\right)\right\} \ge c,$$

for any $\varepsilon \in [0, 1]$, *where* $\mathcal{F}_\alpha = \mathcal{F}_\alpha(M, M_0, M_{\min})$.

Theorems 3.3, 3.4 and 3.5 give minimax rates for the classes of banded and bandable covariance matrices. When $\varepsilon = 0$, the minimax rates of the two classes are given in [9]. Both rates are achieved by a tapered sample covariance estimator when there is no contamination. In comparison, when $\varepsilon > 0$, we achieve the minimax rate by incorporating the structural assumption into the definition of the matrix depth function.

3.4. *Sparse covariance matrix.* We consider sparse covariance matrices in this section. For a subset of coordinates $S \subset [p]$, define $\mathcal{G}(S) = \{G = (g_{ij}) \in \mathbb{R}^{p\times p} : g_{ij} = 0 \text{ if } i \notin S \text{ or } j \notin S\}$. Define $\mathcal{G}(s) = \bigcup_{S\subset[p]:|S|\le s} \mathcal{G}(S)$. Then the sparse covariance class is

$$\mathcal{F}_s = \big\{\Sigma \succeq 0 : \Sigma - \operatorname{diag}(\Sigma) \in \mathcal{G}(s)\big\}.$$

In other words, there are s covariates in a block that are correlated with each other. The remaining covariates are independent from this block and from each other. Such sparsity structure has been extensively studied in the problem of sparse principal component analysis [5, 33, 37, 53], and is different from the notion of degree sparsity studied in [1, 10]. Estimating the whole covariance matrix under such sparsity was considered by [6].

To take advantage of the sparsity structure, we define a subset $\mathcal{U}_s \subset S^{p-1}$ for the matrix depth function. For any $S \subset [p]$, define $\mathcal{V}_S = \{u = (u_i) \in S^{p-1} : u_i = 0 \text{ if } i \notin S\}$. The depth function is defined relatively to the following subset:

$$\mathcal{U}_s = \bigcup_{S\subset[p]:|S|=2s} \mathcal{V}_S.$$

A robust sparse covariance matrix estimator is defined by

$$\hat{\Gamma} = \arg\max_{\Gamma\in\mathcal{F}_s} \mathcal{D}_{\mathcal{U}_s}\big(\Gamma, \{X_i\}_{i=1}^n\big). \tag{9}$$

An estimator for Σ is $\hat{\Sigma} = \hat{\Gamma}/\beta$ as in (6).

The error of $\hat{\Sigma}$ is studied in the class $\mathcal{F}_s(M) = \mathcal{F}_s \cap \mathcal{F}(M)$ under the ε-contamination model.

THEOREM 3.6. *Assume that $\varepsilon < 1/5$. Then there exist absolute constants $C, C_1 > 0$, such that for any $\delta \in (0, 1/2)$ satisfying $C_1 \frac{s\log\frac{ep}{s} + \log(1/\delta)}{n} < 1$, we have*

$$\|\hat{\Sigma} - \Sigma\|_{\mathrm{op}}^2 \le C\bigg(\Big(\frac{s\log\frac{ep}{s}}{n} \vee \varepsilon^2\Big) + \frac{\log(1/\delta)}{n}\bigg),$$

with $\mathbb{P}_{(\varepsilon,\Sigma,Q)}$-probability at least $1 - 2\delta$ uniformly over all Q and $\Sigma \in \mathcal{F}_s(M)$.

The next theorem shows that the upper bound in Theorem 3.6 is optimal under the ε-contamination model.

THEOREM 3.7. *There are some absolute constants $C, C_1, c > 0$ such that as long as $\frac{s\log\frac{ep}{s}}{n} \le C_1$ holds, then*

$$\inf_{\hat{\Sigma}} \sup_{\Sigma\in\mathcal{F}_s(M)} \sup_{Q} \mathbb{P}_{(\varepsilon,\Sigma,Q)}\left\{\|\hat{\Sigma}-\Sigma\|_{\mathrm{op}}^2 \ge C\left(\frac{s\log\frac{ep}{s}}{n}\vee\varepsilon^2\right)\right\} \ge c,$$

for any $\varepsilon \in [0,1]$.

Theorems 3.6 and 3.7 together show that the minimax rate of the covariance matrix class $\mathcal{F}_s(M)$ under the ε-contamination model is $\frac{s\log\frac{ep}{s}}{n}\vee\varepsilon^2$. When $\varepsilon = 0$, the rate $\frac{s\log\frac{ep}{s}}{n}$ is obtained by [6] for a closely related matrix class that is a subset of $\mathcal{F}_s(M)$.

3.5. *Sparse principal component analysis.* As an application of Theorem 3.6, we consider sparse principal component analysis. We adopt the spiked covariance model [3, 33]. That is,

$$\Sigma = V\Lambda V^T + I_p,$$

where $V \in \mathbb{R}^{p\times r}$ is an orthonormal matrix and Λ is a diagonal matrix with elements $\lambda_1 \ge \lambda_2 \ge \cdots \ge \lambda_r > 0$. When V has s nonzero rows [5, 6], Σ is in the class $\mathcal{F}_s$. The goal is to estimate the subspace projection matrix VV^T. We propose a robust estimator by applying singular value decomposition to $\hat{\Gamma}$ in (9). That is, define $\hat{V} \in O(p,r)$ to be the matrix whose lth column is the lth eigenvector of $\hat{\Gamma}$. Then $\hat{V}\hat{V}^T$ is a robust estimator of VV^T.

To study the convergence rate of $\hat{V}$, define the covariance matrix class as

$$\mathcal{F}_{s,\lambda}(M,r) = \big\{\Sigma = V\Lambda V^T + I_p : \lambda \le \lambda_r \le \cdots \le \lambda_1 \le M,\ V \in O(p,r),\ \big|\mathrm{supp}(V)\big| \le s\big\},$$

where $O(p,r)$ is the class of $p\times r$ orthonormal matrices and $\mathrm{supp}(V)$ is the set of nonzero rows of V. The rank r is assumed to be bounded by a constant.

THEOREM 3.8. *Assume that $\varepsilon < 1/5$. Then there exist absolute constants $C, C_1, C_2 > 0$, such that for any $\delta \in (0,1/2)$ satisfying $C_1((\frac{s\log\frac{ep}{s}}{n\lambda^2}\vee\frac{\varepsilon^2}{\lambda^2}) + \frac{\log(1/\delta)}{n\lambda^2}) \le 1$ and $r \le C_2$, we have*

$$\|\hat{V}\hat{V} - VV^T\|_{\mathrm{F}}^2 \le C\left(\left(\frac{s\log\frac{ep}{s}}{n\lambda^2}\vee\frac{\varepsilon^2}{\lambda^2}\right) + \frac{\log(1/\delta)}{n\lambda^2}\right),$$

with $\mathbb{P}_{(\varepsilon,\Sigma,Q)}$-probability at least $1-2\delta$ uniformly over all $\Sigma \in \mathcal{F}_{s,\lambda}(M,r)$ and Q.

According to Theorem 3.8, the convergence rate for principal subspace estimation is $\frac{s\log\frac{ep}{s}}{n\lambda^2}\vee\frac{\varepsilon^2}{\lambda^2}$. We have the rate ε^2/λ^2 instead of the usual ε^2 to account for the outliers in the previous cases. As shown in the next theorem, the rate ε^2/λ^2 is in fact optimal for sparse principal component analysis.

THEOREM 3.9. *There exist some absolute constants $C, c, c' > 0$ such that*

$$\inf_{\hat{\Sigma}} \sup_{\Sigma \in \mathcal{F}_{s,\lambda}(M,r)} \sup_{Q} \mathbb{P}_{(\varepsilon,\Sigma,Q)} \left\{ \| \hat{V}\hat{V} - VV^T \|_{\mathrm{F}}^2 \geq C \left(\frac{s \log \frac{ep}{s}}{n\lambda^2} \vee \frac{\varepsilon^2}{\lambda^2} \right) \wedge c' \right\} \geq c,$$

for any $\varepsilon \in [0, 1]$.

Note that Theorems 3.8 and 3.9 imply that the minimax rate of sparse PCA under the ε-contamination model is $\frac{s \log \frac{ep}{s}}{n\lambda^2} \vee \frac{\varepsilon^2}{\lambda^2}$. When $\varepsilon = 0$, our minimax rate reduces to the case without contamination, which was previously obtained by [5, 6]. It is interesting that for this class, the term in the minimax rate that characterizes the influence of contamination is $\frac{\varepsilon^2}{\lambda^2}$, compared with ε^2 in all the previous theorems. We will explain this curious fact by a unified lower bound argument in Section 5.

