Marchenko-Pastur Law for Tyler's and Maronna's M-estimators

Teng Zhang
The Program in Applied and Computational Mathematics (PACM), Princeton University
Princeton, New Jersey 08544, USA
tengz@princeton.edu

Xiuyuan Cheng
Département d'informatique, École normale supérieure, 45 rue d'Ulm, Paris 75005, France
xiuyuan.cheng@ens.fr

Amit Singer
Department of Mathematics and PACM, Princeton University
Princeton, New Jersey 08544, USA
amits@math.princeton.edu

January 16, 2014

Abstract

This paper studies the limiting behavior of Tyler's and Maronna's M-estimators, in the regime that the number of samples n and the dimension p both go to infinity, and p/n converges to a constant y with $0 < y < 1$. We prove that when the data samples are identically and independently generated from the Gaussian distribution $N(\mathbf{0}, \mathbf{I})$, the difference between the sample covariance matrix and a scaled version of Tyler's M-estimator or Maronna's M-estimator tends to zero in spectral norm, and the empirical spectral densities of both estimators converge to the Marchenko-Pastur distribution. We also extend this result to elliptical-distributed data samples for Tyler's M-estimator and non-isotropic Gaussian data samples for Maronna's M-estimator.

1 Introduction

Many statistical estimators and signal processing algorithms are based on the sample covariance matrix of the input, which is defined to be $\boldsymbol{S}_n = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{x}_i \boldsymbol{x}_i^T$ when the input data points are $\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_n \in \mathbb{R}^p$. Due to the importance of the sample covariance matrix, its asymptotic spectral properties at the limit of infinite number of samples have been well studied. A noticeable example is the case of the sample covariance matrix of n i.i.d Gaussian random vectors in $\mathbb{R}^p$. Denoting the eigenvalues of $\boldsymbol{S}_n$ by $\lambda_1(\boldsymbol{S}_n), \lambda_2(\boldsymbol{S}_n), \cdots, \lambda_n(\boldsymbol{S}_n)$, the Marchenko-Pastur law [13] states that the distribution of the eigenvalues of empirical covariance matrix, i.e. the empirical spectral density

$$f(\lambda) = \frac{1}{n}\sum_{i=1}^{n} \delta_{\lambda_i(\boldsymbol{S}_n)}(\lambda)$$

converges in distribution to a deterministic distribution, known as the Marchenko-Pastur distribution, when $p, n \to \infty$ and $p/n \to y$.

In many applications, one needs to use robust estimators for data sets sampled from distributions with heavy tails or outliers. A commonly used robust estimator of covariance is Maronna's M-estimator [12], which is defined as the solution to the equation

$$\Sigma = \frac{1}{n}\sum_{i=1}^{n} u(\boldsymbol{x}_i^T \Sigma^{-1} \boldsymbol{x}_i)\boldsymbol{x}_i \boldsymbol{x}_i^T, \text{ where } u : (0,\infty) \to [0,\infty) \tag{1}$$

Another interesting robust covariance estimator is Tyler's M-estimator [16], which is a special case of Maronna's M-estimator with the choice $u(x) = \frac{p}{x}$. It is shown to be the most robust estimator of the covariance matrix of an elliptical distribution in the sense of minimizing the maximum asymptotic variance. Therefore, Tyler's M-estimator has been used to replace the empirical sample covariance in many applications such as anomaly detection in wireless sensor networks [4], antenna array processing [14] and radar detection [15].

The limiting empirical spectral density of Maronna's M-estimator when both $p, n \to \infty$ and $p/n \to y$ has been analyzed in two recent works [5, 6], which prove that a properly scaled Maronna's M-estimator converges to the sample covariance matrix in terms of operator norm under some assumptions of $u(x)$ and the distribution of data samples.

For Tyler's M-estimator, the original work by Tyler studied the case when p is fixed and n goes to infinity [16, Theorem 3.2, Theorem 4.2], which is the standard setting in classical statistics. Some later works focused on the case $p, n \to \infty$ and $p/n \to 0$: Dümbgen [7] showed that the conditional number of Tyler's estimator converges to $1 + O(\sqrt{p/n})$. Frahm and Glombek [8] showed that the empirical spectral distribution of $\sqrt{n/p}(\hat{\boldsymbol{\Sigma}} - \mathbf{I})$ converges to the a semicircle distribution. However, modern applications involve high-dimensional data for which n and p are of the same order. Yet, no result for the setting $p, n \to \infty$ and $p/n \to y$ has been obtained, although it has been conjectured that the

empirical density distribution follows the Marchenko-Pastur distribution [9, 6]. We note that Tyler's M-estimator was not included in the analysis of [5, 6] because their method depends on the strict monotonicity of $xu(x)$, which is a constant for Tyler's M-estimator since $xu(x) = p$.

The main contribution of this paper are Theorem 3.5 and Corollary 3.7 that prove the conjecture that as $p, n \to \infty$ and $p/n \to y$, $0 < y < 1$, the empirical spectral density of a properly scaled Tyler's M-estimator converges to the Marchenko-Pastur distribution $\rho_{\mathrm{MP}}(x)$, defined by

$$\rho_{\mathrm{MP}}(x) = \frac{1}{2\pi} \frac{y\sqrt{(y_+ - x)(x - y_-)}}{x} \mathbf{1}_{[y_-, y_+]}, \text{ where } y_\pm = (1 \pm \sqrt{y})^2. \quad (2)$$

Our paper and [5, 6] are similar in the sense that the proofs are based on the representation of M-estimator as a weighted sum of $\boldsymbol{x}_i \boldsymbol{x}_i^T$, and the uniform convergence of these weights. However, we give a different proof for the convergence of the weights, by considering the weights as the solution to a system of equations, while the proofs of [5, 6] are based on an iteratively reweighted algorithm. In comparison, our approach can handle Tyler's M-estimator and some Maronna's M-estimators (i.e., some functions $u(x)$) that are not covered in [5, 6]. We remark that while some Lemmas and technical proofs are also covered in [5, 6] (for example, Lemma 5.2 and the analysis in the proof of Theorem 3.2 are similar to [5, Lemma 2], [6, Lemma 6] and proof of Theorem 1 in [5]), we still include them for the completeness of the paper.

Based on the properties of Tyler's and Maronna's M-estimators, this paper also analyzes the empirical spectral density when data samples are i.i.d. drawn from other distributions such as elliptical distributions. In addition, we give estimates for the convergence rates of the empirical density function and the largest eigenvalue of the Tyler's M-estimator as $p, n \to \infty$, $p/n \to y$.

The rest of the paper is organized as follows. In Section 2 we provide the definition of Tyler's and Maronna's M-estimators and state some of their properties, such as existence and uniqueness, and we introduce their representations by linear combinations of $\boldsymbol{x}_i \boldsymbol{x}_i^T$. In Section 3 we present the main results that when data set is i.i.d. sampled from Gaussian distribution $N(\mathbf{0}, \mathbf{I})$, properly scaled Tyler's and Maronna's M-estimators converge to the sample covariance in operator norm, and the limiting empirical spectral density of Tyler's M-estimator follows the Marchenko-Pastur law. We also extend the result to elliptical distributions for Tyler's M-estimator and non-isotropic Gaussian distributions for Maronna's M-estimator. The technical proofs are given in Section 5.

As for notations, we will use c, c', C, C' to denote any fixed constants as $p, n \to \infty$ (though they may depend on y). Depending on the context, they might denote different values in different equations.

2 Properties of Tyler's and Maronna's M-estimators

By the fixed-point algorithm [19, (1.2)], Tyler's M-estimator can also be defined by any $\mathbf{\Sigma}$ satisfying

$$\sum_{i=1}^{n} \frac{\boldsymbol{x}_i \boldsymbol{x}_i^T}{\boldsymbol{x}_i^T \mathbf{\Sigma}^{-1} \boldsymbol{x}_i} = c\mathbf{\Sigma}, \text{ for some } c > 0. \tag{3}$$

When $\mathrm{span}(\{\boldsymbol{x}_i\}_{i=1}^n) = \mathbb{R}^p$, Tyler's M-estimator exists and is unique up to a scaling [19, Theorem 1.1]: it is easy to verify that for any solution to (3), its scaled version is another solution. For the rest of the paper we denote Tyler's M-estimator by $\hat{\Sigma}$ and ensure it uniqueness by fixing its trace to be 1, that is, we assume $\mathrm{tr}(\hat{\Sigma}) = 1$.

As for Maronna's M-estimator, for the convenience of analysis we define it slightly different from the literature by removing the factor $1/n$ from (1), or equivalently, replace $u(x)$ by $\frac{1}{n}u(x)$ in (1):

$$\Sigma = \sum_{i=1}^{n} u(\boldsymbol{x}_i^T \Sigma^{-1} \boldsymbol{x}_i) \boldsymbol{x}_i \boldsymbol{x}_i^T, \tag{4}$$

and we note that a similar modification has also been applied in [5, 6].

The existence and uniqueness of Maronna's M-estimator has been analyzed in [11, 20], by analyzing the minimizer of the objective function

$$L(\Sigma) = \sum_{i=1}^{n} \rho(\boldsymbol{x}_i^T \Sigma^{-1} \boldsymbol{x}_i) + \frac{n}{2} \log\det \Sigma, \text{ where } \rho'(x) = nu(x)/2.$$

We remark that the derivative of $L(\Sigma)$ with respect to Σ^{-1} is

$$\frac{\mathrm{d}}{\mathrm{d}\Sigma^{-1}} L(\Sigma) = \sum_{i=1}^{n} \frac{n}{2} u(\boldsymbol{x}_i^T \Sigma^{-1} \boldsymbol{x}_i) \boldsymbol{x}_i \boldsymbol{x}_i^T - \frac{n}{2}\Sigma,$$

whose roots give solutions to (4).

By analyzing the geodesic convexity of $L(\Sigma)$, [20, Theorem 1] states that the uniqueness of the minimizer of $L(\Sigma)$ is guaranteed when $\rho(x)$ is continuous in $(0, \infty)$, nondecreasing and $\rho(e^x)$ is convex [20, Theorem 1], and the minimizer of $L(\Sigma)$ exists when $a_1 = \sup\{a | x^{a/2} \exp(-\rho(x)) \to 0 \text{ as } x \to \infty\}$ is positive [11, Theorem 2.3] (when $\lim_\infty xu(x)$ exists, $a_1 = n \lim_{x\to\infty} xu(x)$),[1] and

$$\frac{|\{\boldsymbol{x}_i\}_{i=1}^n \cap \mathrm{V}|}{n} < 1 - \frac{p - \dim(\mathrm{V})}{a_1} \text{ for any linear subspace } \mathrm{V} \in \mathbb{R}^p. \tag{5}$$

[1] It follows from the comment after [11, Definition 2.1]. We remark that $u(x)$ in [11] should be replaced by $nu(x)$, since we use (4) over the standard definition (1). This also explains our choice of $\rho'(x) = nu/2$ instead of $\rho'(x) = u/2$ used in [11]. We remark that there is a typo after [11, Definition 2.1], where "$\rho'(x) = 2u(x)$" should be replaced by "$\rho'(x) = u(x)/2$".

