Robust Scatter Matrix Estimation for High Dimensional Distributions with Heavy Tails

Junwei Lu*, Fang Han†, and Han Liu‡

May 18, 2016

Abstract

This paper studies large scatter matrix estimation for heavy tailed distributions. The contributions of this paper are twofold. First, we propose and advocate to use a new distribution family, the pair-elliptical, for modeling the high dimensional data. The pair-elliptical is more flexible and easier to check the goodness of fit compared to the elliptical. Secondly, built on the pair-elliptical family, we advocate using quantile-based statistics for estimating the scatter matrix. For this, we provide a family of quantile-based statistics, which balances the efficiency and robustness. In particular, we show that the propose estimators have comparable performance to the moment-based counterparts under the Gaussian assumption. The method is also tuning-free compared to Catoni's M-estimator for covariance matrix estimation. We further apply the method to conduct a variety of statistical methods. The corresponding theoretical properties as well as numerical performances are provided.

Keyword: Heavy-tailed distribution; Pair-Elliptical Distribution; Quantile-based statistics; Scatter matrix.

1 Introduction

Large covariance matrix estimation is a core problem in multivariate statistics. Pearson's sample covariance matrix is widely used for estimation and proves to enjoy certain optimality under the subgaussian assumption (Bickel and Levina, 2008a,b; Cai et al., 2010; Cai and Zhou, 2012a; Lounici, 2014). However, this assumption is not realistic in many real applications where data are heavy-tailed (Chen, 2002; Bradley and Taqqu, 2003; Han and Liu, 2014).

To handle heavy-tailed data, rank-based statistics are proposed. Compared to Pearson's sample covariance, rank-based estimators achieve extra efficiency via exploiting the dataset's geometric structures. Such structures, like symmetry, are naturally involved in the data generating scheme and allow for both efficient and robust inference. Conducting rank-based covariance matrix estimation

*Department of Operations Research and Financial Engineering, Princeton University, Princeton, NJ 08544, USA; e-mail: junweil@princeton.edu

†Department of Biostatistics, Johns Hopkins University, Baltimore, MD 21205, USA; e-mail: fhan@jhsph.edu

‡Department of Operations Research and Financial Engineering, Princeton University, Princeton, NJ 08544, USA; e-mail: hanliu@princeton.edu

includes two steps. The first step is to estimate the (latent) correlation matrix. For this, Liu et al. (2012a), Xue and Zou (2012), Han and Liu (2012a), Han et al. (2013), Han and Liu (2012b), Liu et al. (2012b), and Han and Liu (2013a) exploit Spearman's rho and Kendall's tau estimators. They work under the nonparanormal or the transelliptical distribution family. The second step is to estimate marginal variances. For this, Wang et al. (2014), Fan et al. (2013a), and Fan et al. (2014) exploit Catoni's M-estimator (Catoni, 2012). However, Catoni's estimator requires to tune parameters. Moreover, it is sensitive to outliers and accordingly is not a robust estimator.

In this paper, we strengthen the results in the literature in two directions. First, we propose and advocate to use a new distribution family, the pair-elliptical. The pair-elliptical family is strictly larger and requires less symmetry structure than the elliptical. We provide detailed studies on the relation between the pair-elliptical and several heavy tailed distribution families, including the nonparanormal, elliptical, and transelliptical. Moreover, it is easier to test the goodness of fit for the pair-elliptical. For conducting such a test, we combine the existing results in low dimensions (Li et al., 1997; Koltchinskii and Sakhanenko, 2000; Sakhanenko, 2008; Huffer and Park, 2007; Batsidis et al., 2014) with the familywise error rate controlling techinques including the Bonferonni's correction, the Holm's step-down procedure (Holm, 1979), and the higher criticism method (Donoho and Jin, 2004; Hall and Jin, 2010).

Secondly, built on the pair-elliptical family, we propose a new set of quantile-based statistics for estimating scatter/covariance matrices[1]. We also provide the theoretical properties of the proposed methods. In particular, we show that the proposed quantile-based methods outperform the existing ones for better balancing the robustness and efficiency. As applications, we exploit the proposed estimators for conducting several high dimensional statistical methods, and show the advantages of using the quantile-based statistics both theoretically and empirically.

1.1 Other Related Works

The quantile-based statistics, such as the median absolute deviation (Hampel, 1974) and the Q_n estimators (Rousseeuw and Croux, 1993; Croux and Ruiz-Gazen, 2005), have been used in estimating marginal standard deviations. Their properties in parameter estimation and robustness to outliers are further studied in low dimensions (Huber and Ronchetti, 2009). Moreover, these estimators have been generalized to estimate the dispersions between random variables (Gnanadesikan and Kettenring, 1972; Genton and Ma, 1999; Ma and Genton, 2001; Maronna and Zamar, 2002).

Given these results, we mainly make three contributions: (i) Methodologically, we propose new quantile-based estimators that generalize the existing MAD and Q_n estimators for better balancing the efficiency and robustness and study their statistical properties in the high dimensional setting. (ii) Theoretically, we provide more understandings on the quantile-based methods. They confirm that the quantile-based estimators are also good alternatives to the prevailing moment-based estimators in high dimensions. (iii) We propose a projection method for overcoming the lack of positive semidefiniteness, which is typical in the robust scatter matrix estimation. This approach maintains the efficiency as well as robustness to data contamination, while the prevailing SVD decomposition approach (Maronna and Zamar, 2002) cannot.

[1]The scatter matrix is any matrix proportional to the covariance matrix. See Maronna et al. (2006) for more details.

Of note, the effectiveness of quantile-base methods is being realized in other fields in high dimensional statistics. For example, Wang et al. (2012), Belloni and Chernozhukov (2011), and Wang (2013) provide analysis on the penalized quantile regression and show that it can handle the case that the noise term is very heavy-tailed. Our method, although very different from theirs, shares similar properties.

1.2 Notation System

Let $\mathbf{M} = [M_{jk}] \in \mathbb{R}^{d\times d}$ be a matrix and $\boldsymbol{v} = (v_1, ..., v_d)^T \in \mathbb{R}^d$ be a vector. We denote $\boldsymbol{v}_I$ to be the subvector of $\boldsymbol{v}$ whose entries are indexed by a set $I \subset \{1, \ldots, d\}$. We denote $\mathbf{M}_{I,J}$ to be the submatrix of $\mathbf{M}$ whose rows and columns are indexed by I and J. For $0 < q < \infty$, we define the ℓ_0, ℓ_q, and ℓ_∞ vector (pseudo-)norms to be $||\boldsymbol{v}||_0 := \sum_{j=1}^d I(v_j \neq 0)$, $||\boldsymbol{v}||_q := (\sum_{j=1}^d |v_j|^q)^{1/q}$, and $||\boldsymbol{v}||_\infty := \max_{1\leq j\leq d} |v_j|$. Here $I(\cdot)$ represents the indicator function. For a matrix $\mathbf{M}$, we denote the matrix ℓ_q, $\ell_{\max}$, and Frobenius ℓ_{F}-norms of $\mathbf{M}$ to be $||\mathbf{M}||_q := \max_{\|\boldsymbol{v}\|_q=1} \|\mathbf{M}\boldsymbol{v}\|_q$, $||\mathbf{M}||_{\max} := \max_{jk} |M_{jk}|$, and $||\mathbf{M}||_{\mathrm{F}} := (\sum_{j,k} |M_{jk}|^2)^{1/2}$. For any matrix $\mathbf{M} \in \mathbb{R}^{d\times d}$, we denote $\mathrm{diag}(\mathbf{M})$ to be the diagonal matrix with the same diagonal entries as $\mathbf{M}$, and $\mathbf{I}_d \in \mathbb{R}^{d\times d}$ to be the d by d identity matrix. Let $\lambda_j(\mathbf{M})$ and $\boldsymbol{u}_j(\mathbf{M})$ represent the j-th largest eigenvalue and the corresponding eigenvector of $\mathbf{M}$, and $\langle \mathbf{M}_1, \mathbf{M}_2 \rangle := \mathrm{Tr}(\mathbf{M}_1^T \mathbf{M}_2)$ be the inner product of $\mathbf{M}_1$ and $\mathbf{M}_2$. For any two random vectors $\boldsymbol{X}$ and $\boldsymbol{Y}$, we write $\boldsymbol{X} \overset{D}{=} \boldsymbol{Y}$ if and only if $\boldsymbol{X}$ and $\boldsymbol{Y}$ are identically distributed. Throughout the paper, we let c, C be two generic absolute constants, whose values may vary at different locations.

1.3 Paper Organization

The rest of the paper is organized as follows. In the next section we provide the theoretical evaluation of the impacts of heavy tails on the moment-based estimators. This motivates our work. Section 3 proposes the pair-elliptical family, and reveals the connection among the Gaussian, elliptical, nonparanormal, transelliptical, and pair-elliptical. In Section 4, we introduce the generalized MAD and Q_n estimators for estimating scatter/covariance matrices. Section 5 provides the theoretical results. Section 6 discusses parameter selection. In Section 7, we apply the proposed estimators to conduct multiple multivariate methods. We put experiments on synthetic and real data in Section 8, more discussions in Section 9, and technical proofs in the appendix.

2 Impacts of Heavy Tails on Moment-Based Estimators

This section illustrates the motivation of quantile-based estimators. In particular, we show how moment-based estimators fail for heavy tailed data. These estimators include the sample mean and sample covariance matrix. Such estimators are known to be efficient under stringent moment assumptions (Lounici, 2014). However, their performance drops down when such assumptions are violated (Cai et al., 2011; Liu et al., 2012a).

We characterize the heavy tailedness by the L_p norm. In detail, for any random variable $X \in \mathbb{R}$ and integer $p \geq 1$, we define the L_p norm of X as

$$||X||_{L_p} := (\mathbb{E}|X|^p)^{1/p}.$$

The random variable X is heavy tailed if there exists some $p > 0$ such that

$$||X||_{L_q} \leq K \leq \infty \quad \text{for } q \leq p, \quad \text{and } ||X||_{L_{p+1}} = \infty.$$

The heavy tailedness of X is measured by how large p could be such that the p-th moment exists.

In the following we first provide an upper bound of the sample mean. It illustrates the "optimal rate but sub-optimal scaling" phenomenon.

Theorem 2.1. Suppose $\boldsymbol{X} = (X_1, \ldots, X_d)^T \in \mathbb{R}^d$ is a random vector with the population mean $\boldsymbol{\mu}$. Assume $\boldsymbol{X}$ satisfies $||X_j||_{L_p} \leq K$, where we assume $d = O(n^\gamma)$ and $p = 2 + 2\gamma + \delta$. Letting $\bar{\boldsymbol{\mu}}$ be the sample mean of n independent observations of $\boldsymbol{X}$, we then have, with probability no smaller than $1 - 2d^{-2.5} - (\log d)^{p/2} n^{-\delta/2}$,

$$||\bar{\boldsymbol{\mu}} - \boldsymbol{\mu}||_\infty \leq 12K \cdot \sqrt{\frac{\log d}{n}}.$$

Theorem 2.1 shows that, for preserving the $O_P(\sqrt{\log d/n})$ rate of convergence, p determines how large the dimension d can be compared to n. For example, when at most $(4+\epsilon)$-th moment exists for $\boldsymbol{X}$ for some $\epsilon > 0$, the sample mean attains the optimal rate $O_P(\sqrt{\log d/n})$ under the suboptimal scaling $d = O(n)$.

The results in Theorem 2.1 cannot be improved without adding more assumptions. Via a worst case analysis, the next theorem characterizes the sharpness of Theorem 2.1.

Theorem 2.2. For any fixed constant C, $p = 2 + 2\gamma$ with $\gamma > 0$, and $d = n^{\gamma + \delta_0}$ for some $\delta_0 > 0$, there exists certain random vector $\boldsymbol{X}$, satisfying

$$||X_i||_{L_q} < K, \text{ for some absolute constant } K > 0 \text{ and all } q \leq p,$$

such that, with probability tending to 1, we have

$$||\bar{\boldsymbol{\mu}} - \boldsymbol{\mu}||_\infty \geq \sqrt{\frac{C \log d}{n}}.$$

Theorems 2.1 and Theorem 2.2 together illustrate the constraints of applying moment-based estimators to study heavy tailed distributions. This motivates us to consider alternative methods that are more efficient in handling heavy tailedness.

3 Pair-Elliptical Distribution

In this section, we introduce the pair-elliptical distribution family. We first briefly review several existing distribution families: Gaussian, elliptical, nonparanormal, and transelliptical. Then we elaborate the relations between the pair-elliptical and aforementioned families.

3.1 Multivariate Distribution Families

We start by first introducing the elliptical distribution. The elliptical family contains symmetric but possibly very heavy tailed distributions.

Definition 3.1 (Elliptical distribution, Fang et al. (1990))**.** A d-dimensional random vector $\boldsymbol{X}$ is said to follow an elliptical distribution if and only if there exists a vector $\boldsymbol{\mu} \in \mathbb{R}^d$, a nonnegative random variable $\xi \in \mathbb{R}$, a matrix $\mathbf{A} \in \mathbb{R}^{d\times q}$ ($q \leq d$) of rank q, a random vector $\boldsymbol{U} \in \mathbb{R}^q$ uniformly distributed in q-dimension sphere $\mathbb{S}^{q-1}$ and independent from ξ, such that

$$\boldsymbol{X} \overset{D}{=} \boldsymbol{\mu} + \xi \mathbf{A}\boldsymbol{U}.$$

In this case, we represent $\boldsymbol{X} \sim EC_d(\boldsymbol{\mu}, \mathbf{S}, \xi)$, where $\mathbf{S} := \mathbf{A}\mathbf{A}^T$ is of rank q.

Remark 3.2. An equivalent definition of the elliptical distribution is: Any random vector $\boldsymbol{X} \sim EC_d(\boldsymbol{\mu}, \mathbf{S}, \xi)$ is elliptically distributed if and only if the characteristic function of $\boldsymbol{X}$ is of the form $\exp(\mathrm{i}\boldsymbol{t}^T\boldsymbol{\mu})\phi(\boldsymbol{t}^T\mathbf{S}\boldsymbol{t})$, where i is the imaginary number satisfying $\mathrm{i}^2 = -1$, ϕ is a properly defined characteristic function, and there exists a one to one map between ξ and ϕ. In this case, we represent $\boldsymbol{X} \sim EC_d(\boldsymbol{\mu}, \mathbf{S}, \phi)$. Moreover, when the elliptical distribution is absolutely continuous, the density function is of the form $g((\boldsymbol{x}-\boldsymbol{\mu})^T\mathbf{S}^{-1}(\boldsymbol{x}-\boldsymbol{\mu}))$ for some nonnegative function $g(\cdot)$. In this case, we represent $\boldsymbol{X} \sim EC_d(\boldsymbol{\mu}, \mathbf{S}, g)$.

Although elliptical distributions have been extensively explored in modeling many real world data, including financial (Owen and Rabinovitch, 1983; Berk, 1997; McNeil et al., 2010; Embrechts et al., 2002) and imaging data (Marden and Manolakis, 2004; Frontera-Pons et al., 2012), it can be quite restrictive due to the symmetry constraint (Frahm, 2004). One way to handle asymmetric data is to exploit the copula technique. This results to the transelliptical family (meta-elliptical family) proposed and discussed in Fang et al. (2002) and Han and Liu (2014). Below we give the formal definition of the transelliptical distribution in Han and Liu (2014).

Definition 3.3 (Transelliptical distribution, Han and Liu (2014))**.** A continuous random vector $\boldsymbol{X} = (X_1, \ldots, X_d)^T$ follows a transelliptical distribution, denoted by $\boldsymbol{X} \sim TE_d(\boldsymbol{\Sigma}^0, \xi; f_1, \ldots, f_d)$, if there exist univariate strictly increasing functions $f_1, \ldots, f_d$ such that

$$(f_1(X_1), \ldots, f_d(X_d))^T \sim EC_d(\mathbf{0}, \boldsymbol{\Sigma}^0, \xi), \quad \text{where } \mathrm{diag}(\boldsymbol{\Sigma}^0) = \mathbf{I}_d \text{ and } \mathbb{P}(\xi = 0) = 0. \tag{3.1}$$

In particular, when

$$(f_1(X_1), \ldots, f_d(X_d))^T \sim N_d(\mathbf{0}, \boldsymbol{\Sigma}^0), \quad \text{where } \mathrm{diag}(\boldsymbol{\Sigma}^0) = \mathbf{I}_d,$$

$\boldsymbol{X}$ follows a nonparanormal distribution (Liu et al., 2009, 2012a). Here $\boldsymbol{\Sigma}^0$ is called the latent generalized correlation matrix.

3.2 Pair-Elliptical Distribution

In this section we propose a new distribution family, the pair-elliptical. Compared to the elliptical and transelliptical, the pair-elliptical distribution is of more interest to us. Specifically, it balances the modeling flexibility and interpretability in covariance/scatter matrices estimation.

Definition 3.4. A continuous random vector $\boldsymbol{X} = (X_1, \ldots, X_d)^T$ is said to follow a pair-elliptical distribution, denoted by $\boldsymbol{X} \sim PE_d(\boldsymbol{\mu}, \mathbf{S}, \xi)$, if and only if any pair of random variables $(X_j, X_k)^T$ of $\boldsymbol{X}$ is elliptically distributed. In other words, we have

$$(X_j, X_k)^T \sim EC_2\Big(\boldsymbol{\mu}_{\{j,k\}}, \mathbf{S}_{\{j,k\},\{j,k\}}, \xi\Big) \text{ for all } j \neq k \in \{1, \ldots, d\} \text{ and } \mathbb{P}(\xi = 0) = 0.$$

As a special example, a distribution is said to be pair-normal, written as $PN_d(\boldsymbol{\mu}, \mathbf{S})$, if any pairs of $\boldsymbol{X}$ is bivariate Gaussian distributed.

It is obvious that the pair-elliptical family contains the elliptical distribution family. Moreover, the elliptical is a strict subfamily of the pair-elliptical by considering the following example.

Example 3.5. Let $f(X_1, X_2, X_3)$ be the density function of a three dimensional standard Gaussian distribution with median $\mathbf{0}$ and covariance matrix $\mathbf{I}_3$, and $\boldsymbol{X} = (X_1, X_2, X_3)^T$ be a 3-dimensional random vector with the density function

$$g(X_1, X_2, X_3) = \begin{cases} 2f(X_1, X_2, X_3), & \text{if } X_1X_2X_3 \geq 0, \\ 0, & \text{otherwise.} \end{cases} \tag{3.2}$$

The distribution in Example 3.5 with density in (3.2) is bivariate Gaussian distributed for any pairwise marginal distributions, and therefore belongs to the pair-elliptical family. On the other hand, this distribution is marginally Gaussian distributed but not multivariate Gaussian distributed, and accordingly cannot be elliptically distributed.

Example 3.5 also shows that the pair-elliptical distribution can be asymmetric. Moreover, the pair-elliptical distribution has a naturally defined scatter matrix $\mathbf{S}$, which is proportional to the covariance matrix $\boldsymbol{\Sigma}$ when $\mathbb{E}\xi^2$ exists. This makes the pair-elliptical compatible with many multivariate methods such as principal component analysis and linear discriminant analysis.

The rest of this section focuses on characterizing the relations among the Gaussian, elliptical[2], transelliptical, nonparanormal, pair-elliptical, and pair-normal families. Recall that in this paper we are only interested in the continuous distributions with density existing. It is obvious that the Gaussian family is a strict subfamily of the elliptical, and the elliptical is also a strict subfamily of both the transelliptical and the pair-elliptical. The next proposition shows that the only intersection between the elliptical and the nonparanormal is the Gaussian.

Proposition 3.6 (Liu et al. (2012b))**.** If a random vector is both nonparanormally and elliptically distributed, it must follow a Gaussian distribution.

In the next proposition, we show that the only intersection between the transelliptical and the pair-elliptical is the elliptical.

Proposition 3.7. If a random vector is both transelliptically and pair-elliptically distributed, it must follow an elliptical distribution.

We defer the proof of Preposition 3.7 to the appendix. In the end, let's consider the relation among the pair-normal and all the other distribution families. By definition, the pair-normal is a strict subfamily of the pair-elliptical. On the other hand, the next proposition shows that any random scaled version of the pair-normal is pair-elliptically distributed.

Proposition 3.8. Let $\boldsymbol{Y} \sim PN_d(\boldsymbol{\mu}, \mathbf{S})$ follow a pair-normal distribution. Then for any nonnegative random variable ξ with $\mathbb{P}(\xi = 0) = 0$ and independent of $\boldsymbol{Y}$, we have $\boldsymbol{X} = \boldsymbol{\mu}' + \xi\boldsymbol{Y}$ follows a pair-elliptical distribution.

In the end, we have the following proposition, which characterizes pair-normal's connections to the elliptical and nonparanormal distributions.

[2]In the rest of this section we only focus on the continuous elliptical distributions with $\mathbb{P}(\xi = 0) = 0$. And we are only interested in those whose covariance matrix is not the identity.

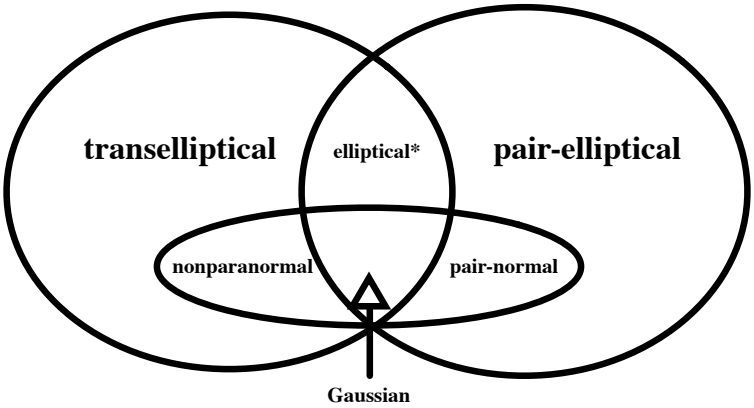

Figure 1: The Venn diagram illustrating the relations of the Gaussian, elliptical, nonparanormal, transelliptical, pair-normal, and pair-elliptical families. Here "elliptical*" represents the continuous elliptical family with $\mathbb{P}(\xi = 0) = 0$.

Proposition 3.9. For the pair-normal, elliptical, and nonparanormal distributions, we have

(i) The only intersection between the pair-normal and elliptical is the Gaussian;

(ii) The only intersection between the pair-normal and nonparanormal is the Gaussian.

In conclusion, the Venn diagram in Figure 1 summarizes the relation among the Gaussian, elliptical, nonparanormal, transelliptical, pair-normal, and pair-elliptical. From the figure, we can see that the Gaussian distribution locates in the central area whose covariance can be well estimated by the sample covariance matrix. The transelliptical covers the left-hand side of the diagram, where we advocate using rank-based estimators to estimate the covariance matrix. The pair-elliptical covers a new regime on the right-hand side of the diagram, where we will introduce the quantile-based estimators for estimating the covariance matrix.

3.3 Goodness of Fit Test of the Pair-Elliptical

This section proposes a goodness of fit test of the pair-elliptical. The pair-elliptical family has its advantages here: Both the transelliptical and elliptical require global geometric constraints over all covariates; In comparison, the pair-elliptical only requires a local pairwise symmetry structure, which could be more easily checked. In this section, we combine the test of elliptical symmetry proposed in Batsidis et al. (2014) with the step-down procedure in Holm (1979) for performing the pair-elliptical goodness of fit test.

Specifically, we propose a test of pair-elliptical:

$$\mathrm{H}_0 : \text{The data are pair-elliptically distributed.} \tag{3.3}$$

The proposed test is in two steps: In the first step, we test the pairwise elliptical symmetry; In the second step, we use the Holm's step-down procedure to control the family-wise error.