To close this section, we briefly discuss the case where M in various covariance matrix classes is not necessarily a constant. For unstructured covariance class $\mathcal{F}(M)$ in Theorem 3.1, banded covariance class $\mathcal{F}_k(M)$ in Theorem 3.3, sparse covariance class $\mathcal{F}_s(M)$ in Theorem 3.6 and spiked covariance class $\mathcal{F}_{s,\lambda}(M, r)$ in Theorem 3.8, all the upper and lower bounds can be readily extended so that the minimax rates with respect to $\|\hat{\Sigma} - \Sigma\|_{\mathrm{op}}$ or $\|\hat{V}\hat{V} - VV^T\|_{\mathrm{F}}$ will include an extra factor of M. For the bandable class $\mathcal{F}_\alpha(M, M_0, M_{\min})$ in Theorem 3.3, we can assume all three values M, M_0, $M_{\min}$ are at the same order and scale together. For this case, all the upper and lower bounds can also be readily extended so that the minimax rates linearly depend on M.

4. Extension to elliptical distributions. In Section 3, we considered estimating the covariance matrix under the Gaussian distribution $P_\Sigma = N(0, \Sigma)$. Though we show that our covariance estimator via matrix depth function is robust to arbitrary outliers, it is not clear whether such property also holds under more general distributions. In real applications, the data may not follow a Gaussian distribution and can have very heavy tails. It is even possible that the distribution may not have finite first or second moment. In this section, we extend the Gaussian setting in Section 3 to general elliptical distributions. We show that at the population level, the scatter matrix of an elliptical distribution achieves the maximum of the matrix depth function. This fact motivates us to use the matrix depth estimator (5) in the elliptical distribution setting. Indeed, all error bounds we prove under the Gaussian distribution continue to hold under the elliptical distributions. Therefore, the proposed estimator is also adaptive to the shape of the distribution. As is pointed out by a referee, the estimator induced by matrix depth is well defined even if the underlying distribution is not elliptical. It can be interpreted as a multivariate analogue to the median absolute deviation and can serve as a robust scale estimator of the distribution.

We start by introducing the definition of an elliptical distribution.

DEFINITION 4.1 ([23]). A random vector $X \in \mathbb{R}^p$ follows an elliptical distribution if and only if it has the representation $X = \mu + \xi AU$, where $\mu \in \mathbb{R}^p$ and $A \in \mathbb{R}^{p\times r}$ are model parameters. The random variable U is distributed uniformly on the unit sphere S^{p-1} and $\xi \geq 0$ is a random variable in $\mathbb{R}$ independent of U. Letting $\Sigma = AA^T$ and we denote $X \sim EC(\mu, \Sigma, \xi)$.

For simplicity, we consider the model with $\mu = 0$. We want to remark two points on this definition. First, the representation $EC(0, \Sigma, \xi)$ is not unique. This is because $EC(0, \Sigma, \xi) = EC(0, a^{-2}\Sigma, a\xi)$ for any $a > 0$. Second, for an elliptical random variable $X \sim EC(0, \Sigma, \xi)$ with $s_{\min}(\Sigma) > 0$, given any unit vector $u \in S^{p-1}$, the distribution of $u^T X/\sqrt{u^T \Sigma u}$ is independent of u. In other words, $\Sigma^{-1/2}X$ is spherically symmetric. Motivated by these two points, we define the canonical representation of an elliptical distribution as follows.

DEFINITION 4.2. For a nondegenerate elliptical distribution $EC(0, \Sigma, \xi)$ in the sense that $s_{\min}(\Sigma) > 0$, $EC(0, \Gamma, \eta)$ is its canonical representation if and only if $\Gamma = a^{-2}\Sigma$ and $\eta = a\xi$ for some $a > 0$, and $P_\Gamma(\frac{|u^T X|^2}{u^T \Gamma u} \leq 1) = \frac{1}{2}$, where $P_\Gamma = EC(0, \Gamma, \eta)$. From now on, whenever we use $P_\Gamma = EC(0, \Gamma, \eta)$, it always denotes the canonical representation.

To guarantee the existence and uniqueness of the canonical representation, we need the following assumption on the marginal distribution. Define the distribution function:

$$G(t) = P_\Gamma\left(\frac{|u^T X|^2}{u^T \Gamma u} \leq t\right). \tag{10}$$

Note that $G(t)$ does not depend on the specific direction $u \in S^{p-1}$ used in the definition. We assume that $G(t)$ is continuous at $t = 1$ and there exist some $\tau \in (0, 1/2)$ and $\alpha, \kappa > 0$ such that

$$\inf_{|t|\geq\alpha} |G(1) - G(1 + t)| \geq \tau \quad \text{and} \quad \inf_{|t|<\alpha} \frac{|G(1) - G(1 + t)|}{|t|} \geq \kappa^{-1/2}. \tag{11}$$

Intuitively speaking, we require $G(\cdot)$ to be strictly increasing in a neighborhood of $t = 1$.

PROPOSITION 4.1. *For an elliptical distribution $EC(0, \Gamma, \eta)$ that satisfies (11), its canonical representation exists and is unique.*

The existence and uniqueness of the canonical representation of $EC(0, \Gamma, \eta)$ imply that the matrix Γ is a well-defined object. We call Γ the scatter matrix. The following proposition shows that the scatter matrix Γ is actually the deepest one with respect to the matrix depth function.

PROPOSITION 4.2. *For any subset* $\mathcal{U} \subset S^{p-1}$, *we have* $\mathcal{D}_{\mathcal{U}}(\Gamma, P_\Gamma) = \frac{1}{2}$.

When $X \sim EC(0, \Gamma, \eta)$ has a density function, it must have the form $p(x) = f(x^T \Gamma^{-1} x)$ for some univariate function $f(\cdot)$ [23]. Examples of elliptical distributions include:

1. *Multivariate Gaussian.* Density function $p(x) \propto \exp(-\beta x^T \Gamma^{-1} x/2)$, where the constant β is defined in (3). Proposition 3.1 implies that $\beta^{-1}\Gamma$ is the Gaussian covariance matrix.
2. *Multivariate Laplace.* Density function $p(x) \propto \exp(-\sqrt{\beta x^T \Gamma^{-1} x})$, where the constant β is determined through the canonical representation. The covariance matrix has formula $(p+1)\beta^{-1}\Gamma$.
3. *Multivariate t.* Density function $p(x) \propto (1 + \beta x^T \Gamma^{-1} x/d)^{-\frac{d+p}{2}}$, where d is the degree of freedom. The constant β is determined through the canonical representation. When $d > 2$, the covariance matrix is $\frac{d}{d-2}\beta^{-1}\Gamma$. Otherwise, the covariance does not exist.
4. *Multivariate Cauchy.* This is a special case of multivariate t distribution when $d = 1$. The density function is $p(x) \propto (1 + \beta x^T \Gamma^{-1} x)^{-\frac{p+1}{2}}$.

PROPOSITION 4.3. *For all the four examples above,* β *is an absolute constant independent of* p. *Moreover, the condition* (11) *holds with absolute constants* τ, α, κ *independent of* p.

Let us proceed to consider estimating the scatter matrix Γ under the ε-contamination model $\mathbb{P}_{(\varepsilon,\Gamma,Q)} = (1-\varepsilon)P_\Gamma + \varepsilon Q$. This requires the estimator to be robust in two senses. First, it should be resistant to the outliers. Second, it should be adaptive to the distribution. Using the property of the scatter matrix spelled out in Proposition 4.2, we show that the depth-induced estimator (5) enjoys optimal rates of convergence in various settings.

THEOREM 4.1. *Consider the estimator* $\hat{\Gamma}$ *defined in* (7). *Assume* $\varepsilon < \tau/3$ *and the distribution* $P_\Gamma = EC(0, \Gamma, \eta)$ *satisfies* (11). *Then, there exist absolute constants* $C, C_1 > 0$, *such that for any* $\delta \in (0, 1/2)$ *satisfying* $C_1 \frac{p+\log(1/\delta)}{n} < 1$, *we have*

$$\|\hat{\Gamma} - \Gamma\|_{\text{op}}^2 \le C\kappa\left(\left(\frac{p}{n} \vee \varepsilon^2\right) + \frac{\log(1/\delta)}{n}\right),$$

with $\mathbb{P}_{(\varepsilon,\Gamma,Q)}$*-probability at least* $1 - 2\delta$ *uniformly over all* Q *and* $\Gamma \in \mathcal{F}(M)$.

THEOREM 4.2. *Consider the estimator* $\hat{\Gamma}$ *defined in* (8). *Assume* $\varepsilon < \tau/3$ *and the distribution* $P_\Gamma = EC(0, \Gamma, \eta)$ *satisfies* (11). *Then there exist absolute constants* $C, C_1 > 0$, *such that for any* $\delta \in (0, 1/2)$ *satisfying* $C_1 \frac{k+\log p+\log(1/\delta)}{n} < 1$,

we have

$$\|\hat{\Gamma} - \Gamma\|_{\mathrm{op}}^2 \le C\kappa\left(\left(\frac{k + \log p}{n} \vee \varepsilon^2\right) + \frac{\log(1/\delta)}{n}\right),$$

with $\mathbb{P}_{(\varepsilon,\Gamma,Q)}$*-probability at least* $1 - 2\delta$ *uniformly over all* Q *and* $\Gamma \in \mathcal{F}_k(M)$.