When the underlying distribution of $\{\boldsymbol{x}\}_{i=1}^n$ does not concentrate on any subspace (i.e., the measure of any subspace is 0), then the LHS of (5) is bounded above by $\frac{\dim(V)}{n}$ almost surely and (5) becomes

$$\frac{d}{n} < 1 - \frac{p-d}{a_1} \text{ for any } 1 \le d \le p-1. \tag{6}$$

Applying $a_1 = n \lim_{x\to\infty} xu(x)$, (6) holds for $p, n \to \infty$ when $\lim_{x\to\infty} xu(x) > y$.

When the above condition holds, the minimizer of $L(\Sigma)$ exists, and the minimizer is also a solution to (4). Due to the geodesic convexity of $L(\Sigma)$ [20, Theorem 1], the minimizer of $L(\Sigma)$ is unique, and any solution to (4) is also a minimizer of $L(\Sigma)$. Therefore, the solution to (4) is also unique. That is, under the above assumptions on $\rho(x)$, we have the existence and uniqueness of the solution to (4).

Since uniqueness and existence of both Maronna's M-estimator and Tyler's M-estimator requires $\text{span}(\{\boldsymbol{x}_i\}_{i=1}^n) = \mathbb{R}^p$, we let $y < 1$ throughout the paper.

The analysis for Tyler's and Maronna's M-estimators in this paper is based on the following representations, whose proofs are deferred to Section 5:

Lemma 2.1. *Tyler's M-estimator can be written as*

$$\hat{\Sigma} = \sum_{i=1}^n \hat{w}_i \boldsymbol{x}_i \boldsymbol{x}_i^T \Big/ \text{tr}\Big(\sum_{i=1}^n \hat{w}_i \boldsymbol{x}_i \boldsymbol{x}_i^T\Big), \tag{7}$$

where $\{\hat{w}_i\}_{i=1}^n$ are uniquely defined by

$$(\hat{w}_1, \hat{w}_2, \cdots, \hat{w}_n) = \underset{\sum_{i=1}^n w_i = 1}{\arg\min} -\sum_{i=1}^n \log w_i + \frac{n}{p} \log\det\big(\sum_{i=1}^n w_i \boldsymbol{x}_i \boldsymbol{x}_i^T\big). \tag{8}$$

Lemma 2.2. *When Maronna's M-estimator exists and is unique, any $\{\bar{w}_i\}_{i=1}^n$ satisfying*

$$\bar{w}_j = \boldsymbol{x}_j^T \Big(\sum_{i=1}^n u(\bar{w}_i) \boldsymbol{x}_i \boldsymbol{x}_i^T\Big)^{-1} \boldsymbol{x}_j, \text{ for } j = 1, 2, \cdots, n \tag{9}$$

gives Maronna's M-estimator by

$$\bar{\Sigma} = \sum_{i=1}^n u(\bar{w}_i) \boldsymbol{x}_i \boldsymbol{x}_i^T . \tag{10}$$

3 Main Results

In this section we present the main results: we prove the convergence of Tyler's and Maronna's M-estimators to the sample covariance matrix under the Gaussian model $N(\mathbf{0}, \mathbf{I})$ in terms of the operator norm in Section 3.1, and then extend the result to elliptical distributions/non-isotropic Gaussian distributions in Section 3.2. Based on the convergence, we obtain the limiting empirical density distributions of Tyler's and Maronna's M-estimators in Section 3.3.

3.1 Isotropic Gaussian Distribution

3.1.1 Tyler's M-estimator

In this section, we assume that $\{\boldsymbol{x}_i\}_{i=1}^n \subset \mathbb{R}^p$ are i.i.d. drawn from $N(\mathbf{0}, \mathbf{I})$. The main result, Theorem 3.2, characterizes the convergence and convergence rate of Tyler's M-estimator to the sample covariance in terms of the operator norm. Its proof applies Lemma 3.1, whose proof is rather technical and therefore in Section 5.

Lemma 3.1. *If $\boldsymbol{x}_i \sim N(\mathbf{0}, \mathbf{I})$ for all $1 \le i \le n$, then $\max_{1\le i\le n} |n\,\hat{w}_i - 1|$ converges to 0 almost surely as $p, n \to \infty$. In particular, there exists $C, c, c' > 0$ such that for any $\varepsilon < c'$,*

$$\Pr\left(\max_{1\le i\le n} |n\,\hat{w}_i - 1| \le \varepsilon\right) \ge 1 - Cne^{-c\varepsilon^2 n}. \tag{11}$$

Theorem 3.2. *Suppose that $\{\boldsymbol{x}_i\}_{i=1}^n$ are i.i.d. sampled from $N(\mathbf{0}, \mathbf{I})$, $p, n \to \infty$ and $p/n = y$, where $0 < y < 1$, and $\boldsymbol{x}_i \sim N(\mathbf{0}, \mathbf{I})$ for all $1 \le i \le n$, then a scaled Tyler's M-estimator converges to the sample covariance in operator norm almost surely, and there exist $C, c, c' > 0$ such that for any $\varepsilon < c'$,*

$$\Pr\left(\left\| p\,\hat{\Sigma} - \frac{1}{n}\sum_{i=1}^n \boldsymbol{x}_i\boldsymbol{x}_i^T \right\| \le \varepsilon\right) \ge 1 - Cne^{-c\varepsilon^2 n}. \tag{12}$$

The strategy of the proof for Theorem 3.2 is as follows. According to Lemma 2.1, a scaled Tyler's M-estimator is a linear combination of $\boldsymbol{x}_i\boldsymbol{x}_i^T$, i.e., it can be written as $\sum_{i=1}^n \hat{w}_i\boldsymbol{x}_i\boldsymbol{x}_i^T$ (up to scale). Then Lemma 3.1 shows that $n\hat{w}_i$ converges to 1 uniformly, and based on the following matrix analysis, Theorem 3.2 can be concluded.

Proof of Theorem 3.2. We first prove that for $\varepsilon < c'$,

$$\Pr\left(\left\| \sum_{i=1}^n \hat{w}_i\boldsymbol{x}_i\boldsymbol{x}_i^T - \frac{1}{n}\sum_{i=1}^n \boldsymbol{x}_i\boldsymbol{x}_i^T \right\| \le \varepsilon\right) \ge 1 - Cne^{-c\varepsilon^2 n}. \tag{13}$$

Let $\boldsymbol{B}_n = \sum_{i=1}^n (\hat{w}_i - \frac{1}{n})\boldsymbol{x}_i\boldsymbol{x}_i^T = \sum_{i=1}^n \hat{w}_i\boldsymbol{x}_i\boldsymbol{x}_i^T - \frac{1}{n}\sum_{i=1}^n \boldsymbol{x}_i\boldsymbol{x}_i^T$, then

$$\begin{aligned}\|\boldsymbol{B}_n\| &= \sup_{\|\boldsymbol{v}\|=1} \boldsymbol{v}^T\boldsymbol{B}_n\boldsymbol{v} = \sup_{\|\boldsymbol{v}\|=1} \sum_{i=1}^n (\hat{w}_i - \frac{1}{n})(\boldsymbol{v}^T\boldsymbol{x}_i)^2 \\ &\le \sup_{\|\boldsymbol{v}\|=1} \sum_{i=1}^n \left\|\hat{\boldsymbol{w}} - \frac{1}{n}\mathbf{1}\right\|_\infty (\boldsymbol{v}^T\boldsymbol{x}_i)^2 \le \|n\hat{\boldsymbol{w}} - \mathbf{1}\|_\infty \|\frac{1}{n}\sum_{i=1}^n \boldsymbol{x}_i\boldsymbol{x}_i^T\|.\end{aligned}$$

Since $\|n\hat{\boldsymbol{w}} - \mathbf{1}\|_\infty \to 0$ with probability estimated in (11), and $\|\frac{1}{n}\sum_{i=1}^n \boldsymbol{x}_i\boldsymbol{x}_i^T\|$ is bounded above by $(1 + 2\sqrt{y})^2$ with probability $1 - C\exp(-cn)$ [10, Theorem II.13], (13) is proved.

Second, since $\|\sum_{i=1}^n \hat{w}_i \boldsymbol{x}_i \boldsymbol{x}_i^T\| \le \left\|\sum_{i=1}^n \hat{w}_i \boldsymbol{x}_i \boldsymbol{x}_i^T - \frac{1}{n}\sum_{i=1}^n \boldsymbol{x}_i \boldsymbol{x}_i^T\right\| + \left\|\frac{1}{n}\sum_{i=1}^n \boldsymbol{x}_i \boldsymbol{x}_i^T\right\|$,

$$\Pr\left(\|\sum_{i=1}^n \hat{w}_i \boldsymbol{x}_i \boldsymbol{x}_i^T\| < C'\right) > 1 - Cn\exp(-cn). \tag{14}$$

Besides, $\mathrm{tr}(\sum_{i=1}^n \hat{w}_i \boldsymbol{x}_i \boldsymbol{x}_i^T) = \sum_{i=1}^n \hat{w}_i \boldsymbol{x}_i^T \boldsymbol{x}_i \to p$ in the same rate as in (14): applying the concentration of high-dimensional Gaussian measure on the sphere [2, Corollary 2.3], we have

$$\max\left(\Pr\left(\sum_{i=1}^n \hat{w}_i \boldsymbol{x}_i^T \boldsymbol{x}_i < p(1-\varepsilon)\right), \Pr\left(\sum_{i=1}^n \hat{w}_i \boldsymbol{x}_i^T \boldsymbol{x}_i > p/(1-\varepsilon)\right)\right) \tag{15}$$
$$\le \max\left(\Pr\left(\min_{1\le i\le n} \|\boldsymbol{x}_i\|^2 < p(1-\varepsilon)\right), \Pr\left(\max_{1\le i\le n} \|\boldsymbol{x}_i\|^2 > p/(1-\varepsilon)\right)\right) < ne^{-\varepsilon^2 p/4}.$$

Combining (14), (15) and (7),

$$\left\|\sum_{i=1}^n \hat{w}_i \boldsymbol{x}_i \boldsymbol{x}_i^T - p\hat{\Sigma}\right\| = \left\|\sum_{i=1}^n \hat{w}_i \boldsymbol{x}_i \boldsymbol{x}_i^T\right\| \left(1 - p/\mathrm{tr}(\sum_{i=1}^n \hat{w}_i \boldsymbol{x}_i \boldsymbol{x}_i^T)\right) \tag{16}$$

converges in the same rate as specified in (13). (12) is then proved by combining (13), (16) and the triangle inequality. □

From the probabilistic estimation (12) we obtain a convergence rate of $O(\sqrt{\log n/n})$. In simulations we observe a rate of $O(1/\sqrt{n})$, which means our estimation might be off by a factor of $\sqrt{\log n}$.

