Journal of Multivariate Analysis 149 (2016) 114–123

Contents lists available at ScienceDirect

Journal of Multivariate Analysis

journal homepage: www.elsevier.com/locate/jmva

Marčenko–Pastur law for Tyler's M-estimator

Teng Zhang[a,*], Xiuyuan Cheng[b], Amit Singer[c]

[a] Department of Mathematics, University of Central Florida, Orlando, FL 32816, USA
[b] Applied Mathematics Program, Yale University, New Haven, CT 06511, USA
[c] Department of Mathematics and PACM, Princeton University, Princeton, NJ 08544, USA

ARTICLE INFO

Article history:
Received 2 June 2015
Available online 12 April 2016

AMS subject classifications:
15B52
62H10

Keywords:
Covariance estimation
Robust statistics
Random matrix theory

ABSTRACT

This paper studies the limiting behavior of Tyler's M-estimator for the scatter matrix, in the regime that the number of samples n and their 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 $\boldsymbol{x}_1, \dots, \boldsymbol{x}_n$ are identically and independently generated from the Gaussian distribution $\mathcal{N}(\mathbf{0}, \mathbf{I})$, the operator norm of the difference between a properly scaled Tyler's M-estimator and $\sum_{i=1}^n \boldsymbol{x}_i\boldsymbol{x}_i^\top/n$ tends to zero. As a result, the spectral distribution of Tyler's M-estimator converges weakly to the Marčenko–Pastur distribution.

1. Introduction

Many statistical estimators and signal processing algorithms require the estimation of the covariance matrix of the data samples. When the underlying distribution of the data samples $\boldsymbol{x}_1, \dots, \boldsymbol{x}_n \in \mathbb{R}^p$ is assumed to have zero mean, a commonly used estimator is the sample covariance matrix $\boldsymbol{S}_n = \sum_{i=1}^n \boldsymbol{x}_i\boldsymbol{x}_i^\top/n$.

However, the estimator $\boldsymbol{S}_n$ is sensitive to outliers, and performs poorly in terms of statistical efficiency (i.e., it has a large variance) for heavy-tailed distributions, e.g., when the tail decays slower than the Gaussian tail.

A popular robust covariance estimator is an M-estimator introduced by Tyler [20], denoted by $\hat{\Sigma}$, which is the unique solution to

$$\hat{\Sigma} = \frac{p}{n}\sum_{i=1}^n \frac{\boldsymbol{x}_i\boldsymbol{x}_i^\top}{\boldsymbol{x}_i^\top\hat{\Sigma}^{-1}\boldsymbol{x}_i}, \quad \operatorname{tr}(\hat{\Sigma}) = 1. \tag{1}$$

Tyler's M-estimator gives the "shape" of the covariance, but is missing its magnitude. However, for many applications the "shape" of the covariance suffices, for example, the principal components can be obtained from the "shape".

Compared with the sample covariance estimator, Tyler's M-estimator is more robust to heavy-tailed elliptical distributions. The density function of elliptical distributions in $\mathbb{R}^p$ takes the form

$$f(\boldsymbol{x}; \Sigma, \mu) = |\Sigma|^{-1/2} g\{(\boldsymbol{x}-\mu)^\top\Sigma^{-1}(\boldsymbol{x}-\mu)\},$$

where g is some nonnegative function such that $\int_0^\infty x^{p-1}g(x)\,\mathrm{d}x$ is finite. This family of distributions is a natural generalization of the Gaussian distribution by allowing heavier or lighter tails while maintaining the elliptical geometry of the equidensity contours. Elliptical distributions are considered important in portfolio theory and financial data, and we

[*] Corresponding author.
E-mail addresses: teng.zhang@ucf.edu (T. Zhang), xiuyuan.cheng@yale.edu (X. Cheng), amits@math.princeton.edu (A. Singer).

http://dx.doi.org/10.1016/j.jmva.2016.03.010
0047-259X/© 2016 Elsevier Inc. All rights reserved.

refer to the work by El Karoui [11, Section 4] for further discussion. Besides, elliptical distributions are used by Ollila and Tyler [19] in modeling radar data, where the empirical distributions are heavy-tailed because of outliers.

Tyler [20] showed that when a data set follows an unknown elliptical distribution (with mean zero), Tyler's M-estimator is the most robust covariance estimator in the sense of minimizing the maximum asymptotic variance. This property suggests that Tyler's M-estimator should be more accurate than the sample covariance estimator for elliptically distributed data. Empirically, it has been shown to outperform the sample covariance estimator in applications such as finance in the work by Frahm and Jaekel [13], anomaly detection in wireless sensor networks by Chen et al. [4], antenna array processing by Ollila and Koivunen [18], and radar detection by Ollila and Tyler [19].

1.1. Asymptotic analysis in a high-dimensional setting

Many scientific domains customarily deal with sets of high dimensional data samples, and therefore it is increasingly common to work with data sets where the number of variables, p, is of the same order of magnitude as the number of observations, n. Under this high-dimensional setting, the asymptotic spectral properties of $\boldsymbol{S}_n$ at the limit of infinite number of samples and infinite dimensions have been well studied by Johnstone [15]. A noticeable example is the convergence of the spectral distribution. Denoting the eigenvalues of a matrix $\boldsymbol{A}$ by $\lambda_1(\boldsymbol{A}), \dots, \lambda_n(\boldsymbol{A})$, its spectral distribution is a discrete probability measure

$$P = P(\cdot|\boldsymbol{A}) = \frac{1}{n}\sum_{i=1}^{n} \delta_{\lambda_i(\boldsymbol{A})}$$

with δ_s denoting Dirac measure at $s \in \mathbb{R}$. Marčenko and Pastur [16] showed that when the entries of $\{\boldsymbol{x}_i\}_{i=1}^n$ are Gaussian independent identically distributed random variables with mean 0 and variance 1, $p, n \to \infty$ and $p/n \to y$, where $0 < y \le 1$, the spectral distribution of the eigenvalues of $\boldsymbol{S}_n$ converges weakly to the Marčenko–Pastur distribution defined by

