Performance Analysis of Tyler's Covariance Estimator

Ilya Soloveychik and Ami Wiesel,

Rachel and Selim Benin School of Computer Science and Engineering, The Hebrew University of Jerusalem,
Israel

Abstract—This paper analyzes the performance of Tyler's Mestimator of the scatter matrix in elliptical populations. We focus on the non-asymptotic setting and derive the estimation error bounds depending on the number of samples n and the dimension p. We show that under quite mild conditions the squared Frobenius norm of the error of the inverse estimator decays like p^2/n with high probability.

Index Terms—Elliptical distribution shape matrix estimation, scatter matrix M-estimators, Tyler's scatter estimator, concentration bounds.

I. INTRODUCTION

Estimation of large covariance matrices, particularly in situations where the data dimension p and the sample size n are of close magnitudes has recently attracted considerable attention. Estimators in this field can be classified based on the underlying distribution and the additional structure assumptions. Most of the research is traditionally devoted to the multivariate Gaussian setting which is currently well understood. Various algorithms based on different structures and their performance analysis have been derived for the Gaussian distributions, see e.g. [1–3]. This allows for higher reliability and better tuning of regularization parameters. Recently, similar challenges have appeared in the more ambitious setting of non-Gaussian and robust estimation. A prominent approach in this area is Tyler's scatter estimator [4]. The goal of this paper is to analyze its non-asymptotic behavior as defined below more rigorously. We start by first reviewing the state of the art in both Gaussian and Tyler's covariance estimation, and then introduce our result.

Most of the works on covariance estimation address the Gaussian scenario. When the number of samples is greater than the dimension, the Maximum Likelihood Estimator (MLE) of the covariance exists with probability one and coincides with the Sample Covariance Matrix (SCM). Recently, there has emerged a great amount of literature on regularized versions of this estimator, their parameter tuning and performance analysis, such as shrinkage-estimator, see e.g. [1, 3].

Roughly speaking, covariance estimation performance analyses can be divided into three regimes. The first is classical asymptotic analysis. This approach assumes that the dimension of the underlying sample space p is fixed and that the number of samples grows $n \to \infty$. Typical results in this regime

concern asymptotic consistency and computation of asymptotic variance of an estimator in comparison to the Cramer-Rao lower bound. The second regime is also asymptotic and is based on Random Matrix Theory (RMT). It treats the case where both $n \to \infty$ and $p \to \infty$ while their ratio tends to a fixed number $p/n \to \zeta \in [0;\infty)$. One of the most important results of this kind concerning SCM is the Marchenko-Pastur law [5] defining the asymptotic distribution of the eigenvalues of the properly defined sequence of SCM. Other typical results include contributions on the bulk and edge regions of the spectrum of SCM, see e.g. [6–9]. These works led to numerous theoretical and practical breakthroughs in covariance estimation and its applications [10–15].

The third kind of results stems from the non-asymptotic analysis of the SCM, see e.g. [16–18]. In contrast to the previously described regimes, the non-asymptotic results answer the question what error bound can one obtain given the values of n and p. They are usually formulated in the form of concentration of measure results:

$$\left\|\widetilde{\mathbf{\Theta}} - \mathbf{\Theta}_0 \right\| \leq f(n,p)$$
 with probability at least $1 - e^{g(n,p)}$,

where the particular estimator $\widetilde{\Theta}$ and norm should be specified and the functions f and g should satisfy some properties. Unlike the asymptotic regimes, such kinds of results describe the speed of convergence of the estimator to the true covariance matrix Θ_0 .

Recently non-asymptotic analysis of SCM in Gaussian distributions has become popular among signal processing society due to advances in high dimensional statistics. In particular, [2, 3, 19] consider regularized covariance estimation and include its performance analysis. A common thread to all of these works is that the estimators are defined as the solutions of convex optimization problems, and the analysis is directly related to the notion of strong convexity.