3.1.2 Maronna's M-estimator

In this section we first state our assumptions for $u(x)$ in (4):

A1. $u : [0,\infty) \to (0,\infty)$ is nonnegative, $\psi(x) = xu(x)$ is increasing and $\lim_{x\to\infty} \psi(x) > y$.

A2. $u(x)$ is twice differentiable, and $xu'(x) < u(x)$.

We require assumption **A1** to ensure the existence and uniqueness of Maronna's M-estimator so that Lemma 2.2 can be applied. When $u(x)$ is nonnegative and $\psi(x) = xu(x)$ is increasing, the uniqueness condition in Section 2, i.e., $\rho(x)$ is non-decreasing and $\rho(e^x)$ is convex, are guaranteed (recall $\rho'(x) = nu(x)/2$). And the condition $\lim_{x\to\infty} \psi(x) > y$ guarantees the existence of Maronna's M-estimator as $p, n \to \infty$, as discussed in Section 2.

We require assumption **A2** for some technical steps in our proof, though we conjecture that our results about Maronna's M-estimator in this paper will still hold without this assumption.

Here we compare our assumption of $u(x)$ with the assumption in [5, 6]. Since $\big(Z, u(x)\big)$ in [5, 6] is equivalent to $\big(p\bar{\Sigma}, yu(x)\big)$ in our setting, their assumptions of $u(x)$ can be translated to:

- $u(x)$ is nonnegative, continuous and increasing.

- $\psi(x)$ is increasing and bounded, and $\frac{1}{y} < \lim_{x\to\infty} \psi(x) < \frac{1}{y^2}$.

There are three main differences between the assumptions of $u(x)$, and our assumptions allow some $u(x)$ that was not covered in their work. First, our assumption of $\lim_{x\to\infty} \psi(x)$ is less restrictive and allows it to be infinity. As a consequence, our theory allows some commonly used $u(x)$ such as $u(x) = x^\beta$ (see [20]). Second, our assumption is less restrictive in the sense that we replaced the assumption "$u(x)$ is nonincreasing" (i.e., $u'(x) \le 0$) by $xu'(x) < u(x)$. However, our assumption on the twice differentiability of $u(x)$ is more restrictive.

Based on these assumptions, we obtain the convergence of Maronna's M-estimator to a scaled version of the sample covariance matrix in operator norm.

Lemma 3.3. *If $\boldsymbol{x}_i \sim N(\mathbf{0}, \mathbf{I})$ for all $1 \le i \le n$, let $\psi(x) = xu(x)$, then there exists a solution $\{\bar{w}_i\}_{i=1}^n$ to (9) $\max_{1\le i\le n} |\bar{w}_i - \psi^{-1}(\frac{1}{y})|$ converges to 0 almost surely as $p, n \to \infty$. In particular, there exists $C, c, c' > 0$ such that for any $\varepsilon < c'$,*

$$\Pr\left(\max_{1\le i\le n} |\bar{w}_i - \psi^{-1}(1/y)| \le \varepsilon\right) \ge 1 - Cne^{-c\varepsilon^2 n}. \tag{17}$$

Theorem 3.4. *Suppose that $\{\boldsymbol{x}_i\}_{i=1}^n$ are i.i.d. sampled from $N(\mathbf{0}, \mathbf{I})$, $p, n \to \infty$ and $p/n \to y$, where $0 < y < 1$, and $\boldsymbol{x}_i \sim N(\mathbf{0}, \mathbf{I})$ for all $1 \le i \le n$, then a scaled Maronna's M-estimator converges to the sample covariance matrix in operator norm almost surely, and there exist $C, c, c' > 0$ such that for any $\varepsilon < c'$,*

$$\Pr\left(\left\| \frac{1}{n\,\psi^{-1}(1/y)}\,\bar{\Sigma} - \frac{1}{n}\sum_{i=1}^n \boldsymbol{x}_i\boldsymbol{x}_i^T \right\| \le \varepsilon\right) \ge 1 - Cne^{-c\varepsilon^2 n}. \tag{18}$$

3.2 More General Distributions

3.2.1 Tyler's M-estimator

In this section, we extend Theorem 3.2 from the setting of the normal distribution $N(\mathbf{0}, \mathbf{I})$ to elliptical distributions. We say that μ_p is an elliptical distribution, if μ_p can be characterized by $\mu_p(\boldsymbol{x}) = C(g_p)\det(\boldsymbol{T}_p)^{-1/2} g_p(\boldsymbol{x}^T\boldsymbol{T}_p^{-1}\boldsymbol{x})$, where $\boldsymbol{T}_p$ is a positive definite matrix in $\mathbb{R}^{p\times p}$, $g_p : [0,\infty)$ to $[0,\infty)$ satisfies $\int_0^\infty g_p(x)x^{p-1} < \infty$, and $C(g_p)$ is a normalization parameter that only depends on g_p.

When $\boldsymbol{T}_p$ is a scalar matrix, the distribution is isotropic and we call μ_p spherically symmetric distribution.

Our analysis is based on Theorem 3.2 and two properties of Tyler's M-estimator: 1. Tyler's M-estimator is invariant to the scaling of data set, i.e., if $\{\boldsymbol{x}_i\}_{i=1}^n$ are replaced by $\{c_i\boldsymbol{x}_i\}_{i=1}^n$ and $\{c_i\}_{i=1}^n$ are arbitrary numbers in $\mathbb{R}$, then $\hat{\Sigma}$ remains the same. 2. For any non-singular linear operator $\boldsymbol{T} : \mathbb{R}^p \to \mathbb{R}^p$, if Tyler's M-estimator for $\{\boldsymbol{x}_i\}_{i=1}^n$ is $\hat{\Sigma}$, then Tyler's M-estimator for $\{\boldsymbol{T}\boldsymbol{x}_i\}_{i=1}^n$ is $\boldsymbol{T}\hat{\Sigma}\boldsymbol{T}/\mathrm{tr}(\boldsymbol{T}\hat{\Sigma}\boldsymbol{T})$.

Both properties can be obtained by verifying (3). To prove the first propriety, note that the LHS of (3) is unchanged if $\{\boldsymbol{x}_i\}_{i=1}^n$ is replaced by $\{c_i\boldsymbol{x}_i\}_{i=1}^n$. To

prove the second property, one can show that (3) still holds when $\{\boldsymbol{x}_i\}_{i=1}^n$ and $\hat{\Sigma}$ are replaced by $\{\boldsymbol{T}\boldsymbol{x}_i\}_{i=1}^n$ and $\boldsymbol{T}\hat{\Sigma}\boldsymbol{T}/\mathrm{tr}(\boldsymbol{T}\hat{\Sigma}\boldsymbol{T})$.

Theorem 3.5. *If $\{\boldsymbol{x}_i\}_{i=1}^n$ are i.i.d. sampled from elliptical distribution $\mu_p(\boldsymbol{x}) = C(g_p)\det(\boldsymbol{T}_p)^{-1/2}g_p(\boldsymbol{x}^T\boldsymbol{T}_p^{-1}\boldsymbol{x})$, then we have the following property for Tyler's M-estimator: there exist $c, C, c' > 0$ such that for any $\varepsilon < c'$,*

$$\Pr\left(\left\|p\boldsymbol{T}_p^{-1/2}\hat{\Sigma}\boldsymbol{T}_p^{-1/2}/\mathrm{tr}(\boldsymbol{T}_p^{-1/2}\hat{\Sigma}\boldsymbol{T}_p^{-1/2}) - \frac{p}{n}\sum_{i=1}^n \boldsymbol{y}_i\boldsymbol{y}_i^T\right\| \le \varepsilon\right) \ge 1 - Cne^{-c\varepsilon^2 n} \tag{19}$$

for $\boldsymbol{y}_i = \boldsymbol{T}_p^{-1/2}\boldsymbol{x}_i/\|\boldsymbol{T}_p^{-1/2}\boldsymbol{x}_i\|$.

Proof. First, since $\boldsymbol{x}_i$ are i.i.d. sampled from an elliptical distribution with covariance matrix $\boldsymbol{T}_p$, $\boldsymbol{y}_i = \boldsymbol{T}_p^{-1/2}\boldsymbol{x}_i/\|\boldsymbol{T}_p^{-1/2}\boldsymbol{x}_i\|$ are uniformly distributed over the $p-1$-dimensional unit sphere.