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

Tyler's M-estimator is closely related to and can be considered as a special case of Maronna's M-estimator, which is defined by

$$\bar{\Sigma} = \frac{1}{n}\sum_{i=1}^{n} u(\boldsymbol{x}_i^\top \bar{\Sigma}^{-1}\boldsymbol{x}_i)\boldsymbol{x}_i\boldsymbol{x}_i^\top \tag{3}$$

for a nonnegative function $u : [0, \infty) \to [0, \infty)$. The properties of Maronna's M-estimator in the high-dimensional regime when $p, n \to \infty$, $p/n \to y$ and $0 < y < 1$ have been analyzed in recent works by Couillet et al. [7,8], which obtained convergence results for a properly scaled Maronna's M-estimator under the assumptions that $u(x)$ is nonnegative, nonincreasing and continuous; $xu(x)$ is nondecreasing and bounded and $\sup_x xu(x) > 1$. Moreover, spiked random matrix models were also studied by Couillet [5]. However, these results do not apply to Tyler's M-estimator, although Frahm and Jaekel [13] have conjectured that the spectral distribution converges weakly to the Marčenko–Pastur distribution. Some works focused on the performance of Tyler's M-estimator for the case $p, n \to \infty$ and $p/n \to 0$: Dümbgen [10] showed that the condition number of Tyler's estimator is $1 + 4\sqrt{p/n} + o(\sqrt{p/n})$, and Frahm and Glombek [12] showed that the spectral distribution of $\sqrt{n/p}(\bar{\Sigma} - \mathbf{I})$ converges weakly to a semicircle distribution.

1.2. Main results

In this paper, we analyze Tyler's M-estimator in the high-dimensional setting. Our main results, Theorem 2.3 and Corollary 2.4, show that as $p, n \to \infty$ and $p/n \to y$, $0 < y < 1$, the spectral distribution of a properly scaled Tyler's M-estimator converges weakly to the Marčenko–Pastur distribution $\rho_{\mathrm{MP},y}(x)$. Based on the properties of Tyler's M-estimator, this paper analyzes the spectral distribution when data samples are i.i.d. drawn from other distributions, such as elliptical distributions.

When data samples are generated from elliptical distributions, the spectral distribution of the sample covariance estimator has been studied by El Karoui [11, Theorem 2]. Compared to Corollary 2.4, the limiting spectral distribution of $\boldsymbol{S}_n$ is much more complicated, and therefore our result might be more applicable in practice.

High-dimensional analysis of Maronna's M-estimator of the covariance are generally obtained by showing that the operator norm of the difference between M-estimator and a standard Wishart matrix (or sample covariance matrix) tends to 0: Dümbgen [10] proved it by a linear expansion of the M-estimator, and Couillet et al. [7,8] proved it by representing Maronna's M-estimator as a weighted sum of $\boldsymbol{x}_i\boldsymbol{x}_i^\top$ and proved the uniform convergence of the weights. We follow the same direction while giving an alternate proof for the convergence of the weights, by considering the weights as the solution to an optimization problem, which can handle Tyler's M-estimator that is not covered by the results in Couillet et al. [7,8]. We remark that this approach can also be applied to Maronna's M-estimator to prove some of the results in Couillet et al. [7,8].

The rest of the paper is organized as follows. In Section 2 we introduce the representation of Tyler's M-estimator as a linear combination of $\boldsymbol{x}_i\boldsymbol{x}_i^\top$ and present the main result that when the data set is i.i.d. sampled from the Gaussian distribution $\mathcal{N}(\mathbf{0},\mathbf{I})$, a properly scaled Tyler's is asymptotically equivalent to $\boldsymbol{S}_n$ in the sense that $\|p\hat{\Sigma}-\boldsymbol{S}_n\|\to 0$. As a result, the spectral distribution of Tyler's M-estimator converges weakly to the Marčenko–Pastur distribution. We also extend the result to elliptical distributions. The technical proofs are given in Section 3. While some Lemmas and technical proofs are also used by Couillet et al. [7,8] (for example, Lemma 3.2 and the analysis in the proof of Theorem 2.3 are similar to Couillet et al. [7, Lemma 2, Theorem 1] and Couillet et al. [8, Lemma 6]), we still include them for the completeness of the paper.

As for notations, we 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. Tyler's *M*-estimator in the high-dimensional regime

We introduce the representation of Tyler's M-estimator as a linear combination of $\boldsymbol{x}_i\boldsymbol{x}_i^\top$ in Section 2.1, and present the main result in Section 2.2 that when the data set is i.i.d. sampled from the Gaussian distribution $\mathcal{N}(\mathbf{0},\mathbf{I})$, $\|p\hat{\Sigma}-\boldsymbol{S}_n\|$ converges to 0 almost surely. Based on this observation, we prove that the spectral distribution of Tyler's M-estimator converges weakly the Marčenko–Pastur distribution in Section 2.3. The generalization of the results to more general settings is also discussed in Section 2.3.

2.1. Properties of Tyler's M-estimator

The analysis for Tyler's M-estimator in this paper is based on the following representation, whose proof is deferred to Section 3. We remark that Eq. (5) in Lemma 2.1 has appeared in the work by Wiesel [21, (27)] and Hardt and Moitra [14, Section A] as "covariance estimation in scaled Gaussian distributions" and "Barthe's convex program", but its connection to Tyler's M-estimator has not been rigorously justified yet.