In many applications the underlying multivariate distribution is actually non-Gaussian and robust covariance estimation methods are required. This occurs whenever the probability distribution of the measurements is heavy-tailed or a small proportion of the samples represents outlier behavior, [20, 21]. A common robust estimator of scatter is due to Tyler [4]. Given n independent, identically distributed (i.i.d.) measurements $\mathbf{x}_i \in \mathbb{R}^p, i = 1, \dots, n$, Tyler's shape matrix estimator is defined as the solution to the fixed point equation

$$\mathbf{T} = \frac{p}{n} \sum_{i=1}^{n} \frac{\mathbf{x}_i \mathbf{x}_i^T}{\mathbf{x}_i^T \mathbf{T}^{-1} \mathbf{x}_i}.$$
 (1)

When \mathbf{x}_i are Generalized Elliptically (GE) distributed [22], their shape matrix Θ_0 is positive definite and n > p, Tyler's estimator exists with probability one and is a consistent estimator of Θ_0 up to a positive scaling factor. The GE family includes as particular cases generalized Gaussian distribution, compound Gaussian, elliptical and many others [22]. Therefore, it has been successfully used to replace the SCM in many applications such as anomaly detection in wireless sensor networks [23], antenna array processing [24] and radar detection [25–28].

Performance analysis of Tyler's estimator dates back to the original work in classical robust statistics literature [4]. In the classical asymptotic regime where p is fixed and $n \to \infty$ it was shown that, when properly scaled, Tyler's estimator is strongly consistent with the true covariance matrix, if it exists, and is asymptotically normally distributed. Its asymptotic variance, which coincides with the Cramer-Rao bound, was analyzed in [29, 30]. RMT regime performance results were also reported. For the case $n, p \to \infty, p/n \to 0$ it was shown in [31] that the condition number of Tyler's estimator multiplied by the inverse true shape matrix of the underlying distribution converges to $1 + O(\sqrt{p/n})$. In [32] it was demonstrated that the empirical spectral distribution of $\sqrt{n/p}(\mathbf{T} - \mathbf{I})$ converges to the semicircle law in probability, when the underlying population is white and Tyler's estimator is properly scaled.

In the paper [33] the authors proved that Tyler's M-estimator converges in operator norm to the SCM matrix as $n,p\to\infty$ and $p/n\to\zeta\in(0,1)$, when data samples follow the standard normal distribution. The authors extended this result to elliptical distributions and proved that the empirical spectral density of Tyler's M-estimator converges to the Marchenko-Pastur distribution. The paper [33] also quantified the non-asymptotic behavior of Tyler's M-estimator, but in a different way with emphases on the RMT regime, the Marchenko-Pastur law, and comparison to the SCM. The paper [34] analyzed the asymptotic behavior of regularized Tyler's estimator in the RMT regime, which allowed optimal parameter tuning. Additional RMT results on other M-estimators were developed in [35, 36].

We focus on the the non-asymptotic analysis of Tyler's estimator for moderate values of n and p based on the concentration of measure phenomenon. Note that the estimator is not given in closed form and has to be iteratively computed using the fixed point iteration (1). In order to exploit the optimization based machinery discussed above, we rely on an alternative derivation of Tyler's estimator. In particular, the estimator can also be obtained as an MLE of normalized GE distributed vectors defined as $\frac{\mathbf{x}_i}{\|\mathbf{x}_i\|}$, [22]. Our method is therefore based on the analysis of the negative log-likelihood function of this distribution. We present high probability error bounds on the deviation of a properly scaled Tyler's estimator from the true underlying shape matrix. In particular, we prove that as long as n is larger than p the Frobenius norm of the error in inverse matrices decays like $\frac{p}{\sqrt{n}}$ with high probability. We also show that our performance bounds exhibit correct asymptotic behavior in the RMT regime. This contribution basically complements the previous classical asymptotic and RMT analysis of Tyler's method. The derivation generally

follows the optimization based approach due to [3, 19].

The paper is organized as following: first we introduce notations, state the problem, the main result and provide a discussion of it. Then we outline the strategy of the proof and bring in a few auxiliary results. After this we prove the main theorem. The body of the article contains the sketch of the proof with the statements of the most significant lemmas. Finally, we provide numerical simulations illustrating the obtained results. The proofs of lemmas and the rest of the auxiliary statements are left for the Appendices.