If we consider $\boldsymbol{y}_i$ as the projections of $\boldsymbol{x}_i \sim N(\boldsymbol{0}, \mathbf{I})$, the concentration of $N(\boldsymbol{0}, \mathbf{I})$ on the sphere with radius $\sqrt{p}$ [2, Corollary 2.3] and the boundedness of $\|\frac{1}{n}\sum_{i=1}^n \boldsymbol{x}_i\boldsymbol{x}_i^T\|$ from above [10, Theorem II.13] gives that for $\varepsilon < c'$,

$$\Pr\left(\left\|\frac{p}{n}\sum_{i=1}^n \boldsymbol{y}_i\boldsymbol{y}_i^T - \frac{1}{n}\sum_{i=1}^n \boldsymbol{x}_i\boldsymbol{x}_i^T\right\| < \varepsilon\right) <\geq 1 - Cne^{-c\varepsilon^2 n}. \tag{20}$$

Assuming the Tyler's M-estimator for $\{\boldsymbol{y}_i\}_{i=1}^n$ is $\hat{\Sigma}_y$, then Theorem 3.2 and (20) gives

$$\Pr\left(\left\|p\,\hat{\Sigma}_y - \frac{p}{n}\sum_{i=1}^n \boldsymbol{y}_i\boldsymbol{y}_i^T\right\| \le \varepsilon\right) \ge 1 - Cne^{-c\varepsilon^2 n}. \tag{21}$$

Applying Property 1 (scale invariance) of Tyler's M-estimator, $\hat{\Sigma}_y$ is also the Tyler's M-estimator for the set $\{\boldsymbol{T}_p^{-1/2}\boldsymbol{x}_i\}_{i=1}^n$. Applying Property 2,

$$\hat{\Sigma}_y = \boldsymbol{T}_p^{-1/2}\hat{\Sigma}\boldsymbol{T}_p^{-1/2}/\mathrm{tr}(\boldsymbol{T}_p^{-1/2}\hat{\Sigma}\boldsymbol{T}_p^{-1/2}). \tag{22}$$

Combining (21) and (22), Theorem 3.5 is proved. □

3.2.2 Maronna's M-estimator

In this section, we extend Theorem 3.4 from the setting of the normal distribution $N(\boldsymbol{0}, \mathbf{I})$ to non-isotropic Gaussian distributions. The model is more restrictive than the model of Tyler's M-estimator, since Maronna's M-estimator lacks Property 1 (scale invariance) of Tyler's M-estimator. We extend Theorem 3.4 to non-isotropic Gaussian distributions by applying a similar property to the Property 2 of Tyler's M-estimator: For any non-singular linear operator $\boldsymbol{T} : \mathbb{R}^p \to \mathbb{R}^p$, if Maronna's M-estimator for $\{\boldsymbol{x}_i\}_{i=1}^n$ is $\bar{\Sigma}$, then Tyler's M-estimator for $\{\boldsymbol{T}\boldsymbol{x}_i\}_{i=1}^n$ is $\boldsymbol{T}\hat{\Sigma}\boldsymbol{T}$.

Corollary 3.6. *If $\{\boldsymbol{x}_i\}_{i=1}^n$ are i.i.d. sampled from $N(\mathbf{0}, \boldsymbol{T}_p)$, where $\boldsymbol{T}_p$ is a positive definite matrix in $\mathbb{R}^{p\times p}$,. Then we have the following property for Tyler's M-estimator: there exist $c, C, c' > 0$ such that for any $\varepsilon < c'$,*

$$\Pr\left(\left\|\boldsymbol{T}_p^{-1/2}\Big(\frac{1}{n\,\psi^{-1}(1/y)}\,\bar{\Sigma} - \frac{1}{n}\sum_{i=1}^{n}\boldsymbol{x}_i\boldsymbol{x}_i^T\Big)\boldsymbol{T}_p^{-1/2}\right\| \le \varepsilon\right) \ge 1 - Cne^{-c\varepsilon^2 n}. \quad (23)$$

The distributions for data samples in this section can be compared to the model given in [5, Section II] and [6, Assumption 2]. For the simplicity of the discussion we only discuss [6, Assumption 2], which assumes that $\boldsymbol{x}_i \in \mathbb{R}^p$ is defined by $\sqrt{\tau_i}\boldsymbol{A}_N\boldsymbol{y}_i$, where $\boldsymbol{y}_i$ has independent entries with zero mean and unit variance, and τ_i follows from some distribution.

While [6] covers more models than Corollary 3.6, we note that our proof only depends Lemma 5.2. That is, our proof can be applied to any distribution that satisfies Lemma 5.2. Since the distribution in [6] satisfies [6, Lemma 6], which is equivalent to Lemma 5.2, our proof can also be applied to their models.

3.3 Empirical Spectral Density

3.3.1 Tyler's M-estimator

This section investigates the distribution of the eigenvalues of Tyler's M-estimator, i.e., its empirical spectral density. We follow the setting of previous sections and present two corollaries, where the first corollary proves the conjecture proposed in [9] that the empirical spectral density converges to the Marchenko-Pastur distribution when $\{\boldsymbol{x}_i\}_{i=1}^n$ are drawn from $N(\mathbf{0}, \mathbf{I})$, and the second corollary gives the limiting distribution under the setting of elliptical distributions.

Corollary 3.7. *If $\{\boldsymbol{x}_i\}_{i=1}^n$ are i.i.d. sampled from spherically symmetric distributions $C(g_p)g_p(\|\boldsymbol{x}\|^2)$, then the empirical spectral density of $p\hat{\Sigma}$ converges to the following Marchenko-Pastur distribution.*

To visualize Corollary 3.7, we simulated the case $n = 20000$ and $p = 4000$ with Gaussian distribution $N(\mathbf{0}, \mathbf{I})$, and Figure 1 shows that the empirical spectral density of $p\hat{\Sigma}$ is well approximated by the corresponding Marchenko-Pastur distribution.

Lemma 3.8. *Assume a set of matrices $\{\boldsymbol{A}_n\}_{n\ge1}$ with size $k_n \times k_n$, and with empirical spectral density converging to a continuous distribution ρ, and another sequence of matrices $\{\boldsymbol{B}_n\}_{n\ge1}$ such that $\boldsymbol{B}_n$ is also of size $k_n \times k_n$ and $\|\boldsymbol{B}_n\| \to 0$. Then the empirical spectral density of $\{\boldsymbol{A}_n + \boldsymbol{B}_n\}_{n\ge1}$ also converges to ρ.*

Proof of Corollary 3.7. The proof follows from Theorem 3.2 and Lemma 3.8, and the proof of Lemma 3.8 will be given later in Section 5.

First, due to Property 1 in Section 3.2, it suffices to consider the case $\boldsymbol{x}_i \sim N(\mathbf{0}, \mathbf{I})$. Then Corollary 3.7 is proved by combining Theorem 3.2, Lemma 3.8 and the fact that the empirical density of $\frac{1}{n}\sum_{i=1}^n \boldsymbol{x}_i\boldsymbol{x}_i^T$ converges to (2) as $p, n \to \infty$ [13]. □

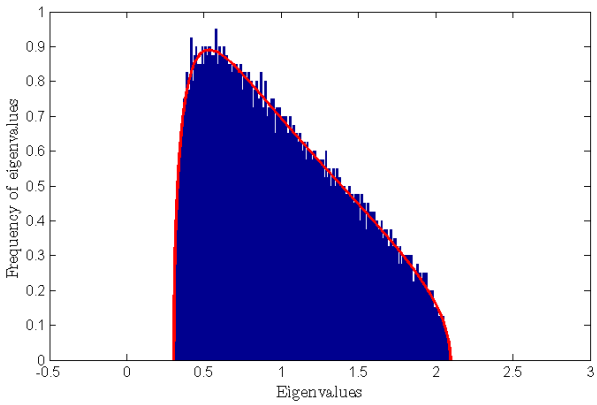

Figure 1: The empirical spectral density when $n = 20000$ and $p = 4000$, and $\{\boldsymbol{x}_i\}_{i=1}^n$ are drawn from $N(\mathbf{0}, \mathbf{I})$ The red ind represents the Marchenko-Pastur distribution for $y = 1/5$.

Next we extend the analysis to general elliptical distributions.

Corollary 3.9. *Suppose $\{\boldsymbol{x}_i\}_{i=1}^n$ are i.i.d. sampled from elliptical distribution $C(g_p)g_p(\boldsymbol{x}^T\boldsymbol{T}_p^{-1}\boldsymbol{x})$, and the empirical spectral density of $\boldsymbol{T}_p$ converges to H. Then the empirical spectral density of $\mathrm{tr}(\boldsymbol{T}_p)\hat{\Sigma}$ converges to ρ, whose Stieltjes transform $s(z)$ satisfies*

$$s(z) = \int \frac{1}{t(1 - y - yz\, s(z)) - z}\, \mathrm{d}H(t). \tag{24}$$

Proof. Let

$$\boldsymbol{B}_p = p\boldsymbol{T}_p^{-1/2}\hat{\Sigma}\boldsymbol{T}_p^{-1/2}/\mathrm{tr}(\boldsymbol{T}_p^{-1/2}\hat{\Sigma}\boldsymbol{T}_p^{-1/2}) - \frac{1}{n}\sum_{i=1}^n \boldsymbol{z}_i\boldsymbol{z}_i^T, \tag{25}$$

where $\boldsymbol{z}_i = h_i \cdot \boldsymbol{y}_i = h_i \cdot \boldsymbol{T}_p^{-1/2}\boldsymbol{x}_i/\|\boldsymbol{T}_p^{-1/2}\boldsymbol{x}_i\|$, and $h_i \sim \sqrt{\chi_p^2}$. Then Theorem 3.5 and the convergence of h_i to $\sqrt{p}$ implies

$$\hat{\Sigma} = \frac{\boldsymbol{T}_p^{1/2}(\frac{1}{n}\sum_{i=1}^n \boldsymbol{z}_i\boldsymbol{z}_i^T + \boldsymbol{B}_p)\boldsymbol{T}_p^{1/2}}{\mathrm{tr}\left(\boldsymbol{T}_p^{1/2}(\frac{1}{n}\sum_{i=1}^n \boldsymbol{z}_i\boldsymbol{z}_i^T + \boldsymbol{B}_p)\boldsymbol{T}_p^{1/2}\right)},$$

where $\|\boldsymbol{B}_p\| \to 0$.

Since $\|\boldsymbol{B}_p\| \to 0$ and $\boldsymbol{z}_i$ are i.i.d from $N(\mathbf{0}, \mathbf{I})$, $\mathrm{tr}\left(\boldsymbol{T}_p^{1/2}(\frac{1}{n}\sum_{i=1}^n \boldsymbol{z}_i\boldsymbol{z}_i^T + \boldsymbol{B}_p)\boldsymbol{T}_p^{1/2}\right) \to \mathrm{tr}(\boldsymbol{T}_p)$ almost surely. Therefore we only need to prove that the empirical spectral density of $\mathrm{tr}\left(\boldsymbol{T}_p^{1/2}(\frac{1}{n}\sum_{i=1}^n \boldsymbol{z}_i\boldsymbol{z}_i^T + \boldsymbol{B}_p)\boldsymbol{T}_p^{1/2}\right)\hat{\Sigma} = \boldsymbol{T}_p^{1/2}(\frac{1}{n}\sum_{i=1}^n \boldsymbol{z}_i\boldsymbol{z}_i^T + \boldsymbol{B}_p)\boldsymbol{T}_p^{1/2}$ converges to ρ.