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

$$\hat{\Sigma}=\sum_{i=1}^n \hat{w}_i\boldsymbol{x}_i\boldsymbol{x}_i^\top \Big/ \operatorname{tr}\Big(\sum_{i=1}^n \hat{w}_i\boldsymbol{x}_i\boldsymbol{x}_i^\top\Big), \tag{4}$$

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

$$(\hat{w}_1,\ldots,\hat{w}_n)=\underset{w_i>0,\sum_{i=1}^n w_i=1}{\arg\min}\; -\sum_{i=1}^n \ln w_i+\frac{n}{p}\ln\det\Big(\sum_{i=1}^n w_i\boldsymbol{x}_i\boldsymbol{x}_i^\top\Big). \tag{5}$$

2.2. Isotropic Gaussian distribution

In this subsection, we assume that $\{\boldsymbol{x}_i\}_{i=1}^n\subset\mathbb{R}^p$ are i.i.d. drawn from $\mathcal{N}(\mathbf{0},\mathbf{I})$. The main result, Theorem 2.3, characterizes the convergence and convergence rate of Tyler's M-estimator to $\boldsymbol{S}_n$ in terms of the operator norm. Its proof applies Lemma 2.2, whose proof is rather technical and deferred to Section 3.

Tyler's M-estimator does not exist when $p > n$ (see the argument by Zhang [22, Theorem III.1]) and it is not unique when $p = n$ (one may check that when $\boldsymbol{x}_i = \boldsymbol{e}_i$ for all $1 \le i \le p$, all diagonal matrices with trace 1 satisfy (1)). As a result, throughout the paper we assume $y < 1$.

Lemma 2.2. *If* $\{\boldsymbol{x}_i\}_{i=1}^n$ *are i.i.d. sampled from* $\mathcal{N}(\mathbf{0},\mathbf{I})$*, then* $\max_{1\le i\le n}|n\,\hat{w}_i-1|$ *converges to* 0 *almost surely as* $p, n\to\infty$*: There exist* $C, c, c' > 0$ *such that for any* $\varepsilon < c'$,

$$\Pr\left(\max_{1\le i\le n}|n\,\hat{w}_i-1|\le\varepsilon\right)\ge 1-Cne^{-c\varepsilon^2 n}. \tag{6}$$

Theorem 2.3. *Suppose that* $\{\boldsymbol{x}_i\}_{i=1}^n$ *are i.i.d. sampled from* $\mathcal{N}(\mathbf{0},\mathbf{I})$, $p, n\to\infty$ *and* $p/n=y$*, where* $0<y<1$*, then the operator norm of the difference between* $\boldsymbol{S}_n$ *and a scaled Tyler's M-estimator converges to* 0 *almost surely, and there exist* $C, c, c' > 0$ *such that for any* $\varepsilon < c'$,

$$\Pr\left(\left\|p\,\hat{\Sigma}-\frac{1}{n}\sum_{i=1}^n\boldsymbol{x}_i\boldsymbol{x}_i^\top\right\|\le\varepsilon\right)\ge 1-Cne^{-c\varepsilon^2 n}. \tag{7}$$

Theorem 2.3 implies that all first order properties of the sample covariance matrix extend to Tyler's estimator. The strategy of the proof for Theorem 2.3 is as follows. According to Lemma 2.1, a scaled Tyler's M-estimator is a linear combination of $\boldsymbol{x}_i\boldsymbol{x}_i^\top$, i.e., it can be written as $\sum_{i=1}^n \hat{w}_i\boldsymbol{x}_i\boldsymbol{x}_i^\top$ (up to a scaling). Then Lemma 2.2 shows that $n\hat{w}_i$ converges to 1 uniformly. Based on the following matrix analysis, Theorem 2.3 is concluded.

Proof of Theorem 2.3. We first prove that for $\varepsilon < c'$,

$$\Pr\left(\left\|\sum_{i=1}^n \hat{w}_i\boldsymbol{x}_i\boldsymbol{x}_i^\top - \frac{1}{n}\sum_{i=1}^n \boldsymbol{x}_i\boldsymbol{x}_i^\top\right\| \leq \varepsilon\right) \geq 1 - Cne^{-c\varepsilon^2 n}. \tag{8}$$

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

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

Since $\|n\hat{\boldsymbol{w}} - \mathbf{1}\|_\infty \to 0$ with probability estimated in (6), and Davidson and Szarek [9, Theorem II.13] showed that $\|\sum_{i=1}^n\boldsymbol{x}_i\boldsymbol{x}_i^\top/n\|$ is bounded above by $(1+2\sqrt{y})^2$ with probability $1 - C\exp(-cn)$, (8) is proved.

Second, since

$$\left\|\sum_{i=1}^n\hat{w}_i\boldsymbol{x}_i\boldsymbol{x}_i^\top\right\| \leq \left\|\sum_{i=1}^n\hat{w}_i\boldsymbol{x}_i\boldsymbol{x}_i^\top - \sum_{i=1}^n\boldsymbol{x}_i\boldsymbol{x}_i^\top/n\right\| + \left\|\sum_{i=1}^n\boldsymbol{x}_i\boldsymbol{x}_i^\top/n\right\|,$$

$$\Pr\left(\left\|\sum_{i=1}^n\hat{w}_i\boldsymbol{x}_i\boldsymbol{x}_i^\top\right\| < C'\right) > 1 - Cn\exp(-cn). \tag{9}$$

Besides, $\mathrm{tr}(\sum_{i=1}^n\hat{w}_i\boldsymbol{x}_i\boldsymbol{x}_i^\top) = \sum_{i=1}^n\hat{w}_i\boldsymbol{x}_i^\top\boldsymbol{x}_i \to p$ in the same rate as in (9): applying the concentration of high-dimensional Gaussian measure on the sphere by Barvinok [2, Corollary 2.3], we have

$$\begin{aligned}&\max\left[\Pr\left\{\sum_{i=1}^n\hat{w}_i\boldsymbol{x}_i^\top\boldsymbol{x}_i < p(1-\varepsilon)\right\}, \Pr\left\{\sum_{i=1}^n\hat{w}_i\boldsymbol{x}_i^\top\boldsymbol{x}_i > p/(1-\varepsilon)\right\}\right] \\ &\quad\leq \max\left[\Pr\left\{\min_{1\leq i\leq n}\|\boldsymbol{x}_i\|^2 < p(1-\varepsilon)\right\}, \Pr\left\{\max_{1\leq i\leq n}\|\boldsymbol{x}_i\|^2 > p/(1-\varepsilon)\right\}\right] < ne^{-\varepsilon^2 p/4}.\end{aligned} \tag{10}$$