In the first step, we apply the statistic proposed in Batsidis et al. (2014) for testing pairwise elliptical symmetry. Let $\overline{\boldsymbol{Z}}$ and $\widehat{\boldsymbol{\Sigma}}$ be the sample mean and sample covariance of $\{\boldsymbol{Z}_i\}_{i=1}^n$. We standardize the data by letting $\boldsymbol{Y}_i := \widehat{\boldsymbol{\Sigma}}^{-1/2}(\boldsymbol{Z}_i - \overline{\boldsymbol{Z}})$ and $t(\boldsymbol{Z}_i) := \sqrt{2}\widetilde{\boldsymbol{Y}}_i/w_i$ where $\widetilde{\boldsymbol{Y}}_i := (\boldsymbol{Y}_{i1} + \boldsymbol{Y}_{i2})/2$

and $w_i^2 := \sum_{j=1}^2 (\boldsymbol{Y}_{ij} - \widetilde{\boldsymbol{Y}}_i)^2$ for $i = 1, \ldots, n$. Under H_0, we have $t(\boldsymbol{Z}_i) \xrightarrow{D} t_1$ for $i = 1, \ldots, n$, where t_1 is the t distribution with degree of freedom 1. To study the goodness-of-fit of the t-distribution, we define $M := [\sqrt{n}]$, where $[\cdot]$ represents the integer part of a real number and $E := n/M$. Let T_ℓ be the $\ell/M \times 100\%$ quantile of the t_1 distribution for $\ell = 0, \ldots, M$, where $T_0 := -\infty$ and $T_M := +\infty$. We also denote the observed frequency $O_\ell := |\{t(\boldsymbol{Z}_i) : T_{\ell-1} < t(\boldsymbol{Z}_i) \leq T_\ell\}|$ for $1 \leq \ell \leq M$. Batsidis et al. (2014) consider the following Pearson's chi-squared test statistic:

$$Z(\{\boldsymbol{Z}_i\}) := \sum_{\ell=1}^{M} \frac{(O_\ell - E)^2}{E}.$$

By its nature, $Z(\{\boldsymbol{Z}_i\})$ is asymptotically chi-squared distributed with degrees of freedom $M - 1$.

In the second step, we screen the data to find whether there is any pair $\{X_j, X_k\}$ that does not follow an elliptical distribution. Considering the following null hypothesis for any $1 \leq j, k \leq d$:

$$\mathrm{H}_{jk} : \{X_j, X_k\} \text{ are elliptically distributed,} \tag{3.4}$$

we use the Holm's step-down procedure (Holm, 1979) to control the family-wise error rate. Denote the p-values of $Z(\{(X_{ij}, X_{ik})\})$ as π_{jk} and let m_{jk} be the rank statistic of π_{jk} such that

$$m_{jk} := |\{\pi_{j'k'} \,|\, \pi_{j'k'} < \pi_{jk}, 1 \leq j', k' \leq d\}|.$$

The Holm's adjusted p-values are defined as

$$\pi_{jk}^H := \max\left\{1 - (1 - \pi_{j'k'})^{t_{j'k'}} \,|\, m_{j'k'} \leq m_{jk}, 1 \leq j', k' \leq d\right\}, \text{ where } t_{jk} = 1 - \frac{2(m_{jk} - 1)}{d(d-1)}. \tag{3.5}$$

Applying the adjusted p-values, we reject H_{jk} if π_{jk}^H is smaller than the level of significance α. Let $\omega_0 := \{(j,k) \,|\, H_{jk} \text{ in (3.4) is true}, 1 \leq j \neq k \leq d\}$, Holm (1979) shows that we can control the family-wise error rate as

$$\mathbb{P}_{\omega_0}\left(\pi_{jk}^H \leq \alpha \text{ for some } (j,k) \in \omega_0\right) \leq \alpha.$$

Under the setting of goodness of fit test and H_0 in (3.3), we have $\omega_0 = \{(j,k) \,|\, 1 \leq j \neq k \leq d\}$ and therefore

$$\mathbb{P}_{H_0}\left(\pi_{jk}^H \leq \alpha \text{ for some } j \neq k\right) \leq \alpha.$$

4 Quantile-Based Scatter Matrix Estimation

This section introduces the quantile-based scatter matrix estimators. To this end, we first briefly review the existing quantile-based estimators, including MAD and Q_n proposed in Hampel (1974) and Rousseeuw and Croux (1993). Secondly, we generalize these two estimators for estimating scatter matrices. Thirdly, we introduce the projection idea for constructing a positive semidefinite scatter matrix estimator.

4.1 Robust Scale Estimation

This section briefly reviews the robust estimators of the true marginal standard deviation. To this end, we first define the population and sample versions of the quantile function. For any random variable $Z \in \mathbb{R}$ and fixed value $r \in [0,1]$, let $Q(Z;r)$ represent the $r \times 100\%$ quantile of Z:

$$Q(Z;r) := \inf\{z : r \le \mathbb{P}(Z \le z)\}.$$

The $r \times 100\%$ quantile is said to be unique if and only if there exists one and only one $z \in \mathbb{R}$ such that $\mathbb{P}(Z \le z) = r$. Letting $Z^{(1)} \le \ldots \le Z^{(n)}$ be the ordered statistics of i.i.d. data $Z_1, \ldots, Z_n \overset{D}{=} Z$, we define the empirical version of $Q(Z;r)$ as

$$\widehat{Q}(\{Z_i\};r) := Z^{(k^*)}, \quad \text{where } k^* := \underset{i\in\{1,\ldots,n\}}{\arg\min} \Big\{\frac{i}{n} \ge r\Big\}. \tag{4.1}$$

We then introduce the standard deviation estimators based on the quantiles. These include the median absolute deviation (MAD) and Q_n estimators. MAD is defined as follows:

$$\text{(MAD estimator)} \qquad c^{\mathrm{MAD}} \cdot \widehat{Q}\Big(\Big\{\Big|Z_i - \widehat{Q}\Big(\{Z_i\};\frac{1}{2}\Big)\Big|\Big\}_i;\frac{1}{2}\Big),$$

where c^{MAD} is the constant making its population counterpart the standard deviation. In general, c^{MAD} is different for different distributions. MAD is robust to outliers with $1/2$ breakdown points (Huber and Ronchetti, 2009), but has relatively low efficiency compared to the sample standard deviation under the Gaussian model.

To improve the efficiency while preserving the robustness, the Q_n estimator is proposed:

$$(Q_n \text{ estimator}) \qquad c^{Q_n} \cdot \widehat{Q}\Big(\Big\{|Z_i - Z_{i'}|\Big\}_{i<i'};1/4\Big),$$

where c^{Q_n} is another constant comparable to c^{MAD}. Q_n is known to be more efficient than MAD.

4.2 Robust Scatter Matrix Estimators

In this section, we propose our approach for estimating the scatter matrix. This is via generalizing the aforementioned quantile-based scale estimators.

Assume $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ are n independent observations of a d-dimensional random vector $\boldsymbol{X} = (X_1, \ldots, X_d)^T$ with $\boldsymbol{X}_i = (X_{i1}, \ldots, X_{id})^T$. We first propose to estimate the marginal standard deviations by generalizing the MAD and Q_n estimators. Specifically, we define generalized median absolute deviation (gMAD) as follows: For the j-th entry of $\boldsymbol{X}$, the population and sample versions of gMAD are:

$$\begin{aligned} \text{(gMAD)} \qquad & \sigma^{\mathrm{M}}(X_j;r) := Q(|X_j - Q(X_j;1/2)|;r), \\ & \widehat{\sigma}^{\mathrm{M}}(X_j;r) := \widehat{Q}(|X_{ij} - \widehat{Q}(\{X_{ij}\}_{i=1}^n;1/2)|;r). \end{aligned} \tag{4.2}$$

Here the median is replaced by the $r \times 100\%$ quantile[3]. We then define the generalized Q_n estimator (gQNE) using the same idea. The population and sample versions of the gQNE for the j-th entry

[3]Later we will show that using the r-th quantile instead of the median can potentially increase the efficiency of the estimator, in the cost of losing some robustness though.

of $\boldsymbol{X}$ are:

$$\text{(gQNE)} \qquad \begin{aligned} \sigma^{\mathrm{Q}}(X_j; r) &:= Q(|X_j - \widetilde{X}_j|; r), \\ \widehat{\sigma}^{\mathrm{Q}}(X_j; r) &:= \widehat{Q}(\{|X_{ij} - X_{i'j}|\}_{i<i'}; r), \end{aligned} \tag{4.3}$$

where $\widetilde{\boldsymbol{X}} := (\widetilde{X}_1, \ldots, \widetilde{X}_d)^T$ is an independent copy of $\boldsymbol{X}$. It is easy to check that, when setting $r = 1/2$ and $r = 1/4$ in (4.2) and (4.3), we recover the median absolute deviation (MAD) and Q_n estimators. This explains why we call them the generalized MAD and Q_n estimators. Of note, for any $j \in \{1, \ldots, d\}$, we have median$(X_j - \widetilde{X}_j) = 0$. Therefore, gQNE is a generalization to the gMAD estimator without requiring estimating the medians.

For estimating the scatter matrix, besides estimating the marginal scales, we also need to estimate the covariance between any two random variables. For this, we follow the idea in Gnanadesikan and Kettenring (1972). We first remind that

$$\mathrm{Cov}(X, Y) = \frac{1}{4}\Big[\{\sigma(X+Y)\}^2 - \{\sigma(X-Y)\}^2\Big],$$

where for any random variable Z, $\sigma(Z)$ represents the populational standard deviation of Z. We then define the robust estimators of the dispersion between X and Y based on gMAD and gQNE as follows:

$$\begin{aligned} \sigma^{\mathrm{M}}(X, Y; r) &:= \frac{1}{4}\Big[\{\sigma^{\mathrm{M}}(X+Y; r)\}^2 - \{\sigma^{\mathrm{M}}(X-Y; r)\}^2\Big]; \\ \sigma^{\mathrm{Q}}(X, Y; r) &:= \frac{1}{4}\Big[\{\sigma^{\mathrm{Q}}(X+Y; r)\}^2 - \{\sigma^{\mathrm{Q}}(X-Y; r)\}^2\Big]. \end{aligned}$$

Let $\widehat{\sigma}^{\mathrm{M}}(X, Y)$ and $\widehat{\sigma}^{\mathrm{Q}}(X, Y; r)$ be the corresponding empirical versions. For any d-dimensional random vector $\boldsymbol{X} = (X_1, \ldots, X_d)^T$, we then define the d by d robust gMAD and gQNE scatter matrices $\mathbf{R}^{\mathrm{M};r} = [\mathbf{R}^{\mathrm{M};r}_{jk}]$ and $\mathbf{R}^{\mathrm{Q};r} = [\mathbf{R}^{\mathrm{Q};r}_{jk}]$ as follows: For any $j \in \{1, \ldots, d\}$ and $k < j$, we write

$$\begin{aligned} \mathbf{R}^{\mathrm{M};r}_{jj} &= (\sigma^{\mathrm{M}}(X_j; r))^2, \qquad & \mathbf{R}^{\mathrm{M};r}_{jk} &= \mathbf{R}^{\mathrm{M};r}_{kj} = \sigma^{\mathrm{M}}(X_j, X_k; r); \\ \mathbf{R}^{\mathrm{Q};r}_{jj} &= (\sigma^{\mathrm{Q}}(X_j; r))^2, \qquad & \mathbf{R}^{\mathrm{Q};r}_{jk} &= \mathbf{R}^{\mathrm{Q};r}_{kj} = \sigma^{\mathrm{Q}}(X_j, X_k; r). \end{aligned}$$

In the later section we will show that $\mathbf{R}^{\mathrm{M};r}$ and $\mathbf{R}^{\mathrm{Q};r}$ are indeed scatter matrices under the pair-elliptical family. Let $\widehat{\mathbf{R}}^{\mathrm{M};r}$ and $\widehat{\mathbf{R}}^{\mathrm{Q};r}$ be the empirical versions of $\mathbf{R}^{\mathrm{M};r}$ and $\mathbf{R}^{\mathrm{Q};r}$ via replacing $\sigma^{\mathrm{M}}(\cdot)$ and $\sigma^{\mathrm{Q}}(\cdot)$ by $\widehat{\sigma}^{\mathrm{M}}(\cdot)$ and $\widehat{\sigma}^{\mathrm{Q}}(\cdot)$. $\widehat{\mathbf{R}}^{\mathrm{M};r}$ and $\widehat{\mathbf{R}}^{\mathrm{Q};r}$ are the proposed robust scatter matrix estimators.

There are two remarks. First, we do not discuss how to select r in this section, which will be studied in more details in Section 6. Secondly, we note that $\widehat{\mathbf{R}}^{\mathrm{M};r}$ and $\widehat{\mathbf{R}}^{\mathrm{Q};r}$ are both symmetric matrices by definition. However, they are not necessarily positive semidefinite. We will discuss this issue in the next section.

4.3 Projection Method

In this section we introduce the projection idea to overcome the lack of positive semidefiniteness (PSD) in robust covariant matrix estimation. It is known that the robust covariance matrix estimator might be non-PSD (Maronna et al., 2006), especially when the dimension is close to or

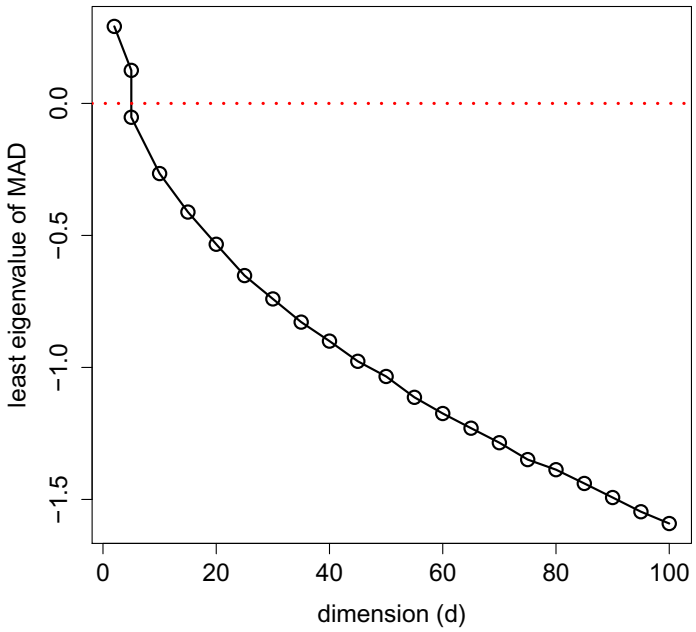

Figure 2: The plot of the averaged least eigenvalues of a MAD scatter matrix (i.e., $\widehat{\mathbf{R}}^{\mathrm{M};1/2}$) against the dimension d ranging from 2 to 200. Here the $n = 50$ observations are coming from the standard Gaussian distribution with dimension d, and the simulations are conducted with 100 repetitions.

higher than the sample size. To illustrate this, Figure 2 shows the averaged least eigenvalue of the MAD scatter matrix estimator under the standard multivariate Gaussian model, with the sample size n fixed to be 50 and the dimension d increasing from 2 to 200.

The lack of PSD can cause problems for many high dimensional multivariate methods. To solve it, we propose a general projection method. In detail, for arbitrary non-PSD matrix estimator $\widehat{\mathbf{R}}$, we consider the projection of $\widehat{\mathbf{R}}$ to the positive semidefinite matrix cone:

$$\widetilde{\mathbf{R}} = \underset{\mathbf{M} \succeq \mathbf{0}}{\arg\min} \, ||\mathbf{M} - \widehat{\mathbf{R}}||, \tag{4.4}$$

where $\mathbf{M} \succeq \mathbf{0}$ represents that $\mathbf{M}$ is PSD and $||\cdot||$ is a certain matrix norm of interest. For any given norm $||\cdot||$, a computationally efficient algorithm to solve (4.4) is given in Supplementary Material Section D.

Due to reasons that will become clearer later, we are interested in the projection with regard to the matrix element wise supremum norm $||\cdot||_{\max}$ in (4.4). Of note, $\widetilde{\mathbf{R}}$ and $\widehat{\mathbf{R}}$ have the same breakdown point because $\widetilde{\mathbf{R}}$ is independent of the data conditioning on $\widehat{\mathbf{R}}$. Moreover, we have the following property about $\widehat{\mathbf{R}}$.

Lemma 4.1. Let $\widetilde{\mathbf{R}}$ be the solution to (4.4) with certain matrix norm $||\cdot||$ of interest. We have, for any $t \geq 0$ and $\mathbf{M} \in \mathbb{R}^{d \times d}$ with $\mathbf{M} \succeq \mathbf{0}$,

$$\mathbb{P}\Big(\big\|\widetilde{\mathbf{R}} - \mathbf{R}\big\| \geq t\Big) \leq \mathbb{P}\Big(\big\|\widehat{\mathbf{R}} - \mathbf{R}\big\| \geq \frac{t}{2}\Big).$$

Of note, Maronna and Zamar (2002) propose an alternative approach to solve the non-PSD problem. Their method exploits the SVD decomposition of any given non-PSD matrix. However,

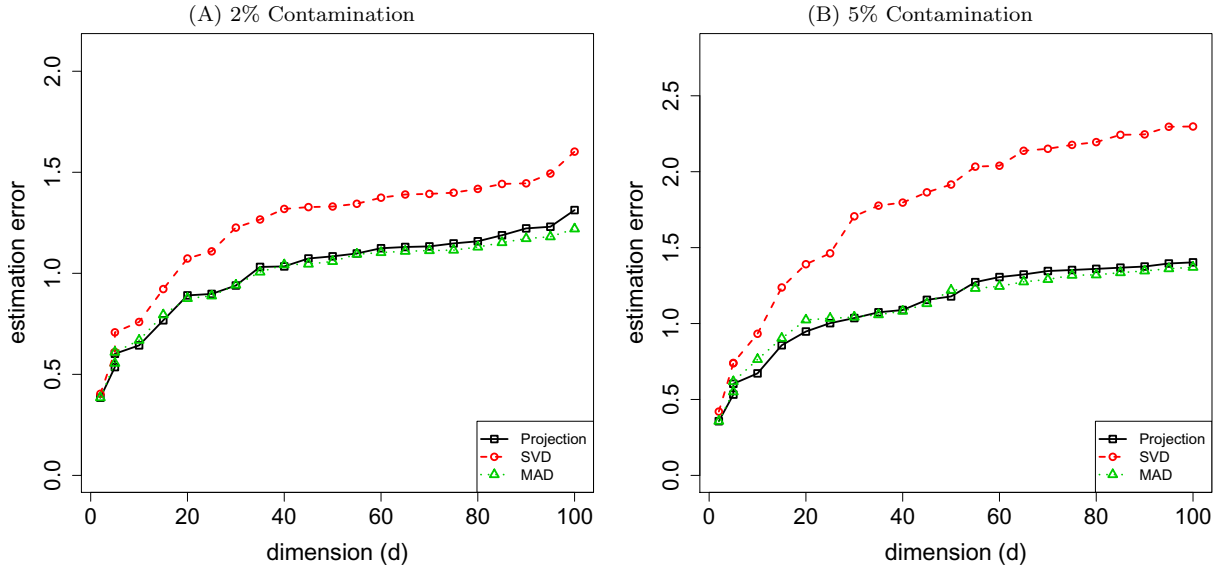

Figure 3: The plot of the averaged estimation erros of the MAD ("MAD") and PSD scatter matrix estimators using the projection and SVD decomposition ideas (denoted by "Projection" and "SVD"). The distances are calculated based on the $||\cdot||_{\max}$ norm and are plotted against the dimension d ranging from 2 to 100. Here the sample size is 50 and observations are coming from a standard Gaussian distributed data with dimension d, and 2% and 5% data points are randomly chosen and replaced by $+N(3,3)$ or $-N(3,3)$. The results are obtained based on 200 repetitions.

Maronna's method is not a robust procedure and is sensitive to outliers. More specifically, Figure 3 shows the averaged distance between the population scatter matrix and three different scatter matrix estimators: the possibly non-PSD MAD estimator (denoted by "MAD"), the PSD estimator calculated by using Maronna's SVD decomposition idea (denoted by "SVD"), and the PSD estimator calculated by our projection idea with regard to the $||\cdot||_{\max}$ norm (denoted by "Projection"). Figures 3 (A) and (B) illustrate the results regarding a standard Gaussian distributed data (i.e., the data follow a $N_d(\mathbf{0}, \mathbf{I}_d)$ distribution) with 2% and 5% points being randomly chosen and replaced by $+N(3,3)$ or $-N(3,3)$. It shows that the PSD estimator obtained by projection is as insensitive as the MAD estimator (and their estimation accuracy is very close). On the other hand, Maronna's method is very sensitive to such data contamination.

5 Theoretical Results

This section provides the theoretical results of the proposed quantile-based gQNE and gMAD scatter matrix estimators. The section is divided into two parts: In the first part, under the pair-elliptical family, we characterize the relations among the population gQNE, gMAD statistics and Pearson's covariance matrix; In the second part, we provide the theoretical analysis for gQNE and gMAD estimators.

5.1 Quantile-Based Estimators under the Pair-Elliptical

In this section we show that the population gMAD and gQNE statistics, $\mathbf{R}^{\mathrm{M};r}$ and $\mathbf{R}^{\mathrm{Q};r}$, are scatter matrices of $\boldsymbol{X}$ when $\boldsymbol{X}$ is pair-elliptically distributed.

We first focus on gMAD. The next theorem characterizes a sufficient condition under which $\mathbf{R}^{\mathrm{M};r}$ is proportional to the covariance matrix. It also quantifies the scale constant $c^{\mathrm{M};r}$ that connects $\mathbf{R}^{\mathrm{M};r}$ to the covariance matrix.

Theorem 5.1. Suppose that $\boldsymbol{X} = (X_1, \ldots, X_d)^T$ is a d-dimensional random vector with the covariance matrix $\boldsymbol{\Sigma} \in \mathbb{R}^{d\times d}$. Then there exists some constant $c^{\mathrm{M};r}$ such that

$$\mathbf{R}^{\mathrm{M};r} = c^{\mathrm{M};r}\boldsymbol{\Sigma},$$

if for any $j \neq k \in \{1, \ldots, d\}$,

$$\begin{aligned}\sqrt{c^{M;r}} = Q\Big(\Big|\frac{X_j - Q(X_j, 1/2)}{\sigma(X_j)}\Big|; r\Big) = Q\Big(\Big|\frac{X_j + X_k - Q(X_j + X_k; 1/2)}{\sigma(X_j + X_k)}\Big|; r\Big)\\ = Q\Big(\Big|\frac{X_j - X_k - Q(X_j - X_k; 1/2)}{\sigma(X_j - X_k)}\Big|; r\Big),\end{aligned} \tag{5.1}$$

and the above quantiles are all unique.

We then study gQNE. The next theorem gives a sufficient condition under which $\mathbf{R}^{\mathrm{Q};r}$ is proportional to the covariance matrix and again quantifies the scale constant $c^{\mathrm{Q};r}$.

Theorem 5.2. Suppose that $\boldsymbol{X} = (X_1, \ldots, X_d)^T$ is a d-dimensional random vector with the covariance matrix $\boldsymbol{\Sigma} \in \mathbb{R}^{d\times d}$. Let $\widetilde{\boldsymbol{X}}$ be an independent copy of $\boldsymbol{X}$ and $\boldsymbol{Z} = (Z_1, \ldots, Z_d)^T := \boldsymbol{X} - \widetilde{\boldsymbol{X}}$. Then there exists some constant $c^{\mathrm{Q};r}$ such that

$$\mathbf{R}^{\mathrm{Q};r} = c^{\mathrm{Q};r}\boldsymbol{\Sigma},$$

if for any $j \neq k \in \{1, \ldots, d\}$,

$$\sqrt{c^{\mathrm{Q};r}/2} = Q\Big(\Big|\frac{Z_j}{\sigma(Z_j)}\Big|; r\Big) = Q\Big(\Big|\frac{Z_j + Z_k}{\sigma(Z_j + Z_k)}\Big|; r\Big) = Q\Big(\Big|\frac{Z_j - Z_k}{\sigma(Z_j - Z_k)}\Big|; r\Big), \tag{5.2}$$

and the above quantiles are all unique.

For any random variable $X \in \mathbb{R}$, Y is said to be the normalized version of X if $Y = (X - Q(X, 1/2))/\sigma(X)$. Accordingly, we have that (5.1) holds if the normalized versions of $X_j, X_j + X_k$, and $X_j - X_k$ are all identically distributed, and (5.2) holds if the normalized versions of $Z_j, Z_j + Z_k$, and $Z_j - Z_k$ are all identically distributed.

The next theorem shows that (5.1) and (5.2) hold under the pair-elliptical family.