Since $\boldsymbol{T}_p^{1/2}(\frac{1}{n}\sum_{i=1}^n \boldsymbol{z}_i\boldsymbol{z}_i^T + \boldsymbol{B}_p)\boldsymbol{T}_p^{1/2} - \boldsymbol{T}_p^{1/2}(\frac{1}{n}\sum_{i=1}^n \boldsymbol{z}_i\boldsymbol{z}_i^T - \|\boldsymbol{B}_p\|\mathbf{I})\boldsymbol{T}_p^{1/2}$ is positive definite, the eigenvalues of $\boldsymbol{T}_p^{1/2}(\frac{1}{n}\sum_{i=1}^n \boldsymbol{z}_i\boldsymbol{z}_i^T + \boldsymbol{B}_p)\boldsymbol{T}_p^{1/2}$ is bounded below by $\boldsymbol{T}_p^{1/2}(\frac{1}{n}\sum_{i=1}^n \boldsymbol{z}_i\boldsymbol{z}_i^T - \|\boldsymbol{B}_p\|\mathbf{I})\boldsymbol{T}_p^{1/2}$; similarly it is bounded above by $\boldsymbol{T}_p^{1/2}(\frac{1}{n}\sum_{i=1}^n \boldsymbol{z}_i\boldsymbol{z}_i^T + \|\boldsymbol{B}_p\|\mathbf{I})\boldsymbol{T}_p^{1/2}$. Combining it with the fact that the eigenvalues of $\frac{1}{n}\sum_{i=1}^n \boldsymbol{z}_i\boldsymbol{z}_i^T$ are almost surely bounded by $[1-\sqrt{y}-t, 1+\sqrt{y}+t]$ for any $t > 0$ [17, Corollary 5.35], the eigenvalues of $\boldsymbol{T}_p^{1/2}(\frac{1}{n}\sum_{i=1}^n \boldsymbol{z}_i\boldsymbol{z}_i^T + \boldsymbol{B}_p)\boldsymbol{T}_p^{1/2}$ are bounded below and above by $\boldsymbol{T}_p^{1/2}(1 - \frac{\|\boldsymbol{B}_p\|}{1-\sqrt{y}-t})(\frac{1}{n}\sum_{i=1}^n \boldsymbol{z}_i\boldsymbol{z}_i^T)\boldsymbol{T}_p^{1/2}$ and $\boldsymbol{T}_p^{1/2}(1 + \frac{\|\boldsymbol{B}_p\|}{1-\sqrt{y}-t})(\frac{1}{n}\sum_{i=1}^n \boldsymbol{z}_i\boldsymbol{z}_i^T)\boldsymbol{T}_p^{1/2}$ almost surely. By the convergence of the ESD of $\boldsymbol{T}_p^{1/2}(\frac{1}{n}\sum_{i=1}^n \boldsymbol{z}_i\boldsymbol{z}_i^T)\boldsymbol{T}_p^{1/2}$ to ρ [1, (6.1.2)] and the convergence of $\|\boldsymbol{B}_p\|$ to 0, Corollary 3.9 is proved.

□

3.3.2 Maronna's M-estimator

This section investigates the distribution of the eigenvalues of Maronna's M-estimator, when data are sample from Gaussian distribution. The analysis follows from the proof of Theorem 3.5 and Corollary 3.6.

Corollary 3.10. *Suppose $\{\boldsymbol{x}_i\}_{i=1}^n$ are i.i.d. sampled from Gaussian distribution $N(\mathbf{0}, \boldsymbol{T}_p)$, where $\boldsymbol{T}_p$ is a positive definite matrix in $\mathbb{R}^{p\times p}$, and the empirical spectral density of $\boldsymbol{T}_p$ converges to H. Then the empirical spectral density of $\frac{1}{n\,\psi^{-1}(1/y)}\bar{\Sigma}$ converges to ρ, whose Stieltjes transform $s(z)$ satisfies*

$$s(z) = \int \frac{1}{t(1-y-yz\,s(z)) - z}\,\mathrm{d}H(t). \tag{26}$$

In particular, if $\boldsymbol{T}_p = \mathbf{I}$ for all p, $\frac{1}{n\,\psi^{-1}(1/y)}\bar{\Sigma}$ converges to the Marchenko-Pastur distribution in (2).

4 Summary

We established that Maronna's M-estimator and Tyler's M-estimator converge in operator norm to the sample covariance matrix as $p, n \to \infty$ and $p/n \to y$, $0 < y < 1$, where data samples follow the distribution of $N(\mathbf{0}, \mathbf{I})$. We also extended the result to elliptical distribution for Tyler's M-estimator and non-isotropic Gaussian distribution for Maronna's M-estimator, and proved the conjecture that the empirical spectral density of Tyler's M-estimator converges to the Marchenko-Pasture distribution.

There are several possible future directions of this work. First, we would like to know if a more careful analysis can prove the convergence of Maronna's estimator without the assumption **A2**. Second, in simulations we observe the rate of M-estimator's convergence to the sample covariance matrix is $1/\sqrt{n}$, while the current theoretical analysis only gives the order of $O(\sqrt{\log n/n})$, and we would like to find an approach that gives the better empirical rate.

5 Proof of Lemmas

5.1 Proof of Lemma 2.1

We first show the uniqueness of the solution to (8). It follows from the equivalence to the following convex problem:

$$(\hat{z}_1, \hat{z}_2, \cdots, \hat{z}_n) = \underset{\sum_{i=1}^n z_i = 1}{\arg\min} \log\det(\sum_{i=1}^n e^{z_i} \boldsymbol{x}_i \boldsymbol{x}_i^T), \tag{27}$$

The equivalence can be proved by plugging $w_i = ne^{z_i}/(\sum_{i=1}^n e^{z_i})$ to (8) and plugging $z_i = \log w_i - (\sum_{i=1}^n \log w_i - 1)/n$ to (27), and the uniqueness of the solution to (27) follows from its convexity, which is proved in [18, Lemma 4].

Next we will verify (7). We start the proof by verifying that $\hat{w}_i = w_i'$, where $w_i' = c_0/(\boldsymbol{x}_i^T \hat{\Sigma}^{-1} \boldsymbol{x}_i)$ and c_0 is a constant such that $\sum_{i=1}^n w_i' = 1$.

According to the equivalence between (8) and (27), it is enough to show that $z_i' = \log(w_i') + c_1$ (c_1 chosen such that $\sum_{i=1}^n z_i' = 1$) is the unique minimizer of (27). Indeed, applying the iterative algorithm (3), $\hat{\Sigma} = c_2 \sum_{i=1}^n w_i' \boldsymbol{x}_i \boldsymbol{x}_i^T$ for some c_2. Combining it with the definition of w_i', we have

$$w_i' = \frac{c_0/c_2}{\boldsymbol{x}_i^T (\sum_{i=1}^n w_i' \boldsymbol{x}_i \boldsymbol{x}_i^T)^{-1} \boldsymbol{x}_i}. \tag{28}$$

Now we are ready to prove that the directional derivative of the objective function in (27) is 0 at $(z_1', z_2', \cdots, z_n')$: assuming the direction is from $(z_1', z_2', \cdots, z_n')$ to $(z_1' + \delta_1, z_2' + \delta_2, \ldots, z_n' + \delta_n)$, then the directional derivative is

$$\sum_{i=1}^n e^{z_i'} \boldsymbol{x}_i^T (\sum_{i=1}^n e^{z_i'} \boldsymbol{x}_i \boldsymbol{x}_i^T)^{-1} \boldsymbol{x}_i \delta_i = \sum_{i=1}^n w_i' \boldsymbol{x}_i^T (\sum_{i=1}^n w_i' \boldsymbol{x}_i \boldsymbol{x}_i^T)^{-1} \boldsymbol{x}_i \delta_i = \sum_{i=1}^n w_i' \frac{c_0}{c_2 w_i'} \delta_i = 0,$$

where the second equality follows from (28) and the last equality follows from $\sum_{i=1}^n \delta_i = 0$.

Due to the convexity of the objective function in (27), its stationary point is also its minimizer, therefore $\hat{z}_i = z_i'$ and $\hat{w}_i = w_i'$.

Combining $\hat{w}_i = c_0/(\boldsymbol{x}_i^T \hat{\boldsymbol{\Sigma}}^{-1} \boldsymbol{x}_i)$ and the definition of $\hat{\boldsymbol{\Sigma}}$ in (3), (7) is proved.

5.2 Proof of Lemma 2.2

Proof. It follows from the definition in (4) that $\bar{w}_i = \frac{1}{n} \boldsymbol{x}_i^T \bar{\Sigma}^{-1} \boldsymbol{x}_i$ satisfies (10). Plug (10) in the definition of $\bar{w}_i$, we obtain (9).

Since any $\{\bar{w}_i\}_{i=1}^n$ that satisfy (9) give the solution of (4) by (10), and the solution of (4), i.e., Maronna's M-estimator, exists and is unique, therefore any solution to (9) gives the same $\bar{\Sigma}$ by (10). □

5.3 Proof of Lemma 3.1

We start with an outline of the proof, which consists of three parts. First, we rewrite the constrained optimization problem (8) to the problem of finding the root of $g(\boldsymbol{w})$, which will be defined in (29). Since the root of $g(\boldsymbol{w})$ is $n\hat{\boldsymbol{w}} - 1$, we only need to show the convergence of the root of $g(\boldsymbol{w})$. Second, we will show that $g(\mathbf{0})$ converges to $\mathbf{0}$, $\nabla g(\mathbf{0})$ is large and the variation of $\nabla g(\boldsymbol{w})$ is bounded. Finally, we will use a perturbation analysis and the observations on $g(\mathbf{0})$ and $\nabla g(\boldsymbol{w})$ to show that the root of $g(\boldsymbol{w})$ converges to $\mathbf{0}$.

The proof depends on Lemma 5.2, Lemma 5.3 and Lemma 5.1, and their proofs are postponed to subsequent sections.