Combining (9), (10) and (4),

$$\left\|\sum_{i=1}^n\hat{w}_i\boldsymbol{x}_i\boldsymbol{x}_i^\top - p\,\hat{\Sigma}\right\| = \left\|\sum_{i=1}^n\hat{w}_i\boldsymbol{x}_i\boldsymbol{x}_i^\top\right\|\left\{1 - p/\mathrm{tr}\left(\sum_{i=1}^n\hat{w}_i\boldsymbol{x}_i\boldsymbol{x}_i^\top\right)\right\} \tag{11}$$

converges in the same rate as specified in (8). (7) is then proved by combining (8), (11) and the triangle inequality. □

From the probabilistic estimation (7) we obtain a convergence rate of $O(\sqrt{\ln n/n})$. The logarithmic factor is due to a “max” bound of $\{\hat{w}_i\}_{i=1}^n$ in Lemma 2.2, while in fact, an “average” bound is expected. As a result, we conjecture that this $\sqrt{\ln n}$ factor could be possibly removed by a more rigorous argument.

2.3. More general distributions and spectral distribution

We remark that Theorem 2.3 can be extended from the setting of the normal distribution $\mathcal{N}(\mathbf{0},\mathbf{I})$ to any elliptical distribution μ_p, which is characterized by the probability density function $\mu_p(\boldsymbol{x}) = C(g_p)\det(\boldsymbol{T}_p)^{-1/2}g_p(\boldsymbol{x}^\top\boldsymbol{T}_p^{-1}\boldsymbol{x})$, 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. Then $\|\mathrm{tr}(\boldsymbol{T}_p)\hat{\Sigma} - \sum_{i=1}^n\boldsymbol{x}_i\boldsymbol{x}_i^\top/n\| \to 0$ almost surely as $p, n \to \infty$. The analysis is based on Theorem 2.3, the affine equivariance property of Tyler's M-estimator, and the fact that Tyler's M-estimator is unchanged if $\{\boldsymbol{x}_i\}_{i=1}^n$ are replaced by $\{c_i\boldsymbol{x}_i\}_{i=1}^n$.

Another direction of generalization of Theorem 2.3 is the model by Couillet et al. [7]: The elements of $\{\boldsymbol{x}_i\}_{i=1}^n$ are i.i.d. sampled from an either real or circularly symmetric complex distribution with $\mathrm{E}(x_{ij}) = 0$, $\mathrm{E}(x_{ij}^2) = 1$, and $\mathrm{E}(|x_{ij}|^{8+\eta}) < \alpha$ for some $\eta, \alpha > 0$. Then, following the proof in this paper (while replacing Lemma 3.2 by [7, Lemma 2]), one can show that $\left\|p\,\hat{\Sigma} - \sum_{i=1}^n\boldsymbol{x}_i\boldsymbol{x}_i^\top/n\right\| \to 0$ almost surely as $p, n \to \infty$.

We have the following results on the weak convergence of the spectral distribution of Tyler's M-estimator, where the first part proves the conjecture by Frahm and Jaekel [13].

Corollary 2.4.
- *If $\{\boldsymbol{x}_i\}_{i=1}^n$ are i.i.d. sampled from $\mathcal{N}(\mathbf{0},\mathbf{I})$, then the spectral measure $P(\cdot|p\hat{\Sigma})$ converges weakly to the Marčenko–Pastur distribution.*
- *If $\{\boldsymbol{x}_i\}_{i=1}^n$ are i.i.d. sampled from an elliptical distribution $C(g_p)g_p(\boldsymbol{x}^\top \boldsymbol{T}_p^{-1}\boldsymbol{x})$ such that the spectral measure of $\boldsymbol{T}_p$ converges weakly to a distribution H on $\mathbb{R}$. Then the spectral measure $P(\cdot|\mathrm{tr}(\boldsymbol{T}_p)\hat{\Sigma})$ converges weakly to a probabilistic measure ρ whose Stieltjes transform $s(z)=\int 1/(x-z)\rho(\,\mathrm{d}x)$ $(z\in\mathbb{C}\setminus\mathbb{R})$ is given implicitly by*

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

This corollary can be proved by combining $\|\mathrm{tr}(\boldsymbol{T}_p)\hat{\Sigma}-\sum_{i=1}^n \boldsymbol{x}_i\boldsymbol{x}_i^\top/n\|\to 0$, the analysis on the perturbation of eigenvalues by Bhatia [3, Corollary III.4.2], the spectral measure of $\sum_{i=1}^n \boldsymbol{x}_i\boldsymbol{x}_i^\top/n$ by Marčenko and Pastur [16], Bai and Silverstein [16,1, (6.1.2)] and Slutsky's Lemma.

3. Proof of lemmas

3.1. Proof of Lemma 2.1

We start with the definition

$$(\hat{z}_1,\ldots,\hat{z}_n)=\mathop{\arg\min}_{\sum_{i=1}^n z_i=1}\ln\det\Big(\sum_{i=1}^n e^{z_i}\boldsymbol{x}_i\boldsymbol{x}_i^\top\Big) \tag{12}$$

and

$$\hat{\Sigma}_z=\sum_{i=1}^n e^{\hat{z}_i}\boldsymbol{x}_i\boldsymbol{x}_i^\top. \tag{13}$$

The solution to (12) is unique, which follows from the convexity of the objective function (see Wiesel [21, Lemma 4]). Besides, noticing the equivalence between (12) and (5) (by plugging $w_i=e^{z_i}/(\sum_{i=1}^n e^{z_i})$ and $z_i=\ln w_i-(\sum_{i=1}^n \ln w_i-1)/n$), there exists $c_1>0$ such that $\hat{\Sigma}_z=c_1\hat{\Sigma}$.