A. Notations

Denote by S(p) the linear space of $p \times p$ symmetric real matrices and by $\mathcal{P}(p) \subset S(p)$ the closed cone of positive semi-definite matrices. I stands for the identity matrix of a proper dimension.

We endow S(p) with the scalar product $(\mathbf{A}, \mathbf{B}) = \operatorname{Tr}(\mathbf{A}\mathbf{B})$, which induces the Frobenius norm on it. $\|\cdot\|$ will denote the Euclidean norm for vectors, $\|\cdot\|_F$ - the Frobenius norm and $\|\cdot\|_2$ - the spectral norm for matrices. The linear space \mathbb{R}^p is treated as a column vector space with the standard inner product. Given a matrix \mathbf{A} we denote by $\operatorname{vec}(\mathbf{A})$ a column vector obtained by stacking the columns of \mathbf{A} .

For a matrix $\mathbf{A} \in \mathcal{P}(p)$ denote by $\lambda_{\min}(\mathbf{A})$ and $\lambda_{\max}(\mathbf{A})$ its minimal and maximal eigenvalues, correspondingly. When $\lambda_{\min}(\mathbf{A}) > 0$ we write $\mathbf{A} \succ 0$ and denote by $\kappa(\mathbf{A}) = \frac{\lambda_{\max}(\mathbf{A})}{\lambda_{\min}(\mathbf{A})}$ its condition number. $|\mathbf{A}|$ stands for the determinant of \mathbf{A} .

Let Q be a quadratic form over a finite dimensional Euclidean space V, define its minimal and maximal eigenvalues as

$$\lambda_{\min}(Q) = \inf_{\mathbf{a} \in \mathcal{V}, \|\mathbf{a}\| = 1} Q(\mathbf{a}), \quad \lambda_{\max}(Q) = \sup_{\mathbf{a} \in \mathcal{V}, \|\mathbf{a}\| = 1} Q(\mathbf{a}).$$
(2)

The norm of Q is defined as

$$||Q||_2 = \max(|\lambda_{\min}(Q)|, |\lambda_{\max}(Q)|).$$

For n instances a_1, \ldots, a_n of scalars, vectors, matrices or functions we denote by \widehat{a} their arithmetic average, when the index of summation is obvious from the context.

Matrices are denoted by Capital bold letters M, column vectors by non-capital bold v, scalars by non-capital r, operators and quadratic forms by Capital T letters.

II. TYLER'S ESTIMATOR AS A MLE

We define Tyler's estimator as an MLE of a shape matrix parameter of a specific real spherical p-dimensional distribution. The likelihood function of this distribution is later used to derive error bounds of the estimator using its curvature properties.

Definition 1. Assume $\Theta_0 \in \mathcal{P}(p), \Theta_0 \succ 0$, then the function

$$p(\mathbf{x}) = \frac{\Gamma(p/2)}{2\sqrt{\pi^p}} \frac{1}{\sqrt{|\mathbf{\Theta}_0|} (\mathbf{x}^T \mathbf{\Theta}_0^{-1} \mathbf{x})^{p/2}}$$
(3)

is a probability density function of a vector $\mathbf{x} \in \mathbb{R}^p$ lying on a unit sphere. This distribution is usually referred to as the Angular Central Gaussian (ACG) distribution on a sphere [37]

and we denote it as $\mathbf{x} \sim \mathcal{U}(\boldsymbol{\Theta}_0)$. The matrix $\boldsymbol{\Theta}_0$ is referred to as a shape matrix of the distribution and is a multiple of the covariance matrix of \mathbf{x} .

The ACG distribution is closely related to the class of GE distributions, which includes Gaussian, compound Gaussian, elliptical, skew-elliptical, ACG and other distributions, [38]. An important property of the GE family is that the shape matrix of a population does not change when the vector is divided by its Euclidean norm [22, 38]. After normalization, any GE vector becomes ACG distributed. This allows us to treat all these distributions together using a single robust estimator.