Lemma 5.1. *For a function $f(\boldsymbol{w}) : \mathbb{R}^p \to \mathbb{R}^p$, assume that $\nabla f(\mathbf{0}) = \mathbf{I}$, and $\|\nabla f(\boldsymbol{w}) - \nabla f(\mathbf{0})\|_\infty = \max_{i \le i \le p} \|\nabla f_i(\boldsymbol{w}) - \nabla f_i(\mathbf{0})\|_\infty < C_5 \|\boldsymbol{w}\|_\infty$ for $\|\boldsymbol{w}\|_\infty \le 1$, and $\|f(\mathbf{0})\|_\infty < \min(1/9C_5, 1/3)$. Then there exists $\tilde{\boldsymbol{w}}$ such that $\|\tilde{\boldsymbol{w}}\|_\infty < 3\|f(\mathbf{0})\|_\infty$ and $f(\tilde{\boldsymbol{w}}) = \mathbf{0}$.*

Lemma 5.2. *If $\boldsymbol{x}_i \sim N(\mathbf{0}, \mathbf{I})$ for all $1 \le i \le n$, and $\boldsymbol{S} = \frac{1}{n}\sum_{i=1}^n \boldsymbol{x}_i \boldsymbol{x}_i^T$, then there exists $c, C, c' > 0$ such that for any $\varepsilon < c'$,*

$$\Pr\left(\max_{1 \le i \le n} |\frac{1}{p}\boldsymbol{x}_i^T \boldsymbol{S}^{-1} \boldsymbol{x}_i - 1| < \varepsilon\right) \ge 1 - Cne^{-c\varepsilon^2 n}.$$

Lemma 5.3. *For the $n \times n$ matrix $\boldsymbol{A}$ defined by $\boldsymbol{A}_{ij} = \frac{1}{n\,p}(\boldsymbol{x}_i^T \boldsymbol{S}^{-1} \boldsymbol{x}_j)^2$, (a) $\|\boldsymbol{A}\|_\infty < 2$ with probability $1 - Cn\exp(-cn)$.*

(b) There exists $c = c(p, n) > 0$ and $C_2 = C_2(y) > 0$ such that $\|(\mathbf{I} - \boldsymbol{A} + c\mathbf{1}\mathbf{1}^T)^{-1}\|_\infty < C_2$ with probability $1 - Cn\exp(-cn)$.

We start the first part of the proof with the construction of $g(\boldsymbol{w})$. We let

$$g(\boldsymbol{w}) = \nabla G(\boldsymbol{w} + \mathbf{1}), \tag{29}$$

where

$$G(\boldsymbol{w}) = -\sum_{i=1}^n \log w_i + \frac{n}{p}\log\det(\sum_{i=1}^n w_i \boldsymbol{x}_i \boldsymbol{x}_i^T) + \frac{c_0}{2}(\sum_{i=1}^n w_i - n)^2, \tag{30}$$

and the constant c_0 will be specified later before (46).

It is easy to prove that the minimizer of $G(\boldsymbol{w})$ and the zeros of $\nabla G(\boldsymbol{w})$ must satisfies $\sum_{i=1}^n w_i = n$ (otherwise $n\boldsymbol{w}/(\sum_{i=1}^n w_i)$ is a better minimizer and $\nabla G(\boldsymbol{w})$ is nonzero). Therefore minimizing (30) is equivalent to minimizing $-\sum_{i=1}^n \log w_i + \frac{n}{p}\log\det(\sum_{i=1}^n w_i \boldsymbol{x}_i \boldsymbol{x}_i^T)$ with constraint $\sum_{i=1}^n w_i = n$, which is the same as (8) except for the constraint. Noticing that a scaling of $\boldsymbol{w}$ increases $-\sum_{i=1}^n \log w_i + \frac{n}{p}\log\det(\sum_{i=1}^n w_i \boldsymbol{x}_i \boldsymbol{x}_i^T)$ by a constant only depending on the scale, the minimizer of (30) is unique and it is $n\hat{\boldsymbol{w}}$, where $\hat{\boldsymbol{w}}$ is defined in (8). By the convexity of its equivalent problem (27), the root of $g(\boldsymbol{w})$ is also unique and it is $n\hat{\boldsymbol{w}} - 1$.

For the second part of the proof, we start by proving that $g(\mathbf{0})$ is small. By calculation, the i-th component of function $g(\boldsymbol{w})$ is

$$g_i(\boldsymbol{w}) = -\frac{1}{w_i+1} + \frac{n}{p}\boldsymbol{x}_i^T(n\boldsymbol{S} + \sum_{i=1}^n w_i\boldsymbol{x}_i\boldsymbol{x}_i^T)^{-1}\boldsymbol{x}_i + c_0(\sum_{i=1}^n w_i).$$

Applying Lemma 5.2,

$$\Pr\left(\|g(\mathbf{0})\|_\infty < \varepsilon\right) \geq 1 - Cne^{-c\varepsilon^2 n}. \tag{31}$$

Now we will prove that $\nabla g(\mathbf{0})$ is bounded from below. By calculation, its (i,j)-th entry is

$$\big(\nabla g(\boldsymbol{w})\big)_{i,j} = I(i=j)\frac{1}{(w_i+1)^2} - \frac{n}{p}\Big(\boldsymbol{x}_i^T\big(n\boldsymbol{S} + \sum_{i=1}^n w_i\boldsymbol{x}_i\boldsymbol{x}_i^T\big)^{-1}\boldsymbol{x}_j\Big)^2 + c_0.$$

Applying Lemma 5.3,

$$\|(\nabla g(\mathbf{0}))^{-1}\|_\infty < C_2 \text{ with probability } 1 - Cne^{-cn}. \tag{32}$$

Now we bound the variation of $\nabla g(\boldsymbol{w})$ in the region $\|\boldsymbol{w}\|_\infty < 1/2$. Apply $|\frac{1}{(w_i+1)^2} - 1| < 3|w_i - 1| \leq 3\|\boldsymbol{w}\|_\infty$ and coordinatewise comparison,

$$|\nabla_{i,j}g(\boldsymbol{w}) - \nabla_{i,j}g(\mathbf{0})| \leq I(i=j)\left(3\|\boldsymbol{w}\|_\infty\right) + 3\|\boldsymbol{w}\|_\infty \cdot \frac{n}{p}|\boldsymbol{A}_{ij}|.$$

Therefore, the variation of $\nabla g(\boldsymbol{w})$ is bounded by

$$\|\nabla g(\boldsymbol{w}) - \nabla g(\mathbf{0})\|_\infty < (3 + 3n\|\boldsymbol{A}\|_\infty/p)\|\boldsymbol{w}\|_\infty. \tag{33}$$

At last we finish the third part of the proof of Lemma 3.1 by applying Lemma 5.1 to $f(\boldsymbol{w}) = (\nabla g(\mathbf{0}))^{-1}g(\boldsymbol{w}/2)$. It is easy to verify that $\nabla f(\mathbf{0}) = \mathbf{I}$. Due to (31) and (32), $\|f(\mathbf{0})\|_\infty \leq \|(\nabla g(\mathbf{0}))^{-1}\|_\infty\|g(\mathbf{0})\|_\infty \to 0$ in the same rate as in (31) and $\|f(\mathbf{0})\|_\infty < \min(1/9C_5, 1/3)$ holds with probability $1 - Cne^{-cn}$. Due to (31), (33), and the boundedness of $\|\boldsymbol{A}\|_\infty$ (Lemma 5.3), $\|\nabla f(\boldsymbol{w}) - \nabla(\mathbf{0})\|_\infty < C_5\|\boldsymbol{w}\|_\infty$ also holds with probability $1 - Cne^{-cn}$. Therefore the assumption in Lemma 5.1 holds with probability $1 - Cne^{-cn}$ and there exists $\tilde{\boldsymbol{w}}$ such that $f(\tilde{\boldsymbol{w}}) = 0$ and

$$\|\tilde{\boldsymbol{w}}\|_\infty < 3\|f(\mathbf{0})\|_\infty. \tag{34}$$

When $f(\tilde{\boldsymbol{w}}) = 0$, we have $g(2\tilde{\boldsymbol{w}}) = 0$ and by previous discussion $2\tilde{\boldsymbol{w}} = n\hat{\boldsymbol{w}} - 1$. therefore (34) gives

$$\|n\hat{\boldsymbol{w}} - 1\|_\infty < 6\|f(\mathbf{0})\|_\infty.$$

Since $\|f(\mathbf{0})\|_\infty$ converges to 0 in the rate as in (31), $\|n\hat{\boldsymbol{w}} - 1\|_\infty$ converges in the same rate and Lemma 3.1 is proved.

5.4 Proof of Lemma 3.3

Proof. Define $\bar{g}(\boldsymbol{w}) : \mathbb{R}^n \to \mathbb{R}^n$ by $\bar{g}_j(\boldsymbol{w}) = w_j - \boldsymbol{x}_j^T \big(\sum_{i=1}^n u(w_i) \boldsymbol{x}_i \boldsymbol{x}_i^T \big)^{-1} \boldsymbol{x}_j$ for all $1 \le j \le n$, then $\bar{\boldsymbol{w}} = (\bar{w}_1, \bar{w}_2, \cdots, \bar{w}_n)$ is a root of $\bar{g}(\boldsymbol{w})$. Let $\tilde{\boldsymbol{w}} = \psi^{-1}(1/y)\mathbf{1}$, we will prove that

1. $\Pr(\|\bar{g}(\tilde{\boldsymbol{w}})\|_\infty < \varepsilon) > 1 - Cne^{-c\varepsilon^2 n}$ for any $\varepsilon < c'$.
2. $\Pr(\|\big(\nabla \bar{g}(\tilde{\boldsymbol{w}})\big)^{-1}\|_\infty < C') > 1 - Cne^{-cn}$.
3. $\Pr(\|\nabla \bar{g}(\tilde{\boldsymbol{w}}) - \nabla \bar{g}(\boldsymbol{w})\|_\infty < C' \|\boldsymbol{w} - \tilde{\boldsymbol{w}}\|_\infty) > 1 - Cne^{-cn}$ for any $\|\boldsymbol{w} - \tilde{\boldsymbol{w}}\|_\infty < c'$.

The first point can be proved by applying Lemma 5.2. As for the second point, we have

$$\nabla_{ij} \bar{g}(\tilde{\boldsymbol{w}}) = \mathbf{I} + \frac{u'(\psi^{-1}(1/y))}{u(\psi^{-1}(1/y))^2} \big(\boldsymbol{x}_i (n\boldsymbol{S})^{-1} \boldsymbol{x}_j \big)^2.$$

Since $\sum_{j=1}^n \big(\boldsymbol{x}_i (n\boldsymbol{S})^{-1} \boldsymbol{x}_j \big)^2 = \boldsymbol{x}_i (n\boldsymbol{S})^{-1} \boldsymbol{x}_i$, and $\boldsymbol{x}_i (n\boldsymbol{S})^{-1} \boldsymbol{x}_i$ converges to $1/y$ with probability (Lemma 5.2), we have

$$\begin{aligned} &\sum_{j=1}^n \frac{u'(\psi^{-1}(1/y))}{u(\psi^{-1}(1/y))^2} \big(\boldsymbol{x}_i (n\boldsymbol{S})^{-1} \boldsymbol{x}_j \big)^2 \to \frac{u'(\psi^{-1}(1/y))}{u(\psi^{-1}(1/y))^2} \cdot 1/y \\ =&\frac{u'(\psi^{-1}(1/y))}{u(\psi^{-1}(1/y))^2} \cdot u(\psi^{-1}(1/y)) \psi^{-1}(1/y) = \frac{u'(\psi^{-1}(1/y))\, \psi^{-1}(1/y)}{u(\psi^{-1}(1/y))}. \end{aligned}$$

Since $|u'(x)x/u(x)| < 1$, we have $\Big\| \frac{u'(\psi^{-1}(1/y))}{u(\psi^{-1}(1/y))^2} \big(\boldsymbol{x}_i (n\boldsymbol{S})^{-1} \boldsymbol{x}_j \big)^2 \Big\|_\infty < 1$ with exponential probability, and by

$$\begin{aligned} &\Big\| \mathbf{I} + \frac{u'(\psi^{-1}(1/y))}{u(\psi^{-1}(1/y))^2} \big(\boldsymbol{x}_i (n\boldsymbol{S})^{-1} \boldsymbol{x}_j \big)^2 \Big\|_\infty \le \sum_{i=0}^{\infty} \Big\| \frac{u'(\psi^{-1}(1/y))}{u(\psi^{-1}(1/y))^2} \big(\boldsymbol{x}_i (n\boldsymbol{S})^{-1} \boldsymbol{x}_j \big)^2 \Big\|_\infty^i \\ =&\frac{1}{1 - \Big\| \frac{u'(\psi^{-1}(1/y))}{u(\psi^{-1}(1/y))^2} \big(\boldsymbol{x}_i (n\boldsymbol{S})^{-1} \boldsymbol{x}_j \big)^2 \Big\|_\infty}, \end{aligned}$$

the second point is proved.