Theorem 5.3. For any pair-elliptically distributed random vector $\boldsymbol{X} \sim PE_d(\boldsymbol{\mu}, \mathbf{S}, \xi)$, we have both $\mathbf{R}^{\mathrm{M};r}$ and $\mathbf{R}^{\mathrm{Q};r}$ are proportional to $\mathbf{S}$. In particular, when $\mathbb{E}\xi^2 < \infty$, we have both $\mathbf{R}^{\mathrm{M};r}$ and $\mathbf{R}^{\mathrm{Q};r}$ are proportional to the covariance matrix $\mathrm{Cov}(\boldsymbol{X})$ and

$$c^{\mathrm{M};r} = \Big(Q\Big(X_0; \frac{1+r}{2}\Big)\Big)^2 \quad \text{and} \quad c^{\mathrm{Q};r} = 2\Big(Q\Big(Z_0; \frac{1+r}{2}\Big)\Big)^2, \tag{5.3}$$

where X_0 and Z_0 are the normalized versions of X_1 and Z_1.

Remark 5.4. Theorem 5.3 shows that, under the pair-elliptical family, $\mathbf{R}^{\mathrm{M};r}$ and $\mathbf{R}^{\mathrm{Q};r}$ are both proportional to $\mathrm{Cov}(\boldsymbol{X})$ when the covariance exists. Of note, by Theorems 5.1 and 5.2, $\mathbf{R}^{\mathrm{M};r}$ or $\mathbf{R}^{\mathrm{Q};r}$ is proportional to $\mathrm{Cov}(\boldsymbol{X})$ as long as (5.1) or (5.2) holds, and therefore can be applied to study potentially much larger family than the pair-elliptical.

5.2 Theoretical Properties of gMAD and gQNE

This section studies the estimation accuracy for the proposed scatter matrix estimators $\widehat{\mathbf{R}}^{\mathrm{M};r}$ and $\widehat{\mathbf{R}}^{\mathrm{Q};r}$. We show that the proposed methods are capable of handling heavy-tailed distributions, and shed light towards robust alternatives to many multivariate methods in high dimensions.

Before proceeding to the main results, we first introduce some extra notation. For any random vector $\boldsymbol{X} = (X_1, \ldots, X_d)^T$ and any $j \neq k \in \{1, \ldots, d\}$, we denote $F_{1;j}, \bar{F}_{1;j}, F_{2;j,k}, \bar{F}_{2;j,k}, F_{3;j,k}$, and $\bar{F}_{3;j,k}$ to be the distribution functions of $X_j, |X_j - Q(X_j; 1/2)|, X_j + X_k, |X_j + X_k - Q(X_j + X_k; 1/2)|, X_j - X_k$, and $|X_j - X_k - Q(X_j - X_k; 1/2)|$. We suppose that, for some constants κ_1 and η_1 that might scale with n, the following assumption holds:

$$
\begin{aligned}
\text{(A1).}\quad & \min_{\{j, |y - Q(F_{1;j}; 1/2)| < \kappa_1\}} \frac{d}{dy} F_{1;j}(y) \geq \eta_1, && \min_{\{j, |y - Q(\bar{F}_{1;j}; r)| < \kappa_1\}} \frac{d}{dy} \bar{F}_{1;j}(y) \geq \eta_1, \\
& \min_{\{j \neq k, |y - Q(F_{2;j,k}; 1/2)| < \kappa_1\}} \frac{d}{dy} F_{2;j,k}(y) \geq \eta_1, && \min_{\{j \neq k, |y - Q(\bar{F}_{2;j,k}; r)| < \kappa_1\}} \frac{d}{dy} \bar{F}_{2;j,k}(y) \geq \eta_1, \\
& \min_{\{j \neq k, |y - Q(F_{3;j,k}; 1/2)| < \kappa_1\}} \frac{d}{dy} F_{3;j,k}(y) \geq \eta_1, && \min_{\{j \neq k, |y - Q(\bar{F}_{3;j,k}; r)| < \kappa_1\}} \frac{d}{dy} \bar{F}_{3;j,k}(y) \geq \eta_1,
\end{aligned}
$$

where for the random variable X with distribution function F, we denote $Q(F; r) := Q(X; r)$. Assumption (A1) requires that the density functions do not degenerate around median or the r-th quantiles. It is easy to check that Assumption (A1) is satisfied when we choose $\eta_1^{-1} = O(\sqrt{\|\boldsymbol{\Sigma}\|_{\max}})$ for Gaussian distribution. Based on Assumption (A1), we have the following theorem, characterizing the estimation accuracy of the gMAD estimator.

Theorem 5.5 (gMAD concentration)**.** Suppose that Assumption (A1) holds and κ_1 is lower bounded by a positive absolute constant. Then we have, for n large enough, with probability no smaller than $1 - 24\alpha^2$,

$$
\begin{aligned}
&\|\widehat{\mathbf{R}}^{\mathrm{M};r} - \mathbf{R}^{\mathrm{M};r}\|_{\max} \leq \\
&\quad \max\Big\{ \frac{6}{\eta_1^2} \Big(\sqrt{\frac{\log d + \log(1/\alpha)}{n}} + \frac{1}{n} \Big)^2, \frac{4\sqrt{\|\mathbf{R}^{\mathrm{M};r}\|_{\max}}}{\eta_1} \Big(\sqrt{\frac{\log d + \log(1/\alpha)}{n}} + \frac{1}{n} \Big) \Big\}.
\end{aligned}
$$

In particular, when $\boldsymbol{X}$ is pair-elliptically distributed with the covariance matrix $\boldsymbol{\Sigma}$ existing, we have, with probability no smaller than $1 - 24\alpha^2$,

$$
\begin{aligned}
&\|\widehat{\mathbf{R}}^{\mathrm{M};r} - c^{\mathrm{M};r}\boldsymbol{\Sigma}\|_{\max} \leq \\
&\quad \max\Big\{ \frac{6}{\eta_1^2} \Big(\sqrt{\frac{\log d + \log(1/\alpha)}{n}} + \frac{1}{n} \Big)^2, \frac{4\sqrt{\|c^{\mathrm{M};r}\boldsymbol{\Sigma}\|_{\max}}}{\eta_1} \Big(\sqrt{\frac{\log d + \log(1/\alpha)}{n}} + \frac{1}{n} \Big) \Big\}.
\end{aligned}
$$

Theorem 5.5 shows that, when $\kappa_1, \eta_1, \|\boldsymbol{\Sigma}\|_{\max}$, and $c^{\mathrm{M};r}$ are upper and lower bounded by positive absolute constants, the convergence rate of $\widehat{\mathbf{R}}^{\mathrm{M};r}$ with regard to the $\|\cdot\|_{\max}$ is $O_P(\sqrt{\log d/n})$. This

is comparable to the existing results under subgaussian settings (See, for example, Theorem 1 in Cai and Zhou (2012a) and the discussion therein).

We then proceed to quantify the estimation accuracy of the gQNE estimator $\widehat{\mathbf{R}}^{\mathrm{Q};r}$. Let $\widetilde{\boldsymbol{X}} = (\widetilde{X}_1, \ldots, \widetilde{X}_d)^T$ be an independent copy of $\boldsymbol{X}$. For any $j \neq k \in \{1, \ldots, d\}$, let $G_{1;j}, G_{2;j,k}$, and $G_{3;j,k}$ be the distribution functions of $|X_j - \widetilde{X}_j|, |X_j + X_k - (\widetilde{X}_j + \widetilde{X}_k)|$, and $|X_j - X_k - (\widetilde{X}_j - \widetilde{X}_k)|$. Suppose that for some constants κ_2 and η_2 that might scale with n, the following assumption holds:

$$\text{(A2).} \quad \min_{\{j, |y - Q(G_{1;j};r)| < \kappa_2\}} \frac{d}{dy} G_{1;j}(y) \geq \eta_2, \quad \min_{\{j \neq k, |y - Q(G_{2;j,k};r)| < \kappa_2\}} \frac{d}{dy} G_{2;j,k}(y) \geq \eta_2,$$
$$\min_{\{j \neq k, |y - Q(G_{3;j,k};r)| < \kappa_2\}} \frac{d}{dy} G_{3;j,k}(y) \geq \eta_2.$$

Provided that Assumption (A2) holds, we have the following theorem. It gives the rate of convergence for $\widehat{\mathbf{R}}^{\mathrm{Q};r}$ with regard to the element-wise supremum norm.

Theorem 5.6 (gQNE concentration)**.** Suppose that Assumption (A2) holds and κ_2 is lower bounded by a positive absolute constant. Then we have, for n large enough, with probability no smaller than $1 - 8\alpha$,

$$||\widehat{\mathbf{R}}^{\mathrm{Q};r} - \mathbf{R}^{\mathrm{Q};r}||_{\max} \leq$$
$$\max\left\{\frac{2}{\eta_2^2}\left(\sqrt{\frac{2\log d + \log(1/\alpha)}{n}} + \frac{1}{n}\right)^2, \frac{2\sqrt{||\mathbf{R}^{\mathrm{Q};r}||_{\max}}}{\eta_2}\left(\sqrt{\frac{2\log d + \log(1/\alpha)}{n}} + \frac{1}{n}\right)\right\}.$$

In particular, when $\boldsymbol{X}$ is pair-elliptically distributed with the covariance matrix $\boldsymbol{\Sigma}$ existing, we have, with probability no smaller than $1 - 8\alpha$,

$$||\widehat{\mathbf{R}}^{\mathrm{Q};r} - c^{\mathrm{Q};r}\boldsymbol{\Sigma}||_{\max} \leq$$
$$\max\left\{\frac{2}{\eta_2^2}\left(\sqrt{\frac{2\log d + \log(1/\alpha)}{n}} + \frac{1}{n}\right)^2, \frac{2\sqrt{||c^{\mathrm{Q};r}\boldsymbol{\Sigma}||_{\max}}}{\eta_2}\left(\sqrt{\frac{2\log d + \log(1/\alpha)}{n}} + \frac{1}{n}\right)\right\}.$$

Similar to Theorem 5.5, when $\kappa_2, \eta_2, ||\boldsymbol{\Sigma}||_{\max}$, and $c^{\mathrm{Q};r}$ are upper and lower bounded by positive absolute constants, the convergence rate of $\widehat{\mathbf{R}}^{\mathrm{Q};r}$ is $O_P(\sqrt{\log d/n})$. Theorems 5.5 and 5.6 imply that, under the pair-elliptical family, the quantile-based estimators $\widehat{\mathbf{R}}^{\mathrm{M};r}$ and $\widehat{\mathbf{R}}^{\mathrm{Q};r}$ can be good alternatives to the sample covariance matrix.

Remark 5.7. Consider the Gaussian distribution with the diagonal values of $\boldsymbol{\Sigma}$ lower bounded by an absolute constant. Then, for any fixed $r \in (0, 1)$ and lower bounded $\kappa_i, i = 1, 2$, Assumption (A1) and (A2) are satisfied with $\eta_1^{-1}, \eta_2^{-1} = O(\sqrt{||\boldsymbol{\Sigma}||_{\max}})$. This implies that

$$||\widehat{\mathbf{R}}^{\mathrm{M};r} - c^{\mathrm{M};r}\boldsymbol{\Sigma}||_{\max} = O_P\left(||\boldsymbol{\Sigma}||_{\max}\sqrt{\log d/n}\right) \quad \text{and} \quad ||\widehat{\mathbf{R}}^{\mathrm{Q};r} - c^{\mathrm{Q};r}\boldsymbol{\Sigma}||_{\max} = O_P\left(||\boldsymbol{\Sigma}||_{\max}\sqrt{\log d/n}\right).$$

Let $\widetilde{\mathbf{R}}^{\mathrm{M};r}$ and $\widetilde{\mathbf{R}}^{\mathrm{Q};r}$ be the solutions to (4.4). According to Lemma 4.1, we can also establish the concentration for $\widetilde{\mathbf{R}}^{\mathrm{M};r}$ and $\widetilde{\mathbf{R}}^{\mathrm{Q};r}$.

Corollary 5.8. Under Assumptions (A1) and (A2), we have with probability no smaller than $1-24\alpha^2$,

$$||\widetilde{\mathbf{R}}^{\mathrm{M};r}-\mathbf{R}^{\mathrm{M};r}||_{\max}\leq \max\Big\{\frac{3}{\eta_1^2}\Big(\sqrt{\frac{\log d+\log(1/\alpha)}{n}}+\frac{1}{n}\Big)^2,\frac{2\sqrt{||\mathbf{R}^{\mathrm{M};r}||_{\max}}}{\eta_1}\Big(\sqrt{\frac{\log d+\log(1/\alpha)}{n}}+\frac{1}{n}\Big)\Big\};$$

and with probability no smaller than $1-8\alpha$,

$$||\widetilde{\mathbf{R}}^{\mathrm{Q};r}-\mathbf{R}^{\mathrm{Q};r}||_{\max}\leq \max\Big\{\frac{1}{\eta_2^2}\Big(\sqrt{\frac{2\log d+\log(1/\alpha)}{n}}+\frac{1}{n}\Big)^2,\frac{\sqrt{||\mathbf{R}^{\mathrm{Q};r}||_{\max}}}{\eta_2}\Big(\sqrt{\frac{2\log d+\log(1/\alpha)}{n}}+\frac{1}{n}\Big)\Big\}.$$

6 Selection of the Parameter r

Theorems 5.5 and 5.6 show that the estimation accuracy of gMAD and gQNE estimators depends on the selection of the parameter r. In particular, the estimation error in estimating $\mathbf{R}^{\mathrm{M};r}$ and $\mathbf{R}^{\mathrm{Q};r}$ is related to η_1,η_2 and $c^{\mathrm{M};r},c^{\mathrm{Q};r}$. On the other hand, r determines the breakdown points of $\mathbf{R}^{\mathrm{M};r}$ and $\mathbf{R}^{\mathrm{Q};r}$. Accordingly, the parameter r reflects the tradeoff between efficiency and robustness.

This section focuses on selecting the parameter r. The idea is to explore the parameter r that makes the corresponding estimator attain the highest statistical efficiency. Using Theorems 5.5 and 5.6, we have $||\widehat{\mathbf{R}}^{\mathrm{M};r}/c^{\mathrm{M};r}-\mathbf{\Sigma}||_{\max}$ and $||\widehat{\mathbf{R}}^{\mathrm{Q};r}/c^{\mathrm{Q};r}-\mathbf{\Sigma}||_{\max}$ are small when $\eta_1\sqrt{c^{\mathrm{M};r}}$ and $\eta_2\sqrt{c^{\mathrm{Q};r}}$ are large. Therefore, we aim at finding a parameter r such that the first derivatives of $\{\bar{F}_{1;j},\bar{F}_{2;j,k},\bar{F}_{3;j,k}\}$ or $\{G_{1;j},G_{2;j,k},G_{3;j,k}\}$ in a small interval around r times $\sqrt{c^{\mathrm{M};r}}$ or $\sqrt{c^{\mathrm{Q};r}}$ is the highest.

To this end, we separately estimate the derivatives and the scale parameters $\sqrt{c^{\mathrm{M};r}}$ and $\sqrt{c^{\mathrm{Q};r}}$. First, we estimate the derivatives of $\{\bar{F}_{1;j},\bar{F}_{2;j,k},\bar{F}_{3;j,k}\}$ or $\{G_{1;j},G_{2;j,k},G_{3;j,k}\}$ using the kernel density estimator (Tsybakov, 2009). For example for calculating the derivate of $\bar{F}_{1;j}$, we propose to use the data points:

$$|X_{1j}-\widehat{Q}(\{X_{ij}\}_{i=1}^n,1/2)|,|X_{2j}-\widehat{Q}(\{X_{ij}\}_{i=1}^n,1/2)|,\ldots,|X_{nj}-\widehat{Q}(\{X_{ij}\}_{i=1}^n,1/2)|.$$

After obtaining the density estimators $\{\widehat{\bar{f}}_{1;j},\widehat{\bar{f}}_{2;j,k},\widehat{\bar{f}}_{3;j,k}\}$ or $\{\widehat{g}_{1;j},\widehat{g}_{2;j,k},\widehat{g}_{3;j,k}\}$, we calculate the estimators $\widehat{c}^{\mathrm{M};r}$ and $\widehat{c}^{\mathrm{Q};r}$ of $c^{\mathrm{M};r}$ and $c^{\mathrm{Q};r}$ by comparing the scale of the standard deviation and its robust alternative (either gMAD or gQNE) for any chosen entry. We denote the empirical cumulative densities to be $\{\widehat{\bar{F}}_{1;j},\widehat{\bar{F}}_{2;j,k},\widehat{\bar{F}}_{3;j,k}\}$ and $\{\widehat{G}_{1;j},\widehat{G}_{2;j,k},\widehat{G}_{3;j,k}\}$. We also denote their inverse function as $\{\widehat{\bar{F}}_{1;j}^{-1},\widehat{\bar{F}}_{2;j,k}^{-1},\widehat{\bar{F}}_{3;j,k}^{-1}\}$ and $\{\widehat{G}_{1;j}^{-1},\widehat{G}_{2;j,k}^{-1},\widehat{G}_{3;j,k}^{-1}\}$.

In the end, we define the statistics

$$q^{\mathrm{M};r}=\sqrt{\widehat{c}^{\mathrm{M};r}}\min_{j,k}\Big\{\widehat{\bar{f}}_{1;j}\Big(\widehat{\bar{F}}_{1;j}^{-1}(r)\Big),\widehat{\bar{f}}_{2;j,k}\Big(\widehat{\bar{F}}_{2;j,k}^{-1}(r)\Big),\widehat{\bar{f}}_{3;j,k}\Big(\widehat{\bar{F}}_{3;j,k}^{-1}(r)\Big)\Big\},$$
$$q^{\mathrm{Q};r}=\sqrt{\widehat{c}^{\mathrm{Q};r}}\min_{j,k}\Big\{\widehat{g}_{1;j}(\widehat{G}_{1;j}^{-1}(r)),\widehat{g}_{2;j,k}(\widehat{G}_{2;j,k}^{-1}(r)),\widehat{g}_{3;j,k}(\widehat{G}_{2;j,k}^{-1}(r))\Big\}.$$

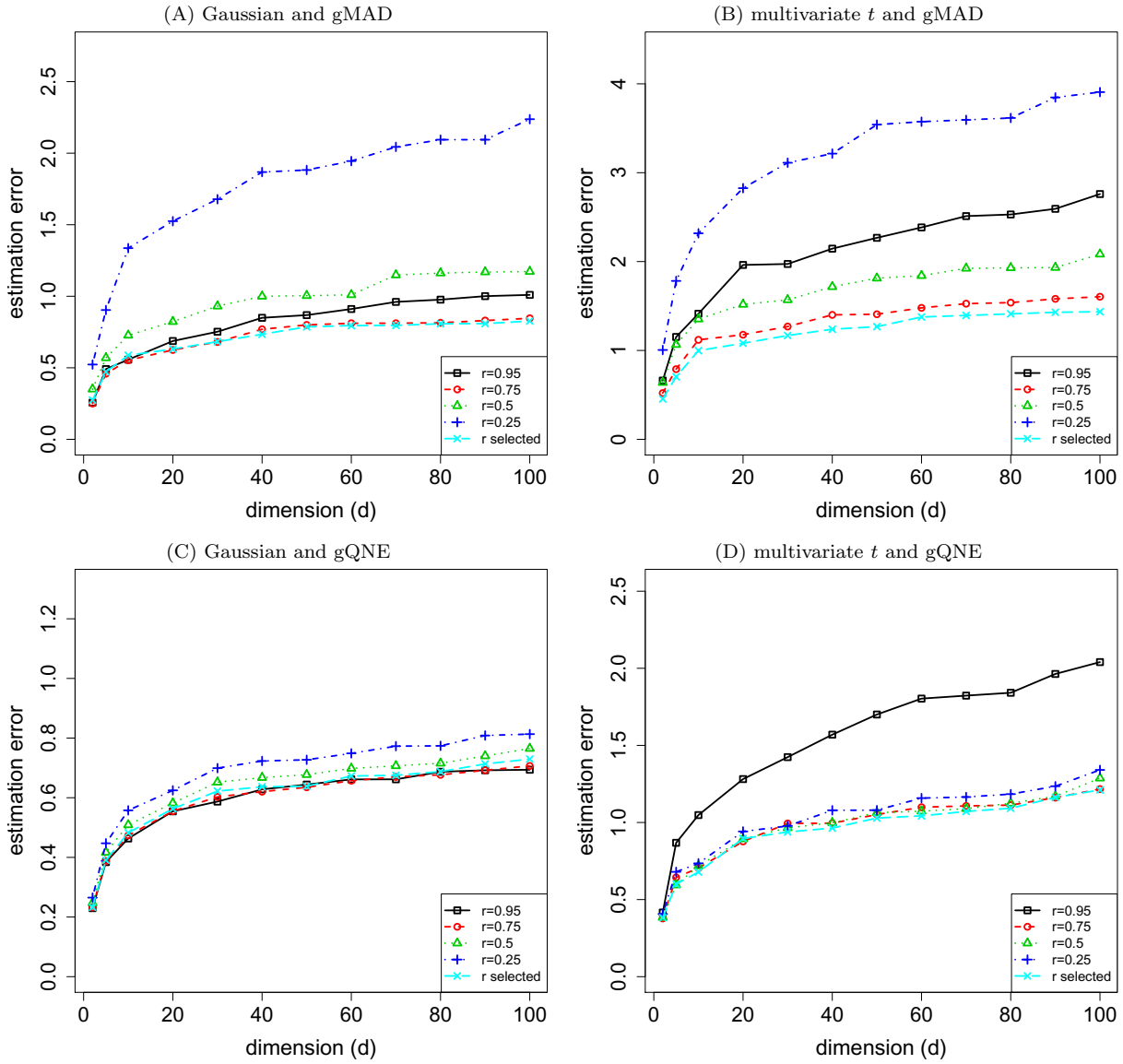

Figure 4: The plot of the averaged estimation accuracy for the gMAD and gQNE estimators (with $r = 0.25$ to 0.95 and a selected r using the procedure described in Section 6). The distances are calculated based on the $||\cdot||_{\max}$ norm and are plotted against the dimension d ranging from 2 to 100. Here the $n = 50$ observations are coming from standard Gaussian or multivariate t distributed data with dimension d. The results are obtained based on 200 repetitions.

The estimators $\widehat{r}^{\mathrm{M}}$ and $\widehat{r}^{\mathrm{Q}}$ are then obtained as:

$$\widehat{r}^{\mathrm{M}} := \mathrm{Truncate}\Big(\operatorname*{arg\,max}_{r\in[0,1]} q^{\mathrm{M};r}; \delta_1, \delta_2\Big), \widehat{r}^{\mathrm{Q}} := \mathrm{Truncate}\Big(\operatorname*{arg\,max}_{r\in[0,1]} q^{\mathrm{Q};r}; \delta_1, \delta_2\Big), \tag{6.1}$$

where δ_1, δ_2 are two pre-defined constants and for any $v \in [0,1]$, we write

$$\mathrm{Truncate}(v; \delta_1, \delta_2) := \begin{cases} \delta_1, & \text{if } v \le \delta_1, \\ v, & \text{if } \delta_1 < v < \delta_2, \\ \delta_2, & \text{if } v \ge \delta_2. \end{cases}$$

Here we control the range of $r \in [\delta_1, \delta_2]$. This is for making the procedure robust and stable. Based on the empirical results, we recommend setting $\delta_1 = 0.25$ and $\delta_2 = 0.75$. In practice, we can also compare different values of $q^{\mathrm{M};r}$ and $q^{\mathrm{Q};r}$, and then choose one that achieve the best balance between estimation accuracy and robustness.

To illustrate the power of the proposed selection procedure, let's consider the following cases: Each time we randomly draw $n = 50$ i.i.d. standard Gaussian or multivariate t (with degrees of freedom 5) distributed data points with dimension d ranging from 2 to 100. We explore the gMAD and gQNE estimators with different quantile parameter r and scale all the scatter matrices such that their population versions are the covariance matrix. We repeat the experiments for 200 times. Figure 4 illustrates the averaged estimation errors with regard to the $||\cdot||_{\max}$ norm. It shows that the procedure using the selected parameter r has the averagely smallest estimation error.

7 Applications

We now turn to various consequences of Theorems 5.5 and 5.6 in conducting multiple multivariate statistical methods in high dimensions. We mainly focus on the methods based on the gQNE estimator, while noting that the analysis for the methods based on the gMAD estimator is similar.