Assuming $\Theta \in \mathcal{P}(p), \Theta \succ 0$ and given n > p i.i.d. copies of a vector $\mathbf{x} \sim \mathcal{U}(\Theta) \colon \mathbf{x}_i, 1 = 1, \dots, n$ we derive the MLE estimator of the shape matrix. For this sake introduce the scaled negative log-likelihood function:

$$\widetilde{f}(\mathbf{\Theta}; \mathbf{x}) = \log|\mathbf{\Theta}| + p\log(\mathbf{x}^T \mathbf{\Theta}^{-1} \mathbf{x}).$$
 (4)

The function (4) is non-convex in Θ . Nevertheless, its critical point given as the solution to (1) provides the global minima with probability one, [39–41].

The negative log-likelihood (4) is invariant under multiplication of the shape matrix by a positive constant, thus we are only interested in the estimation of the shape matrix up to a positive scalar factor. In order to obtain a unique MLE we fix the scale of the estimator by assuming that $\operatorname{Tr}\left(\boldsymbol{\Theta}_{0}^{-1}\right)$ of the true covariance matrix is known (or arbitrarily fixed). Specifically, we define Tyler's estimator to be the solution to the program

$$\mathbf{T} = \arg \begin{cases} \min_{\boldsymbol{\Theta}} & \frac{1}{n} \sum_{i=1}^{n} \widetilde{f}(\boldsymbol{\Theta}; \mathbf{x}_{i}) \\ \text{subject to} & \operatorname{Tr}(\boldsymbol{\Theta}^{-1}) = \operatorname{Tr}(\boldsymbol{\Theta}_{0}^{-1}). \end{cases}$$
 (5)

III. THE MAIN RESULT

In this section we introduce the main result of the paper and compare it with the similar results and related works. Denote $\Omega_0 = \Theta_0^{-1}, \underline{\lambda} := \lambda_{\min}(\Theta_0) = \lambda_{\max}^{-1}(\Omega_0) > 0$ and

$$\cos \phi_0 = \frac{\operatorname{Tr}(\mathbf{\Omega}_0)}{\|\mathbf{I}\|_F \|\mathbf{\Omega}_0\|_F} > 0.$$
 (6)

Theorem 1. Assume we are given n > p i.i.d. copies of $\mathbf{x} \sim \mathcal{U}(\mathbf{\Theta}_0)$, then for $\theta \geq 0$ with probability at least

$$1 - 2\exp\left(\frac{-\theta^2}{2(1+1.7\frac{\theta}{\sqrt{n}})}\right) - 2p^2\exp\left(-\frac{n\cos^2\phi_0}{77\ln(7p)(1+\frac{2}{p})}\right)\left(1 + \frac{15\cdot10^3(1+\frac{2}{p})^4}{n^2\cos^8\phi_0}\right)$$

Tyler's estimator (5) satisfies

$$\|\mathbf{T}^{-1} - \mathbf{\Theta}_0^{-1}\|_F \le \theta \frac{10}{\lambda \cos^2 \phi_0} \frac{p+2}{\sqrt{n}}.$$
 (8)

We note that for the Gaussian populations the same technique provides a similar up to a constant factor bound for the SCM estimator, suggesting that both estimators are bounded by a multiple of $\frac{p}{\sqrt{n}}$ with high probability.

A. Identity Covariance

The value of $\cos \phi_0$ is close to 1 if the condition number $\kappa(\Theta_0)$ is close to 1 and gets close to 0 if the matrix Θ_0 has many small eigenvalues and few large ones. As we know the estimation of the matrix becomes less stable in the latter case, as the theorem suggests. To make the statement of Theorem 1 easier to grasp let us treat the case of $\Theta_0 = \mathbf{I}$ in the following

Corollary 1. Assume we are given n > p i.i.d. copies of $\mathbf{x} \sim \mathcal{U}(\mathbf{I}), \theta < \frac{\sqrt{n}}{4}$, then

$$\begin{split} & \mathbb{P}\left(\left\|\mathbf{T}^{-1} - \mathbf{I}\right\|_F \geq 10\theta \frac{p+2}{\sqrt{n}}\right) \leq 2 \exp\left(\frac{-\theta^2}{3}\right) \\ & + 2p^2 \exp\left(-\frac{n}{77\ln(7p)(1+\frac{2}{p})}\right) \left(1 + \frac{15 \cdot 10^3(1+\frac{2}{p})^4}{n^2}\right). \end{split}$$

This corollary illustrates the essential behavior of the bound: as n and p get large the second probability term vanishes and we get a large deviation-type bound.