Next we will prove that $\hat{\Sigma}_z$ satisfies

$$\sum_{i=1}^n \frac{\boldsymbol{x}_i\boldsymbol{x}_i^\top}{\boldsymbol{x}_i^\top\hat{\Sigma}_z^{-1}\boldsymbol{x}_i}=c\hat{\Sigma}_z,\quad \text{for some } c>0. \tag{14}$$

By checking the directional derivative of the objective function in (12), for any $(\delta_1,\ldots,\delta_n)$ with $\sum_{i=1}^n\delta_i=0$,

$$\sum_{i=1}^n \delta_i e^{\hat{z}_i}\boldsymbol{x}_i^\top\hat{\Sigma}_z^{-1}\boldsymbol{x}_i=0.$$

Therefore, there exists c_2 such that

$$e^{\hat{z}_i}\boldsymbol{x}_i^\top\hat{\Sigma}_z^{-1}\boldsymbol{x}_i=c_2,\quad \text{for all } 1\le i\le n. \tag{15}$$

Therefore, (14) is proved by applying (15) and (13):

$$\sum_{i=1}^n \frac{\boldsymbol{x}_i\boldsymbol{x}_i^\top}{\boldsymbol{x}_i^\top\hat{\Sigma}_z^{-1}\boldsymbol{x}_i}=\sum_{i=1}^n e^{\hat{z}_i}\boldsymbol{x}_i\boldsymbol{x}_i^\top/c_2=\hat{\Sigma}_z/c_2,$$

Since $\hat{\Sigma}_z=c_1\hat{\Sigma}$, (14) also holds when $\hat{\Sigma}_z$ is replaced by $\hat{\Sigma}$:

$$\sum_{i=1}^n \frac{\boldsymbol{x}_i\boldsymbol{x}_i^\top}{\boldsymbol{x}_i^\top\hat{\Sigma}^{-1}\boldsymbol{x}_i}=c\hat{\Sigma},\quad \text{for some } c>0. \tag{16}$$

At last, we will prove that $\hat{\Sigma}$ satisfies the definition of Tyler's M-estimator in (1), that is, the constant c in (16) is given by $c=n/p$. For the objective function

$$F(\Sigma)=\sum_{i=1}^n \ln(\boldsymbol{x}_i^\top\Sigma^{-1}\boldsymbol{x}_i)+c\ln\det(\Sigma),$$

its derivative with respect to Σ^{-1} is given by

$$\sum_{i=1}^{n} \boldsymbol{x}_i^\top (\boldsymbol{x}_i^\top \Sigma^{-1} \boldsymbol{x}_i)^{-1} \boldsymbol{x}_i - c\Sigma.$$

Therefore, $\hat{\Sigma}$ is a stationary point of $F(\Sigma)$. Since $F(\Sigma)$ is geodesically convex (argument follows directly from Wiesel [21] and Zhang [22]), $\hat{\Sigma}$ is the global minimizer of $F(\Sigma)$.

However, the minimizer of $F(\Sigma)$ exists only when $c = n/p$. Since $F(a\mathbf{I}) = \sum_{i=1}^n \ln(\boldsymbol{x}_i^\top \boldsymbol{x}_i) - n \ln a + c\, p \ln a$, we have

$$F(a\mathbf{I}) \to -\infty \begin{cases} \text{as } a \to 0, & \text{if } c > n/p \\ \text{as } a \to \infty, & \text{if } c < n/p. \end{cases}$$

Therefore, the constant c in (16) is given by $c = n/p$, and Lemma 2.1 is proved.

3.2. Proof of Lemma 2.2

We start with an outline of the proof, which consists of three parts. First, we rewrite the constrained optimization problem (5) to the problem of finding the root of $g(\boldsymbol{w})$, which will be defined in (17). 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 Lemmas 3.2, 3.3 and 3.1, and their proofs are postponed to subsequent sections.

Lemma 3.1. *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 3.2. *If $\boldsymbol{x}_i \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ for all $1 \le i \le n$, and $\boldsymbol{S} = \sum_{i=1}^n \boldsymbol{x}_i \boldsymbol{x}_i^\top / n$, then there exists $c, C, c' > 0$ such that for any $\varepsilon < c'$,*

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

Lemma 3.3. *For the $n \times n$ matrix $\boldsymbol{A}$ defined by $\boldsymbol{A}_{ij} = (\boldsymbol{x}_i^\top \boldsymbol{S}^{-1} \boldsymbol{x}_j)^2/(n\,p)$,* (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* $\|(\mathbf{I} - \boldsymbol{A} + c\mathbf{1}\mathbf{1}^\top)^{-1}\|_\infty < C_2$ *with probability* $1 - Cn\exp(-cn)$.

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

$$g(\boldsymbol{w}) = \nabla G(\boldsymbol{w} + \mathbf{1}), \tag{17}$$

where

$$G(\boldsymbol{w}) = -\sum_{i=1}^n \ln w_i + \frac{n}{p} \ln\det\left(\sum_{i=1}^n w_i \boldsymbol{x}_i \boldsymbol{x}_i^\top\right) + \frac{c_0}{2}\left(\sum_{i=1}^n w_i - n\right)^2, \tag{18}$$

and the constant c_0 will be specified later before (34).