THEOREM 4.3. *Consider the estimator* $\hat{\Gamma}$ *defined in* (8) *with* $k = \lceil n^{\frac{1}{2\alpha+1}} \rceil \wedge p$. *Assume* $\varepsilon < \tau/3$ *and the distribution* $P_\Gamma = EC(0, \Gamma, \eta)$ *satisfies* (11). *Then there exist absolute constants* $C, C_1 > 0$, *such that for any* $\delta \in (0, 1/2)$ *satisfying* $C_1 \frac{\min(n^{2\alpha+1} + \log p, p) + \log(1/\delta)}{n} < 1$, *we have*

$$\|\hat{\Gamma} - \Gamma\|_{\mathrm{op}}^2 \le C\kappa\left(\left(\min\left\{n^{-\frac{2\alpha}{2\alpha+1}} + \frac{\log p}{n}, \frac{p}{n}\right\} \vee \varepsilon^2\right) + \frac{\log(1/\delta)}{n}\right),$$

with $\mathbb{P}_{(\varepsilon,\Gamma,Q)}$*-probability at least* $1 - 2\delta$ *uniformly over all* $\Gamma \in \mathcal{F}_\alpha(M, M_0, M_{\min})$ *and* Q.

THEOREM 4.4. *Consider the estimator* $\hat{\Gamma}$ *defined in* (9). *Assume* $\varepsilon < \tau/3$ *and the distribution* $P_\Gamma = EC(0, \Gamma, \eta)$ *satisfies* (11). *Then there exist absolute constants* $C, C_1 > 0$, *such that for any* $\delta \in (0, 1/2)$ *satisfying* $C_1 \frac{s \log \frac{ep}{s} + \log(1/\delta)}{n} < 1$, *we have*

$$\|\hat{\Gamma} - \Gamma\|_{\mathrm{op}}^2 \le C\kappa\left(\left(\frac{s \log \frac{ep}{s}}{n} \vee \varepsilon^2\right) + \frac{\log(1/\delta)}{n}\right),$$

with $\mathbb{P}_{(\varepsilon,\Gamma,Q)}$*-probability at least* $1 - 2\delta$ *uniformly over all* Q *and* $\Gamma \in \mathcal{F}_s(M)$.

THEOREM 4.5. *Consider* $\hat{\Gamma}$ *defined in* (9), *and define* $\hat{V} \in O(p, r)$ *to be the matrix whose* l*th column is the* l*th eigenvector of* $\hat{\Gamma}$. *Assume the distribution* $P_\Gamma = EC(0, \Gamma, \eta)$ *satisfies* (11). *Then there exist absolute constants* $C, C_1, C_2 > 0$, *such that for any* $\delta \in (0, 1/2)$ *satisfying* $C_1\kappa((\frac{s \log \frac{ep}{s}}{n\lambda^2} \vee \frac{\varepsilon^2}{\lambda^2}) + \frac{\log(1/\delta)}{n\lambda^2}) \le 1$ *and* $r \le C_2$, *we have*

$$\|\hat{V}\hat{V} - VV^T\|_{\mathrm{F}}^2 \le C\kappa\left(\left(\frac{s \log \frac{ep}{s}}{n\lambda^2} \vee \frac{\varepsilon^2}{\lambda^2}\right) + \frac{\log(1/\delta)}{n\lambda^2}\right),$$

with $\mathbb{P}_{(\varepsilon,\Gamma,Q)}$*-probability at least* $1 - 2\delta$ *uniformly over all* Q *and* $\Gamma \in \mathcal{F}_{s,\lambda}(M, r)$.

REMARK 4.1. Theorem 4.5 requires the scatter matrix Γ to belong to $\mathcal{F}_{s,\lambda}(M, r)$, which means that $\Gamma = V\Lambda V^T + I_p$. While the I_p part has a clear meaning for covariance matrix, it may not be a suitable way of modeling the scatter matrix. However, we may consider a more general space which contains $\Gamma = V\Lambda V^T + \sigma^2 I_p$ for some absolute constant σ^2 bounded in some interval $[M^{-1}, M]$. Then the result of Theorem 4.5 still holds.

REMARK 4.2. The problem of finding the leading principal subspace for $EC(0, \Gamma, \eta)$ was coined as elliptical component analysis by [28]. While [28] extended sparse principal component analysis to the elliptical distributions, the influence of outliers was not investigated. In comparison, we show that our estimator is robust to both heavy-tailed distributions and the presence of outliers.

REMARK 4.3. Theorems 4.1–4.5 identify a linear dependence on the number κ in the error bounds. This dependence was previously revealed in the literature when $\varepsilon = 0$ and $p = 1$. In this case, our proposed estimator is the median absolute deviation that enjoys asymptotic normality $\sqrt{n}(\hat{\gamma} - \gamma) \rightsquigarrow N(0, \frac{1}{4|G'(1)|^2})$ (see Example 5.24 in [49]). Given the fact that $|G(1) - G(1+t)|/|t| \approx |G'(1)|$ when t is small, κ plays a similar role as $|G'(1)|^{-2}$.

To close this section, we remark that the estimators via matrix depth function does not require the knowledge of the exact elliptical distribution. They are adaptive to all $EC(0, \Gamma, \eta)$ that satisfy the condition (11). Since the class of elliptical distributions includes multivariate Gaussian as a special case, the lower bounds in Section 3 imply that the convergence rates obtained in this section are optimal.

5. A general minimax lower bound. In this section, we provide a general minimax theory for ε-contamination model. Given a general statistical experiment $\{P_\theta : \theta \in \Theta\}$, recall the notation $\mathbb{P}_{(\varepsilon,\theta,Q)} = (1-\varepsilon)P_\theta + \varepsilon Q$. If we denote the minimax rate for the class $\{\mathbb{P}_{(\varepsilon,\theta,Q)} : \theta \in \Theta, Q\}$ under some loss function $L(\theta_1, \theta_2)$ by $\mathcal{M}(\varepsilon)$, then most rates we obtained in Section 2 and Section 3 can be written as $\mathcal{M}(\varepsilon) \asymp \mathcal{M}(0) \vee \varepsilon^2$. The only exception is $\mathcal{M}(\varepsilon) \asymp \mathcal{M}(0) \vee (\varepsilon^2/\lambda^2)$ for sparse principal component analysis. Therefore, a natural question is whether we can have a general theory for the ε-contamination model that governs those minimax rates. The answer for this question lies in a key quantity called modulus of continuity, whose definition goes back to the seminal works of Dohono and Liu [18] and Donoho [16].

The modulus of continuity for the ε-contamination model is defined as

$$\omega(\varepsilon, \Theta) = \sup\{L(\theta_1, \theta_2) : \mathsf{TV}(P_{\theta_1}, P_{\theta_2}) \le \varepsilon/(1-\varepsilon); \theta_1, \theta_2 \in \Theta\}. \tag{12}$$

The quantity $\omega(\varepsilon, \Theta)$ measures the ability of the loss $L(\theta_1, \theta_2)$ to distinguish two distributions P_{θ_1} and P_{θ_2} that are close in total variation at the order of ε. A high level interpretation is that two distributions P_{θ_1} and P_{θ_2} as close as $\varepsilon/(1-\varepsilon)$ under total variation distance cannot be distinguished at the presence of arbitrary contamination distribution with proportion ε. Thus, an error at the order of $\omega(\varepsilon, \Theta)$ cannot be avoided for the loss $L(\cdot, \cdot)$. A general minimax lower bound depending on the modulus of continuity is stated in the following theorem.

THEOREM 5.1. *Suppose there is some* $\mathcal{M}(0)$ *such that for* $\varepsilon = 0$

$$\inf_{\hat{\theta}} \sup_{\theta\in\Theta} \sup_{Q} \mathbb{P}_{(\varepsilon,\theta,Q)}\big\{L(\hat{\theta},\theta) \geq \mathcal{M}(\varepsilon)\big\} \geq c \tag{13}$$

holds. Then for any $\varepsilon \in [0,1]$, (13) *holds for* $\mathcal{M}(\varepsilon) \asymp \mathcal{M}(0) \vee \omega(\varepsilon,\Theta)$.

Theorem 5.1 shows that the quantity $\omega(\varepsilon,\Theta)$ is the price of robustness one has to pay in the minimax rate. To illustrate this result, let us consider the location model in Section 2 where $P_\theta = N(\theta, I_p)$. Since $\|\theta_1 - \theta_2\|^2 = 2D(P_{\theta_1} \parallel P_{\theta_2}) \geq 4\mathsf{TV}(P_{\theta_1}, P_{\theta_2})^2$, we have $\omega(\varepsilon,\Theta) \gtrsim \varepsilon^2$. Besides, it is well known that $\mathcal{M}(0) \asymp p/n$ for the location model, and thus we obtain the rate $(p/n) \vee \varepsilon^2$ as the lower bound, which implies Theorem 2.2. Similar calculation can also be done for the covariance model. In particular, for sparse principal component analysis, we get $\omega(\varepsilon,\Theta) \asymp (\varepsilon/\lambda)^2$. The details of derivation are given in the Supplementary Material [11].