7.1 Sparse Covariance Matrix Estimation

We begin with estimating the sparse covariance matrix in high dimensions. Suppose we have

$$\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_n \overset{i.i.d.}{\sim} \boldsymbol{X} \in \mathbb{R}^d \text{ with covariance matrix } \boldsymbol{\Sigma}.$$

Our target is to estimate the covariance matrix $\boldsymbol{\Sigma}$ itself. Pearson's sample covariance matrix performs poorly in high dimensions (Bai and Yin, 1993; Johnstone, 2001). One common remedy is to assume that the covariance matrix is sparse. This motivates different regularization procedures (Bickel and Levina, 2008b; El Karoui, 2008; Rothman et al., 2009; Cai et al., 2010; Xue et al., 2012; Liu et al., 2013).

We focus on the method in Xue et al. (2012) to illustrate the power of quantile-based statistics. This method directly produces a positive definite estimator and does not require any extra structure for the covariance matrix except for sparsity. The model we focus on is:

$$\mathcal{M}^{\mathrm{COV-Q}}(\boldsymbol{\Sigma}; s) := \{\boldsymbol{X} \in \mathcal{M}^{\mathrm{Q}}(\boldsymbol{\Sigma}) : \lambda_d(\boldsymbol{\Sigma}) > 0 \text{ and } \mathrm{card}(\{(j,k) : \boldsymbol{\Sigma}_{jk} \neq 0, j \neq k\}) \le s\}.$$

Motivated by the above model, we solve the following optimization equation for obtaining the estimator $\widetilde{\mathbf{\Sigma}}$:

$$\widetilde{\mathbf{\Sigma}} = \underset{\mathbf{M}=\mathbf{M}^T}{\arg\min} \frac{1}{2}||\widehat{\mathbf{\Sigma}} - \mathbf{M}||_{\mathrm{F}}^2 + \lambda \sum_{j \neq k} |\mathbf{M}_{j,k}|, \quad \text{subject to } \lambda_d(\mathbf{M}) \geq \epsilon, \tag{7.1}$$

where $\widehat{\mathbf{\Sigma}}$ can be any positive semidefinite matrix based on the quantile-based scatter matrix estimator for approximating the covariance matrix, λ is a tuning parameter inducing sparsity, and ϵ is a small enough constant. Here we focus on using the gQNE estimator for generating the positive semidefinite matrix $\widehat{\mathbf{\Sigma}}$. For estimating the covariance instead of the scatter matrix, we need an extra efficient estimator of the scale parameter $c^{\mathrm{Q};r}$. Here $c^{\mathrm{Q};r}$ is described in Theorem 5.2. When the exact distribution of the pair-elliptical is known, the value $c^{\mathrm{Q};r}$ could be theoretically calculated using (5.3). On the other hand, when the exact distribution is unknown, we note that estimating $c^{\mathrm{Q};r}$ is equivalent to estimating the marginal standard deviation for at least one entry X_j of $\boldsymbol{X}$. The proposed procedure then is in four steps:

1. Calculate the positive semidefinite matrix $\widetilde{\mathbf{R}}^{\mathrm{Q};r}$:

$$\widetilde{\mathbf{R}}^{\mathrm{Q};r} := \underset{\mathbf{R} \succeq \mathbf{0}}{\arg\min} ||\mathbf{R} - \widehat{\mathbf{R}}^{\mathrm{Q};r}||_{\max}. \tag{7.2}$$

   Equation (7.2) can be solved by a matrix maximum norm projection algorithm. See Supplementary Material Section D for details.

2. Choose the j-th entry with the lowest empirical fourth centered moments and use the sample standard deviation $\widehat{\sigma}_j$ to estimate $\sigma(X_j)$.

3. Estimate $c^{\mathrm{Q};r}$ by $\widehat{\mathbf{R}}_{jj}^{\mathrm{Q};r}/\widehat{\sigma}_j^2$, and estimate $\mathbf{\Sigma}$ by

$$\widehat{\mathbf{\Sigma}}^{\mathrm{Q};r} = \frac{\widehat{\sigma}_j^2}{\widehat{\mathbf{R}}_{jj}^{\mathrm{Q};r}} \cdot \widetilde{\mathbf{R}}^{\mathrm{Q};r}. \tag{7.3}$$

4. Produce the final sparse matrix estimator $\widetilde{\mathbf{\Sigma}}^{\mathrm{Q};r}$ by plugging $\widehat{\mathbf{\Sigma}}^{\mathrm{Q};r}$ into (7.1):

$$\widetilde{\mathbf{\Sigma}}^{\mathrm{Q};r} = \underset{\mathbf{M}=\mathbf{M}^T}{\arg\min} \frac{1}{2}||\widehat{\mathbf{\Sigma}}^{\mathrm{Q};r} - \mathbf{M}||_{\mathrm{F}}^2 + \lambda \sum_{j \neq k} |\mathbf{M}_{j,k}|, \quad \text{subject to } \lambda_d(\mathbf{M}) \geq \epsilon,$$

   where we prefix ϵ to be 10^{-5} as recommended in Xue et al. (2012).

We then provide the statistical properties of the quantile-based estimator $\widetilde{\mathbf{\Sigma}}^{\mathrm{Q};r}$ as follows.

Corollary 7.1. Suppose that $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ are n independent observations of $\boldsymbol{X} \in \mathcal{M}^{\mathrm{COV-Q}}(\mathbf{\Sigma}; s)$ and the assumptions in Theorem 4.2 hold. We denote

$$\zeta := \max\Big\{\frac{4}{\eta_2^2}\Big(\sqrt{\frac{4\log d + \log(1/\alpha)}{n}} + \frac{1}{n}\Big)^2, \frac{4\sqrt{||c^{\mathrm{Q};r}\mathbf{\Sigma}||_{\max}}}{\eta_2}\Big(\sqrt{\frac{4\log d + \log(1/\alpha)}{n}} + \frac{1}{n}\Big)\Big\},$$

and

$$\zeta_j := \max\Big\{\frac{2}{\eta_2^2}\Big(\sqrt{\frac{\log(1/\alpha)}{n}} + \frac{1}{n}\Big)^2, \frac{2\sqrt{c^{\mathrm{Q};r}\mathbf{\Sigma}_{jj}}}{\eta_2}\Big(\sqrt{\frac{\log(1/\alpha)}{n}} + \frac{1}{n}\Big)\Big\}.$$

(i) We have, with probability no smaller than $1-8\alpha$,

$$||\widetilde{\mathbf{R}}^{\mathrm{Q};r} - \mathbf{R}^{\mathrm{Q};r}||_{\max} \le \zeta. \tag{7.4}$$

(ii) If $\mathbb{E}X_j^4 \le K$ for some absolute constant K, then there exists an absolute constant c_1 only depending on K, such that when n is large enough, with probability no smaller than $1-n^{-2\xi}-12\alpha$,

$$||\widehat{\mathbf{\Sigma}}^{\mathrm{Q};r} - \mathbf{\Sigma}||_{\max} \le \Big(\frac{c_1 n^{-1/2+\xi}}{\mathbf{R}_{jj}^{\mathrm{Q};r} - \zeta_j} + \frac{1}{c^{\mathrm{Q};r}} \cdot \frac{\zeta_j}{\mathbf{R}_{jj}^{\mathrm{Q};r} - \zeta_j}\Big) \cdot (||\mathbf{R}^{\mathrm{Q};r}||_{\max} + \zeta) + \frac{\zeta}{c^{\mathrm{Q};r}}. \tag{7.5}$$

(iii) If λ takes the same value as the right-hand side of (7.5), then with probability no smaller than $1-n^{-2\xi}-12\alpha$, we have

$$||\widetilde{\mathbf{\Sigma}}^{\mathrm{Q};r} - \mathbf{\Sigma}||_{\mathrm{F}} \le 5\lambda(s+d)^{1/2}.$$

In the following, for notation simplicity, we denote

$$\psi(r,j,\xi,\alpha) := \Big(\frac{c_1 n^{-1/2+\xi}}{\mathbf{R}_{jj}^{\mathrm{Q};r} - \zeta_j} + \frac{1}{c^{\mathrm{Q};r}} \cdot \frac{\zeta_j}{\mathbf{R}_{jj}^{\mathrm{Q};r} - \zeta_j}\Big) \cdot (||\mathbf{R}^{\mathrm{Q};r}||_{\max} + \zeta) + \frac{\zeta}{c^{\mathrm{Q};r}}. \tag{7.6}$$

Corollary 7.1 shows that, when $\eta_1, c^{\mathrm{Q};r}, ||\mathbf{\Sigma}||_{\max}$, and $\min_j \mathbf{\Sigma}_{jj}$ are upper and lower bounded by positive absolute constants, via picking $\lambda > \psi(r,j,\xi,\alpha) = O(\sqrt{\log d/n})$, $\widetilde{\mathbf{\Sigma}}^{\mathrm{Q};r}$ approximates $\mathbf{\Sigma}$ in the rate

$$||\widetilde{\mathbf{\Sigma}}^{\mathrm{Q};r} - \mathbf{\Sigma}||_{\mathrm{F}} = O_P\Big(\sqrt{\frac{(s+d)\log d}{n}}\Big),$$

which is the minimax optimal rate shown in Cai and Zhou (2012b) and the parametric rate obtained in Xue et al. (2012).

Remark 7.2. Following the steps introduced in this section, we could similarly calculate $\widetilde{\mathbf{R}}^{\mathrm{Q};r}$, $\widehat{\mathbf{\Sigma}}^{\mathrm{Q};r}$, and $\widetilde{\mathbf{\Sigma}}^{\mathrm{Q};r}$ for the gMAD estimator $\widehat{\mathbf{R}}^{\mathrm{Q};r}$.

Remark 7.3. For handling heavy tailed data, the quantile-based and rank-based methods are intrinsically different. The rank-based estimator first calculates the correlation matrix, then estimate the d marginal variances via employing Catoni's M-estimator. However, for the quantile-based estimator, as (7.3) shows, we only need to estimate one marginal variance. Moreover, the tuning in Catoni M-estimator is unnecessary.

7.2 Inverse Covariance Matrix Estimation

This section studies estimating inverse covariance matrix. Suppose $\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_n$ are i.i.d. copies of $\boldsymbol{X} \in \mathbb{R}^d$ with covariance matrix $\mathbf{\Sigma}$. We are interested in estimating the precision matrix $\mathbf{\Theta} := \mathbf{\Sigma}^{-1}$. Precision matrix is closely related to graphical models and has been widely studied in the literature: Meinshausen and Bühlmann (2006) propose a neighborhood pursuit method; Yuan and Lin (2007), Friedman et al. (2008), Banerjee et al. (2008), and Lam and Fan (2009) apply penalized likelihood methods to obtain the estimators; Yuan (2010) and Cai et al. (2011) propose graphical Dantzig selector and CLIME estimator; Xue and Zou (2012) and Liu et al. (2012a,b) generalize the Gaussian graphical model to the nonparanormal distribution, and Liu et al. (2012b) further generalize it to the transelliptical distribution.

In this section, we plug the covariance estimator $\widehat{\mathbf{\Sigma}}^{\mathrm{Q};r}$ in (7.3) into the formulation of CLIME proposed by Cai et al. (2011) and obtain the precision matrix estimator:

$$\widehat{\mathbf{\Theta}}^{\mathrm{Q};r} = \underset{\mathbf{\Theta}\in\mathbb{R}^{d\times d}}{\arg\max} \sum_{j,k} |\mathbf{\Theta}_{jk}|, \quad \text{subject to } \|\widehat{\mathbf{\Sigma}}^{\mathrm{Q};r}\mathbf{\Theta} - \mathbf{I}_d\|_{\max} \le \lambda. \tag{7.7}$$

The CLIME estimator in (7.7) can be reformulated as d linear programming (Cai et al., 2011). Let $\mathbf{A}$ be a symmetric matrix. For $q \in [0,1)$, we define $\|\mathbf{A}\|_{q,\infty} := \max_i \sum_j |\mathbf{A}_{ij}|^q$. Considering the set

$$\mathcal{S}_d(q,s,M) := \{\mathbf{\Theta} : \|\mathbf{\Theta}\|_{1,\infty} \le M \text{ and } \|\mathbf{\Theta}\|_{q,\infty} \le s\},$$

the following result gives the estimation accuracy of the precision matrix.

Corollary 7.4. Assume that $\mathbf{\Theta} = \mathbf{\Sigma}^{-1} \in \mathcal{S}_d(q,s,M)$ for some $q \in [0,1)$ and the tuning parameter in (7.7) satisfies

$$\lambda \ge \|\mathbf{\Theta}\|_{1,\infty}\psi(r,j,\xi,\alpha),$$

where $\psi(r,j,\xi,\alpha)$ is defined in (7.6). Then there exist constants C_1, C_2 such that

$$\|\widehat{\mathbf{\Theta}}^{\mathrm{Q};r} - \mathbf{\Theta}\|_2 \le C_1 s\lambda^{1-q}, \|\widehat{\mathbf{\Theta}}^{\mathrm{Q};r} - \mathbf{\Theta}\|_{\max} \le \|\mathbf{\Theta}\|_{1,\infty}\lambda \text{ and } \|\widehat{\mathbf{\Theta}}^{\mathrm{Q};r} - \mathbf{\Theta}\|_{\mathrm{F}}^2 \le C_2 sd\lambda^{2-q}.$$

If $\eta_1, c^{\mathrm{Q};r}, ||\mathbf{\Sigma}||_{\max}$, and $\min_j \mathbf{\Sigma}_{jj}$ are all upper and lower bounded by absolute constants, we choose $\lambda = O(\sqrt{\log d/n})$, and Corollary 7.4 gives us the rate

$$\|\widehat{\mathbf{\Theta}}^{\mathrm{Q};r} - \mathbf{\Theta}\|_2 = O_P\Big(s\Big(\frac{\log d}{n}\Big)^{\frac{1-q}{2}}\Big), \quad \|\widehat{\mathbf{\Theta}}^{\mathrm{Q};r} - \mathbf{\Theta}\|_{\max} = O_P\Big(\sqrt{\frac{\log d}{n}}\Big),$$

and

$$\frac{1}{d}\|\widehat{\mathbf{\Theta}}^{\mathrm{Q};r} - \mathbf{\Theta}\|_{\mathrm{F}}^2 \le O_P\Big(s\Big(\frac{\log d}{n}\Big)^{\frac{2-q}{2}}\Big).$$

According to the minimax estimation rate of precision matrix estimators established in Cai et al. (2014), the estimation rates in terms of all the above three matrix norms are optimal.

7.3 Sparse Principal Component Analysis

In this section we consider conducting principal component analysis (PCA) and sparse PCA. Recall that in (sparse) principal component analysis, we have

$$\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_n \overset{i.i.d.}{\sim} \boldsymbol{X} \in \mathbb{R}^d \text{ with covariance matrix } \mathbf{\Sigma},$$

and our target is to estimate the eigenspace spanned by the m leading eigenvectors $\boldsymbol{u}_1, \ldots, \boldsymbol{u}_m$ of $\mathbf{\Sigma}$. The conventional PCA uses the leading eigenvectors of the sample covariance matrix for estimation. In high dimensions where d can be much larger than n, a sparsity constraint on $\boldsymbol{u}_1, \ldots, \boldsymbol{u}_m$ is sometimes recommended (Johnstone and Lu, 2009), motivating a series of methods referred to as sparse PCA.

In this section, let $\mathbf{U}_m := (\boldsymbol{u}_1, \ldots, \boldsymbol{u}_m) \in \mathbb{R}^{d\times m}$ represent the combination of eigenvectors of interest. We aim to estimate the projection matrix $\mathbf{\Pi}_m := \mathbf{U}_m\mathbf{U}_m^T$. The model we focus on is:

$$\mathcal{M}^{\mathrm{PCA-Q}}(\mathbf{\Sigma};s,m) := \Big\{\boldsymbol{X} \in \mathcal{M}^{\mathrm{Q}}(\mathbf{\Sigma}) : \sum_{j=1}^{d} I\Big(\sum_{k=1}^{m} \boldsymbol{u}_{kj}^2 \ne 0\Big) \le s, \lambda_m(\mathbf{\Sigma}) - \lambda_{m+1}(\mathbf{\Sigma}) > 0\Big\},$$

where $\boldsymbol{u}_{kj}$ is the j-th entry of $\boldsymbol{u}_k$ and $\lambda_m(\boldsymbol{\Sigma})$ is the m-th largest eigenvalue of $\boldsymbol{\Sigma}$. Motivated from the above model, via exploiting the gQNE estimator, we propose the quantile-based (sparse) PCA estimators (Q-PCA) as the optimum to the following Fantope projection problem (Vu et al., 2013):

$$\widehat{\boldsymbol{\Pi}}_m^{\mathrm{Q};r} = \underset{\boldsymbol{\Pi}\in\mathbb{R}^{d\times d}}{\arg\max}\, \mathrm{Tr}(\boldsymbol{\Pi}^T\widehat{\mathbf{R}}^{\mathrm{Q};r}) - \lambda\|\boldsymbol{\Pi}\|_{1,1}, \quad \text{subject to } 0 \preceq \boldsymbol{\Pi} \preceq \mathbf{I}_d \text{ and } \mathrm{Tr}(\boldsymbol{\Pi}) = m, \tag{7.8}$$

where $\|\boldsymbol{\Pi}\|_{1,1} = \sum_{1\le j,k\le d}|\boldsymbol{\Pi}_{jk}|$ and $\mathbf{A} \preceq \mathbf{B}$ implies $\mathbf{B}-\mathbf{A}$ is a positive semidefinite matrix. Intrinsically $\widehat{\boldsymbol{\Pi}}_m^{\mathrm{Q};r}$ is the estimator of the eigenspace spanned by the m leading eigenvectors of $\mathbf{R}^{\mathrm{Q};r}$. Because the scatter matrix shares the same eigenspace with the covariance matrix, under the model $\mathcal{M}^{\text{PCA-Q}}(\boldsymbol{\Sigma}; s, m)$, $\widehat{\boldsymbol{\Pi}}_m^{\mathrm{Q};r}$ is also an estimator of $\boldsymbol{\Pi}_m$. We then have the following corollary, stating that the Q-PCA estimator achieves the parametric rate of convergence in estimating $\boldsymbol{\Pi}_m$ under the model $\mathcal{M}^{\text{PCA-Q}}(\boldsymbol{\Sigma}; s, m)$.

Corollary 7.5. Suppose that $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ are n independent observations of $\boldsymbol{X} \in \mathcal{M}^{\mathrm{PCA-Q}}(\boldsymbol{\Sigma}; s, m)$ and the assumptions in Theorem 5.6 hold. If the tuning parameter in (7.8) satisfies $\lambda \ge \psi(r, j, \xi, \alpha)$ (reminding that $\psi(r, j, \xi, \alpha)$ is defined in (7.6)), we have

$$\|\widehat{\boldsymbol{\Pi}}_m^{\mathrm{Q};r} - \boldsymbol{\Pi}_m\|_{\mathrm{F}} \le \frac{4s\lambda}{\lambda_m(\boldsymbol{\Sigma}) - \lambda_{m+1}(\boldsymbol{\Sigma})}.$$

Corollary 7.5 shows that, under appropriate conditions, the convergence rate of the Q-PCA estimator is $O_P(s\sqrt{\log d/n})$, which is the parametric rate in Vu et al. (2013).

7.4 Discriminant Analysis

In this section, we consider the linear discriminant analysis for conducting high dimensional classification (Guo et al., 2005; Fan and Fan, 2008; Shao et al., 2011; Cai and Liu, 2011; Fan et al., 2012; Han et al., 2013). Let data points $(\boldsymbol{X}_1, Y_1), \ldots, (\boldsymbol{X}_n, Y_n)$ be independently drawn from a joint distribution of $(\boldsymbol{X}, Y)$, where $\boldsymbol{X} \in \mathbb{R}^d$ and $Y \in \{1, 2\}$ is the binary label. We denote $I_1 = \{i : Y_i = 1\}$, $I_2 = \{i : Y_i = 2\}$, and $n_1 = |I_1|, n_2 = |I_2|$. Define $\pi = \mathbb{P}(Y = 1)$, $\boldsymbol{\mu}_1 = \mathbb{E}(\boldsymbol{X} \mid Y = 1)$, $\boldsymbol{\mu}_2 = \mathbb{E}(\boldsymbol{X} \mid Y = 2)$, $\boldsymbol{\Sigma} = \mathrm{Cov}(\boldsymbol{X} \mid Y = 1) = \mathrm{Cov}(\boldsymbol{X} \mid Y = 2)$. If the classifier is defined as $h(\mathbf{x}) = I(f(\mathbf{x}) < c) + 1$ for some function f and constant c, we measure the quality of classification by employing the Rayleigh quotient (Fan et al., 2013a):

$$\mathrm{Rq}(f) = \frac{\mathrm{Var}\{\mathbb{E}[f(\boldsymbol{X}) \mid Y]\}}{\mathrm{Var}\{f(\boldsymbol{X}) - \mathbb{E}[f(\boldsymbol{X}) \mid Y]\}}.$$

For the linear functions $f(\mathbf{x}) = \boldsymbol{\beta}^T\mathbf{x} + c$, the Rayleigh quotient has the formulation

$$\mathrm{Rq}(\boldsymbol{\beta}) = \pi(1-\pi)\frac{[\boldsymbol{\beta}^T(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)]^2}{\boldsymbol{\beta}^T\boldsymbol{\Sigma}\boldsymbol{\beta}}.$$

The Rayleigh quotient is minimized when $\boldsymbol{\beta} = \boldsymbol{\beta}^* = \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)$. When $\boldsymbol{X}|(Y = 1)$ and $\boldsymbol{X}|(Y = 2)$ are multivariate Gaussian distribution, it matches the Fisher's linear discriminant rule $h_{\mathrm{F}}(\mathbf{x}) = I(\mathbf{x}^T\boldsymbol{\beta}^* + c^*) + 1$, where $c^* = (\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2)^T\boldsymbol{\beta}^*/2$. In order to estimate $\boldsymbol{\beta}^*$ and c^*, we apply the estimator proposed in Cai and Liu (2011):

$$\widehat{\boldsymbol{\beta}}^{\mathrm{Q};r} = \underset{\boldsymbol{\beta}\in\mathbb{R}^d}{\arg\max}\, \|\boldsymbol{\beta}\|_1, \quad \text{subject to } \|\widehat{\boldsymbol{\Sigma}}^{\mathrm{Q};r}\boldsymbol{\beta} - (\widehat{\boldsymbol{\mu}}_1 - \widehat{\boldsymbol{\mu}}_2)\|_{\max} \le \lambda, \tag{7.9}$$

where $\widehat{\boldsymbol{\mu}}_1, \widehat{\boldsymbol{\mu}}_2$ are some estimators of $\boldsymbol{\mu}_1, \boldsymbol{\mu}_2$.

In the following, we suggest two kinds of estimators of the mean vectors. For simplicity, we only describe the estimator for $\boldsymbol{\mu}_1$ and similar procedures can also be applied to estimate $\boldsymbol{\mu}_2$. Suppose that $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_{n_1}$ correspond to $Y = 1$. The first estimator for $\boldsymbol{\mu}_1$ is the sample median $\widehat{\mu}_M = (\widehat{\mu}_{M1}, \ldots, \widehat{\mu}_{Md})^T$ where

$$\widehat{\mu}_{M_j} = \text{median}(\{\boldsymbol{X}_{1j}, \ldots, \boldsymbol{X}_{n_1 j}\}).$$

Due to the Hoeffding's inequality, we can derive that there exists some constant c_m that $\mathbb{P}(|\widehat{\mu}_{M_j} - \mu_j| > c_m\sqrt{\log(1/\delta)/n}) \leq \delta$ for any $j = 1, \ldots, d$.

Catoni (2012) proposes an alternative M-estimator for $\boldsymbol{\mu}_1$. Considering a strictly increasing function h such that $-\log(1 - y + y^2/2) \leq h(y) \leq \log(1 + y + y^2/2)$ and some $\delta \in (0, 1)$ satisfying $n > 2\log(1/\delta)$, $v \geq \max\{\sigma_1^2, \ldots, \sigma_d^2\}$, we define

$$\alpha_\delta^2 = \frac{2\log(1/\delta)}{n_1\big(v + \frac{2v\log(1/\delta)}{n - 2\log(1/\delta)}\big)}.$$

The estimator $\widehat{\boldsymbol{\mu}}_C = (\widehat{\mu}_{C1}, \ldots, \widehat{\mu}_{Cd})^T$ is obtained by solving the equation:

$$\sum_{i=1}^{n_1} h(\alpha_\delta(\mathbf{x}_{ij} - \widehat{\mu}_{Cj})) = 0.$$

It is shown in Catoni (2012) that there exists some constant C such that with probability at least $1 - (n \vee d)^{-1}$, $\mathbb{P}\big(\|\widehat{\mu}_{Cj} - \mu_{Cj}\|^2 \geq \frac{2v\log(1/\delta)}{n - 2\log(1/\delta)}\big) \leq \delta$ for all $j = 1, \ldots, d$. By union bound, for both estimators we have

$$\|\widehat{\boldsymbol{\mu}}_M - \boldsymbol{\mu}_1\|_\infty \vee \|\widehat{\boldsymbol{\mu}}_C - \boldsymbol{\mu}_1\|_\infty = O_P\Big(\sqrt{\frac{\log d}{n}}\Big). \tag{7.10}$$

Therefore, we have the following result on how well the Rayleigh quotient of estimated classifier can approximate the optimal one.