It is easy to prove that the minimizer of $G(\boldsymbol{w})$ and the zeros of $\nabla G(\boldsymbol{w})$ must satisfy $\sum_{i=1}^n w_i = n$ (otherwise $n\boldsymbol{w}/(\sum_{i=1}^n w_i)$ is a better minimizer and $\nabla G(\boldsymbol{w})$ is nonzero). Therefore minimizing (18) is equivalent to minimizing $-\sum_{i=1}^n \ln w_i + n/p \cdot \ln\det(\sum_{i=1}^n w_i \boldsymbol{x}_i \boldsymbol{x}_i^\top)$ with constraint $\sum_{i=1}^n w_i = n$, which is the same as (5) except for the constraint. Noticing that a scaling of $\boldsymbol{w}$ increases $-\sum_{i=1}^n \ln w_i + n/p \cdot \ln\det(\sum_{i=1}^n w_i \boldsymbol{x}_i \boldsymbol{x}_i^\top)$ by a constant only depending on the scale, the minimizer of (18) is unique and it is $n\hat{\boldsymbol{w}}$, where $\hat{\boldsymbol{w}}$ is defined in (5). By the convexity of its equivalent problem (12), 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 ith component of function $g(\boldsymbol{w})$ is

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

Applying Lemma 3.2,

$$\Pr\left(\|g(\mathbf{0})\|_\infty < \varepsilon\right) \ge 1 - Cne^{-c\varepsilon^2 n}. \tag{19}$$

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

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

Applying Lemma 3.3,

$$\|\{\nabla g(\mathbf{0})\}^{-1}\|_\infty < C_2 \quad \text{with probability } 1 - Cne^{-cn}. \tag{20}$$

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

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

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

$$\|\nabla g(\boldsymbol{w}) - \nabla g(\mathbf{0})\|_\infty < (3 + 3n\|\boldsymbol{A}\|_\infty/p)\|\boldsymbol{w}\|_\infty. \tag{21}$$

At last we finish the third part of the proof of Lemma 2.2 by applying Lemma 3.1 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 (19) and (20), $\|f(\mathbf{0})\|_\infty \le \|(\nabla g(\mathbf{0}))^{-1}\|_\infty \|g(\mathbf{0})\|_\infty \to 0$ in the same rate as in (19) and $\|f(\mathbf{0})\|_\infty < \min(1/9C_5, 1/3)$ holds with probability $1 - Cne^{-cn}$. Due to (19), (21), and the boundedness of $\|\boldsymbol{A}\|_\infty$ (Lemma 3.3), $\|\nabla f(\boldsymbol{w}) - \nabla(\mathbf{0})\|_\infty < C_5\|\boldsymbol{w}\|_\infty$ also holds with probability $1 - Cne^{-cn}$. Therefore the assumption in Lemma 3.1 holds with probability $1 - Cne^{-cn}$ and there exists $\tilde{\boldsymbol{w}}$ such that $f(\tilde{\boldsymbol{w}}) = 0$ and

$$\|\tilde{\boldsymbol{w}}\|_\infty < 3\|f(\mathbf{0})\|_\infty. \tag{22}$$

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 (22) 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 (19), $\|n\hat{\boldsymbol{w}} - 1\|_\infty$ converges in the same rate and Lemma 2.2 is proved.

3.2.1. *Proof of Lemma 3.1*

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

$$\begin{aligned} f_j(\boldsymbol{w}) - f_j(\mathbf{0}) &= \int_{t=0}^{1} \left\langle \boldsymbol{e}_j \boldsymbol{w}^\top, \nabla f(t\,\boldsymbol{w})\right\rangle \,\mathrm{d}t \\ &= \int_{t=0}^{1} \left\langle \boldsymbol{e}_j \boldsymbol{w}^\top, \nabla f(t\,\boldsymbol{w}) - \nabla f(\mathbf{0}) + \mathbf{I}\right\rangle \,\mathrm{d}t = w_j + \int_{t=0}^{1} \boldsymbol{w}^\top\{\nabla f(t\,\boldsymbol{w}) - \nabla f(\mathbf{0})\}\boldsymbol{e}_j \,\mathrm{d}t \\ &\le w_j + \left\| \int_{t=0}^{1} \boldsymbol{w}^\top\{\nabla f(t\,\boldsymbol{w}) - \nabla f(\mathbf{0})\} \right\|_\infty \le w_j + C_5\|\boldsymbol{w}\|_\infty^2. \end{aligned} \tag{23}$$

Similarly

$$f_j(\boldsymbol{w}) - f_j(\mathbf{0}) \ge -C_5\|\boldsymbol{w}\|_\infty^2 + w_j. \tag{24}$$

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

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

1. $|w_i| < 2\eta$. Then apply (23) and (24) (they are applicable since for any $\boldsymbol{w} \in \mathcal{A}$, $\|\boldsymbol{w}\|_\infty \le 1$), we have $|f_i(\boldsymbol{w})| < |f_i(\mathbf{0})| + C_5\|\boldsymbol{w}\|_\infty^2 + |w_i| \le \eta + C_5(3\eta)^2 + 3\eta < (4+9C_5)\eta$ ($\eta^2 < \eta$ 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 (24),

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

   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 Lemma 3.1 is proved with $\tilde{\boldsymbol{w}}$ being the fixed point. □

3.2.2. *Proof of* Lemma 3.2

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

$$\boldsymbol{X}\boldsymbol{S}^{-1}\boldsymbol{X} = (\boldsymbol{U}\Sigma\boldsymbol{V}^\top)\left(\frac{1}{n}\boldsymbol{V}\Sigma^2\boldsymbol{V}^\top\right)^{-1}(\boldsymbol{U}\Sigma\boldsymbol{V}^\top), \tag{25}$$

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

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 $\{g_j\}_{j=1}^n \sim \mathcal{N}(0, 1)$. Applying the concentration result by Barvinok [2, Corollary 2.3], we have

$$\Pr\left\{\sum_{i=1}^n g_i^2 \ge \frac{n}{1-\varepsilon}\right\} \le e^{-\varepsilon^2 n/4} \tag{26}$$

and

$$\Pr\left\{\sum_{i=1}^n g_i^2 \le n(1-\varepsilon)\right\} \le e^{-\varepsilon^2 n/4}, \tag{27}$$

therefore

$$\begin{aligned}\Pr\left\{\frac{p(1-\varepsilon)^2}{n} \le \|\boldsymbol{u}_1\|^2 \le \frac{p}{n(1-\varepsilon)^2}\right\} &\ge \Pr\left\{p(1-\varepsilon) \le \sum_{i=1}^p g_i^2 \le \frac{p}{1-\varepsilon}\right\} \\ &+ \Pr\left\{n(1-\varepsilon) \le \sum_{i=1}^n g_i^2 \le \frac{n}{1-\varepsilon}\right\} \ge 1 - 2e^{-\varepsilon^2 p/4} - 2e^{-\varepsilon^2 n/4}.\end{aligned}$$