6. Discussion.

6.1. *Impact of contamination on convergence rates*. For all the problems we consider in this paper, the minimax rate under the ε-contamination model has the expression $\mathcal{M}(\varepsilon) \asymp \mathcal{M}(0) \vee \omega(\varepsilon,\Theta)$. Define

$$\varepsilon^* = \sup\big\{\varepsilon : \omega(\varepsilon,\Theta) \leq \mathcal{M}(0)\big\}.$$

Then ε^* is the maximal proportion of outliers under which the minimax rate obtained without outliers can still be preserved. Thus, $n\varepsilon^*$ is the maximal expected number of outliers for an optimal procedure to achieve the minimax rate as if there is no contamination.

Compared to the minimax rate, consistency is easier to achieve. Suppose $\mathcal{M}(0) = o(1)$, then the necessary and sufficient condition for consistency is $\omega(\varepsilon,\Theta) = o(1)$. In most cases where $\omega(\varepsilon,\Theta) \asymp \varepsilon^2$, the condition reduces to $\varepsilon = o(1)$, meaning that as long as the expected number of outliers is at a smaller order of n, the optimal procedure is consistent under the ε-contamination model.

6.2. *Noncentered observations*. In previous sections, we assume that the observations are sampled from a centered distribution. This is essential for the proposed matrix depth method to work. It is important to extend our method to noncentered data in order to make it more practical.

For the Gaussian case, our inspiration is from the simple fact that $\frac{1}{\sqrt{2}}(X_1 - X_2) \sim N(0,\Sigma)$, where $X_1, X_2 \sim N(\theta,\Sigma)$ are independent observations with with $\theta \in \mathbb{R}^p$ being an arbitrary mean vector. This motivates the following definition of a U-version empirical matrix depth function. That is,

$$\bar{\mathcal{D}}_{\mathcal{U}}(\Gamma, \{X_i\}_{i=1}^n) = \min_{u\in\mathcal{U}} \min\Bigg\{ \frac{1}{\binom{n}{2}} \sum_{i<j} \mathbb{I}\big\{\big|u^T(X_i - X_j)\big|^2 \leq 2u^T\Gamma u\big\},$$

$$\frac{1}{\binom{n}{2}} \sum_{i<j} \mathbb{I}\big\{\big|u^T(X_i - X_j)\big|^2 \geq 2u^T\Gamma u\big\}\Bigg\}.$$

Then a covariance matrix estimator $\hat{\Sigma}$ is defined through (5) and (6) with $\mathcal{D}_{\mathcal{U}}(\Gamma, \{X_i\}_{i=1}^n)$ replaced by $\bar{\mathcal{D}}_{\mathcal{U}}(\Gamma, \{X_i\}_{i=1}^n)$. A similar pairwise difference trick was used by [21] in a different setting. It turns out that all the non-asymptotic bounds in Section 3 continue to hold for this new estimator. Due to limited space, we provide more details in Section A of the Ssupplementary Material [11], including the extension to the noncentered elliptical distributions, based on an extension of the concentration inequality for suprema of some empirical process to its corresponding U-process.

6.3. *Connection with δ-breakdown point.* The notion of breakdown point [25] has been widely used to quantify the influence of outliers for a given estimator. Its relation to the ε-contamination model was previously explored through the notion of maximum bias in the context of robust covariance matrix estimation; see, for example, [58]. In this section, we discuss the connection between a population variation of the breakdown point and Huber's ε-contamination model. Let us start by the definition given in [14, 15, 17]. Consider the observations $\{X_i\}_{i=1}^n$ that consist of two parts $\{Y_i\}_{i=1}^{n_1}$ and $\{Z_i\}_{i=1}^{n_2}$ with $n_1 + n_2 = n$. We view $\{Z_i\}_{i=1}^{n_2}$ as the outliers. Then a robust estimator $\hat{\theta}(\cdot)$ should not be influenced much by the outliers if the proportion $n_2/(n_1 + n_2)$ is small. The breakdown point of $\hat{\theta}$ with respect to $\mathcal{Y}$ is defined as

$$\varepsilon(\hat{\theta}, \mathcal{Y}) = \min\Big\{\frac{n_2}{n_1+n_2} : \sup_{\{Y_i\}_{i=1}^{n_1} \in \mathcal{Y}} \sup_{\{Z_i\}_{i=1}^{n_2}} \big\|\hat{\theta}\big(\{Y_i\}_{i=1}^{n_1}\big) - \hat{\theta}\big(\{X_i\}_{i=1}^{n}\big)\big\| = \infty\Big\}, \tag{14}$$

where $\|\cdot\|$ is some norm. In its original form, the supremum of $\{Y_i\}_{i=1}^{n_1}$ over $\mathcal{Y}$ does not appear in the definition. However, $\{Y_i\}_{i=1}^{n_1}$ are usually assumed to be in a general position or follow some distribution. Thus, it is natural to apply this modification. Now let us consider the ε-contamination model $\mathbb{P}_{(\varepsilon,\theta,Q)} = (1-\varepsilon)P_\theta + \varepsilon Q$. For i.i.d. observations $X_1, \ldots, X_n \sim \mathbb{P}_{(\varepsilon,\theta,Q)}$, it can be decomposed into two parts $\{Y_i\}_{i=1}^{n_1}$ and $\{Z_i\}_{i=1}^{n_2}$, where $n_2 \sim \text{Binomial}(n, \varepsilon)$ and $n_1 = n - n_2$. Conditioning on n_1, $Y_1, \ldots, Y_{n_1} \sim P_\theta$ and $Z_1, \ldots, Z_{n_2} \sim Q$. Observe that $\frac{n_2}{n_1+n_2} \approx \varepsilon$, which means the ε in the contamination model plays a similar role to the ratio $\frac{n_2}{n_1+n_2}$ in (14). Motivated by this fact, we introduce a population variation of (14). Given an estimator $\hat{\theta}$, its δ-breakdown point with respect to some parameter space Θ is defined as

$$\begin{aligned}\varepsilon(\hat{\theta}, \Theta, \delta) = \min\Big\{\varepsilon : \sup_{\theta\in\Theta} \sup_{Q} \mathbb{P}_{(\varepsilon,\theta,Q)}\big\{L\big(\hat{\theta}\big(\{Y_i\}_{i=1}^{n_1}\big),\\ \hat{\theta}\big(\{X_i\}_{i=1}^{n}\big)\big) > \delta\big\} > c\Big\},\end{aligned} \tag{15}$$

where $L(\cdot,\cdot)$ is some loss function, and $c \in (0,1)$ is some small constant. We may view (16) as the population variation of (14) because $\sup_Q$ corresponds to $\sup_{\{Z_i\}_{i=1}^{n_2}}$, $\sup_{\theta\in\Theta}$ corresponds to $\sup_{\{Y_i\}_{i=1}^{n_1}\in\mathcal{Y}}$ and ε corresponds to $\frac{n_2}{n_1+n_2}$. We

allow δ to be a sequence of n instead of ∞ because $L(\cdot,\cdot)$ can be a bounded loss such as the one considered in the PCA problem in this paper. When $\delta = \infty$ for an unbounded loss and the bias term dominates the loss, the δ-breakdown point becomes the lower bound of the contamination level ε for which the ε-maxbias is infinite; see, for example, [58]. In general, the δ-breakdown point means the minimal ε such that an estimator $\hat{\theta}$ is influenced at least by the level of δ under the ε-contamination model. In fact, $\varepsilon(\hat{\theta}, \Theta, \delta)$ is a quantity directly related to the lower bound of the convergence rate of $\hat{\theta}$ under the ε-contamination model. This is rigorously stated in the following theorem.

THEOREM 6.1. *Assume the loss function is symmetric and satisfies*

$$L(\theta_1,\theta_2) \le A\big(L(\theta_1,\theta_3) + L(\theta_2,\theta_3)\big) \qquad \forall \theta_1,\theta_2,\theta_3 \in \Theta \text{ with some } A > 0, \tag{16}$$

$$\begin{aligned} &\sup_{\theta\in\Theta} P_\theta^n \left\{ L(\hat{\theta},\theta) > \frac{1}{2} c_1 A^{-1} \delta \right\} \\ &\quad \ge \sup_{\theta\in\Theta} P_\theta^{n'} \left\{ L(\hat{\theta},\theta) > \frac{1}{2} A^{-1} \delta \right\} \qquad \forall n' \ge \frac{n}{3}, \end{aligned} \tag{17}$$

with some constant $c_1 \in (0,1)$. Then, for $\varepsilon = \varepsilon(\hat{\theta}, \Theta, \delta) < \frac{1}{2}$, we have

$$\sup_{\theta\in\Theta} \sup_{Q} \mathbb{P}_{(\varepsilon,\theta,Q)} \left\{ L(\hat{\theta},\theta) > \frac{1}{2} c_1 A^{-1} \delta \right\} > \frac{1}{3} c,$$

for some $c > 0$ in (16) and sufficiently large n.

Before discussing the implications of Theorem 6.1, we remark on assumption (17). The notation P_θ^n means the estimator $\hat{\theta}(\cdot)$ takes a random argument $\hat{\theta}(\{Y_i\}_{i=1}^n)$ with distribution $Y_1,\dots,Y_n \sim P_\theta$. Thus, assumption (17) simply means when the sample sizes n, n' are at the same order, the lower bounds remain at the same order. In most cases including all the examples considered in this paper, (17) automatically holds.