B. Choice of the Norm

Our result and the corresponding results of [3, 19] for the Gaussian settings are formulated in terms of the inverse of the estimated matrix. Actually the necessity to formulate the convergence rates in terms of inverse matrices is twofold. First of all, in many applications one is mostly interested in the inverse covariance matrix estimation, e.g. in regression and denoising. Secondly, this choice is also dictated by the properties of the negative log-likelihood function. It is more convenient for the analysis to parametrize this function by the inverse covariance matrix.

In addition, the fact that the Frobenius and not the spectral norm is used in the bounds also deserves explanation. In the Gaussian case stronger bounds can be obtained with the spectral norm. In fact it can be shown that the rate of convergence of the SCM matrix to the true covariance is of order $\sqrt{\frac{p}{n}}$ in the spectral norm, see [16] and references therein. In our case the problem of obtaining spectral norm non-asymptotic bounds is much more involved and remains an open question.

IV. THE PROOF STRATEGY

The proof of Theorem 1 follows a technique similar to that of [3, 19, 42]. The method of obtaining sample complexity rates proposed by these papers consists of finding the smallest ball around the true inverse covariance that contains the estimator with high probability. In its turn, this is done by considering the second order Taylor expansion of the sample average negative log-likelihood and estimating the maximal radius of the ball, for which the increase of this function on the boundary of the ball is positive with high probability.

The reasoning is as following: assume we are given a continuous function g over a finite dimensional Euclidean space which is known to have a unique local minimum. If we also know that $g(\mathbf{x}) < g|_{\partial \mathcal{B}(\mathbf{x};\rho)}$, where $\mathcal{B}(\mathbf{x};\rho)$ is a closed ball of radius ρ around \mathbf{x} , then the unique minumum of the function belongs to the open ball $\mathcal{B}(\mathbf{x};\rho) \setminus \partial \mathcal{B}(\mathbf{x};\rho)$. Indeed,

g is continuous over a compact set $\mathcal{B}(\mathbf{x}; \rho)$, thus reaches its extrema on it. The minimum of g cannot lie on the boundary since $g|_{\partial \mathcal{B}(\mathbf{x}; \rho)} > g(\mathbf{x})$, so it is strictly inside the ball.

Strong convexity of g is a sufficient condition for it having a unique local minimum. For this purpose any notion of convexity is suitable, since the uniqueness of the minimum does not depend on the metric. The second condition $g(\mathbf{x}) < g|_{\partial \mathcal{B}(\mathbf{x};\rho)}$ is usually demonstrated using strong local convexity of g in the vicinity of the true parameter value. Due to the scale invariance, the Hessian of Tyler's objective has a zero eigenvalue in the vicinity of the true parameter. The trace constraint addresses this invariance and ensures that the Hessian is positive definite.

V. PRELIMINARY RESULTS FOR THE ACG DISTRIBUTION

The analysis becomes easier if the negative log-likelihood is parametrized by the inverse shape matrix. We denote $\Omega = \Theta^{-1}$ and write

$$f(\mathbf{\Omega}; \mathbf{x}) = \widetilde{f}(\mathbf{\Omega}^{-1}; \mathbf{x}) = -\log|\mathbf{\Omega}| + p\log(\mathbf{x}^T \mathbf{\Omega} \mathbf{x}).$$
 (9)

Slightly abusing the notations, below we refer to (9) as the negative log-likelihood of the ACG population.