For $\varepsilon \le 0.1$, we have

$$\Pr\left\{\max_{1\le i\le n}\left|\frac{1}{p}\boldsymbol{x}_i^\top\boldsymbol{S}^{-1}\boldsymbol{x}_i - 1\right| \le \varepsilon\right\} \ge 1 - n\Pr\left\{\left|\|\boldsymbol{u}_1\|^2 - \frac{p}{n}\right| > \frac{p}{n}\varepsilon\right\}$$

$$\ge 1 - n\left[1 - \Pr\left\{\frac{p(1-\varepsilon/3)^2}{n} \le \|\boldsymbol{u}_1\|^2 \le \frac{p}{n(1-\varepsilon/3)^2}\right\}\right] \tag{28}$$

$$\ge 1 - 2ne^{-\varepsilon^2 p/36} - 2ne^{-\varepsilon^2 n/36}, \tag{29}$$

where the second inequality follows from $1 - 3\varepsilon \le (1-\varepsilon)^2$ and $1/(1-\varepsilon)^2 \le 1 + 3\varepsilon$.

3.2.3. *Proof of* Lemma 3.3

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

$$\sum_{1\le j\le n}\boldsymbol{A}_{ij} = \sum_{1\le j\le n}\frac{1}{np}\boldsymbol{x}_i^\top\boldsymbol{S}^{-1}\boldsymbol{x}_j\boldsymbol{x}_j^\top\boldsymbol{S}^{-1}\boldsymbol{x}_i = \boldsymbol{x}_i^\top\boldsymbol{S}^{-1}\left(\sum_{1\le j\le n}\boldsymbol{x}_j\boldsymbol{x}_j^\top\right)\boldsymbol{S}^{-1}\boldsymbol{x}_i/np \tag{30}$$

$$= \boldsymbol{x}_i^\top\boldsymbol{S}^{-1}(n\boldsymbol{S})\boldsymbol{S}^{-1}\boldsymbol{x}_i/np = \boldsymbol{x}_i^\top\boldsymbol{S}^{-1}\boldsymbol{x}_i/p, \tag{31}$$

it follows from (29) 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

$$\|\boldsymbol{A} - c_0\mathbf{1}\mathbf{1}^\top\|_\infty \le C_3 < 1 \quad \text{with probability } 1 - Cn\exp(-cn). \tag{32}$$

We start with the proof of (32) with another lemma:

Lemma 3.4. *There exists a* $c_4 > 0$ *such that with probability* $1 - C\exp(-cn)$,

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

Davidson and S. Szarek [9, Theorem II.13] showed that there exists $C_4 = C_4(y)$ such that $\|\boldsymbol{S}\| < C_4$ with probability $1 - Cn\exp(-cn)$. Therefore $\boldsymbol{x}_i^\top \boldsymbol{S}^{-1}\boldsymbol{x}_j \geq \boldsymbol{x}_i^\top \boldsymbol{x}_j / C_4$ and Lemma 3.4 implies that for any $1 \leq i \leq n$:

$$\sum_{j=1}^{n} I\left(\boldsymbol{x}_i^\top \boldsymbol{S}^{-1}\boldsymbol{x}_j > c_4\sqrt{p}/C_4\right) > 0.75 \quad \text{with probability } 1 - C\exp(-cn). \tag{33}$$

Let $c_0 = (c_4/C_4)^2/n$, then (33) implies

$$\sum_{1\leq j\leq n} |\boldsymbol{A}_{i,j} - c| \leq \sum_{1\leq j\leq n} |\boldsymbol{A}_{i,j}| - 0.25c\,n \leq \boldsymbol{x}_i^\top \boldsymbol{S}^{-1}\boldsymbol{x}_i/p - 0.25(c_4/C_4)^2, \tag{34}$$

where the last step follows from (31).

Applying the estimation of $\boldsymbol{x}_i^\top \boldsymbol{S}^{-1}\boldsymbol{x}_i/p$ in (29) and a union bound argument over all $1 \leq i \leq n$ to (34), (32) is proved for $C_3 = 1 + \eta - 0.25(c_4/C_4)^2$.

Lemma 3.3(b) follows from (32) with $C_2 = 1/(1 - C_3)$, where the expansion of $(\mathbf{I} - \boldsymbol{A} + c\mathbf{1}\mathbf{1}^\top)^{-1}$ exists since $\|\boldsymbol{A} + c\mathbf{1}\mathbf{1}^\top\| \leq \|\boldsymbol{A} + c\mathbf{1}\mathbf{1}^\top\|_\infty < 1$. Applying $\|\boldsymbol{B}_1\boldsymbol{B}_2\|_\infty \leq \|\boldsymbol{B}_1\|_\infty \|\boldsymbol{B}_2\|_\infty$, we have

$$\|(\mathbf{I} - \boldsymbol{A} + c\mathbf{1}\mathbf{1}^\top)^{-1}\|_\infty = \left\|\sum_{k=0}^{\infty} (c\mathbf{1}\mathbf{1}^\top - \boldsymbol{A})^k\right\|_\infty \leq \sum_{k=0}^{\infty} \|c\mathbf{1}\mathbf{1}^\top - \boldsymbol{A}\|_\infty^k \leq \sum_{k=0}^{\infty} C_3^i = \frac{1}{1 - C_3}. \tag{35}$$

3.2.4. Proof of Lemma 3.4

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

$$\mathrm{E}\left\{I\left(|\boldsymbol{x}_1^\top \boldsymbol{x}_2| > c_4\sqrt{p}\right)\right\} \geq 0.85. \tag{36}$$

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

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 (36) is proved by combining the estimations on $|g_1|$, $\boldsymbol{x}_1$ and $|\boldsymbol{x}_1^\top \boldsymbol{x}_2| = |g_1|\,\|\boldsymbol{x}_1\|$.