Corollary 7.6. We assume that $n_1 \asymp n_2$, $\Delta_d = (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^T\boldsymbol{\beta}^* \geq M$ for some constant $M > 0$. If we choose the tuning parameter in (7.9) that $\lambda \geq \|\widehat{\boldsymbol{\mu}}_1 - \widehat{\boldsymbol{\mu}}_2 - \boldsymbol{\mu}_1 + \boldsymbol{\mu}_2\|_\infty + \psi(r, j, \xi, \alpha)\|\boldsymbol{\beta}^*\|_1$, we have

$$\frac{\text{Rq}(\widehat{\boldsymbol{\beta}}^{\text{Q;r}})}{\text{Rq}(\boldsymbol{\beta}^*)} \geq 1 - 2M^{-1}\|\boldsymbol{\beta}^*\|_1^2\|\psi(r, j, \xi, \alpha) - 20M^{-1}\lambda\|\boldsymbol{\beta}^*\|_1.$$

According to the rate in (7.10), if $\eta_1, c^{\text{Q;}r}, ||\boldsymbol{\Sigma}||_{\max}$, and $\min_j \boldsymbol{\Sigma}_{jj}$ are universally bounded from above and below by positive absolute constants and $\|\boldsymbol{\beta}^*\|_1$ is also upper bounded by a constant independent of d and n, we choose the tuning parameter $\lambda = O(\sqrt{\log d/n})$ and obtain the rate through Corollary 7.6 as

$$\frac{\text{Rq}(\widehat{\boldsymbol{\beta}}^{\text{Q;}r})}{\text{Rq}(\boldsymbol{\beta}^*)} \geq 1 - O\Big(\sqrt{\frac{\log d}{n}}\Big).$$

The above result matches the parametric rate in Cai and Liu (2011).

8 Experiments

This section compares the empirical performance of quantile-based estimators to these based on the Pearson, Kendall's tau, and Spearman's rho covariance estimators for both synthetic and real data. Since for synthetic data analysis, conducting inverse covariance estimation and sparse PCA has reflected the quality of matrix estimation very well, these two methods are the main focus for synthetic data. For the real data, on the other hand, we aim to apply our methods to the classification of genomic data.

8.1 Synthetic Data

This section focuses on conducting inverse covariance matrix estimation and sparse PCA. Let $\{\widehat{\sigma}_1, \ldots, \widehat{\sigma}_d\}$ be the sample standard deviations of $\boldsymbol{X}$. Let $\widehat{\mathbf{S}}^{\mathrm{P}}$, $\widehat{\mathbf{S}}^{\mathrm{K}}$ and $\widehat{\mathbf{S}}^{\mathrm{S}}$ be the Pearson's, Kendall's tau and Spearman's rho sample correlation matrices. We compare $\widehat{\mathbf{\Sigma}}^{\mathrm{Q},r}$ and $\widehat{\mathbf{\Sigma}}^{\mathrm{M},r}$ introduced in Section 7.1[4] to three covariance matrix estimators:

1. Pearson's sample covariance matrix: $\widehat{\mathbf{\Sigma}}^{\mathrm{P}}_{ij} = \widehat{\sigma}_i \widehat{\sigma}_j \widehat{\mathbf{S}}^{\mathrm{P}}_{ij}$, for $i, j = 1, \ldots, d$.;
2. Kendall's tau covariance matrix: $\widehat{\mathbf{\Sigma}}^{\mathrm{K}}_{ij} = \widehat{\sigma}_i \widehat{\sigma}_j \widehat{\mathbf{S}}^{\mathrm{K}}_{ij}$, for $i, j = 1, \ldots, d$.;
3. Spearman's rho covariance matrix: $\widehat{\mathbf{\Sigma}}^{\mathrm{S}}_{ij} = \widehat{\sigma}_i \widehat{\sigma}_j \widehat{\mathbf{S}}^{\mathrm{S}}_{ij}$, for $i, j = 1, \ldots, d$.

Moreover, given a covariance matrix $\mathbf{\Sigma}$, we consider the following three schemes for data generation:

Scheme 1: $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ are i.i.d. copies of $\boldsymbol{X} \sim N_d(\mathbf{0}, \mathbf{\Sigma})$;

Scheme 2: $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ are i.i.d. copies of $\boldsymbol{X}$ following the multivariate t-distribution with the degree of freedom 5 and $\mathrm{Cov}(\boldsymbol{X}) = \mathbf{\Sigma}$;

Scheme 3: $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ are i.i.d. copies of $\boldsymbol{X} \sim N_d(\mathbf{0}, \mathbf{\Sigma})$, but with randomly sampled 10% of data points replaced by i.i.d. copies of the mixture Gaussian $0.5N(3, 3) + 0.5N(-3, 3)$.

In the following, we plug the five covariance matrix estimators into the inverse covariance estimation procedure discussed in Section 7.2 and the sparse PCA procedure in Section 7.3.

8.1.1 Inverse Covariance Matrix Estimation

We consider the numerical performance using the setting in Cai et al. (2011). Let $\mathbf{\Omega} = (\omega_{ij})_{1 \le i,j \le d} := \mathbf{\Sigma}^{-1}$ represent the inverse covariance matrix. We consider $\mathbf{\Omega} = \mathbf{B} + \delta \mathbf{I}_d$. Each off-diagonal entry of $\mathbf{B}$ is independent and of the value 0.5 with probability 0.1 and the value 0 with probability 0.9. The value δ is chosen such that the condition number of $\mathbf{\Omega}$ is d. The diagonal of $\mathbf{\Sigma}$ is renormalized to be ones.

We generate the data under the three schemes with dimensions $d = 90, 120, 200$, sample size $n = 100$, and repetition time 100. We measure the estimation error by Frobenius norm $\|\widehat{\mathbf{\Sigma}} - \mathbf{\Sigma}\|_{\mathrm{F}}$. The numerical results are presented in Table 1.

[4]Here the parameter r is set as in Section 6 with δ_1 and δ_2 to be 0.25 and 0.75.

Table 1: Comparison of the gMAD, gQNE, Pearson's, Kendall's tau and Spearman's rho covariance estimators combined with the CLIME and sparse PCA (SPCA) algorithms. The estimation errors of CLIME estimators are in terms of Frobenius norms. The estimation errors of SPCA estimators are in terms of $\|\widehat{\mathbf{\Pi}}_1 - \mathbf{\Pi}_1\|_{\mathrm{F}}$. The errors are averaged over 100 repetitions with standard deviations in the parentheses.

		Scheme 1			**Scheme 2**			**Scheme 3**		
	d	100	200	500	100	200	500	100	200	500
CLIME	gMAD	5.47	6.29	8.82	7.73	9.06	11.09	**6.83**	7.70	9.48
		(0.22)	(0.28)	(0.60)	(0.64)	(0.80)	(0.92)	(0.24)	(0.27)	(0.30)
	gQNE	**4.75**	5.42	7.53	**7.58**	**8.76**	**10.95**	6.84	**7.59**	**9.41**
		(0.20)	(0.19)	(0.22)	(0.68)	(0.79)	(1.03)	(0.22)	(0.27)	(0.30)
	Pearson	4.75	**5.38**	**7.39**	7.77	8.95	11.37	7.25	8.02	9.76
		(0.22)	(0.22)	(0.26)	(0.82)	(1.05)	(1.45)	(0.12)	(0.15)	(0.12)
	Kendall	5.38	6.35	8.84	8.47	10.01	13.34	7.58	8.80	11.83
		(0.23)	(0.25)	(0.29)	(0.55)	(0.62)	(0.66)	(0.13)	(0.15)	(0.16)
	Spearman	4.86	5.49	7.60	7.99	9.17	11.33	7.18	7.97	9.52
		(0.20)	(0.20)	(0.30)	(0.64)	(0.78)	(1.04)	(0.14)	(0.19)	(0.21)
SPCA	gMAD	0.35	0.41	0.48	0.48	0.53	0.60	**0.42**	0.51	**0.55**
		(0.23)	(0.27)	(0.26)	(0.25)	(0.25)	(0.27)	(0.24)	(0.25)	(0.23)
	gQNE	0.19	0.25	0.34	**0.37**	**0.43**	**0.47**	0.42	**0.48**	**0.55**
		(0.16)	(0.24)	(0.26)	(0.26)	(0.26)	(0.29)	(0.25)	(0.25)	(0.24)
	Pearson	**0.17**	**0.20**	**0.31**	0.78	0.80	0.91	0.80	0.85	0.84
		(0.14)	(0.20)	(0.28)	(0.27)	(0.23)	(0.16)	(0.17)	(0.13)	(0.11)
	Kendall	0.20	0.23	0.32	0.45	0.53	0.58	0.53	0.61	0.69
		(0.15)	(0.20)	(0.28)	(0.30)	(0.28)	(0.30)	(0.23)	(0.22)	(0.21)
	Spearman	0.21	0.25	0.34	0.46	0.54	0.57	0.53	0.61	0.70
		(0.16)	(0.23)	(0.28)	(0.32)	(0.28)	(0.29)	(0.23)	(0.22)	(0.21)

8.1.2 Sparse PCA

We use the setting in Han and Liu (2012b) to investigate the numerical performance of the sparse principal component analysis. We consider the spike covariance $\mathbf{\Sigma} = \lambda_1 v_1 v_1^T + \lambda_2 v_2 v_2^T + \mathbf{I}_d$, where $\lambda_1 = 5$ and $\lambda_2 = 2$, $v_{1j} = 1/\sqrt{10}$ for $j = 1 \le j \le 10$, $v_{1j} = 0$ for $j > 10$ and $v_{2j} = 1/\sqrt{10}$ for $11 \le j \le 20$, $v_{2j} = 0$ otherwise. The data sample size is $n = 50$ for $d = 90, 120, 200$ with repetition 100 times. We measure the difference between the true projection matrix $\mathbf{\Pi}_1 = v_1 v_1^T$ and its estimator $\widehat{\mathbf{\Pi}}_1$ through (7.8) by the Frobenius norm $\|\widehat{\mathbf{\Pi}}_1 - \mathbf{\Pi}_1\|_{\mathrm{F}}$. The results are given in Table 1.

8.1.3 Summary of Synthetic Data Results

From the numerical results summarized in Table 1, the gMAD and gQNE covariance estimators perform better than other estimators for both CLIME and sparse PCA under Scheme 2 and Scheme 3. This implies that gMAD and gQNE have better performance in studying heavy-tailed distributions and contaminated data. This is due to the advantage that only one variance is required to be estimated in (7.3), while the other covariance estimators demand variance estimators for d covariates. Moreover, gQNE outperforms gMAD estimator. This matches the assertion that gQNE is more efficient than gMAD. Under the Scheme 1, since the synthetic data follow Gaussian distribution, the sample covariance estimator is efficient and outperforms the gMAD and gQNE covariance estimators. However, the performance is still comparable even under this scheme.

8.2 Genomic Data Analysis

In this section, we apply the method introduced in Section 7.4 to the large-scale microarray dataset, GPL96 (McCall et al., 2010). The dataset contains 20,263 probes with 8,124 samples belonging to 309 tissue types. The probes correspond to 12,679 genes and we average the probes from same genes. The tissue types include prostate tumor (148 samples), B cell lymphoma (213 samples), and many others. The target of our real data analysis is to compare the performance of classifying the prostate tumor from the B cell lymphoma by using gMDA, gQNE, and other estimators. We only focus on the top 1,000 genes of the largest p-values in performing the marginal two-sample t-test.

First, we consider the goodness of fit test of the pair-elliptical. Our goal is to select genes such that their corresponding data from both tissue types are approximately pair-elliptical separately. Then we apply the linear discriminative classifier in (7.9) to the selected genes for classification. Specifically, let $\{\boldsymbol{X}_i\}_{i=1}^n$ be the dataset corresponding to the prostate tumor. We employ the goodness of fit test proposed in Section 3.3 by calculating the statistic $Z(\{(X_{ij}, X_{ik})\})$ for each pair of $1 \le j \ne k \le d$. The Holm's adjusted p-values $\{\pi_{jk}^H\}_{j,k=1}^d$ are evaluated according to (3.5). We delete any gene j if there exists another gene k such that the adjusted p-value $\pi_{jk}^H < \alpha = 0.05$. All the left ones are the selected genes for the prostate tumor. Same procedures are also applied to the dataset of B cell lymphoma. The final screened genes are the intersection of the selected genes for both categories. In the end, the number of selected genes is 225.

Figure 5 illustrates the results for the above procedure corresponding to the prostate tumor. In detail, Figure 5(A) reports the histogram of the test statistics $Z(\{(X_{ij}, X_{ik})\})$ for all pairs of genes corresponding to the prostate tumor. The curve in the figure is the density of χ^2 distribution with degrees of freedom $[\sqrt{n}] - 1 = 11$. Figure 5(A) shows that the empirical distribution of the statistics is close to the χ^2 distribution. Figure 5(B) illustrates the histogram of the Holm's adjusted p-values for all pairs of genes. Figure 5(B) shows that we do not reject the goodness of fit test of the pair-elliptical for most pairs of genes in the dataset. We select the pair of genes with the largest p-value and denote the pair as (j^*, k^*). Figure 5(C) reports the estimated density function of the t-statistics from the selected pair $\{t((X_{ij^*}, X_{ik^*}))\}_{i=1}^n$ versus the density of the t-distribution with degree of freedom 1. The density function is estimated by kernel density estimator with bandwidth $h = 0.4$. We see that these two distributions are close to each other.

Secondly, we plug the samples of these selected genes to the linear discriminative classifier in (7.9). To compare the classification errors for different covariance estimators, each time we randomly

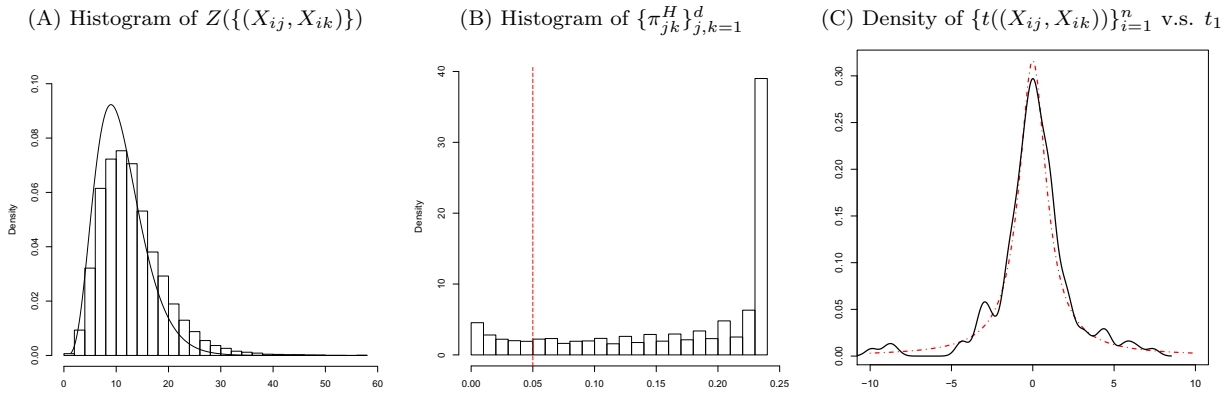

Figure 5: (A) The histogram of $Z(\{(X_{ij}, X_{ik})\})$ for the pairs of genes for the prostate tumor and the curve is the density function of χ^2_{11}; (B) The histogram of the Holm's adjusted p-values for the prostate tumor and the red dashed line is the significance level $\alpha = 5\%$; (C) The estimated density of $\{t((X_{ij^*}, X_{ik^*}))\}^n_{i=1}$ (black solid line) versus the density of t_1 (red dashed line) for the pair of genes with largest p-value of the statistic $Z(\{(X_{ij}, X_{ik})\})$.

Table 2: Means (standard deviations in the parentheses) of classification errors on GPL96 for prostate tumor and B cell lymphoma by applying the gMAD, gQNE, Pearson's, Kendall's tau and Spearman's rho covariance estimators to the linear discriminative classifier with 100 replications.

gMAD	gQNE	Pearson	Kendall	Spearman
0.055	0.059	0.134	0.099	0.098
(0.003)	(0.004)	(0.003)	(0.007)	(0.005)

select 74 samples from the prostate tumor and 74 from the B cell lymphoma as the training dataset. For the training dataset, we divide each category into two parts randomly: One is to derive the classifiers from different covariance estimators by (7.9) and the other is for tuning the parameters of the first part by minimizing its classification error. The rest samples are applied as the testing dataset to calculate the final classification error. The above steps are repeated for 100 times. We summarize the classification errors in Table 2. The errors reported in Table 2 demonstrate that the gMAD and the gQNE estimators significantly outperform the other estimators in the GPL96 dataset. This indicates the power of quantile-based statistics in high dimensions.

9 Discussion

This paper studies estimating large scatter matrices for high dimensional data with possibly very heavy tails. We propose a new distribution family, the pair-elliptical, for modeling such data. The pair-elliptical is more flexible than the elliptical. We also characterize the relation between the pair-elliptical and several popular distribution families in high dimensions. Built on the pair-elliptical family, we advocate using the quantile-based approaches for estimating large scatter and covariance matrices. These procedures are both statistically efficient and robust compared to their Gaussian alternatives.

In the future, it is of interest to investigate the performance of the quantile-based methods when the data are not independent. Studies of moment-based and rank-based approaches under high dimensional dependent data include Fan et al. (2013b) and Han and Liu (2013b). However, their proof techniques cannot be directly applied to analyze the quantile-based estimators.

All the results in this paper are confined in parameter estimation and focus on nonasymptotic analysis. In the future, we are also interested in studying the asymptotic properties of quantile-based estimators under the regime that both d and n go to infinity.

A Supporting Lemmas

Lemma A.1. For any continuous random variable $X \in \mathbb{R}$ with the distribution function F and any n independent observations $X_1, \ldots, X_n$ of X, we have for any $t > 0$,

$$\begin{aligned}&\mathbb{P}(|\widehat{Q}(\{X_i\}; r) - Q(X; r)| \geq t)\\ &\quad \leq \exp\Big(- 2n[F\{F^{-1}(r) + t\} - r - n^{-1}]^2\Big) + \exp\Big(- 2n[r - F\{F^{-1}(r) - t\}]^2\Big),\end{aligned}$$

whenever $F\{F^{-1}(r) + t\} > r + n^{-1}$.

Proof. Let F_n denote the empirical distribution of $X_1, \ldots, X_n$ and $F_n^{-1}(r) := \widehat{Q}(\{X_i\}; r)$. By the definition of $\widehat{Q}(\cdot; \cdot)$ in (4.1), we have for any $\epsilon \in [0, 1]$,

$$\epsilon \leq F_n(F_n^{-1}(\epsilon)) \leq \epsilon + \frac{1}{n}.$$

This implies that

$$\begin{aligned}\mathbb{P}\Big\{\widehat{Q}(\{X_i\}; r) - Q(X; r) \geq t\Big\} &= \mathbb{P}\Big\{F_n^{-1}(r) - F^{-1}(r) \geq t\Big\}\\ &\leq \mathbb{P}\Big[r + \frac{1}{n} \geq F_n\{t + F^{-1}(r)\}\Big]\\ &= \mathbb{P}\Big[\frac{1}{n}\sum_{i=1}^n I\{X_i \leq F^{-1}(r) + t\} \leq r + \frac{1}{n}\Big].\end{aligned}$$

We have

$$\begin{aligned}&\mathbb{P}\Big[\frac{1}{n}\sum_{i=1}^n I\{X_i \leq F^{-1}(r) + t\} \leq r + \frac{1}{n}\Big]\\ =&\mathbb{P}\Big[\frac{1}{n}\sum_{i=1}^n I\{X_i \leq F^{-1}(r) + t\} - F\{F^{-1}(r) + t\} \leq r + \frac{1}{n} - F\{F^{-1}(r) + t\}\Big].\end{aligned}$$

Because $\mathbb{E}I\{X_i \leq F^{-1}(r)+t\} = \mathbb{P}(X \leq F^{-1}(r)+t) = F\{F^{-1}(r)+t\}$ and $I\{X_i \leq F^{-1}(r)+t\} \in [0, 1]$ is a bounded random variable, by Hoeffding's inequality, we have

$$\mathbb{P}\Big\{\widehat{Q}(\{X_i\}; r) - Q(X; r) > t\Big\} \leq \exp\Big(- 2n[F\{F^{-1}(r) + t\} - r - n^{-1}]^2\Big), \tag{A.1}$$

as long as t is large enough such that $F\{F^{-1}(r) + t\} > r + n^{-1}$.

On the other hand, we have

$$\begin{aligned}\mathbb{P}\Big\{\widehat{Q}(\{X_i\};r)-Q(X;r)\le -t\Big\} &= \mathbb{P}\Big\{F_n^{-1}(r)-F^{-1}(r)\le -t\Big\}\\ &\le \mathbb{P}\Big[r\le F_n\{F^{-1}(r)-t\}\Big]\\ &= \mathbb{P}\Big[\frac{1}{n}\sum_{i=1}^{n} I\{X_i\le F^{-1}(r)-t\}\ge r\Big].\end{aligned}$$

Then again, exploiting the Hoeffding's inequality as above, we have

$$\mathbb{P}\Big\{\widehat{Q}(\{X_i\};r)-Q(X;r)\le -t\Big\}\le \exp\Big(-2n[r-F\{F^{-1}(r)-t\}]^2\Big). \tag{A.2}$$

Combining (A.1) and (A.2), we have the desired result. □

Lemma A.2 (gMAD concentration inequality)**.** Letting $X\in\mathbb{R}$ be a one dimensional continuous random variable. We denote F_1 and F_2 to be the distribution functions of X and $|X-Q(X;1/2)|$. Then we have for any $t>0$,

$$\begin{aligned}&\mathbb{P}\Big(|\widehat{\sigma}^{\mathrm{M}}(X;r)-\sigma^{\mathrm{M}}(X;r)|>t\Big)\le 2\exp\Big(-2n\Big[F_1\{F_1^{-1}(1/2)+t/2\}-\frac{1}{2}-\frac{1}{n}\Big]^2\Big)+\\ &\quad 2\exp\Big(-2n\Big[\frac{1}{2}-F_1\{F_1^{-1}(1/2)-t/2\}\Big]^2\Big)+\exp\Big(-2n\Big[F_2\{F_2^{-1}(r)+t/2\}-r-\frac{1}{n}\Big]^2\Big)\\ &\quad +\exp\Big(-2n\Big[r-F_2\{F_2^{-1}(r)-t/2\}\Big]^2\Big),\end{aligned}$$

whenever $F_1\{F_1^{-1}(1/2)+t/2\}-1/2>1/n$ and $F_2\{F_2^{-1}(r)+t/2\}-r>1/n$.