As for the third point, we note that the j-th component of the gradient of $\bar{g}_k$ is

$$\nabla_j \big(\bar{g}_k(\tilde{\boldsymbol{w}}) \big) = \delta(j = k) + u'(w_j) \Big(\boldsymbol{x}_k^T \big(\sum_{i=1}^n u(w_i) \boldsymbol{x}_i \boldsymbol{x}_i^T \big)^{-1} \boldsymbol{x}_j \Big)^2,$$

u is twice differentiable, and $\|\boldsymbol{A}\|_\infty$ is bounded with high probability, therefore the third point holds.

Apply Lemma 5.1 with $f(\boldsymbol{w}) = \big(\nabla \bar{g}(\tilde{\boldsymbol{w}}) \big)^{-1} \bar{g}(\boldsymbol{w})$, then Lemma 3.3 follows the same procedure as in the third part of the proof of Lemma 3.1. □

5.4.1 Proof of Lemma 5.1

Proof. When $\|\boldsymbol{w}\|_\infty \le 1$,

$$f_j(\boldsymbol{w}) - f_j(\boldsymbol{0}) = \int_{t=0}^{1} \left\langle \boldsymbol{e}_j \boldsymbol{w}^T, \nabla f(t\,\boldsymbol{w}) \right\rangle \mathrm{d}t \tag{35}$$

$$= \int_{t=0}^{1} \left\langle \boldsymbol{e}_j \boldsymbol{w}^T, \nabla f(t\,\boldsymbol{w}) - \nabla f(\boldsymbol{0}) + \mathbf{I} \right\rangle \mathrm{d}t = w_j + \int_{t=0}^{1} \boldsymbol{w}^T (\nabla f(t\,\boldsymbol{w}) - \nabla f(\boldsymbol{0})) \boldsymbol{e}_j \,\mathrm{d}t$$

$$\le w_j + \| \int_{t=0}^{1} \boldsymbol{w}^T (\nabla f(t\,\boldsymbol{w}) - \nabla f(\boldsymbol{0}))\|_\infty \le w_j + C_5 \|\boldsymbol{w}\|_\infty^2.$$

Similarly

$$f_j(\boldsymbol{w}) - f_j(\boldsymbol{0}) \ge -C_5 \|\boldsymbol{w}\|_\infty^2 + w_j. \tag{36}$$

To prove it, we consider the continuous mapping $h(\boldsymbol{w}) = \boldsymbol{w} - f(\boldsymbol{w})/(4+9C_5)$ and will prove that h maps $\mathcal{A}$ to itself, where

$$\mathcal{A} = \{\boldsymbol{w} : \boldsymbol{w} \in [-3\eta, 3\eta]^n\} \text{ and } \eta = \|f(\boldsymbol{0})\|_\infty.$$

1. $|w_i| < 2\eta$. Then apply (35) and (36) (they are applicable since for any $\boldsymbol{w} \in \mathcal{A}$, $\|\boldsymbol{w}\|_\infty \le 1$), we have $|f_i(\boldsymbol{w})| < |f_i(\boldsymbol{0})| + C_5\|\boldsymbol{w}\|_\infty^2 + |w_i| \le \eta + C_5(3\eta)^2 + 3\eta < (4+9C_5)\eta$ ($\eta^2 < \eta$ since $\eta < 1$). Therefore, $|h_i(\boldsymbol{w})| \le |w_i| + |f_i(\boldsymbol{w})|/(4+9C_5) \le 3\eta$.

2. $w_i > 2\eta$, then applying (36),

$$f_i(\boldsymbol{w}) \ge -|f_i(\boldsymbol{0})| + w_i - C_5\|\boldsymbol{w}\|_\infty^2 \ge -\eta + 2\eta - C_5(3\eta)^2.$$

Since $\eta < 1/9C_5$, we have $f_i(\boldsymbol{w}) < 0$ and therefore $h_i(\boldsymbol{w}) \le w_i \le 3\eta$.

Similar to case 1 we can prove that $h_i(\boldsymbol{w}) \ge -3\eta$. Therefore $|h_i(\boldsymbol{w})| < 3\eta$.

3. Similar to case 2, when $w_i < -2\eta$, $|h_i(\boldsymbol{w})| < 3\eta$.

Therefore the continuous mapping h maps the convex, compact set $\mathcal{A}$ to itself. By Schauder fixed point theorem $h(\boldsymbol{x})$ has a fixed point in $\mathcal{A}$ and therefore Lemma 5.1 is proved with $\tilde{\boldsymbol{w}}$ being the fixed point. □

5.4.2 Proof of Lemma 5.2

Assuming the SVD decomposition of $\boldsymbol{X}$ is $\boldsymbol{X} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^T$, where $\boldsymbol{U} \in \mathbb{R}^{n\times p}$ and $\boldsymbol{U}^T\boldsymbol{U} = \mathbf{I}$. Since $\boldsymbol{x}_i \sim N(\boldsymbol{0}, \mathbf{I})$ for all $1 \le i \le n$, $\boldsymbol{U}$ is uniformly distributed over the space of all orthogonal $n \times p$ matrices. Since

$$\boldsymbol{X}\boldsymbol{S}^{-1}\boldsymbol{X} = (\boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^T)(\frac{1}{n}\boldsymbol{V}\boldsymbol{\Sigma}^2\boldsymbol{V}^T)^{-1}(\boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^T), \tag{37}$$

if we write the row of $\boldsymbol{U}$ by $\boldsymbol{u}_1, \boldsymbol{u}_2, \cdots, \boldsymbol{u}_n$, then $\frac{1}{n}\boldsymbol{x}_i\boldsymbol{S}^{-1}\boldsymbol{x}_i = \boldsymbol{u}_i^T\boldsymbol{u}_i = \|\boldsymbol{u}_i\|^2$.

Since $\boldsymbol{U}$ can be considered as the first p columns of a random $n\times n$ orthogonal matrix (with haar measure over the set of all $n \times n$ orthogonal matrices), $\boldsymbol{u}_i$ can be considered as the first p entries from a random vector of length n that is sampled from the uniform sphere in $\mathbb{R}^n$.

Therefore, $\|\boldsymbol{u}_i\|^2 \sim \sum_{j=1}^p g_j^2 / \sum_{j=1}^n g_j^2$ for i.i.d. random variables $\{g_j\}_{j=1}^n \sim N(0,1)$. Applying [2, Corollary 2.3], we have

$$\Pr\left(\sum_{i=1}^n g_i^2 \geq \frac{n}{1-\varepsilon}\right) \leq e^{-\varepsilon^2 n/4} \tag{38}$$

and

$$\Pr\left(\sum_{i=1}^n g_i^2 \leq n(1-\varepsilon)\right) \leq e^{-\varepsilon^2 n/4}, \tag{39}$$

therefore

$$\Pr\left(\frac{p(1-\varepsilon)^2}{n} \leq \|\boldsymbol{u}_1\|^2 \leq \frac{p}{n(1-\varepsilon)^2}\right) \geq \Pr\left(p(1-\varepsilon) \leq \sum_{i=1}^p g_i^2 \leq \frac{p}{1-\varepsilon}\right)$$
$$+ \Pr\left(n(1-\varepsilon) \leq \sum_{i=1}^n g_i^2 \leq \frac{n}{1-\varepsilon}\right) \geq 1 - 2e^{-\varepsilon^2 p/4} - 2e^{-\varepsilon^2 n/4}.$$

For $\varepsilon \leq 0.1$, we have

$$\Pr\left(\max_{1\leq i\leq n} |\frac{1}{p}\boldsymbol{x}_i^T \boldsymbol{S}^{-1}\boldsymbol{x}_i - 1| \leq \varepsilon\right) \geq 1 - n\Pr\left(|\|\boldsymbol{u}_1\|^2 - \frac{p}{n}| > \frac{p}{n}\varepsilon\right)$$
$$\geq 1 - n\left(1 - \Pr\left(\frac{p(1-\varepsilon/3)^2}{n} \leq \|\boldsymbol{u}_1\|^2 \leq \frac{p}{n(1-\varepsilon/3)^2}\right)\right) \tag{40}$$
$$\geq 1 - 2ne^{-\varepsilon^2 p/36} - 2ne^{-\varepsilon^2 n/36}, \tag{41}$$

where the second inequality follows from $1 - 3\varepsilon \leq (1-\varepsilon)^2$ and $\frac{1}{(1-\varepsilon)^2} \leq 1 + 3\varepsilon$.

5.4.3 Proof of Lemma 5.3

(a) Since $\|\boldsymbol{A}\|_\infty = \max_{1\leq i\leq n}(\sum_{1\leq j\leq n} \boldsymbol{A}_{ij})$, and

$$\sum_{1\leq j\leq n} \boldsymbol{A}_{ij} = \sum_{1\leq j\leq n} \frac{1}{np}\boldsymbol{x}_i^T \boldsymbol{S}^{-1}\boldsymbol{x}_j\boldsymbol{x}_j^T \boldsymbol{S}^{-1}\boldsymbol{x}_i = \boldsymbol{x}_i^T \boldsymbol{S}^{-1}(\sum_{1\leq j\leq n} \boldsymbol{x}_j\boldsymbol{x}_j^T)\boldsymbol{S}^{-1}\boldsymbol{x}_i/np \tag{42}$$

$$= \boldsymbol{x}_i^T \boldsymbol{S}^{-1}(n\boldsymbol{S})\boldsymbol{S}^{-1}\boldsymbol{x}_i/np = \boldsymbol{x}_i^T \boldsymbol{S}^{-1}\boldsymbol{x}_i/p, \tag{43}$$

it follows from (41) with $\varepsilon = 0.1$ that $\|\boldsymbol{A}\|_\infty < 2$ holds with probability $1 - Cn\exp(-cn)$.