A general lower bound based on the notion of δ-breakdown point is provided by Theorem 6.1. Given an estimator $\hat{\theta}$ and an ε-contamination model, the solution δ to the equation

$$\varepsilon(\hat{\theta}, \Theta, \delta) = \varepsilon \tag{18}$$

lower bounds its rate of convergence. When $\hat{\theta}$ is a minimax optimal estimator with rate $\mathcal{M}(\varepsilon)$, we obtain $\mathcal{M}(\varepsilon) \gtrsim \delta$. In other words, the convergence rate δ under the ε-contamination model automatically implies a δ-breakdown point with the same ε.

6.4. *A unified framework of robustness and rate of convergence.* Huber's ε-contamination model is very classical in robust statistics, and allows for a deeper investigation than the breakdown point alone. For example, it has been well studied how much bias an estimator wound suffer under the contamination model via the concept of maxbias in various models, including [58]. In this paper, we demonstrate that Huber's ε-contamination model allows a simultaneous joint study of robustness and rate of convergence of an estimator in the minimax sense. There are some important works that studied such properties of robust estimators under ε-contamination model. We mention [4, 31, 32] among others. However, such results in high-dimensional settings are rarely explored. This is our major reason to develop the minimax rate optimality theory of robust covariance matrix estimation under this framework. We illustrate the importance of this view by revisiting the componentwise median studied in Section 2. Without contamination, the componentwise median is a location estimator with minimax rate under Gaussian distribution. It is also robust because of its high breakdown point [17]. However, Proposition 2.1 shows that its performance under the presence of contamination is not optimal. In contrast, Tukey's multivariate median shows its advantage over the componentwise median by obtaining optimality under the ε-contamination model. This example suggests that the rate optimality and the robustness property of an estimator should be studied together rather than separately.

Recently, Donoho and Montanari [19] have studied Huber's M-estimator under the ε-contamination model in a regression setting where p/n converges to a constant. They find a critical ε^* that determines the variance breakdown point. The setting of ε-contamination model plays a critical role in their work to illustrate both efficiency and robustness of Huber's M-estimator in a unified way.

7. Proofs of main results. This section provides proofs for the results in Section 3.

7.1. *Auxiliary lemmas.* For i.i.d. data $\{X_i\}_{i=1}^n$ from a contaminated distribution $(1-\varepsilon)P + \varepsilon Q$, it can be written as $\{Y_i\}_{i=1}^{n_1} \cup \{Z_i\}_{i=1}^{n_2}$. Marginally, we have $n_2 \sim \text{Binomial}(n, \varepsilon)$ and $n_1 = n - n_2$. Conditioning on n_1 and n_2, $\{Y_i\}_{i=1}^{n_1}$ are i.i.d. from P and $\{Z\}_{i=1}^{n_2}$ are i.i.d. from Q. The following lemmas control the ratio n_2/n_1 and characterize an important property, respectively. Their proofs are given in the Supplementary Material [11].

LEMMA 7.1. *Assume $\varepsilon < 1/5$. For any $\delta > 0$ satisfying $\sqrt{\frac{1}{2n}\log(1/\delta)} < 1/5$, we have*

$$\frac{n_2}{n_1} \le \frac{\varepsilon}{1-\varepsilon} + \frac{25}{12}\sqrt{\frac{1}{2n}\log(1/\delta)}, \tag{19}$$

with probability at least $1-\delta$. *Moreover, assume* $\varepsilon^2 > 1/n$, *and then we have*

$$\frac{n_2}{n_1} > c'\varepsilon, \tag{20}$$

with probability at least $1/2$ *for some constant* $c' > 0$.

LEMMA 7.2. *Consider any parametric family* $\{P_\theta : \theta \in \Theta\}$. *Then*

$$\big\{(1-\varepsilon_1)P_\theta + \varepsilon_1 Q : \theta \in \Theta, Q\big\} \subset \big\{(1-\varepsilon_2)P_\theta + \varepsilon_2 Q : \theta \in \Theta, Q\big\},$$

holds for any $0 \le \varepsilon_1 < \varepsilon_2 \le 1$.

Recall that for any $S \subset [p]$, $\mathcal{V}_S = \{u = (u_i) \in S^{p-1} : u_i = 0 \text{ if } i \notin S\}$. In particular, if $S = \{l_1, \ldots, l_2\}$, then $\mathcal{V}_S = \mathcal{V}_{[l_1,l_2]}$ defined in Section 3.3. Moreover, $\mathcal{V}_S = S^{p-1}$ if $S = \{1, \ldots, p\}$. Define a subset $IH_{u,t}$ of $\mathbb{R}^p$ as $IH_{u,t} = \{y : |u^T y| \le t\}$. Finally, we need the following concentration inequality for suprema of the empirical process indexed by these subsets $IH_{u,t}$, where $u \in \mathcal{V}_S$ and $t \in \mathbb{R}$. Its proof is given in the Supplementary Material [11] by using Dudley's entropy integral [20] and VC classes [50].

LEMMA 7.3. *For i.i.d. real-valued data* $X_1, \ldots, X_n$ *from distribution* $\mathbb{P}$, *we have for any* $S \subset [p]$, *with probability at least* $1-\delta$,

$$\sup_{u \in \mathcal{V}_S, t \in \mathbb{R}} \big|\mathbb{P}(IH_{u,t}) - \mathbb{P}_n(IH_{u,t})\big| \le \sqrt{\frac{1440e\pi}{1-e^{-1}}}\sqrt{\frac{3+2|S|}{n}} + \sqrt{\frac{\log(1/\delta)}{2n}},$$

where $\mathbb{P}_n$ *denotes the empirical distribution of* $\{X_i\}_{i=1}^n$.

7.2. *Proofs of upper bounds in Section 3.* We first prove the following master theorem.

THEOREM 7.1. *For some index subsets* $S_1, \ldots, S_m \subset [p]$ *with* $\max_i |S_i| \le s$, *consider the estimator* $\hat{\Sigma}$ *defined in* (6) *with* $\mathcal{U} = \bigcup_{i=1}^m \mathcal{V}_{S_i}$. *Assume* $\varepsilon < 1/5$. *Then there exist absolute constants* $C, C_1 > 0$, *such that for any* $\delta \in (0, 1/2)$ *satisfying* $C_1 \frac{1+s+\log(m/\delta)}{n} < 1$, *we have*

$$\sup_{u \in \mathcal{U}} |u^T \hat{\Sigma} u - u^T \Sigma u| \le C\bigg(\varepsilon + \sqrt{\frac{1+s+\log(m/\delta)}{n}}\bigg),$$

$\mathbb{P}_{(\varepsilon,\Sigma,Q)}$*-probability at least* $1-2\delta$ *uniformly over all* Q *and* $\Sigma \in \mathcal{F}(M)$ *with* $\beta\Sigma \in \mathcal{F}$, *where constant* β *is defined in* (3).

PROOF. By Lemma 7.1, we decompose the data $\{X_i\}_{i=1}^n = \{Y_i\}_{i=1}^{n_1} \cup \{Z_i\}_{i=1}^{n_2}$. The following analysis is conditioning on the set of (n_1, n_2) that satisfies (19) with probability at least $1-\delta$. To facilitate the proof, define

$$\mathcal{D}_u(\Gamma, P_\Sigma) = \min\{P_\Sigma(|u^T Y|^2 \le u^T \Gamma u), P_\Sigma(|u^T Y|^2 > u^T \Gamma u)\},$$

$$\mathcal{D}_u(\Gamma, \{Y_i\}_{i=1}^{n_1}) = \min\left\{\frac{1}{n_1}\sum_{i=1}^{n_1} \mathbb{I}\{|u^T Y_i|^2 \le u^T \Gamma u\}, \frac{1}{n_1}\sum_{i=1}^{n_1} \mathbb{I}\{|u^T Y_i|^2 > u^T \Gamma u\}\right\},$$

for each $u \in S^{p-1}$. Then we have $\mathcal{D}_{\mathcal{U}}(\Gamma, P_\Sigma) = \inf_{u \in \bigcup_{i=1}^m \mathcal{V}_{S_i}} \mathcal{D}_u(\Gamma, P_\Sigma)$ and $\mathcal{D}_{\mathcal{U}}(\Gamma, \{Y_i\}_{i=1}^{n_1}) = \min_{u \in \bigcup_{i=1}^m \mathcal{V}_{S_i}} \mathcal{D}_u(\Gamma, \{Y_i\}_{i=1}^{n_1})$. Observe that

$$\begin{aligned}
&\sup_{\Gamma \in \mathcal{F}} \left|\mathcal{D}_{\mathcal{U}}(\Gamma, P_\Sigma) - \mathcal{D}_{\mathcal{U}}(\Gamma, \{Y_i\}_{i=1}^{n_1})\right| \\
&\quad \le \sup_{u \in \bigcup_{i=1}^m \mathcal{V}_S, t \in \mathbb{R}} \left|\frac{1}{n_1}\sum_{i=1}^{n_1} \mathbb{I}\{|u^T Y_i|^2 \le t\} - P_\Sigma(|u^T Y|^2 \le t)\right| \\
&\quad = \sup_{u \in \bigcup_{i=1}^m \mathcal{V}_S, t \in \mathbb{R}} \left|P_\Sigma(IH_{u,t}) - \mathbb{P}_{n_1}(IH_{u,t})\right|.
\end{aligned}$$