Proof. By definition, we have for any $t>0$,

$$\mathbb{P}\Big(\widehat{\sigma}^{\mathrm{M}}(X;r)-\sigma^{\mathrm{M}}(X;r)>t\Big)=\mathbb{P}\Big(\widehat{Q}\Big[\Big\{\Big|X_i-\widehat{Q}\Big\{\{X_i\};\frac{1}{2}\Big\}\Big|\Big\};r\Big]-Q\Big\{\Big|X-Q\Big(X;\frac{1}{2}\Big)\Big|;r\Big\}>t\Big).$$

We denote $\nu:=Q(X;1/2)$ and $\widehat{\nu}:=\widehat{Q}(\{X_i\};1/2)$. We then have

$$\begin{aligned}\mathbb{P}\Big(\widehat{\sigma}^{\mathrm{M}}(X;r)-\sigma^{\mathrm{M}}(X;r)>t\Big) &= \mathbb{P}\Big\{\widehat{Q}\Big(\{|X_i-\widehat{\nu}|\},r\Big)-Q\Big(|X-\nu|,r\Big)>t\Big\}\\ &\le \mathbb{P}\Big\{\widehat{Q}\Big(\{|X_i-\nu|\},r\Big)+|\widehat{\nu}-\nu|-Q\Big(|X-\nu|,r\Big)>t\Big\}\\ &\le \underbrace{\mathbb{P}\Big\{\widehat{Q}\Big(\{|X_i-\nu|\},r\Big)-Q\Big(|X-\nu|,r\Big)>t/2\Big\}}_{A_1}\\ &\quad +\underbrace{\mathbb{P}(|\widehat{\nu}-\nu|>t/2)}_{B}.\end{aligned}$$

On the other hand, we have

$$\begin{aligned}\mathbb{P}\Big(\widehat{\sigma}^{\mathrm{M}}(X;r)-\sigma^{\mathrm{M}}(X;r)<-t\Big) &= \mathbb{P}\Big\{\widehat{Q}\Big(\{|X_i-\widehat{\nu}|\},r\Big)-Q\Big(|X-\nu|,r\Big)<-t\Big\}\\ &\le \mathbb{P}\Big\{\widehat{Q}\Big(\{|X_i-\nu|\},r\Big)-|\widehat{\nu}-\nu|-Q\Big(|X-\nu|,r\Big)<-t\Big\}\\ &\le \underbrace{\mathbb{P}\Big\{\widehat{Q}\Big(\{|X_i-\nu|\},r\Big)-Q\Big(|X-\nu|,r\Big)<-t/2\Big\}}_{A_2}\\ &\quad+\underbrace{\mathbb{P}(|\widehat{\nu}-\nu|>t/2)}_{B}.\end{aligned}$$

Using Lemma A.1, we have

$$B\le \exp\Big(-2n\Big[F_1\{F_1^{-1}(1/2)+t/2\}-\frac{1}{2}-\frac{1}{n}\Big]^2\Big)+\exp\Big(-2n\Big[\frac{1}{2}-F_1\{F_1^{-1}(1/2)-t/2\}\Big]^2\Big).$$

Moreover, we have

$$\begin{aligned}&A_1+A_2=\mathbb{P}\Big\{\Big|\widehat{Q}\Big(\{|X_i-\nu|\},r\Big)-Q\Big(|X-\nu|,r\Big)\Big|>\frac{t}{2}\Big\}\\ &\le \exp\Big(-2n\Big[F_2\{F_2^{-1}(r)+t/2\}-r-\frac{1}{n}\Big]^2\Big)+\exp\Big(-2n\Big[r-F_2\{F_2^{-1}(r)-t/2\}\Big]^2\Big),\end{aligned}$$

where we remind that F_1 and F_2 represent the distribution functions of X and $|X-Q(X;1/2)|$. Finally, using the fact that

$$\mathbb{P}\Big(\Big|\widehat{\sigma}^{\mathrm{M}}(X;r)-\sigma^{\mathrm{M}}(X;r)\Big|>t\Big)\le A_1+A_2+2B,$$

we have the desired result. □

Lemma A.3 (gQNE concentration)**.** Let $X\in\mathbb{R}$ be a one dimensional continuous random variable and $\widetilde{X}$ be an independent copy of X. Let $G(\cdot)$ be the distribution function of $|X-\widetilde{X}|$. We then have

$$\begin{aligned}&\mathbb{P}(|\widehat{\sigma}^{\mathrm{Q}}(X;r)-\sigma^{\mathrm{Q}}(X;r)|>t)\\ &\quad\le \exp\Big(-n[G\{G^{-1}(r)+t\}-r-n^{-1}]^2\Big)+\exp\Big(-n[r-G\{G^{-1}(r)-t\}]^2\Big),\end{aligned}$$

whenever $G\{G^{-1}(r)+t\}>r+n^{-1}$.

Proof. We denote $Z:=|X-\widetilde{X}|$. By definition, we have

$$\mathbb{P}(|\widehat{\sigma}^{\mathrm{Q}}(X;r)-\sigma^{\mathrm{Q}}(X;r)|>t)=\mathbb{P}(|\widehat{Q}(\{|X_i-X_j|\}_{i<j};r)-Q(Z;r)|>t).$$

Similar as in the proof of Lemma A.1, we have

$$\mathbb{P}\Big\{\widehat{Q}(\{|X_i-X_j|\}_{i<j};r)-Q(Z;r)\ge t\Big\}\le \mathbb{P}\Big[\frac{2}{n(n-2)}\sum_{i<j}I\{|X_i-X_j|\le G^{-1}(r)+t\}\le r+\frac{1}{n}\Big].$$

By Hoeffding's inequality in U-statistics (c.f., Equation (5.7) in Hoeffding (1963)), we have

$$\mathbb{P}\Big\{\widehat{Q}(\{|X_i - X_j|\}_{i<j}; r) - Q(Z; r) > t\Big\} \leq \exp\Big(-n[G\{G^{-1}(r) + t\} - r - n^{-1}]^2\Big), \tag{A.3}$$

where we remind that G represents the distribution function of Z. Similarly, we have

$$\mathbb{P}\Big\{(\widehat{Q}(\{|X_i - X_j|\}_{i<j}; r) - Q(Z; r) \leq -t\Big\} \leq \exp\Big(-n[r - G\{G^{-1}(r) - t\}]^2\Big). \tag{A.4}$$

Combining (A.3) and (A.4), we have the desired result. □

B Proof of Main Results

B.1 Proof of Propositions 3.7, 3.8, and 3.9

Proof of Proposition 3.7. Consider a pair-elliptically distributed random vector $\boldsymbol{X} = (X_1, \ldots, X_d)^T \sim PE_d(\boldsymbol{\mu}, \mathbf{S}, \xi_1)$. If $\boldsymbol{X}$ is also transelliptically distributed, by definition, there exists a set of univariate strictly increasing functions $f = \{f_j\}_{j=1}^d$ such that $f(\boldsymbol{X}) := (f_1(X_1), \ldots, f_d(X_d))^T \sim EC_d(\mathbf{0}, \boldsymbol{\Sigma}^0, \xi_2)$ for some generating variable ξ_2. Because $\boldsymbol{X}$ has the same Kendall's tau correlation matrix as $f(\boldsymbol{X})$, without loss of generality, we can assume $\mathbf{S} = \boldsymbol{\Sigma}^0$. Accordingly, the margins of $\boldsymbol{X}$ follow the same distribution. Combined with the fact that the margins of $f(\boldsymbol{X})$ are identically distributed, we have the transformation functions $f_1 = f_2 = \cdots = f_d = f_0$. The desired result follows by considering the following two cases. If $|\boldsymbol{\Sigma}^0_{jk}| = 1$ for all $j, k \in \{1, \ldots, d\}$, then we have $X_1 = (\pm)X_2 = \cdots = (\pm)X_d$ almost surely. This, combined with the fact that $\boldsymbol{X}$ are pair-elliptically distributed, implies that $\boldsymbol{X}$ is elliptical distributed. Otherwise, if there exists $j \neq k$ such that $|\boldsymbol{\Sigma}^0_{jk}| \in (0, 1)$, then without loss of generality, we assume $j = 1, k = 2$ and let $\rho := \boldsymbol{\Sigma}^0_{12} \neq 0$. Because $(X_1, X_2)^T$ is elliptically distributed, we then have

$$X_2 | X_1 = X_1 \sim EC_1(\rho X_1, 1 - \rho^2, \xi_1'), \quad \text{for some generating variable } \xi_1',$$

which implies that

$$\text{median}(X_2 | X_1 = X_1) = \rho X_1 \quad \text{and} \quad \text{median}(f_0(X_2) | f_0(X_1) = f_0(X_1)) = f_0(\rho X_1). \tag{B.1}$$

On the other hand, because $(f_0(X_1), f_0(X_2))^T$ is elliptically distributed, we also have

$$f_0(X_2) | f_0(X_1) = f_0(X_1) \sim EC_1(\rho f_0(X_1), 10\rho^2, \xi_2'),, \quad \text{for some generating variable } \xi_2',$$

which implies that

$$\text{median}(f_0(X_2) | f_0(X_1) = f_0(X_1)) = \rho f_0(X_1). \tag{B.2}$$

Combining Equations (B.1) and (B.2), we have for all $X_1 \in \mathbb{R}$, $f_0(\rho X_1) = \rho f_0(X_1)$, implying that $f_0(x) = ax$ for some $a \in \mathbb{R}$. This further implies that $\boldsymbol{X} \sim EC_d(\mathbf{0}, a^2\boldsymbol{\Sigma}^0, \xi_2)$ is elliptically distributed. □

Proof of Proposition 3.8. By definition, any pair in $\boldsymbol{Y}$ satisfies

$$(Y_j, Y_k)^T \overset{D}{=} \xi_G \mathbf{A}\boldsymbol{U} \sim N_2(\boldsymbol{\mu}_{\{j,k\}}, \mathbf{S}_{\{j,k\},\{j,k\}}),$$

where $q = \text{rank}(\mathbf{S}_{\{j,k\},\{j,k\}})$, $\mathbf{A} \in \mathbb{R}^{2\times q}$ with $\mathbf{A}\mathbf{A}^T = \mathbf{S}_{\{j,k\},\{j,k\}}$, $\boldsymbol{U} \in \mathbb{R}^q$ uniformly distributed in $\mathbb{S}^{q-1}$, and ξ_G^2 follows a chi square distribution with degree of freedom q. This further implies that any pair in $\boldsymbol{X}$ satisfies

$$(X_j, X_k)^T = \xi(Y_j, Y_k)^T \sim \xi \cdot \xi_G \cdot \mathbf{A}\boldsymbol{U},$$

and accordingly is elliptically distributed. This verifies that $\boldsymbol{X}$ is pair-elliptically distributed. □

Proof of Proposition 3.9. Because the margin of a pair-normal is normally distributed and the only elliptical distribution that has normal margins is the Gaussian, we have the first assertion is true. We then prove the second assertion. Suppose that $\boldsymbol{Y} = (Y_1, \ldots, Y_d)^T$ is pair-normally as well as nonparanormlly distributed. Then for any $j \in \{1, \ldots, d\}$, Y_j is normally distributed. Moreover, because $\boldsymbol{Y}$ is nonparanoramlly distributed, we have $f_j(Y_j) \sim N(0, 1)$. This implies that f_j is a linear transformation. Therefore, $\boldsymbol{Y}$ is Gaussian distributed because $f_1, \ldots, f_d$ are linear. □

B.2 Proofs of Lemma 4.1

Proof. Since $\widetilde{\mathbf{R}}$ is the minimizer of (4.4), we have

$$\|\widetilde{\mathbf{R}} - \widehat{\mathbf{R}}\| \leq \|\widetilde{\mathbf{R}} - \widehat{\mathbf{R}}\|,$$

which implies $\mathbb{P}(\|\widetilde{\mathbf{R}} - \mathbf{R}\| \geq t) \leq \mathbb{P}(\|\widetilde{\mathbf{R}} - \widehat{\mathbf{R}}\| + \|\widehat{\mathbf{R}} - \mathbf{R}\| \geq t) \leq \mathbb{P}(\|\widehat{\mathbf{R}} - \mathbf{R}\| \geq t)$. □

B.3 Proofs of Theorems 2.1

Proof of Theorem 2.1. Let $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ be n independent observations of $\boldsymbol{X} \in \mathbb{R}^d$ with $\boldsymbol{X}_i = (X_{i1}, \ldots, X_{id}^T)$. By Chebychev's inequality, we have

$$\mathbb{P}(|X_{ij}| \geq t) \leq (t/||X_{ij}||_{L_p})^{-p}.$$

Let's cut X_{ij} into two parts:

$$Y_{ij} = X_{ij}\,\mathbb{1}(|X_{ij}| \geq \epsilon) \quad \text{and} \quad Y_{ij}^* = X_{ij}\,\mathbb{1}(|X_{ij}| \leq \epsilon).$$

We have $|Y_{ij}^* - \mathbb{E}Y_{ij}^*|$ is upper bounded by 2ϵ and its variance has the property:

$$\text{Var}(Y_{ij}^*) \leq \mathbb{E}(Y_{ij}^*)^2 \leq ||X_{ij}||_{L_2}^2 \leq ||X_{ij}||_{L_p}^2 = K^2.$$

Accordingly, by Bernstein's inequality, we have

$$\mathbb{P}\Big(\frac{1}{n}(\sum Y_{ij}^* - \mathbb{E}Y_{ij}^*) > t/2\Big) \leq \exp\Big(-\frac{1/8 \cdot nt^2}{K^2 + 1/3 \cdot \epsilon t}\Big).$$

For Y_{ij}, we have

$$\mathbb{P}(Y_{ij} \neq 0) = \mathbb{P}(|X_{ij}| \geq \epsilon) \leq (\epsilon/K)^{-p},$$

and

$$|\mathbb{E}Y_{ij}| = |\mathbb{E}X_{ij}\,\mathbb{1}(|X_{ij}| \geq \epsilon)| \leq \frac{\mathbb{E}|X_{ij}|^p}{\epsilon^{p-1}} = \frac{K^p}{\epsilon^{p-1}}.$$

Accordingly, for any ϵ such that

$$K^p/\epsilon^{p-1} \leq t/2, \tag{B.3}$$

we have for any $j \in \{1, \ldots, d\}$,

$$\begin{aligned}
\mathbb{P}\Big(\Big|\frac{1}{n}\sum X_{ij} - \mu_j\Big| > t\Big) \leq & \mathbb{P}\Big(\frac{1}{n}\Big|\sum (Y_{ij}^* - \mathbb{E}Y_{ij}^*)\Big| > t/2\Big) + \mathbb{P}\Big(\Big|\frac{1}{n}\sum (Y_{ij} - \mathbb{E}Y_{ij})\Big| > t/2\Big) \\
\leq & 2\exp\Big(-\frac{1/8 \cdot nt^2}{K^2 + 1/3 \cdot \epsilon t}\Big) + n\mathbb{P}(Y_{ij} \neq 0) \\
\leq & 2\exp\Big(-\frac{1/8 \cdot nt^2}{K^2 + 1/3 \cdot \epsilon t}\Big) + n(\epsilon/K)^{-p}.
\end{aligned}$$

This yields that

$$\begin{aligned}
\mathbb{P}(||\bar{\boldsymbol{\mu}} - \boldsymbol{\mu}||_\infty > t) \leq & d \cdot \mathbb{P}\Big(\Big|\frac{1}{n}\sum X_{ij} - \mu_j\Big| > t\Big) \\
\leq & d \cdot \Big(2\exp\Big(-\frac{1/8 \cdot nt^2}{K^2 + 1/3 \cdot \epsilon t}\Big) + n(\epsilon/K)^{-p}\Big).
\end{aligned}$$

Taking $t = 12K \cdot \sqrt{\log d/n}$ and $\epsilon = K \cdot \sqrt{n/(\log d)}$, as long as $\log d/n = o(1)$, we have

$$2d\exp\Big(-\frac{1/8 \cdot nt^2}{K^2 + 1/3 \cdot \epsilon t}\Big) \leq 2d^{-2.5}.$$

It is also straightforward to verify that for such chosen t and ϵ, we have $K^p/\epsilon^{p-1} \leq t/2$ as long as $p \geq 2$.

For the second term, we have

$$\begin{aligned}
\log(n\epsilon^{-p} \cdot K^p) = & \log n - p\log\epsilon + p\log K \\
= & \log n - p(\log K + 1/2\log n - 1/2\log\log d)) + p\log K \\
= & (-p/2 + 1)\log n + p/2\log\log d.
\end{aligned}$$

So for large enough p and K, we need to have

$$\log d + (-p/2 + 1)\log n + p/2\log\log d \to -\infty.$$

Thus, letting $p = 2 + 2\gamma + \delta$, for $d = O(n^\gamma)$, with probability no smaller than $1 - 2d^{-2.5} - (\log d)^{p/2}n^{-\delta/2}$, we have

$$||\bar{\boldsymbol{\mu}} - \boldsymbol{\mu}||_\infty \leq 12K\sqrt{\frac{\log d}{n}}.$$

This completes the proof. □

Proof of Theorem 2.2. Consider the distribution $\boldsymbol{X} = (X_1, \ldots, X_d)$ with $X_1, \ldots, X_d$ independent and identically distributed. For reasons that will be clear later, we can set $C = 1$ and consider

$$\mathbb{P}(X_1 = \sqrt{n \log d}) = \frac{1}{2}(n \log d)^{-p/2}, \quad \mathbb{P}(X_1 = -\sqrt{n \log d}) = \frac{1}{2}(n \log d)^{-p/2}, \\ \text{and } \mathbb{P}(X_1 = 0) = 1 - (n \log d)^{-p/2}. \tag{B.4}$$

We then have $\boldsymbol{\mu} = \mathbf{0}$ and $||X_1||_{L_q} \to I(q = p) + \infty \cdot I(q > p)$. Moreover, letting

$$\alpha := \mathbb{P}(|\bar{\mu}_1| \geq \sqrt{\log d/n}) \geq 1 - \mathbb{P}(X_1 = \ldots = X_n = 0) \geq n(n \log d)^{-p/2}(1 + o(1)), \tag{B.5}$$

we have

$$\begin{aligned} \mathbb{P}(||\bar{\boldsymbol{\mu}} - \boldsymbol{\mu}||_\infty \geq \sqrt{\log d/n}) &= d\alpha - \tbinom{d}{2}\alpha^2 + \tbinom{d}{3}\alpha^3 - \cdots \\ &= 1 - (1 - \alpha)^d. \end{aligned}$$

When $p = 2 + 2\gamma$ and $d = n^{\gamma + \delta_0}$ with $\delta_0 > 0$ and some $0 < \delta_1 < \delta_0$, we have

$$(1 - \alpha)^d \leq \Big(1 - n^{-\gamma}(\log d)^{-p/2}\Big)^{n^{\gamma+\delta_0}} \leq \Big(1 - \frac{1}{n^{\gamma+\delta_1}}\Big)^{n^{\gamma+\delta_0}} \leq (1/e)^{n^{\delta_0 - \delta_1}} \to 0.$$

Accordingly, with probability tending to 1, we have

$$||\bar{\boldsymbol{\mu}} - \boldsymbol{\mu}||_\infty \geq \sqrt{\log d/n}.$$

When $C \neq 1$, we can replace all terms $\sqrt{n \log d}$ in Equation (B.4) with $\sqrt{Cn \log d}$ and all the proofs can follow. □

B.4 Proofs of Theorems 5.1 and 5.2

Proof of Theorem 5.1. For $j = 1, \ldots, d$, we denote $\mu_j := Q(X_j; 1/2)$. By definition, we have for $j = 1, \ldots, d$,

$$\sigma^{\mathrm{M}}(X_j; r) = Q(|X_j - Q(X_j; 1/2)|; r) = Q(|X_j - \mu_j|; r).$$

In other words, we have

$$r = \mathbb{P}\{|X_j - \mu_j| \leq \sigma^M(X_j; r)\} = \mathbb{P}\Big\{\frac{|X_j - \mu_j|}{\sigma(X_j)} \leq \frac{\sigma^M(X_j; r)}{\sigma(X_j)}\Big\},$$

because all the quantiles are unique. Accordingly, we have

$$\frac{\sigma^M(X_j; r)}{\sigma(X_j)} = Q\Big(\frac{|X_j - \mu_j|}{\sigma(X_j)}; r\Big) \Rightarrow \sigma^M(X_j; r) = \sigma(X_j) \cdot Q\Big(\frac{|X_j - \mu_j|}{\sigma(X_j)}; r\Big).$$

Using a similar technique as above, we can further derive that

$$\sigma^M(X_j + X_k; r) = \sigma(X_j + X_k) \cdot Q\Big(\frac{|X_j + X_k - Q(X_j + X_k; 1/2)|}{\sigma(X_j + X_k)}; r\Big), \\ \text{and} \quad \sigma^M(X_j - X_k; r) = \sigma(X_j - X_k) \cdot Q\Big(\frac{|X_j - X_k - Q(X_j - X_k; 1/2)|}{\sigma(X_j - X_k)}; r\Big).$$

Therefore, $\mathbf{R}^{\mathrm{M};r} = c^{\mathrm{M};r}\boldsymbol{\Sigma}$ if (5.1) holds and by definition

$$c^{\mathrm{M};r} = \Big\{\frac{\sigma^M(X_j;r)}{\sigma(X_j)}\Big\}^2 = \Big\{Q\Big(\frac{|X_j-\mu_j|}{\sigma(X_j)};r\Big)\Big\}^2.$$

This completes the proof. □

Proof of Theorem 5.2. The proof of Theorem 5.2 is similar as Theorem 5.1, and is accordingly omitted here. □

B.5 Proof of Theorem 5.3

Proof. We first prove the case for the gMAD estimator $\mathbf{R}^{\mathrm{M};r}$. By the definition of the pair-elliptical and the discussions in Remark 3.2, we have any pair in $PE_d(\boldsymbol{\mu},\mathbf{S},\xi)$ can be written as $EC_2(\boldsymbol{\mu}_{\{j,k\}},\mathbf{S}_{\{j,k\},\{j,k\}},\phi)$ where ϕ is a properly defined characteristic function only depending on ξ. On one hand, by Theorem 2.16 in Fang et al. (1990), we have for any $j\in\{1,\ldots,d\}$,

$$\frac{X_j-\mu_j}{\sqrt{\mathbf{S}_{jj}}} \sim EC_1(0,1,\phi). \tag{B.6}$$

Because $\boldsymbol{X}$ is continuous (by the definition of the pair-elliptical) and accordingly all quantiles of $EC_d(0,1,\xi)$ are uniquely defined, using the same proof techniques as exploited in the proof of Theorem 5.1, (B.6) shows that

$$\mathbf{R}^{\mathrm{M};r}_{jj} = \Big(Q\Big(\Big|EC_1(0,1,\phi)\Big|;r\Big)\Big)^2\cdot\mathbf{S}_{jj}$$

for any $j\in\{1,\ldots,d\}$. On the other hand, for those $j\neq k$ such that $\mathbf{S}_{jj}+\mathbf{S}_{kk}+2\mathbf{S}_{jk}\neq 0$ and $\mathbf{S}_{jj}+\mathbf{S}_{kk}-2\mathbf{S}_{jk}\neq 0$, we have

$$\begin{aligned} & X_j+X_k \sim EC_d(\mu_j+\mu_k,\mathbf{S}_{jj}+\mathbf{S}_{kk}+2\mathbf{S}_{jk},\phi),\\ \text{and}\quad & X_j-X_k \sim EC_d(\mu_j-\mu_k,\mathbf{S}_{jj}+\mathbf{S}_{kk}-2\mathbf{S}_{jk},\phi),\end{aligned}$$

and accordingly X_j+X_k and X_j-X_k are both elliptically distributed with the same characterization function ϕ. Therefore, we have

$$\frac{X_j+X_k-Q(X_j+X_k;1/2)}{\sqrt{\mathbf{S}_{jj}+\mathbf{S}_{kk}+2\mathbf{S}_{jk}}} \overset{D}{=} \frac{X_j-X_k-Q(X_j-X_k;1/2)}{\sqrt{\mathbf{S}_{jj}+\mathbf{S}_{kk}-2\mathbf{S}_{jk}}} \sim EC_1(0,1,\phi),$$

and accordingly we have

$$\Big|\frac{X_j-\mu_j}{\sqrt{\mathbf{S}_{jj}}}\Big| \overset{D}{=} \Big|\frac{X_j+X_k-Q(X_j+X_k;1/2)}{\sqrt{\mathbf{S}_{jj}+\mathbf{S}_{kk}+2\mathbf{S}_{jk}}}\Big| \overset{D}{=} \Big|\frac{X_j-X_k-Q(X_j-X_k;1/2)}{\sqrt{\mathbf{S}_{jj}+\mathbf{S}_{kk}-2\mathbf{S}_{jk}}}\Big|.$$

Thus, in parallel to the proof in Theorem 5.1, we have

$$\mathbf{R}^{\mathrm{M};r}_{jk} = \Big(Q\Big(\Big|EC_1(0,1,\phi)\Big|;r\Big)\Big)^2\cdot\mathbf{S}_{jk}.$$

For those $j \neq k \in \{1, \ldots, d\}$ such that $\mathbf{S}_{jj} + \mathbf{S}_{kk} - 2\mathbf{S}_{jk} = 0$, we have $\mathrm{Cov}(X_j - X_k) = 0$, implying that $X_j = X_k$, $a.s.$. Accordingly, we have

$$\sigma^{\mathrm{M}}(X_j, X_k; r) = \frac{1}{4}\{\sigma^{\mathrm{M}}(X_j + X_k; r)\}^2 = (\sigma^{\mathrm{M}}(X_j + X_k; r))^2.$$

Accordingly, $\mathbf{R}^{\mathrm{M};r}_{jk} = \mathbf{R}^{\mathrm{M};r}_{jj} = \Big(Q\Big(\Big|EC_1(0,1,\phi)\Big|; r\Big)\Big)^2 \cdot \mathbf{S}_{jj} = \Big(Q\Big(\Big|EC_1(0,1,\phi)\Big|; r\Big)\Big)^2 \cdot \mathbf{S}_{jk}$. Similarly, for those $j \neq k \in \{1, \ldots, d\}$ such that $\mathbf{S}_{jj} + \mathbf{S}_{kk} + 2\mathbf{S}_{jk} = 0$, we have $\mathbf{R}^{\mathrm{M};r}_{jk} = -\mathbf{R}^{\mathrm{M};r}_{jj} = -\Big(Q\Big(\Big|EC_1(0,1,\phi)\Big|; r\Big)\Big)^2 \cdot \mathbf{S}_{jj} = \Big(Q\Big(\Big|EC_1(0,1,\phi)\Big|; r\Big)\Big)^2 \cdot \mathbf{S}_{jk}$. This completes the proof of the first part.