A. Derivatives and Their Expectations

The negative log-likelihood gradient and Hessian read as

$$\nabla f_{\Omega} = -\Omega^{-1} + p \frac{\mathbf{x} \mathbf{x}^{T}}{\mathbf{x}^{T} \Omega \mathbf{x}},$$

$$\nabla^{2} f_{\Omega} = \Omega^{-1} \otimes \Omega^{-1} - p \frac{\mathbf{x} \mathbf{x}^{T}}{\mathbf{x}^{T} \Omega \mathbf{x}} \otimes \frac{\mathbf{x} \mathbf{x}^{T}}{\mathbf{x}^{T} \Omega \mathbf{x}}.$$
 (10)

These and their expectations can be considered as linear and quadratic forms over S(p) respectively:

$$\nabla f_{\Omega}(\mathbf{U}) = -\text{Tr}\left(\mathbf{\Omega}^{-1}\mathbf{U}\right) + p\frac{\mathbf{x}^{T}\mathbf{U}\mathbf{x}}{\mathbf{x}^{T}\mathbf{\Omega}\mathbf{x}},$$
 (11)

$$\nabla^{2} f_{\Omega}(\mathbf{U}) = \operatorname{Tr}\left(\mathbf{\Omega}^{-1} \mathbf{U} \mathbf{\Omega}^{-1} \mathbf{U}\right) - p \left(\frac{\mathbf{x}^{T} \mathbf{U} \mathbf{x}}{\mathbf{x}^{T} \mathbf{\Omega} \mathbf{x}}\right)^{2}$$
$$= \left\|\mathbf{\Omega}^{-1/2} \mathbf{U} \mathbf{\Omega}^{-1/2}\right\|_{F}^{2} - p \left(\frac{\mathbf{x}^{T} \mathbf{U} \mathbf{x}}{\mathbf{x}^{T} \mathbf{\Omega} \mathbf{x}}\right)^{2}. \tag{12}$$

Let us compute the expectations of the derivatives for $\mathbf{x} \sim \mathcal{U}(\mathbf{\Theta}_0)$. Denote

$$T_{\mathbf{\Omega}} = \mathbb{E}\left(\nabla f_{\mathbf{\Omega}}\right), H_{\mathbf{\Omega}} = \mathbb{E}\left(\nabla^2 f_{\mathbf{\Omega}}\right),$$

$$R^{\nu}(\mathbf{U}, \mathbf{\Omega}; \mathbf{\Theta}_0) = \mathbb{E}\left[\left(\frac{\mathbf{x}^T \mathbf{U} \mathbf{x}}{\mathbf{x}^T \mathbf{\Omega} \mathbf{x}}\right)^{\nu}\right], \nu \in \mathbb{N}, \quad (13)$$

then

$$T_{\Omega}(\mathbf{U}) = -\text{Tr}\left(\Omega^{-1}\mathbf{U}\right) + pR^{1}(\mathbf{U}, \Omega; \boldsymbol{\Theta}_{0}),$$

$$H_{\Omega}(\mathbf{U}) = \text{Tr}\left((\Omega^{-1}\mathbf{U})^{2}\right) - pR^{2}(\mathbf{U}, \Omega; \boldsymbol{\Theta}_{0}).$$
 (14)

In particular, at the true parameter value Ω_0 we obtain

$$T_{\Omega_0}(\mathbf{U}) = 0, \forall \mathbf{U} \in \mathcal{S}(p),$$

$$\begin{split} H_{\mathbf{\Omega}_0}(\mathbf{U}) &= \frac{p \mathrm{Tr} \left((\mathbf{\Theta}_0 \mathbf{U})^2 \right) - \left(\mathrm{Tr} \left(\mathbf{\Theta}_0 \mathbf{U} \right) \right)^2}{p+2} \\ &= \frac{p \left\| \mathbf{\Theta}_0^{1/2} \mathbf{U} \mathbf{\Theta}_0^{1/2} \right\|_F^2 - \left(\mathrm{Tr} \left(\mathbf{\Theta}_0^{1/2} \mathbf{U} \mathbf{\Theta}_0^{1/2} \right) \right)^2}{p+2}. \end{split}$$

These formulas follow directly from (11), (12) and formulas (22), (23) from Appendix A. Our analysis is based on strong convexity of the Hessian. It can be easily shown that $H_{\Omega_0}(\mathbf{U})$ has a one-dimensional kernel spanned by Ω_0 . In what follows we show that the trace constraint in (5) eliminates this direction and makes the Hessian strongly convex in the vicinity of Ω_0 with high probability.