(b) We first prove that there exists $C_3 = C_3(y)$ such that

$$\|\boldsymbol{A} - c_0\boldsymbol{1}\boldsymbol{1}^T\|_\infty \leq C_3 < 1 \text{ with probability } 1 - Cn\exp(-cn). \tag{44}$$

We start with the proof of (44) with another lemma:

Lemma 5.4. *There exists a $c_4 > 0$ such that with probability $1 - C\exp(-cn)$,*

$$\sum_{j=1}^{n} I(\boldsymbol{x}_1^T \boldsymbol{x}_j > c_4\sqrt{p}) > 0.75n.$$

There exists $C_4 = C_4(y)$ such that $\|\boldsymbol{S}\| < C_4$ with probability $1-Cn\exp(-cn)$ [10, Theorem II.13]. Therefore $\boldsymbol{x}_i^T \boldsymbol{S}^{-1}\boldsymbol{x}_j \geq \boldsymbol{x}_i^T \boldsymbol{x}_j/C_4$ and Lemma 5.4 implies that for any $1 \leq i \leq n$:

$$\sum_{j=1}^{n} I(\boldsymbol{x}_i^T \boldsymbol{S}^{-1}\boldsymbol{x}_j > c_4\sqrt{p}/C_4) > 0.75 \text{ with probability } 1 - C\exp(-cn). \quad (45)$$

Let $c_0 = (c_4/C_4)^2/n$, then (45) implies

$$\sum_{1\leq j\leq n} |\boldsymbol{A}_{i,j} - c| \leq \sum_{1\leq j\leq n} |\boldsymbol{A}_{i,j}| - 0.25c\,n \leq \boldsymbol{x}_i^T \boldsymbol{S}^{-1}\boldsymbol{x}_i/p - 0.25(c_4/C_4)^2, \quad (46)$$

where the last step follows from (43).

Applying the estimation of $\boldsymbol{x}_i^T \boldsymbol{S}^{-1}\boldsymbol{x}_i/p$ in (41) and a union bound argument over all $1 \leq i \leq n$ to (46), (44) is proved for $C_3 = 1 + \eta - 0.25(c_4/C_4)^2$.

Lemma 5.3(b) follows from (44) with $C_2 = \frac{1}{1-C_3}$, where the expansion of $(\mathbf{I} - \boldsymbol{A} + c\mathbf{1}\mathbf{1}^T)^{-1}$ is valid since $\|\boldsymbol{A} + c\mathbf{1}\mathbf{1}^T\| \leq \|\boldsymbol{A} + c\mathbf{1}\mathbf{1}^T\|_\infty < 1$:

$$\|(\mathbf{I} - \boldsymbol{A} + c\mathbf{1}\mathbf{1}^T)^{-1}\|_\infty \leq \sum_{i=0}^{\infty} \|\boldsymbol{A} - c\mathbf{1}\mathbf{1}^T\|_\infty^i = \sum_{i=0}^{\infty} C_3^i = \frac{1}{1-C_3}. \quad (47)$$

5.4.4 Proof of Lemma 5.4

We first show that there exists c_4 such that for all p,

$$E(I(|\boldsymbol{x}_1^T \boldsymbol{x}_2| > c_4\sqrt{p})) \geq 0.85. \quad (48)$$

WLOG we rotate $\boldsymbol{x}_1$ such that it is nonzero only at the first coordinate, and $\boldsymbol{x}_2 = (g_1, g_2, ..., g_p)$ where $g_i \sim N(0,1)$. Then $|\boldsymbol{x}_1^T \boldsymbol{x}_2| = |g_1|\,\|\boldsymbol{x}_1\|$.

Notice that $\|\boldsymbol{x}_1\|^2$ is the sum of p independent χ_1^2 distribution and $E\chi_1^2 = 1$, by central limit theorem, $\|\boldsymbol{x}_1\| \leq \sqrt{2p}$ with probability $1 - Ce^{-cn}$. Besides, $\Pr(|g_1| > \sqrt{2}\,c_4) \geq 0.85$ for $c_4 = \Phi^{-1}(1 - 0.85/2)/\sqrt{2}$. Therefore (48) is proved by combining the estimations on $|g_1|$, $\boldsymbol{x}_1$ and $|\boldsymbol{x}_1^T \boldsymbol{x}_2| = |g_1|\,\|\boldsymbol{x}_1\|$.

To obtain Lemma 5.4 from (48), we apply Hoeffding's inequality to the indicator function $I(|\boldsymbol{x}_i^T \boldsymbol{x}_j| > c_4\sqrt{p})$ over all $1 \leq j \leq n, j \neq i$.

5.5 Proof of Lemma 3.8

Denoting the k-th eigenvalue of any matrix $\boldsymbol{A}$ by $\lambda_k(\boldsymbol{A})$, then [3, Corollary III.4.2] gives

$$\lambda_k(\boldsymbol{A}_n + \boldsymbol{B}_n) - \lambda_k(\boldsymbol{A}_n) \leq \|\boldsymbol{B}_n\|, \quad (49)$$

Assuming the empirical spectral density of $\boldsymbol{A}_n$ and $\boldsymbol{A}_n + \boldsymbol{B}_n$ are ρ_n and ρ'_n, then (49) implies

$$\int_a^b \rho'_n(x)\,\mathrm{d}x \leq \int_{a-\|\boldsymbol{B}_n\|}^{b+\|\boldsymbol{B}_n\|} \rho_n(x)\,\mathrm{d}x.$$

Since $\|\boldsymbol{B}_n\| \to 0$, for any $\varepsilon > 0$,

$$\lim_{n\to\infty} \sup \int_a^b \rho'_n(x)\,\mathrm{d}x \leq \int_{a-\varepsilon}^{b+\varepsilon} \rho(x)\,\mathrm{d}x.$$

By the continuity of ρ, $\lim_{n\to\infty} \sup \int_a^b \rho'_n(x)\,\mathrm{d}x \leq \int_a^b \rho(x)$. Similarly we can prove that $\lim_{n\to\infty} \inf \int_a^b \rho'_n(x)\,\mathrm{d}x \geq \int_a^b \rho(x)$, and therefore $\lim_{n\to\infty} \int_a^b \rho'_n(x)\,\mathrm{d}x = \int_a^b \rho(x)$ and Lemma 3.8 is proved.

Acknowledgement

A. Singer was partially supported by Award Number FA9550-12-1-0317 and FA9550-13-1-0076 from AFOSR, by Award Number R01GM090200 from the NIGMS, and by Award Number LTR DTD 06-05-2012 from the Simons Foundation.

References

[1] Z. Bai and J. Silverstein. *Spectral Analysis of Large Dimensional Random Matrices*. Springer series in statistics. Springer, 2009.

[2] A. Barvinok. Math 710: Measure concentration. Lecture notes, Department of Mathematics, University of Michigan, 2005.

[3] R. Bhatia. *Matrix Analysis*. Graduate Texts in Mathematics. Springer Verlag, 1997.

[4] Y. Chen, A. Wiesel, and A. Hero. Robust shrinkage estimation of high-dimensional covariance matrices. *Signal Processing, IEEE Transactions on*, 59(9):4097 –4107, sept. 2011.

[5] R. Couillet, F. Pascal, and J. W. Silverstein. Robust M-estimation for array processing: A random matrix approach. *CoRR*, abs/1204.5320, 2012.

[6] R. Couillet, F. Pascal, and J. W. Silverstein. The random matrix regime of Maronna's M-estimator with elliptically distributed samples. *CoRR*, abs/1311.7034, 2013.

[7] L. Dümbgen. On Tyler's M-functional of scatter in high dimension. *Annals of the Institute of Statistical Mathematics*, 50(3):471–491, 1998.

[8] G. Frahm and K. Glombek. Semicircle law of Tyler's M-estimator for scatter. *Statistics & Probability Letters*, 82(5):959 – 964, 2012.

[9] G. Frahm and U. Jaekel. Tyler's M-estimator, random matrix theory, and generalized elliptical distributions with applications to finance. Technical report, 2007.

[10] K. Davidson. and S. Szarek. Local operator theory, random matrices and banach spaces. Handbook of the Geometry of Banach Spaces, 1:317, 2001.

[11] J. T. Kent and D. E. Tyler. Redescending M-estimates of multivariate location and scatter. *The Annals of Statistics*, 19(4):pp. 2102–2119, 1991.

[12] R. A. Maronna. Robust M-estimators of multivariate location and scatter. *The Annals of Statistics*, 4(1):pp. 51–67, 1976.

[13] V. A. Marčenko and L. A. Pastur. Distribution of eigenvalues for some sets of random matrices. *Mathematics of the USSR-Sbornik*, 1(4):457, 1967.

[14] E. Ollila and V. Koivunen. Robust antenna array processing using M-estimators of pseudo-covariance. In *Personal, Indoor and Mobile Radio Communications, 2003. PIMRC 2003. 14th IEEE Proceedings on*, volume 3, pages 2659–2663 vol.3, 2003.

[15] E. Ollila, D. E. Tyler, V. Koivunen, and H. V. Poor. Complex Elliptically Symmetric Distributions: Survey, New Results and Applications. *IEEE Transactions on Signal Processing*, 60(11):5597–5625, Nov. 2012.

[16] D. E. Tyler. A distribution-free M-estimator of multivariate scatter. *The Annals of Statistics*, 15(1):pp. 234–251, 1987.

[17] R. Vershynin. Introduction to the non-asymptotic analysis of random matrices, 2011. Available at http://arxiv.org/abs/1011.3027v7.

[18] A. Wiesel. Geodesic convexity and covariance estimation. *Signal Processing, IEEE Transactions on*, 60(12):6182 –6189, dec. 2012.

[19] T. Zhang. Robust subspace recovery by geodesically convex optimization. *arXiv preprint arXiv:1206.1386*, 2012.

[20] T. Zhang, A. Wiesel, and M. Greco. Multivariate generalized gaussian distribution: Convexity and graphical models. *Signal Processing, IEEE Transactions on*, 61(16):4141–4148, 2013.