Secondly, we switch to the gQNE estimator. We note that, by Remark 3.2,

$$\mathbb{E}\exp(i\boldsymbol{t}^T(\boldsymbol{X}_{\{j,k\}} - \widetilde{\boldsymbol{X}}_{\{j,k\}})) = \phi^2(\boldsymbol{t}^T\mathbf{S}_{\{j,k\},\{j,k\}}\boldsymbol{t}),$$

and accordingly $\boldsymbol{X} - \widetilde{\boldsymbol{X}} \sim PE_d(\mathbf{0}, 2\mathbf{S}, \phi^2)$ and is continuously distributed. Then by following the same proof as in the first part we have the desired result.

In the end, let's show the scale constant. Using the same argument as above, we have

$$\mathbf{R}^{\mathrm{M};r} = \Big(Q\Big(\Big|X_0\Big|; r\Big)\Big)^2 \cdot \boldsymbol{\Sigma}.$$

Because X_0 is continuous and symmetric, letting $q_r := Q\Big(\Big|X_0\Big|; r\Big)$, we have

$$\mathbb{P}(|X_0| \leq q_r) = 2\mathbb{P}(X_0 \leq q_r) - 1 = r \Rightarrow \mathbb{P}(X_0 \leq q_r) = \frac{1+r}{2} \Rightarrow q_r = Q\Big(\Big|X_0\Big|; \frac{1+r}{2}\Big).$$

Similarly, we have

$$\mathbf{R}^{\mathrm{Q};r} = \Big(Q\Big(\Big|Z_0\Big|; \frac{1+r}{2}\Big)\Big)^2 \cdot \boldsymbol{\Sigma}.$$

This finalizes the proof. □

B.6 Proof of Theorem 5.5

Proof. Using Lemma A.2 and Assumption (A1), we have for any $j \in \{1, \ldots, d\}$,

$$\mathbb{P}(|\widehat{\sigma}^{\mathrm{M}}(X_j; r) - \sigma^{\mathrm{M}}(X_j; r)| > t) \leq 3\exp(-2n(\eta_1 t/2 - 1/n)^2) + 3\exp(-2n(\eta_1 t/2)^2),$$

whenever $t/2 \leq \kappa_1$ and $\eta_1 t/2 > 1/n$. Accordingly, we have

$$\begin{aligned}
\mathbb{P}(|\widehat{\mathbf{R}}^{\mathrm{M};r}_{jj} - \mathbf{R}^{\mathrm{M};r}_{jj}| > t) &= \mathbb{P}(|(\widehat{\sigma}^{\mathrm{M}}(X_j; r) + \sigma^{\mathrm{M}}(X_j; r))(\widehat{\sigma}^{\mathrm{M}}(X_j; r) - \sigma^{\mathrm{M}}(X_j; r))| > t) \\
&\leq \mathbb{P}(|(\widehat{\sigma}^{\mathrm{M}}(X_j; r) - \sigma^{\mathrm{M}}(X_j; r))^2 + 2\sigma^{\mathrm{M}}(X_j; r)(\widehat{\sigma}^{\mathrm{M}}(X_j) - \sigma^{\mathrm{M}}(X_j; r))| > t) \\
&\leq \mathbb{P}\Big(|\widehat{\sigma}^{\mathrm{M}}(X_j; r) - \sigma^{\mathrm{M}}(X_j; r)| > \sqrt{\frac{t}{2}}\Big) + \mathbb{P}\Big(|\widehat{\sigma}^{\mathrm{M}}(X_j; r) - \sigma^{\mathrm{M}}(X_j; r)| > \frac{t}{2\sigma^{\mathrm{M}}(X_j; r)}\Big) \\
&\leq 6\exp(-2n(\eta_1\sqrt{t/2}/2 - 1/n)^2) + 6\exp(-2n(\eta_1 t/(4\sigma^{\mathrm{M}}(X_j; r)))^2). \qquad \text{(B.7)}
\end{aligned}$$

Using Lemma A.2 and Assumption (A1), we have for any $j \neq k$,

$$\mathbb{P}(|\widehat{\sigma}^{\mathrm{M}}(X_j + X_k; r) - \sigma^{\mathrm{M}}(X_j + X_k; r)| > t) \leq 3\exp(-2n(\eta_1 t/2 - 1/n)^2) + 3\exp(-2n(\eta_1 t/2)^2),$$

and

$$\mathbb{P}(|\widehat{\sigma}^{\mathrm{M}}(X_j - X_k; r) - \sigma^{\mathrm{M}}(X_j - X_k; r)| > t) \leq 3\exp(-2n(\eta_1 t/2 - 1/n)^2) + 3\exp(-2n(\eta_1 t/2)^2).$$

And accordingly, letting $\theta_{\max} := 2\sqrt{||\mathbf{R}^{\mathrm{M};r}||_{\max}}$, we have

$$\begin{aligned}
&\mathbb{P}(|\widehat{\mathbf{R}}^{\mathrm{M};r}_{jk} - c^{\mathrm{M};r}\mathbf{\Sigma}_{jk}| > t) \\
\leq&\mathbb{P}(|(\widehat{\sigma}^{\mathrm{M}}(X_j + X_k; r))^2 - (\sigma^{\mathrm{M}}(X_j + X_k; r))^2| > 2t) \\
&+ \mathbb{P}(|(\widehat{\sigma}^{\mathrm{M}}(X_j - X_k; r))^2 - (\sigma^{\mathrm{M}}(X_j - X_k; r))^2| > 2t) \\
\leq&\mathbb{P}\Big(|\widehat{\sigma}^{\mathrm{M}}(X_j + X_k; r) - \sigma^{\mathrm{M}}(X_j + X_k; r)| > \sqrt{t}\Big) \\
&+ \mathbb{P}\Big(|\widehat{\sigma}^{\mathrm{M}}(X_j + X_k; r) - \sigma^{\mathrm{M}}(X_j + X_k; r)| > \frac{t}{\sigma^{\mathrm{M}}(X_j + X_k; r)}\Big) \\
&+ \mathbb{P}\Big(|\widehat{\sigma}^{\mathrm{M}}(X_j - X_k; r) - \sigma^{\mathrm{M}}(X_j - X_k; r)| > \sqrt{t}\Big) \\
&+ \mathbb{P}\Big(|\widehat{\sigma}^{\mathrm{M}}(X_j - X_k; r) - \sigma^{\mathrm{M}}(X_j - X_k; r)| > \frac{t}{\sigma^{\mathrm{M}}(X_j - X_k; r)}\Big) \\
\leq&6\exp(-2n(\eta_1\sqrt{t}/2 - 1/n)^2) + 6\exp(-n\eta_1^2 t/2) + 6\exp(-2n(\eta_1 t/(2\theta_{\max}) - 1/n)^2) \\
&+ 6\exp(-n\eta_1^2 t^2/(2\theta_{\max}^2)) \\
\leq&24\max\{\exp(-2n(\eta_1\sqrt{t}/2 - 1/n)^2), \exp(-2n(\eta_1 t/(2\theta_{\max}) - 1/n)^2)\}. \qquad \text{(B.8)}
\end{aligned}$$

Combining Equations (B.7) and (B.8), we have, with probability $1 - 24\alpha^2$,

$$\begin{aligned}
&||\widehat{\mathbf{R}}^{\mathrm{M};r} - \mathbf{R}^{\mathrm{M};r}||_{\max} \leq \\
&\max\Big\{\underbrace{\frac{6}{\eta_1^2}\Big(\sqrt{\frac{\log d + \log(1/\alpha)}{n}} + \frac{1}{n}\Big)^2}_{T_1}, \underbrace{\frac{4\sqrt{||\mathbf{R}^{\mathrm{M};r}||_{\max}}}{\eta_1}\Big(\sqrt{\frac{\log d + \log(1/\alpha)}{n}} + \frac{1}{n}\Big)}_{T_2}\Big\},
\end{aligned}$$

whenever n is large enough such that $T_1 \leq 8\kappa_1^2$ and $T_2 \leq 2\kappa_1 \cdot \min_{j\neq k}\{2\sigma^{\mathrm{M}}(X_j; r), \sigma^{\mathrm{M}}(X_j + X_k; r), \sigma^{\mathrm{M}}(X_j - X_k; r)\}$. Combining the above inequality with Theorem 5.3, we complete the whole proof. □

B.7 Proof of Theorem 5.6

Proof. Using Lemma A.3 and Assumption (A2), we have for any $j \in \{1, \ldots, d\}$,

$$\mathbb{P}(|\widehat{\sigma}^{\mathrm{Q}}(X; r) - \sigma^{\mathrm{Q}}(X; r)| > t) \leq \exp\Big(-n[\eta_2 t - 1/n]^2\Big) + \exp\Big(-n(\eta_2 t)^2\Big),$$

whenever $t \leq \kappa_2$ and $\eta_2 t > 1/n$. Using a similar proof technique as in the proof of Theorem 5.5, we have

$$\begin{aligned}
\mathbb{P}(|\widehat{\mathbf{R}}^{\mathrm{Q};r}_{jj} - \mathbf{R}^{\mathrm{Q};r}_{jj}| > t) \leq&\mathbb{P}\Big(|\widehat{\sigma}^{\mathrm{Q}}(X_j; r) - \sigma^{\mathrm{Q}}(X_j; r)| > \sqrt{t/2}\Big) + \mathbb{P}\Big(|\widehat{\sigma}^{\mathrm{Q}}(X_j; r) - \sigma^{\mathrm{Q}}(X_j; r)| > \frac{t}{2\sigma^{\mathrm{Q}}(X_j; r)}\Big) \\
\leq&2\exp\Big(-n[\eta_2\sqrt{t/2} - 1/n]^2\Big) + 2\exp\Big(-n(\eta_2 t/(2\sigma^{Q}(X_j; r)) - 1/n)^2\Big). \qquad \text{(B.9)}
\end{aligned}$$

Similarly, we have

$$\begin{aligned}
&\mathbb{P}(|\widehat{\mathbf{R}}^{Q;r}_{jk} - \mathbf{R}^{Q;r}_{jk}| > t) \\
\leq &\mathbb{P}\Big(|\widehat{\sigma}^{\mathrm{Q}}(X_j + X_k; r) - \sigma^{\mathrm{Q}}(X_j + X_k; r)| > \sqrt{t}\Big) + \mathbb{P}\Big(|\widehat{\sigma}^{\mathrm{Q}}(X_j - X_k; r) - \sigma^{\mathrm{Q}}(X_j - X_k; r)| > \sqrt{t}\Big) \\
&+ \mathbb{P}\Big(|\widehat{\sigma}^{\mathrm{Q}}(X_j + X_k; r) - \sigma^{\mathrm{Q}}(X_j + X_k; r)| > \frac{t}{\sigma^{\mathrm{Q}}(X_j + X_k; r)}\Big) \\
&+ \mathbb{P}\Big(|\widehat{\sigma}^{\mathrm{Q}}(X_j - X_k; r) - \sigma^{\mathrm{Q}}(X_j - X_k; r)| > \frac{t}{\sigma^{\mathrm{Q}}(X_j - X_k; r)}\Big) \\
\leq &4\exp(-n[\eta_2\sqrt{t} - 1/n]^2) + 4\exp(-n(\eta_2 t/\zeta_{\max} - 1/n)^2),
\end{aligned} \tag{B.10}$$

where $\zeta_{\max} := 2\sqrt{||\mathbf{R}^{\mathrm{Q};r}||_{\max}}$. Combining (B.9) and (B.10) leads to that, with probability larger than or equal to $1 - 8\alpha$,

$$\begin{aligned}
&||\widehat{\mathbf{R}}^{\mathrm{Q};r} - \mathbf{R}^{\mathrm{Q};r}||_{\max} \leq \\
&\max\Big\{\underbrace{\frac{2}{\eta_2^2}\Big(\sqrt{\frac{2\log d + \log(1/\alpha)}{n}} + \frac{1}{n}\Big)^2}_{T_3}, \underbrace{\frac{2\sqrt{||\mathbf{R}^{\mathrm{Q};r}||_{\max}}}{\eta_2}\Big(\sqrt{\frac{2\log d + \log(1/\alpha)}{n}} + \frac{1}{n}\Big)}_{T_4}\Big\},
\end{aligned}$$

whenever $T_3 \leq 2\kappa_2^2$ and $T_4 \leq 2\kappa_1 \cdot \min_{j\neq k}\{2\sigma^{\mathrm{Q}}(X_j; r), \sigma^{\mathrm{Q}}(X_j + X_k), \sigma^{\mathrm{Q}}(X_j - X_k)\}$. Finally, combining the above inequality with Theorem 5.3, we complete the whole proof. □

C Proofs of Corollaries in Section 7

In this section we provide the proofs of the results presented in Section 7.

C.1 Proof of Corollary 7.1

Proof. We first prove that (7.4) holds. Because $\mathbf{R}^{\mathrm{Q};r}$ is feasible to Equation (7.2), we have

$$||\widehat{\mathbf{R}}^{\mathrm{Q};r} - \widetilde{\mathbf{R}}^{\mathrm{Q};r}||_{\max} \leq ||\widehat{\mathbf{R}}^{\mathrm{Q};r} - \mathbf{R}^{\mathrm{Q};r}||_{\max},$$

implying that

$$\begin{aligned}
\mathbb{P}\Big(\Big|\widetilde{\mathbf{R}}^{\mathrm{Q};r}_{jk} - \mathbf{R}^{\mathrm{Q};r}_{jk}\Big| \geq t\Big) \leq &\mathbb{P}\Big(\Big|\widetilde{\mathbf{R}}^{\mathrm{Q};r}_{jk} - \widehat{\mathbf{R}}^{\mathrm{Q};r}_{jk}\Big| + \Big|\widehat{\mathbf{R}}^{\mathrm{Q};r}_{jk} - \mathbf{R}^{\mathrm{Q};r}_{jk}\Big| \geq t\Big) \\
\leq &\mathbb{P}\Big(||\widetilde{\mathbf{R}}^{\mathrm{Q};r} - \widehat{\mathbf{R}}^{\mathrm{Q};r}||_{\max} + ||\widehat{\mathbf{R}}^{\mathrm{Q};r} - \mathbf{R}^{\mathrm{Q};r}||_{\max} \geq t\Big) \\
\leq &\mathbb{P}(||\widehat{\mathbf{R}}^{\mathrm{Q};r} - \mathbf{R}^{\mathrm{Q};r}||_{\max} \geq t/2).
\end{aligned}$$

Combined with (B.9) and (B.10), the above inequality implies that

$$\mathbb{P}\Big(||\widetilde{\mathbf{R}}^{\mathrm{Q};r}_{jk} - \mathbf{R}^{\mathrm{Q};r}_{jk}||_{\max} \geq t\Big) \leq d^4(4\exp(-n[\eta_2\sqrt{t/4} - 1/n]^2) + 4\exp(-n(\eta_2 t/(2\zeta_{\max}) - 1/n)^2)).$$

Plugging $t = \zeta$ into the above equation, we have the desired result.

Secondly, we prove that (7.5) holds. Because $\mathbb{E}X_j^4 \leq K$, by Chebyshev's inequality, with probability no smaller than $1 - n^{-2\xi}$, we have

$$|\widehat{\sigma}_j^2 - \sigma^2(X_j)| \leq c_1 n^{-1/2+\xi}.$$

Moreover, using (B.9), we have for any given $j \in \{1, \ldots, d\}$, by Markov inequality, with probability larger than or equal to $1 - 4\alpha$,

$$|\widehat{\mathbf{R}}_{jj}^{\mathrm{Q};r} - \mathbf{R}_{jj}^{\mathrm{Q};r}| \leq \zeta_j.$$

For notation simplicity, we denote $\sigma_j := \sigma_j(X_j)$, $r_j := \mathbf{R}_{jj}^{\mathrm{Q};r}$, and $\widehat{r}_j := \widehat{\mathbf{R}}_{jj}^{\mathrm{Q};r}$. Accordingly, we have

$$\begin{aligned}
||\widehat{\mathbf{\Sigma}}^{\mathrm{Q};r} - \mathbf{\Sigma}||_{\max} &= \Big\| \frac{\widehat{\sigma}_j^2}{\widehat{r}_j} \widetilde{\mathbf{R}}^{\mathrm{Q};r} - \frac{\sigma_j^2}{r_j} \mathbf{R}^{\mathrm{Q};r} \Big\|_{\max} \\
&\leq \Big\| \frac{\widehat{\sigma}_j^2}{\widehat{r}_j} \widetilde{\mathbf{R}}^{\mathrm{Q};r} - \frac{\sigma_j^2}{r_j} \widetilde{\mathbf{R}}^{\mathrm{Q};r} \Big\|_{\max} + \frac{\sigma_j^2}{r_j} ||\widetilde{\mathbf{R}}^{\mathrm{Q};r} - \mathbf{R}^{\mathrm{Q};r}||_{\max} \\
&\leq \Big| \frac{\widehat{\sigma}_j^2}{\widehat{r}_j} - \frac{\sigma_j^2}{r_j} \Big| \cdot (||\widetilde{\mathbf{R}}^{\mathrm{Q};r} - \mathbf{R}^{\mathrm{Q};r}||_{\max} + ||\mathbf{R}^{\mathrm{Q};r}||_{\max}) + \frac{1}{c^{\mathrm{Q};r}} ||\widetilde{\mathbf{R}}^{\mathrm{Q};r} - \mathbf{R}^{\mathrm{Q};r}||_{\max}
\end{aligned}$$

while noting that $c^{\mathrm{Q};r} = r_j/\sigma_j^2$. Finally, we have

$$\Big| \frac{\widehat{\sigma}_j^2}{\widehat{r}_j} - \frac{\sigma_j^2}{r_j} \Big| = \Big| \frac{\widehat{\sigma}_j^2 - \sigma_j^2}{\widehat{r}_j} + \sigma_j^2 \Big(\frac{1}{\widehat{r}_j} - \frac{1}{r_j} \Big) \Big| \leq \frac{|\widehat{\sigma}_j^2 - \sigma_j^2|}{r_j - |\widehat{r}_j - r_j|} + \frac{\sigma_j^2}{r_j} \cdot \frac{|\widehat{r}_j - r_j|}{r_j - |\widehat{r}_j - r_j|}.$$

This implies that with probability no smaller than $1 - n^{-2\xi} - 12\alpha$, we have

$$\Big| \frac{\widehat{\sigma}_j^2}{\widehat{r}_j} - \frac{\sigma_j^2}{r_j} \Big| \leq \frac{c_1 n^{-1/2+\xi}}{r_j - \zeta_j} + \frac{1}{c^{\mathrm{Q};r}} \cdot \frac{\zeta_j}{r_j - \zeta_j},$$

which completes the proof of the second part. The third part can then be proved by combining Equation (7.5) and the proof of Theorem 2 in Xue et al. (2012). □

C.2 Proof of Corollary 7.4

Proof. According to Theorem 6 in Cai et al. (2011), if the tuning parameter $\lambda \geq \|\mathbf{\Theta}\|_{1,\infty} \|\widehat{\mathbf{\Sigma}}^{\mathrm{Q};r} - \mathbf{\Sigma}\|_{\max}$, we have there exists two constants C_1, C_2 such that

$$\|\widehat{\mathbf{\Theta}}^{\mathrm{Q};r} - \mathbf{\Theta}\|_2 \leq C_1 s \lambda^{1-q}, \|\widehat{\mathbf{\Theta}}^{\mathrm{Q};r} - \mathbf{\Theta}\|_{\max} \leq \|\mathbf{\Theta}\|_{1,\infty} \lambda \text{ and } \|\widehat{\mathbf{\Theta}}^{\mathrm{Q};r} - \mathbf{\Theta}\|_{\mathrm{F}}^2 \leq C_2 s d \lambda^{2-q}.$$

Combining the above results with Corollary 7.1, we prove the desire rate. □

C.3 Proof of Corollary 7.5

Proof. According to Theorem 3.1 in Vu et al. (2013), if $\lambda \geq \|\widehat{\mathbf{R}} - \mathbf{R}\|_{\max}$ and $\max_{1 \leq k \leq m} ||\boldsymbol{u}_k(\mathbf{\Sigma})||_0 \leq s$, we have

$$\|\widehat{\mathbf{\Pi}}_m^{\mathrm{Q};r} - \mathbf{\Pi}_m\|_{\mathrm{F}} \leq \frac{4s\lambda}{\lambda_1(\mathbf{R}) - \lambda_2(\mathbf{R})} \cdot ||\widehat{\mathbf{R}} - \mathbf{R}||_{\max}. \tag{C.1}$$

where $\mathbf{\Pi}_m^{\mathrm{Q};r}$, $\mathbf{R}$, and $\widehat{\mathbf{R}}$ can be the Q-PCA estimator, population and sample versions of the quantile-based scatter matrix estimators based on the gQNE. Combining the rate of $\|\widehat{\mathbf{R}} - \mathbf{R}\|_{\max}$ in Theorem 5.5 and Theorem 5.6, the theorem is proved.