Applying Lemma 7.3 and union bound with $\max_i |S_i| \le s$, we get

$$\sup_{\Gamma \in \mathcal{F}} \left|\mathcal{D}_{\mathcal{U}}(\Gamma, P_\Sigma) - \mathcal{D}_{\mathcal{U}}(\Gamma, \{Y_i\}_{i=1}^{n_1})\right| \le \sqrt{\frac{1440e\pi}{1-e^{-1}}}\sqrt{\frac{3+2s}{n_1}} + \sqrt{\frac{\log(m/\delta)}{2n_1}}, \tag{21}$$

with probability at least $1-\delta$. We lower bound $\mathcal{D}_{\mathcal{U}}(\hat{\Gamma}, P_\Sigma)$ by

$$\mathcal{D}_{\mathcal{U}}(\hat{\Gamma}, \{Y_i\}_{i=1}^{n_1}) - \sqrt{\frac{1440e\pi}{1-e^{-1}}}\sqrt{\frac{3+2s}{n_1}} - \sqrt{\frac{\log(m/\delta)}{2n_1}} \tag{22}$$

$$\ge \frac{n}{n_1}\mathcal{D}_{\mathcal{U}}(\hat{\Gamma}, \{X_i\}_{i=1}^n) - \frac{n_2}{n_1} - \sqrt{\frac{1440e\pi}{1-e^{-1}}}\sqrt{\frac{3+2s}{n_1}} - \sqrt{\frac{\log(m/\delta)}{2n_1}} \tag{23}$$

$$\ge \frac{n}{n_1}\mathcal{D}_{\mathcal{U}}(\beta\Sigma, \{X_i\}_{i=1}^n) - \sqrt{\frac{1440e\pi}{1-e^{-1}}}\sqrt{\frac{3+2s}{n_1}} - \sqrt{\frac{\log(m/\delta)}{2n_1}} \tag{24}$$

$$\ge \mathcal{D}_{\mathcal{U}}(\beta\Sigma, \{Y_i\}_{i=1}^{n_1}) - \frac{n_2}{n_1} - \sqrt{\frac{1440e\pi}{1-e^{-1}}}\sqrt{\frac{3+2s}{n_1}} - \sqrt{\frac{\log(m/\delta)}{2n_1}} \tag{25}$$

$$\ge \mathcal{D}_{\mathcal{U}}(\beta\Sigma, P_\Sigma) - \frac{n_2}{n_1} - 2\sqrt{\frac{1440e\pi}{1-e^{-1}}}\sqrt{\frac{3+2s}{n_1}} - \sqrt{\frac{2\log(m/\delta)}{n_1}} \tag{26}$$

$$= \frac{1}{2} - \frac{n_2}{n_1} - 2\sqrt{\frac{1440e\pi}{1-e^{-1}}}\sqrt{\frac{3+2s}{n_1}} - \sqrt{\frac{2\log(m/\delta)}{n_1}}. \tag{27}$$

The inequalities (22) and (26) are by (21). The inequalities (23) and (25) are due to the property of depth function that

$$n_1\mathcal{D}_{\mathcal{U}}(\Gamma, \{Y_i\}_{i=1}^{n_1}) \geq n\mathcal{D}_{\mathcal{U}}(\Gamma, \{X_i\}_{i=1}^{n}) - n_2 \geq n_1\mathcal{D}_{\mathcal{U}}(\Gamma, \{Y_i\}_{i=1}^{n_1}) - n_2,$$

for any $\Gamma \in \mathcal{F}$. The inequality (24) is by the definition of $\hat{\Gamma}$ and that $\beta\Sigma \in \mathcal{F}$. Finally, the equality (27) is due to Proposition 3.1. Now let us use Lemma 7.1 so that the right-hand side of (27) can be lower bounded by

$$\frac{1}{2} - \frac{\varepsilon}{1-\varepsilon} - 40\sqrt{\frac{6e\pi}{1-e^{-1}}}\sqrt{\frac{3+2s}{n}} - \frac{7}{2}\sqrt{\frac{\log(m/\delta)}{n}},$$

with probability at least $1-2\delta$. Using the property that $\mathcal{D}_u(\hat{\Gamma}, P_\Sigma) \geq \mathcal{D}_{\mathcal{U}}(\hat{\Gamma}, P_\Sigma)$ for each $u \in \mathcal{U}$, we have shown that uniformly for all $u \in \mathcal{U}$,

$$\mathcal{D}_u(\hat{\Gamma}, P_\Sigma) \geq \frac{1}{2} - \frac{\varepsilon}{1-\varepsilon} - 40\sqrt{\frac{6e\pi}{1-e^{-1}}}\sqrt{\frac{3+2s}{n}} - \frac{7}{2}\sqrt{\frac{\log(m/\delta)}{n}}, \tag{28}$$

with probability at least $1-2\delta$. By Proposition 3.1 and the fact that $\frac{1}{2} - \min(x, 1-x) = |x - 1/2|$ for all $x \in [0,1]$, we get

$$\frac{1}{2} - \mathcal{D}_u(\hat{\Gamma}, P_\Sigma) = 2\left|\Phi(\sqrt{\beta}) - \Phi\left(\sqrt{\frac{u^T\hat{\Gamma}u}{u^T\Sigma u}}\right)\right|.$$

Combining with (28), we have

$$\sup_{u\in\mathcal{U}}\left|\Phi(\sqrt{\beta}) - \Phi\left(\sqrt{\frac{u^T\hat{\Gamma}u}{u^T\Sigma u}}\right)\right| \leq \frac{\varepsilon/2}{1-\varepsilon} + \sqrt{\frac{2400e\pi}{1-e^{-1}}}\sqrt{\frac{3+2s}{n}} + \frac{7}{4}\sqrt{\frac{\log(m/\delta)}{n}},$$

with probability at least $1-2\delta$. Under the assumption that $\varepsilon < 1/5$ and $C_1\frac{1+s+\log(m/\delta)}{n} < 1$ with some absolute constant $C_1 > 0$, we have

$$\sup_{u\in\mathcal{U}}\left|\sqrt{\beta} - \sqrt{\frac{u^T\hat{\Gamma}u}{u^T\Sigma u}}\right| \leq C_2\left(\varepsilon + \sqrt{\frac{1+s+\log(m/\delta)}{n}}\right),$$

for some absolute constant $C_2 > 0$ with probability at least $1-2\delta$. Finally, due to assumption $\Sigma \in \mathcal{F}(M)$, we obtain that $\sup_{u\in\mathcal{U}} u^T\Sigma u \leq \|\Sigma\|_{\mathrm{op}} \leq M$, which implies

$$\sup_{u\in\mathcal{U}}|u^T\hat{\Gamma}u/\beta - u^T\Sigma u| \leq C\left(\varepsilon + \sqrt{\frac{1+s+\log(m/\delta)}{n}}\right),$$

with probability at least $1-2\delta$. Thus, the proof is complete. □

PROOF OF THEOREM 3.1. Since $\mathcal{U} = S^{p-1}$ and $\mathcal{F}$ is taken as the set of all positive semidefinite matrices, the conclusion follows the result of Theorem 7.1 with $m = 1$ and $S_1 = [p]$ by noting $\|\hat{\Sigma} - \Sigma\|_{\mathrm{op}} = \sup_{u\in\mathcal{V}_{S_1}} |u^T\hat{\Sigma}u - u^T\Sigma u|$. □

PROOF OF THEOREM 3.3. Consider the weights

$$w_{ij} = k^{-1}\big((2k - |i-j|)_+ - (k - |i-j|)_+\big).$$

Since $\hat{\Sigma} - \Sigma = (\hat{\sigma}_{ij} - \sigma_{ij}) \in \mathcal{F}_k$, we have $(\hat{\sigma}_{ij} - \sigma_{ij}) = ((\hat{\sigma}_{ij} - \sigma_{ij})w_{ij})$. This means $\hat{\Sigma} - \Sigma$ can also be viewed as a tapered matrix. Then Lemma 2 of [9] implies that $\|\hat{\Sigma} - \Sigma\|_{\mathrm{op}} \le 3\max_{u\in\mathcal{U}_k} |u^T(\hat{\Sigma} - \Sigma)u|$. Using the fact that $\mathcal{U}_k = \bigcup_{l=1}^{p+1-2k} \mathcal{V}_{[l,l+2k-1]}$ for $2k < p$, the conclusion follows by Theorem 7.1 with $m = p + 1 - 2k$, $S_i = [i, i + 2k - 1]$ for $i = 1, \dots, m$ and $s = 2k$. The result holds trivially according to Theorem 7.1 when $2k > p$ since $\mathcal{U}_k = S^{p-1}$. $\square$