In order to explore the convexity properties of $H_{\Omega_0}(\mathbf{U})$ let us state the following simple

Lemma 1. Let V be a Euclidean space and $S \subset V$ be its subspace of codimension one with normal vector \mathbf{n} , then for any $\mathbf{v} \in V$

$$\sin^2(\angle \mathbf{u},\mathbf{v}) \geq \cos^2(\angle \mathbf{n},\mathbf{v}) = \frac{(\mathbf{n},\mathbf{v})}{\|\mathbf{n}\| \|\mathbf{v}\|}, \forall \mathbf{u} \in \mathcal{S}.$$

Proof. Among all the vectors $\mathbf{u} \in \mathcal{S}$ the one that minimizes the angle $(\angle \mathbf{u}, \mathbf{v})$ is coplanar with \mathbf{n} and \mathbf{v} . Now the statement follows from Pythagoras theorem.

Let us now turn to the space S(p) and consider its subspace \mathcal{L} defined by the condition $\operatorname{Tr}(\mathbf{U}) = 0$. Apply to this subspace the following linear transformation:

$$\mathcal{L}' = \mathbf{\Theta}_0^{1/2} \mathcal{L} \mathbf{\Theta}_0^{1/2}. \tag{15}$$

 Ω_0 is a normal vector of \mathcal{L}' , since $(\boldsymbol{\Theta}_0^{1/2}\mathbf{U}\boldsymbol{\Theta}_0^{1/2},\Omega_0) = \operatorname{Tr}(\mathbf{U}) = 0, \forall \mathbf{U} \in \mathcal{L}$. Recall that

$$\begin{aligned} &\operatorname{Tr}\left(\boldsymbol{\Theta}_{0}^{1/2}\mathbf{U}\boldsymbol{\Theta}_{0}^{1/2}\right) = (\boldsymbol{\Theta}_{0}^{1/2}\mathbf{U}\boldsymbol{\Theta}_{0}^{1/2},\mathbf{I}) \\ &= \sqrt{p} \left\|\boldsymbol{\Theta}_{0}^{1/2}\mathbf{U}\boldsymbol{\Theta}_{0}^{1/2}\right\|_{F} \cos(\angle \boldsymbol{\Theta}_{0}^{1/2}\mathbf{U}\boldsymbol{\Theta}_{0}^{1/2},\mathbf{I}), \end{aligned}$$

and set $\mathbf{n} = \mathbf{\Omega}_0, \mathbf{v} = \mathbf{I}$ to apply Lemma 1 and get

$$H_{\mathbf{\Omega}_{0}}(\mathbf{U}) = \frac{p}{p+2} \sin^{2}(\angle \mathbf{\Theta}_{0}^{1/2} \mathbf{U} \mathbf{\Theta}_{0}^{1/2}, \mathbf{I}) \left\| \mathbf{\Theta}_{0}^{1/2} \mathbf{U} \mathbf{\Theta}_{0}^{1/2} \right\|_{F}^{2}$$
$$\geq \frac{p}{p+2} \cos^{2}(\angle \mathbf{I}, \mathbf{\Omega}_{0}) \left\| \mathbf{\Theta}_{0}^{1/2} \mathbf{U} \mathbf{\Theta}_{0}^{1/2} \right\|_{F}^{2}, \forall \mathbf{U} \in \mathcal{L}.$$

For brevity denote

$$\cos\phi_{0}=\cos(\angle\mathbf{I},\mathbf{\Omega}_{0})=\frac{\mathrm{Tr}\left(\mathbf{\Omega}_{0}\right)}{\left\|\mathbf{I}\right\|_{F}\left\|\mathbf{\Omega}_{0}\right\|_{F}}=\frac{\mathrm{Tr}\left(\mathbf{\Omega}_{0}\right)}{\sqrt{p}\left\|\mathbf{\Omega}_{0}\right\|_{F}}\geq\frac{1}{\kappa(\mathbf{\Theta}_{0})},$$

as was already defined in (6). This quantity is closely related to the notion of sphericity, [43]. We see that $\cos \phi_0$ defines the convexity properties of $H_{\Omega_0}(\mathbf{U})$ restricted to $\mathcal L$ and thus plays a crucial role in the bound provided by Theorem 1.