□

C.4 Proof of Corollary 7.6

Proof. We mainly follow the procedure in the proof of Theorem 2 in Cai and Liu (2011). Suppose $\boldsymbol{\delta}^T\boldsymbol{\beta}^* \leq M$. Let $\boldsymbol{\delta} = \boldsymbol{\mu}_1 - \boldsymbol{\mu}_2$, $\widehat{\boldsymbol{\delta}} = \widehat{\boldsymbol{\mu}}_1 - \widehat{\boldsymbol{\mu}}_2$. Since $(\widehat{\boldsymbol{\beta}}^{\mathrm{Q};r})^{\mathrm{Q};r}$ is the solution of (7.9), we have

$$|(\boldsymbol{\beta}^*)^T\widehat{\boldsymbol{\Sigma}}^{\mathrm{Q};r}\widehat{\boldsymbol{\beta}}^{\mathrm{Q};r} - (\boldsymbol{\beta}^*)^T\boldsymbol{\delta}| \leq (\lambda + \|\widehat{\boldsymbol{\delta}} - \boldsymbol{\delta}\|_\infty)\|\boldsymbol{\beta}^*\|_1 \leq 2\lambda\|\boldsymbol{\beta}^*\|_1,$$

and

$$|(\boldsymbol{\beta}^*)^T\widehat{\boldsymbol{\Sigma}}^{\mathrm{Q};r}\widehat{\boldsymbol{\beta}}^{\mathrm{Q};r} - (\widehat{\boldsymbol{\beta}}^{\mathrm{Q};r})^T\boldsymbol{\delta}| \leq (\lambda + \|\widehat{\boldsymbol{\delta}} - \boldsymbol{\delta}\|_\infty)\|\widehat{\boldsymbol{\beta}}^{\mathrm{Q};r}\|_1 \leq 2\lambda\|\boldsymbol{\beta}^*\|_1.$$

Combining the above two equations together, we have $|(\widehat{\boldsymbol{\beta}}^{\mathrm{Q};r} - \boldsymbol{\beta}^*)^T\boldsymbol{\delta}| \leq 4\lambda\|\boldsymbol{\beta}^*\|_1$ which implies

$$\frac{((\widehat{\boldsymbol{\beta}}^{\mathrm{Q};r})^T\boldsymbol{\delta})^2}{(\boldsymbol{\delta}^T\boldsymbol{\beta}^*)^2} \geq 1 - 8M^{-1}\lambda\|\boldsymbol{\beta}^*\|_1. \tag{C.2}$$

On the other hand, we have

$$\|\boldsymbol{\Sigma}\widehat{\boldsymbol{\beta}}^{\mathrm{Q};r} - \boldsymbol{\delta}\|_\infty \leq \|(\boldsymbol{\Sigma} - \widehat{\boldsymbol{\Sigma}}^{\mathrm{Q};r})\widehat{\boldsymbol{\beta}}^{\mathrm{Q};r}\|_\infty + 2\lambda \leq \|\boldsymbol{\beta}^*\|_1\|\boldsymbol{\Sigma} - \widehat{\boldsymbol{\Sigma}}^{\mathrm{Q};r}\|_{\max} + 2\lambda,$$

which implies that

$$\begin{aligned} |(\widehat{\boldsymbol{\beta}}^{\mathrm{Q};r})^T\boldsymbol{\Sigma}\widehat{\boldsymbol{\beta}}^{\mathrm{Q};r} - \boldsymbol{\delta}^T\boldsymbol{\beta}^*| &\leq |(\widehat{\boldsymbol{\beta}}^{\mathrm{Q};r})^T\boldsymbol{\Sigma}\widehat{\boldsymbol{\beta}}^{\mathrm{Q};r} - \boldsymbol{\delta}^T\widehat{\boldsymbol{\beta}}^{\mathrm{Q};r}| + |(\widehat{\boldsymbol{\beta}}^{\mathrm{Q};r} - \boldsymbol{\beta}^*)^T\boldsymbol{\delta}| \\ &\leq \|\boldsymbol{\beta}^*\|_1^2\|\boldsymbol{\Sigma} - \widehat{\boldsymbol{\Sigma}}^{\mathrm{Q};r}\|_{\max} + 6\lambda\|\boldsymbol{\beta}^*\|_1, \end{aligned} \tag{C.3}$$

where the last inequality is due to the definition of $\widehat{\boldsymbol{\beta}}^{\mathrm{Q};r}$. Therefore, for sufficiently large n, we have $(\widehat{\boldsymbol{\beta}}^{\mathrm{Q};r})^T\boldsymbol{\Sigma}\widehat{\boldsymbol{\beta}}^{\mathrm{Q};r} \geq M/2$. We can rewrite (C.3) as

$$\frac{\boldsymbol{\delta}^T\boldsymbol{\beta}^*}{(\widehat{\boldsymbol{\beta}}^{\mathrm{Q};r})^T\boldsymbol{\Sigma}\widehat{\boldsymbol{\beta}}^{\mathrm{Q};r}} \geq 1 - 2M^{-1}\|\boldsymbol{\beta}^*\|_1^2\|\boldsymbol{\Sigma} - \widehat{\boldsymbol{\Sigma}}^{\mathrm{Q};r}\|_{\max} - 12M^{-1}\lambda\|\boldsymbol{\beta}^*\|_1. \tag{C.4}$$

Combining (C.2) and (C.4), we have

$$\frac{\mathrm{Rq}((\widehat{\boldsymbol{\beta}}^{\mathrm{Q};r})^{\mathrm{Q};r})}{\mathrm{Rq}(\boldsymbol{\beta}^*)} = \frac{((\widehat{\boldsymbol{\beta}}^{\mathrm{Q};r})^T\boldsymbol{\delta})^2}{(\widehat{\boldsymbol{\beta}}^{\mathrm{Q};r})^T\boldsymbol{\Sigma}\widehat{\boldsymbol{\beta}}^{\mathrm{Q};r}\boldsymbol{\delta}^T\boldsymbol{\beta}^*} \geq 1 - 2M^{-1}\|\boldsymbol{\beta}^*\|_1^2\|\boldsymbol{\Sigma} - \widehat{\boldsymbol{\Sigma}}^{\mathrm{Q};r}\|_{\max} - 20M^{-1}\lambda\|\boldsymbol{\beta}^*\|_1.$$

This completes the proof. □

D Algorithm for Matrix Maximum Norm Projection

This section describes the algorithm to solve (4.4) for the matrix element-wise supremum norm $\|\cdot\|_{\max}$. In particular, we aim to solve

$$\widetilde{\mathbf{R}} = \underset{\mathbf{R}\succeq\mathbf{0}}{\arg\min}\,||\mathbf{R} - \widehat{\mathbf{R}}||_{\max} \tag{D.1}$$

by the algorithm proposed in Xu and Shao (2012). Since for any matrix $\mathbf{A} \in \mathbb{R}^d$ we can reformulate $\|\mathbf{A}\|_{\max} = \max_{\mathbf{Z}\in B_1}\mathrm{Tr}(\mathbf{Z}^T\mathbf{A})$, where $B_1 = \{\mathbf{Z} \in \mathbb{R}^{d\times d} \mid \mathbf{Z} = \mathbf{Z}^T, \sum_{i,j=1,\ldots,d}|\mathbf{Z}_{ij}| \leq 1\}$. We define $B_2 = \{\mathbf{Z} \in \mathbb{R}^{d\times d} \,|\succeq \mathbf{0}\}$ and (D.1) can be rewritten as a minimax problem

$$\min_{\mathbf{R}\in B_2}\max_{\mathbf{Z}\in B_1}\mathrm{Tr}(\mathbf{Z}^T(\mathbf{R} - \widehat{\mathbf{R}})). \tag{D.2}$$

Algorithm 1 Matrix nearness problem in the maximum norm in (D.1)

$\widetilde{\mathbf{R}} \leftarrow \text{MatrixMaxProj}(\widehat{\mathbf{R}}, \mathbf{Z}^0, \mathbf{R}^0, \gamma, \epsilon, N)$
for $t = 0, \ldots, N$ **do**
  $\mathbf{R}_0^t \leftarrow \mathbf{R}^t - P_{B_2}(\mathbf{R}^t - \mathbf{Z}^t)$
  $\mathbf{Z}_0^t \leftarrow \mathbf{Z}^t - P_{B_1}(\mathbf{R}^t + \mathbf{Z}^t - \widehat{\mathbf{R}})$
  if $\max\{\|\mathbf{R}_0^t\|_{\max}, \|\mathbf{R}_0^t\|_{\max}\} < \epsilon$, **then**
    break
  else
    $\mathbf{R}^{t+1} \leftarrow \mathbf{R}^t - \gamma(\mathbf{R}_0^t - \mathbf{Z}_0^t)/2$
    $\mathbf{Z}^{t+1} \leftarrow \mathbf{Z}^t - \gamma(\mathbf{R}_0^t + \mathbf{Z}_0^t)/2$
  end if
end for
return $\widetilde{\mathbf{R}} = \mathbf{R}^t$

In order to solve the above minimax problem, we need to first study two subproblems on matrix projection. The first is

$$P_{B_2}(\mathbf{A}) = \arg\min_{\mathbf{B} \in B_2} \|\mathbf{A} - \mathbf{B}\|_{\mathrm{F}}^2, \tag{D.3}$$

where $\|\cdot\|_{\mathrm{F}}$ is the Frobenius norm. We can have the closed form solution to D.3 such that $P_{B_2}(\mathbf{A}) = \mathbf{U}\mathbf{\Lambda}_+\mathbf{U}^T$, where $\mathbf{A} = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^T$ is the spectral decomposition of $\mathbf{A}$ and $\mathbf{\Lambda}_+ = \text{diag}(\mathbf{\Lambda}_{11} \vee 0, \ldots, \mathbf{\Lambda}_{dd} \vee 0)$. The second subproblem we are interested in is $P_{B_1}(\mathbf{A}) = \arg\min_{\mathbf{B} \in B_2} \|\mathbf{A} - \mathbf{B}\|_{\mathrm{F}}^2$. This is equivalent to the vectorized problem $P_{B_1}(\mathbf{A}) = \arg\min_{\|\boldsymbol{v}\|_1 \le 1} \|\text{vec}(\mathbf{A}) - \boldsymbol{v}\|_2^2$. We denote $\boldsymbol{a} = \text{vec}(\mathbf{A})$ and $|\boldsymbol{a}| = \text{sign}(\boldsymbol{a}) \circ \boldsymbol{a}$, where $\circ$ is the Hadamard product. Let $T_{|\boldsymbol{a}|}$ be the permutation transformation matrix of $|\boldsymbol{a}|$ such that $T_{|\boldsymbol{a}|}(|\boldsymbol{a}|)$ is in descending order. We now define $\widetilde{\boldsymbol{x}}, \widetilde{y}$ under two kinds of cases: $\|\boldsymbol{a}\|_1 \le 1$ and $\|\boldsymbol{a}\|_1 > 1$.

If $\|\boldsymbol{a}\|_1 \le 1$, we let $(\widetilde{\boldsymbol{x}}, \widetilde{y}) := (\widetilde{\boldsymbol{a}}, 0)$. If $\|\boldsymbol{a}\|_1 > 1$, we define $\Delta\boldsymbol{a} = (\widetilde{a}_1 - \widetilde{a}_2, \ldots, \widetilde{a}_{d-1} - \widetilde{a}_d, \widetilde{a}_d)^T$. Since $\Delta\boldsymbol{a}_i > 0$ for alll $i = 1, \ldots, n$ and $\sum_{i=1}^d i\Delta a_i = \|\boldsymbol{a}\|_1 > 1$. We choose the smallest integer K such that $\sum_{i=1}^K i\Delta a_i \ge 1$. Let

$$\widetilde{y} := \frac{1}{K}\Big(\sum_{i=1}^K \widetilde{a}_i - 1\Big), \widetilde{\boldsymbol{x}} := (\widetilde{a}_1 - \widetilde{y}, \ldots, \widetilde{a}_K - \widetilde{y}, 0, \ldots, 0)^T \in \mathbb{R}^d.$$

According to Xu and Shao (2012), we can write $P_{B_1}(\mathbf{A}) = \text{sign}(\boldsymbol{a}) \circ T_{\|\boldsymbol{a}\|}^{-1}(\widetilde{\boldsymbol{x}})$. We set arbitrary $\mathbf{Z}^0 \in B_1$, $\mathbf{R}^0 \in B_2$ as the initializations of the algorithm, $\gamma \in (0, 2)$ as the step length of each iteration, $\epsilon > 0$ as the tolerance, N as the maximum number of iterations. Algorithm 1 provides the following convergence to the exact solution of (D.1).

Theorem D.1 (Xu and Shao (2012))**.** If $(\mathbf{R}^{\text{opt}}, \mathbf{Z}^{\text{opt}})$ is the solution to (D.2). Let $\mathbf{R}^t, \mathbf{Z}^t, \mathbf{R}_0^t, \mathbf{Z}_0^t$ be the sequence obtained from Algorithm 1. We have

$$\|\mathbf{R}^{t+1} - \mathbf{R}^{\text{opt}}\|_{\mathrm{F}}^2 + \|\mathbf{Z}^{t+1} - \mathbf{Z}^{\text{opt}}\|_{\mathrm{F}}^2 \le \|\mathbf{R}^t - \mathbf{R}^{\text{opt}}\|_{\mathrm{F}}^2 + \|\mathbf{Z}^t - \mathbf{Z}^{\text{opt}}\|_{\mathrm{F}}^2 + \frac{\gamma(2-\gamma)}{2}(\|\mathbf{R}_0^t\|_{\mathrm{F}}^2 + \|\mathbf{Z}_0^t\|_{\mathrm{F}}^2).$$

References

Bai, Z. and Yin, Y. (1993). Limit of the smallest eigenvalue of a large dimensional sample covariance matrix. *The Annals of Probability* **21** 1275–1294.

Banerjee, O., El Ghaoui, L. and d'Aspremont, A. (2008). Model selection through sparse maximum likelihood estimation for multivariate gaussian or binary data. *The Journal of Machine Learning Research* **9** 485–516.

Batsidis, A., Martin, N., Pardo, L. and Zografos, K. (2014). A necessary power divergence-type family of tests for testing elliptical symmetry. *Journal of Statistical Computation and Simulation* **84** 57–83.

Belloni, A. and Chernozhukov, V. (2011). ℓ_1-penalized quantile regression in high-dimensional sparse models. *The Annals of Statistics* **39** 82–130.

Berk, J. B. (1997). Necessary conditions for the CAPM. *Journal of Economic Theory* **73** 245–257.

Bickel, P. J. and Levina, E. (2008a). Covariance regularization by thresholding. *The Annals of Statistics* **36** 2577–2604.

Bickel, P. J. and Levina, E. (2008b). Regularized estimation of large covariance matrices. *The Annals of Statistics* **36** 199–227.
URL http://dx.doi.org/10.1214/009053607000000758

Bradley, B. O. and Taqqu, M. S. (2003). Financial risk and heavy tails. *Handbook of Heavy Tailed Distributions in Finance* 35–103.

Cai, T. and Liu, W. (2011). A direct estimation approach to sparse linear discriminant analysis. *Journal of the American Statistical Association* **106** 1566–1577.

Cai, T., Liu, W. and Luo, X. (2011). A constrained ℓ_1 minimization approach to sparse precision matrix estimation. *Journal of the American Statistical Association* **106** 594–607.

Cai, T. T., Liu, W. and Zhou, H. H. (2014). Estimating sparse precision matrix: Optimal rates of convergence and adaptive estimation. *The Annals of Statistics (to appear)* .

Cai, T. T., Zhang, C.-H. and Zhou, H. H. (2010). Optimal rates of convergence for covariance matrix estimation. *The Annals of Statistics* **38** 2118–2144.

Cai, T. T. and Zhou, H. H. (2012a). Minimax estimation of large covariance matrices under ℓ_1 norm. *Statistica Sinica* **22** 1319–1378.

Cai, T. T. and Zhou, H. H. (2012b). Optimal rates of convergence for sparse covariance matrix estimation. *The Annals of Statistics* **40** 2389–2420.

Catoni, O. (2012). Challenging the empirical mean and empirical variance: a deviation study. *Annales de l'Institut Henri Poincaré, Probabilités et Statistiques* **48** 1148–1185.

Chen, W.-K. (2002). *The Circuits and Filters Handbook*. CRC Press.

Croux, C. and Ruiz-Gazen, A. (2005). High breakdown estimators for principal components: the projection-pursuit approach revisited. *Journal of Multivariate Analysis* **95** 206–226.

Donoho, D. and Jin, J. (2004). Higher criticism for detecting sparse heterogeneous mixtures. *The Annals of Statistics* **32** 962–994.

El Karoui, N. (2008). Operator norm consistent estimation of large-dimensional sparse covariance matrices. *The Annals of Statistics* **36** 2717–2756.
URL http://dx.doi.org/10.1214/07-AOS559

Embrechts, P., McNeil, A. and Straumann, D. (2002). Correlation and dependence in risk management: properties and pitfalls. *Risk Management: Value at Risk and Beyond* 176–223.

Fan, J. and Fan, Y. (2008). High-dimensional classification using features annealed independence rules. *The Annals of Statistics* **36** 2605–2637.
URL http://dx.doi.org/10.1214/07-AOS504

Fan, J., Feng, Y. and Tong, X. (2012). A ROAD to classification in high dimensional space. *Journal of the Royal Statistical Society. Series B, Statistical Methodology* **74** 745–771.

Fan, J., Han, F. and Liu, H. (2014). PAGE: robust pattern guided estimation of large covariance matrix. *technical report* .

Fan, J., Ke, T., Liu, H. and Xia, L. (2013a). QUADRO: A supervised dimension reduction method via rayleigh quotient optimization. *arXiv preprint arXiv:1311.5542* .

Fan, J., Liao, Y. and Mincheva, M. (2013b). Large covariance estimation by thresholding principal orthogonal complements. *Journal of the Royal Statistical Society: Series B, Statistical Methodology* **75** 603–680.

Fang, H.-B., Fang, K.-T. and Kotz, S. (2002). The meta-elliptical distributions with given marginals. *Journal of Multivariate Analysis* **82** 1–16.

Fang, K.-T., Kotz, S. and Ng, K. W. (1990). *Symmetric Multivariate and Related Distributions Monographs on Statistics and Applied Probability*. London: Chapman and Hall Ltd. MR1071174.

Frahm, G. (2004). *Generalized elliptical distributions: theory and applications*. Ph.D. thesis, Universität zu Köln.

Friedman, J., Hastie, T. and Tibshirani, R. (2008). Sparse inverse covariance estimation with the graphical Lasso. *Biostatistics* **9** 432–441.

Frontera-Pons, J., Mahot, M., Ovarlez, J. P., Pascal, F., Pang, S. K. and Chanussot, J. (2012). A class of robust estimates for detection in hyperspectral images using elliptical distributions background. In *Geoscience and Remote Sensing Symposium (IGARSS), 2012 IEEE International*. IEEE.

Genton, M. G. and Ma, Y. (1999). Robustness properties of dispersion estimators. *Statistics and Probability Letters* **44** 343–350.

Gnanadesikan, R. and Kettenring, J. R. (1972). Robust estimates, residuals, and outlier detection with multiresponse data. *Biometrics* **28** 81–124.

Guo, Y., Hastie, T. and Tibshirani, R. (2005). Regularized discriminant analysis and its application in microarrays. *Biostatistics* **1** 1–18.

Hall, P. and Jin, J. (2010). Innovated higher criticism for detecting sparse signals in correlated noise. *The Annals of Statistics* **38** 1686–1732.

Hampel, F. R. (1974). The influence curve and its role in robust estimation. *Journal of the American Statistical Association* **69** 383–393.

Han, F. and Liu, H. (2012a). Semiparametric principal component analysis. In *Advances in Neural Information Processing Systems*.

Han, F. and Liu, H. (2012b). Transelliptical component analysis. In *Advances in Neural Information Processing Systems 25*.

Han, F. and Liu, H. (2013a). Optimal rates of convergence for latent generalized correlation matrix estimation in transelliptical distribution. *arXiv preprint arXiv:1305.6916* .

Han, F. and Liu, H. (2013b). Principal component analysis on non-Gaussian dependent data. In *International Conference on Machine Learning*, vol. 30.

Han, F. and Liu, H. (2014). Scale-invariant sparse PCA on high dimensional Meta-Elliptical data. *Journal of the American Statistical Association* **109** 275–287.

Han, F., Zhao, T. and Liu, H. (2013). CODA: high dimensional copula discriminant analysis. *The Journal of Machine Learning Research* **14** 629–671.

Hoeffding, W. (1963). Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association* **58** 13–30.

Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian journal of statistics* **6** 65–70.

Huber, P. J. and Ronchetti, E. (2009). *Robust Statistics, 2nd Edition*. Wiley.

Huffer, F. W. and Park, C. (2007). A test for elliptical symmetry. *Journal of Multivariate Analysis* **98** 256–281.

Johnstone, I. M. (2001). On the distribution of the largest eigenvalue in principal components analysis. *The Annals of Statistics* **29** 295–327.

Johnstone, I. M. and Lu, A. Y. (2009). On consistency and sparsity for principal components analysis in high dimensions. *Journal of the American Statistical Association* **104** 682–693.

Koltchinskii, V. and Sakhanenko, L. (2000). Testing for ellipsoidal symmetry of a multivariate distribution. In *High Dimensional Probability II*. Springer, 493–510.

Lam, C. and Fan, J. (2009). Sparsistency and rates of convergence in large covariance matrix estimation. *The Annals of Statistics* **37** 4254–4278.
URL http://www.jstor.org/stable/25662231

Li, R.-Z., Fang, K.-T. and Zhu, L.-X. (1997). Some QQ probability plots to test spherical and elliptical symmetry. *Journal of Computational and Graphical Statistics* **6** 435–450.

Liu, H., Han, F., Yuan, M., Lafferty, J. and Wasserman, L. (2012a). High-dimensional semiparametric Gaussian copula graphical models. *The Annals of Statistics* **40** 2293–2326.

Liu, H., Han, F. and Zhang, C.-H. (2012b). Transelliptical graphical models. In *Advances in Neural Information Processing Systems*.

Liu, H., Lafferty, J. and Wasserman, L. (2009). The nonparanormal: Semiparametric estimation of high dimensional undirected graphs. *The Journal of Machine Learning Research* **10** 2295–2328.

Liu, H., Wang, L. and Zhao, T. (2013). Sparse covariance matrix estimation with eigenvalue constraints. *Journal of Computational and Graphical Statistics (to appear)* .

Lounici, K. (2014). High-dimensional covariance matrix estimation with missing observations. *Bernoulli (to appear)* .

Ma, Y. and Genton, M. G. (2001). Highly robust estimation of dispersion matrices. *Journal of Multivariate Analysis* **78** 11–36.

Marden, D. B. and Manolakis, D. G. (2004). Using elliptically contoured distributions to model hyperspectral imaging data and generate statistically similar synthetic data. In *Defense and Security*. International Society for Optics and Photonics.

Maronna, R. A., Martin, R. D. and Yohai, V. J. (2006). *Robust Statistics: Theory and Methods*. J. Wiley.

Maronna, R. A. and Zamar, R. H. (2002). Robust estimates of location and dispersion for high-dimensional datasets. *Technometrics* **44** 307–317.

McCall, M. N., Bolstad, B. M. and Irizarry, R. A. (2010). Frozen robust multiarray analysis (fRMA). *Biostatistics* **11** 242–253.

McNeil, A. J., Frey, R. and Embrechts, P. (2010). *Quantitative Risk Management: Concepts, Techniques, and Tools*. Princeton University Press.

Meinshausen, N. and Bühlmann, P. (2006). High-dimensional graphs and variable selection with the lasso. *The Annals of Statistics* **34** 1436–1462.
URL http://dx.doi.org/10.1214/009053606000000281

Owen, J. and Rabinovitch, R. (1983). On the class of elliptical distributions and their applications to the theory of portfolio choice. *The Journal of Finance* **38** 745–752.

Rothman, A. J., Levina, E. and Zhu, J. (2009). Generalized thresholding of large covariance matrices. *Journal of the American Statistical Association* **104** 177–186.

Rousseeuw, P. J. and Croux, C. (1993). Alternatives to the median absolute deviation. *Journal of the American Statistical Association* **88** 1273–1283.

Sakhanenko, L. (2008). Testing for ellipsoidal symmetry: A comparison study. *Computational Statistics and Data Analysis* **53** 565–581.

Shao, J., Wang, Y., Deng, X. and Wang, S. (2011). Sparse linear discriminant analysis by thresholding for high dimensional data. *The Annals of Statistics* **39** 1241–1265.
URL http://dx.doi.org/10.1214/10-AOS870

Tsybakov, A. B. (2009). *Introduction to Nonparametric Estimation*. Springer New York.

Vu, V. Q., Cho, J., Lei, J. and Rohe, K. (2013). Fantope projection and selection: A near-optimal convex relaxation of sparse PCA. In *Advances in Neural Information Processing Systems*.

Wang, L. (2013). L1 penalized LAD estimator for high dimensional linear regression. *Journal of Multivariate Analysis* **120** 135–151.

Wang, L., Wu, Y. and Li, R. (2012). Quantile regression for analyzing heterogeneity in ultra-high dimension. *Journal of the American Statistical Association* **107** 214–222.

Wang, Z., Liu, H. and Zhang, T. (2014). Optimal computational and statistical rates of convergence for sparse nonconvex learning problems. *The Annals of Statistics* **42** 2164–2201.
URL http://dx.doi.org/10.1214/14-AOS1238

Xu, M. H. and Shao, H. (2012). Solving the matrix nearness problem in the maximum norm by applying a projection and contraction method. *Advances in Operations Research* **2012** 1–15.

Xue, L., Ma, S. and Zou, H. (2012). Positive-definite ℓ_1-penalized estimation of large covariance matrices. *Journal of the American Statistical Association* **107** 1480–1491.

Xue, L. and Zou, H. (2012). Regularized rank-based estimation of high-dimensional nonparanormal graphical models. *The Annals of Statistics* **40** 2541–2571.

Yuan, M. (2010). High dimensional inverse covariance matrix estimation via linear programming. *The Journal of Machine Learning Research* **11** 2261–2286.

Yuan, M. and Lin, Y. (2007). Model selection and estimation in the gaussian graphical model. *Biometrika* **94** 19–35.