To obtain Lemma 3.4 from (36), we apply Hoeffding's inequality to the indicator function $I(|\boldsymbol{x}_i^\top \boldsymbol{x}_j| > c_4\sqrt{p})$ over all $1 \leq j \leq n, j \neq i$.

4. Summary

We showed that Tyler's M-estimator is asymptotically equivalent to $\boldsymbol{S}_n$ in the sense that $\|p\hat{\Sigma} - \boldsymbol{S}_n\| \to 0$ as $p, n \to \infty$ and $p/n \to y$, where $0 < y < 1$ and data samples follow the distribution of $\mathcal{N}(\mathbf{0}, \mathbf{I})$. We also proved the conjecture that the spectral distribution of Tyler's M-estimator converges weakly to the Marčenko–Pastur distribution, and extended the results to elliptical distributions.

There are several possible future directions of this work. First, it would be interesting to analyze the second order statistics of Tyler's M-estimator, considering that Couillet [6] has already investigated Maronna's M-estimators. Second, we would like to theoretically quantify the behavior of Tyler's M-estimator in the spiked covariance model by Couillet [5], which includes the analysis of the distribution of the top eigenvalue for the null cases and the analysis of the non-null case. A recent work by Morales-Jimenez et al. [17] on the non-null case introduced a mixture model that consists of a Gaussian distribution and some deterministic or random outliers, and analyzed the performance of Maronna's M-estimator. Analyzing the performance of Tyler's M-estimator in this model would be another possible future direction.

Acknowledgments

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, J. Silverstein, Spectral Analysis of Large Dimensional Random Matrices, in: Springer Series in Statistics, Springer, 2009.
[2] A. Barvinok, Math 710: Measure Concentration, in: Lecture Notes, Department of Mathematics, University of Michigan, 2005.
[3] R. Bhatia, Matrix Analysis, in: Graduate Texts in Mathematics, Number 169, Springer, New York, 1997.
[4] Y. Chen, A. Wiesel, A. Hero, Robust shrinkage estimation of high-dimensional covariance matrices, IEEE Trans. Signal Process. 59 (9) (2011) 4097–4107.
[5] R. Couillet, Robust spiked random matrices and a robust g-music estimator, J. Multivariate Anal. 140 (2015) 139–161.
[6] R. Couillet, A. Kammoun, F. Pascal, Second order statistics of robust estimators of scatter. application to GLRT detection for elliptical signals, J. Multivariate Anal. 143 (2016) 249–274.

[7] R. Couillet, F. Pascal, J. Silverstein, Robust estimates of covariance matrices in the large dimensional regime, IEEE Trans. Inform. Theory 60 (11) (2014) 7269–7278.
[8] R. Couillet, F. Pascal, J.W. Silverstein, The random matrix regime of Maronna's M-estimator with elliptically distributed samples, J. Multivariate Anal. 139 (2015) 56–78.
[9] K. Davidson, S. Szarek, Local operator theory, random matrices and banach spaces, in: Handbook of the Geometry of Banach Spaces, Vol. 1, 2001, p. 317.
[10] L. Dümbgen, On Tyler's M-functional of scatter in high dimension, Ann. Inst. Statist. Math. 50 (3) (1998) 471–491.
[11] N. El Karoui, Concentration of measure and spectra of random matrices: Applications to correlation matrices, elliptical distributions and beyond, Ann. Appl. Probab. 19 (6) (2009) 2362–2405.
[12] G. Frahm, K. Glombek, Semicircle law of Tyler's M-estimator for scatter, Statist. Probab. Lett. 82 (5) (2012) 959–964.
[13] G. Frahm, U. Jaekel, Tyler's M-estimator, random matrix theory, and generalized elliptical distributions with applications to finance. Technical report, 2007.
[14] M. Hardt, A. Moitra, Algorithms and hardness for robust subspace recovery, in: S. Shalev-Shwartz, I. Steinwart (Eds.), COLT, in: JMLR Proceedings, vol. 30, JMLR.org, 2013, pp. 354–375.
[15] I.M. Johnstone, High dimensional statistical inference and random matrices, in: International Congress of Mathematicians. Vol. I, Eur. Math. Soc., Zürich, 2007, pp. 307–333.
[16] V.A. Marčenko, L.A. Pastur, Distribution of eigenvalues for some sets of random matrices, Math. USSR-Sb. 1 (4) (1967) 457.
[17] D. Morales-Jimenez, R. Couillet, M. McKay, Large dimensional analysis of robust M-estimators of covariance with outliers, IEEE Trans. Signal Process. 63 (21) (2015) 5784–5797.
[18] E. Ollila, V. Koivunen, Robust antenna array processing using M-estimators of pseudo-covariance, in: 14th IEEE Proceedings on Personal, Indoor and Mobile Radio Communications, 2003. PIMRC 2003, volume 3, vol.3, 2003, pp. 2659–2663.
[19] E. Ollila, D. Tyler, Distribution-free detection under complex elliptically symmetric clutter distribution, in: 2012 IEEE 7th Sensor Array and Multichannel Signal Processing Workshop, SAM, June 2012, pp. 413–416.
[20] D.E. Tyler, A distribution-free M-estimator of multivariate scatter, Ann. Statist. 15 (1) (1987) 234–251.
[21] A. Wiesel, Geodesic convexity and covariance estimation, IEEE Trans. Signal Process. 60 (12) (2012) 6182–6189.
[22] T. Zhang, Robust subspace recovery by Tyler's M-estimator, Inf. Inference (2015).