PROOF OF THEOREM 3.4. The main argument of the proof is due to a bias-variance tradeoff. For $\Sigma = (\sigma_{ij}) \in \mathcal{F}_\alpha(M, M_0, M_{\min})$, define $\Sigma_k = (\sigma_{ij}\mathbb{I}\{|i-j| \le k\})$. Then

$$\begin{aligned}
&\big|\mathcal{D}_{\mathcal{U}_k}(\beta\Sigma, P_\Sigma) - \mathcal{D}_{\mathcal{U}_k}(\beta\Sigma_k, P_\Sigma)\big| \\
&\quad\le \max_{u\in\mathcal{U}_k}\big|\mathcal{D}_u(\beta\Sigma, P_\Sigma) - \mathcal{D}_u(\beta\Sigma_k, P_\Sigma)\big| \\
&\quad\le 2\max_{u\in\mathcal{U}_k}\left|\Phi(\sqrt{\beta}) - \Phi\left(\sqrt{\frac{\beta u^T\Sigma_k u}{u^T\Sigma u}}\right)\right| \le \sqrt{\frac{2\beta}{\pi}}\max_{u\in\mathcal{U}_k}\left|1 - \sqrt{\frac{u^T\Sigma_k u}{u^T\Sigma u}}\right| \\
&\quad\le \sqrt{\frac{2\beta}{\pi}}\max_{u\in\mathcal{U}_k}\left|\frac{u^T(\Sigma_k - \Sigma)u}{u^T\Sigma u}\right| \le \sqrt{\frac{2\beta}{\pi}}M_{\min}^{-1}\|\Sigma_k - \Sigma\|_{\mathrm{op}}.
\end{aligned}$$

Recall $\mathcal{U}_k = \bigcup_{l=1}^{p+1-2k} \mathcal{V}_{[l,l+2k-1]}$ when $2k \le p$ and $\mathcal{U}_k = S^{p-1}$ when $2k > p$, where $k = \lceil n^{\frac{1}{2\alpha+1}} \rceil \wedge p$. Using the bias bound above and the fact that $\beta\Sigma_k \in \mathcal{F}_k$, and modifying the arguments (22)–(28) in the proof of Theorem 7.1 with $\mathcal{U} = \mathcal{U}_k$, $m = \max(p + 1 - 2k, 1)$ and $s = (2k) \wedge p$, we obtain

$$\begin{aligned}
\mathcal{D}_u(\hat{\Gamma}, P_\Sigma) \ge \frac{1}{2} - \frac{\varepsilon}{1-\varepsilon} - 40\sqrt{\frac{6e\pi}{1-e^{-1}}}\sqrt{\frac{3+4k}{n}} \\
- \frac{7}{2}\sqrt{\frac{\log(m/\delta)}{n}} - \sqrt{\frac{2\beta}{\pi}}M_{\min}^{-1}\|\Sigma_k - \Sigma\|_{\mathrm{op}},
\end{aligned}$$

uniformly for all $u \in \mathcal{U}_k$ with probability at least $1 - 2\delta$. Repeating the corresponding subsequent argument in the proof of Theorem 7.1, we have

$$\sup_{u\in\mathcal{U}_k}|u^T\hat{\Gamma}u/\beta - u^T\Sigma u| \le C_1 M\left(\varepsilon + \sqrt{\frac{k + \log(m/\delta)}{n}} + M_{\min}^{-1}\|\Sigma_k - \Sigma\|_{\mathrm{op}}\right).$$

A triangle inequality implies

$$\sup_{u\in\mathcal{U}_k}|u^T\hat{\Sigma}u - u^T\Sigma_k u| \le C_2\left(\varepsilon + \sqrt{\frac{k + \log(m/\delta)}{n}} + \|\Sigma_k - \Sigma\|_{\mathrm{op}}\right).$$

By the argument in the proof of Theorem 3.3 and triangle inequality, we get

$$\|\hat{\Sigma} - \Sigma_k\|_{\mathrm{op}} \le C_3\biggl(\varepsilon + \sqrt{\frac{k + \log(m/\delta)}{n}} + \|\Sigma_k - \Sigma\|_{\mathrm{op}}\biggr),$$

$$\|\hat{\Sigma} - \Sigma\|_{\mathrm{op}} \le C\biggl(\varepsilon + \sqrt{\frac{k + \log(m/\delta)}{n}} + \|\Sigma_k - \Sigma\|_{\mathrm{op}}\biggr).$$

A bias argument in [9] implies that $\|\Sigma_k - \Sigma\|_{\mathrm{op}} \le C_4 k^{-\alpha}$. The proof is complete by observing that $k = \lceil n^{\frac{1}{2\alpha+1}} \rceil \wedge p$ and $m = \max(p + 1 - 2k, 1)$. $\square$

PROOF OF THEOREM 3.6. Note that $\hat{\Sigma} - \Sigma \in \mathcal{F}_{2s}$, and thus $\|\hat{\Sigma} - \Sigma\|_{\mathrm{op}} = \max_{|S|=2s} \|(\hat{\Sigma} - \Sigma)_{SS}\|_{\mathrm{op}} = \sup_{u \in \mathcal{U}_s} |u^T(\hat{\Sigma} - \Sigma)u|$. We denote all subsets of $[p]$ with cardinality $2s$ as $S_1, \ldots, S_m$, where $m = \binom{p}{2s} \le \exp(2s \log \frac{ep}{s})$. The proof is complete by applying Theorem 7.1 with these subsets $S_1, \ldots, S_m$, noting that $\mathcal{U}_s = \bigcup_{i=1}^m S_i$. $\square$

PROOF OF THEOREM 3.8. Since $\mathcal{F}_{s,\lambda}(M, r) \subset \mathcal{F}_s(M + 1)$, the result of Theorem 3.6 applies and we get

$$\|\hat{\Gamma}/\beta - \Sigma\|_{\mathrm{op}}^2 \le C\biggl(\frac{s \log \frac{ep}{s}}{n} \vee \varepsilon^2 + \frac{\log(1/\delta)}{n}\biggr),$$

with probability at least $1 - 2\delta$. Weyl's inequality implies $|s_{r+1}(\hat{\Gamma}/\beta) - 1| \le \|\hat{\Gamma}/\beta - \Sigma\|_{\mathrm{op}}$. Under the assumption that the rate is bounded by a small constant, we have $s_r(\Sigma) - s_{r+1}(\hat{\Gamma}/\beta) > c\lambda$ for some constant $c > 0$. By the Davis–Kahan theorem [13], we have $\|\hat{V}\hat{V}^T - VV^T\|_{\mathrm{F}} \le C'\|\hat{\Gamma}/\beta - \Sigma\|_{\mathrm{op}}/\lambda$, and the proof is complete. $\square$

Acknowledgments. Fang Han suggested the U-statistics idea behind Section 6.2. Johannes Schmidt-Hieber suggested a weaker assumption for the main theorems. The authors are grateful to the extensive reviews from an Associate Editor and two referees. Their comments greatly improved the paper. The authors also thank Andrew Barron, John Hartigan, David Pollard and Harrison Zhou for valuable comments in a YPNG seminar at Yale.

SUPPLEMENTARY MATERIAL

Supplement to "Robust covariance and scatter matrix estimation under Huber's contamination model" (DOI: 10.1214/17-AOS1607SUPP; .pdf). In this supplement, we collect the proofs for the remaining main results, provide details on the extension to the noncentered observations and demonstrate numerical studies in low-to-moderate dimensional settings.

REFERENCES

[1] BICKEL, P. J. and LEVINA, E. (2008). Covariance regularization by thresholding. *Ann. Statist.* **36** 2577–2604. MR2485008

[2] BICKEL, P. J. and LEVINA, E. (2008). Regularized estimation of large covariance matrices. *Ann. Statist.* **36** 199–227. MR2387969

[3] BIRNBAUM, A., JOHNSTONE, I. M., NADLER, B. and PAUL, D. (2013). Minimax bounds for sparse PCA with noisy high-dimensional data. *Ann. Statist.* **41** 1055–1084. MR3113803

[4] BUJA, A. (1986). On the Huber–Strassen theorem. *Probab. Theory Related Fields* **73** 149–152. MR0849070

[5] CAI, T. T., MA, Z. and WU, Y. (2013). Sparse PCA: Optimal rates and adaptive estimation. *Ann. Statist.* **41** 3074–3110.

[6] CAI, T. T., MA, Z. and WU, Y. (2015). Optimal estimation and rank detection for sparse spiked covariance matrices. *Probab. Theory Related Fields* **161** 781–815.

[7] CAI, T. T., REN, Z. and ZHOU, H. H. (2013). Optimal rates of convergence for estimating Toeplitz covariance matrices. *Probab. Theory Related Fields* **156** 101–143. MR3055254

[8] CAI, T. T., REN, Z. and ZHOU, H. H. (2016). Estimating structured high-dimensional covariance and precision matrices: Optimal rates and adaptive estimation. *Electron. J. Stat.* **10** 1–59. MR3466172

[9] CAI, T. T., ZHANG, C.-H. and ZHOU, H. H. (2010). Optimal rates of convergence for covariance matrix estimation. *Ann. Statist.* **38** 2118–2144. MR2676885

[10] CAI, T. T. and ZHOU, H. H. (2012). Optimal rates of convergence for sparse covariance matrix estimation. *Ann. Statist.* **40** 2389–2420. MR3097607

[11] CHEN, M., GAO, C. and REN, Z. (2018). Supplement to "Robust covariance and scatter matrix estimation under Huber's contamination model." DOI:10.1214/17-AOS1607SUPP.

[12] DAVIDSON, K. R. and SZAREK, S. J. (2001). Local operator theory, random matrices and Banach spaces. *Handbook of the Geometry of Banach Spaces* **1** 131.

[13] DAVIS, C. and KAHAN, W. M. (1970). The rotation of eigenvectors by a perturbation. III. *SIAM J. Numer. Anal.* **7** 1–46.

[14] DONOHO, D. and HUBER, P. J. (1983). The notion of breakdown point. In *A Festschrift for Erich L. Lehmann*. 157–184. Wadsworth, Belmont, CA. MR0689745

[15] DONOHO, D. L. (1982). Breakdown properties of multivariate location estimators. Technical report, Harvard Univ., Boston. Available at http://www-stat.stanford.edu/~donoho/Reports/Oldies/BPMLE.pdf.

[16] DONOHO, D. L. (1994). Statistical estimation and optimal recovery. *Ann. Statist.* **22** 238–270. MR1272082

[17] DONOHO, D. L. and GASKO, M. (1992). Breakdown properties of location estimates based on halfspace depth and projected outlyingness. *Ann. Statist.* **20** 1803–1827.

[18] DONOHO, D. L. and LIU, R. C. (1991). Geometrizing rates of convergence, III. *Ann. Statist.* **19** 668–701.

[19] 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.

[20] DUDLEY, R. M. (1978). Central limit theorems for empirical measures. *Ann. Probab.* **6** 899–929. MR0512411

[21] DÜMBGEN, L. (1998). On Tyler's M-functional of scatter in high dimension. *Ann. Inst. Statist. Math.* **50** 471–491.

[22] FAN, J., HAN, F. and LIU, H. (2014). PAGE: Robust pattern guided estimation of large covariance matrix. Technical report, Princeton Univ., Princeton, NJ.

[23] FANG, K.-T., KOTZ, S. and NG, K. W. (1990). *Symmetric Multivariate and Related Distributions*. Chapman & Hall, London.

[24] FRISTON, K. J., JEZZARD, P. and TURNER, R. (1994). Analysis of functional MRI time-series. *Human Brain Mapping* **1** 153–171.
[25] HAMPEL, F. R. (1971). A general qualitative definition of robustness. *Ann. Math. Stat.* **42** 1887–1896.
[26] HAN, F. and LIU, H. (2013). Optimal rates of convergence for latent generalized correlation matrix estimation in transelliptical distribution. Preprint. Available at arXiv:1305.6916.
[27] HAN, F. and LIU, H. (2014). Scale-invariant sparse PCA on high-dimensional meta-elliptical data. *J. Amer. Statist. Assoc.* **109** 275–287.
[28] HAN, F. and LIU, H. (2017). ECA: High dimensional elliptical component analysis in non-Gaussian distributions. *J. Amer. Statist. Assoc*. To appear.
[29] HAN, F., LU, J. and LIU, H. (2014). Robust scatter matrix estimation for high dimensional distributions with heavy tails. Technical report, Princeton Univ.
[30] HUBER, P. J. (1964). Robust estimation of a location parameter. *Ann. Math. Stat.* **35** 73–101. MR0161415
[31] HUBER, P. J. (1965). A robust version of the probability ratio test. *Ann. Math. Stat.* **36** 1753–1758. MR0185747
[32] HUBER, P. J. and STRASSEN, V. (1973). Minimax tests and the Neyman–Pearson lemma for capacities. *Ann. Statist.* **1** 251–263. MR0356306
[33] JOHNSTONE, I. M. and LU, A. Y. (2009). On consistency and sparsity for principal components analysis in high dimensions. *J. Amer. Statist. Assoc.* **104** 682–693.
[34] LAM, C. and FAN, J. (2009). Sparsistency and rates of convergence in large covariance matrix estimation. *Ann. Statist.* **37** 4254–4278. MR2572459
[35] LIU, R. Y. (1990). On a notion of data depth based on random simplices. *Ann. Statist.* **18** 405–414.
[36] LIU, R. Y., PARELIUS, J. M. and SINGH, K. (1999). Multivariate analysis by data depth: Descriptive statistics, graphics and inference. *Ann. Statist.* **27** 783–858. MR1724033
[37] MA, Z. (2013). Sparse principal component analysis and iterative thresholding. *Ann. Statist.* **41** 772–801. MR3099121
[38] MARONNA, R. A. (1976). Robust M-estimators of multivariate location and scatter. *Ann. Statist.* **4** 51–67.
[39] MITRA, R. and ZHANG, C.-H. (2014). Multivariate analysis of nonparametric estimates of large correlation matrices. Preprint. Available at arXiv:1403.6195.
[40] MIZERA, I. (2002). On depth and deep points: A calculus. *Ann. Statist.* **30** 1681–1736.
[41] MIZERA, I. and MÜLLER, C. H. (2004). Location-scale depth. *J. Amer. Statist. Assoc.* **99** 949–966.
[42] OJA, H. (1983). Descriptive statistics for multivariate distributions. *Statist. Probab. Lett.* **1** 327–332.
[43] ROUSSEEUW, P. J. and HUBERT, M. (1999). Regression depth. *J. Amer. Statist. Assoc.* **94** 388–402.
[44] SERFLING, R. (2004). Some perspectives on location and scale depth functions. *J. Amer. Statist. Assoc.* **99** 970–973.
[45] TUKEY, J. W. (1974). T6: Order Statistics, in mimeographed notes for Statistics 411. Dept. Statistics, Princeton Univ.
[46] TUKEY, J. W. (1975). Mathematics and the picturing of data. In *Proceedings of the International Congress of Mathematicians* **2** 523–531.
[47] TUKEY, J. W. (1977). Exploratory Data Analysis. Addison-Wesley Series in Behavioral Science: Quantitative Methods. Reading, MA.
[48] TYLER, D. E. (1987). A distribution-free M-estimator of multivariate scatter. *Ann. Statist.* **15** 234–251.
[49] VAN DER VAART, A. W. (2000). *Asymptotic Statistics*. Cambridge Univ. Press, Cambridge.

[50] VAPNIK, V. N. and CHERVONENKIS, A. Y. (1971). On the uniform convergence of relative frequencies of events to their probabilities. *Theory Probab. Appl.* **16** 264–280.
[51] VARDI, Y. and ZHANG, C.-H. (2000). The multivariate ℓ_1-median and associated data depth. *Proc. Natl. Acad. Sci. USA* **97** 1423–1426.
[52] VISSER, H. and MOLENAAR, J. (1995). Trend estimation and regression analysis in climatological time series: An application of structural time series models and the Kalman filter. *J. Climate* **8** 969–979.
[53] VU, V. Q. and LEI, J. (2013). Minimax sparse principal subspace estimation in high dimensions. *Ann. Statist.* **41** 2905–2947.
[54] WEGKAMP, M. and ZHAO, Y. (2016). Adaptive estimation of the copula correlation matrix for semiparametric elliptical copulas. *Bernoulli* **22** 1184–1226.
[55] XUE, L. and ZOU, H. (2013). Optimal estimation of sparse correlation matrices of semiparametric Gaussian copulas. *Stat. Interface* **7** 201–209.
[56] XUE, L. and ZOU, H. (2014). Rank-based tapering estimation of bandable correlation matrices. *Statist. Sinica* **24** 83–100.
[57] ZHANG, J. (2002). Some extensions of Tukey's depth function. *J. Multivariate Anal.* **82** 134–165.
[58] ZUO, Y. and CUI, H. (2005). Depth weighted scatter estimators. *Ann. Statist.* **33** 381–413.
[59] ZUO, Y. and SERFLING, R. (2000). General notions of statistical depth function. *Ann. Statist.* **28** 461–482.
[60] ZUO, Y. and SERFLING, R. (2000). Nonparametric notions of multivariate "scatter measure" and "more scattered" based on statistical depth functions. *J. Multivariate Anal.* **75** 62–78.

M. CHEN
DEPARTMENT OF MEDICINE
UNIVERSITY OF CHICAGO
CHICAGO, ILLINOIS 60637
USA
E-MAIL: mengjiechen@uchicago.edu
URL: http://www.mengjiechen.com

C. GAO
DEPARTMENT OF STATISTICS
UNIVERSITY OF CHICAGO
CHICAGO, ILLINOIS 60637
USA
E-MAIL: chaogao@galton.uchicago.edu
URL: http://www.stat.uchicago.edu/~chaogao

Z. REN
DEPARTMENT OF STATISTICS
UNIVERSITY OF PITTSBURGH
PITTSBURGH, PENNSYLVANIA 15260
USA
E-MAIL: zren@pitt.edu
URL: http://www.pitt.edu/~zren