VI. PROOF OF THE MAIN THEOREM

Denote

$$\widehat{f_{\mathbf{\Omega}}} = \frac{1}{n} \sum_{i=1}^{n} f(\mathbf{\Omega}; \mathbf{x}_i),$$

and let us use the Taylor polynomial formula with remainder in the Lagrange form. The expansion in the vicinity of the true inverse covariance matrix Ω_0 reads as:

$$\widehat{f_{\Omega}} - \widehat{f_{\Omega_0}} = \nabla \widehat{f_{\Omega_0}}(\Delta \Omega) + \frac{1}{2} \nabla^2 \widehat{f_{\overline{\Omega}}}(\Delta \Omega), \tag{16}$$

where $\Omega = \Omega_0 + \Delta\Omega$, $\overline{\Omega} = \Omega_0 + \alpha\Delta\Omega$, $\alpha \in [0, 1]$ and $\Delta\Omega \in \mathcal{L}$. The sample average negative log-likelihood function (16) is zero for $\Delta\Omega = 0$ and we want to show that with high probability it is positive on the boundary of some ball.

Note that in spite of the fact the negative log-likelihood (9) is not globally convex in Ω , it is geodesically convex as shown in [39] and thus has a unique global minimizer no matter what metric over $\mathcal{P}(p)$ is considered. This justifies application of the technique described in section IV.

In order to make the proof more transparent and instructive, and the notations more concise let us note that the distributions of the functions in both sides of (16) are invariant under the linear map ' defined in (15)

$$\mathcal{L}' = \mathbf{\Theta}_0^{1/2} \mathcal{L} \mathbf{\Theta}_0^{1/2},$$

applied to matrices Ω_0 , $\Delta\Omega$ followed by a change of random vectors from $\mathbf{x}_i \sim \mathcal{U}(\mathbf{\Theta}_0)$ to $\mathbf{x}_i' \sim \mathcal{U}(\mathbf{I})$ for $i = 1, \dots, n$. This linear transformation relocates the domain under consideration from the vicinity of Ω_0 into a vicinity of \mathbf{I} and (16) reads as

$$\widehat{f_{\Omega'}} - \widehat{f_{\mathbf{I}}} = \nabla \widehat{f_{\mathbf{I}}}(\Delta \Omega') + \frac{1}{2} \nabla^2 \widehat{f_{\overline{\Omega'}}}(\Delta \Omega'),$$

where $\Omega' = I + \Delta\Omega'$, $\overline{\Omega'} = I + \alpha\Delta\Omega'$, $\Delta\Omega' \in \mathcal{L}'$, $\alpha \in [0, 1]$. Since ' is a tension-compression map, the maximal and minimal distance changes are known and we can use an inequality

$$\underline{\lambda} \left\| \mathbf{T}^{-1} - \mathbf{\Theta}_0^{-1} \right\|_F \le \left\| \mathbf{T}'^{-1} - \mathbf{\Theta}_0'^{-1} \right\|_F \tag{17}$$

to establish high probability bounds for inverse Tyler's estimator given the respective bounds on its '-image.

A. Gradient and Hessian Bounds

Lemma 2. Uniformly over $U' \in \mathcal{L}'$

$$\mathbb{P}\left(\left|\nabla \widehat{f}_{\mathbf{I}}(\mathbf{U}')\right| \geq tp \left\|\mathbf{U}'\right\|_{F}\right) \leq 2 \exp\left(\frac{-nt^{2}}{2(1+1.7t)}\right), \forall t \geq 0.$$

Proof. Is provided in Appendix B.

We proceed by establishing concentration properties of the sample mean Hessian restricted to \mathcal{L}' at the point Ω' .

Lemma 3. Let $\Omega' = \mathbf{I} + \Delta \Omega'$, $\Delta \Omega' \in \mathcal{L}'$ and $\varepsilon = \|\Delta \Omega'\|_2 < 1$, then for $\frac{1}{2\sqrt{2}} \frac{1+1/p^2}{\ln(32\sqrt{2}p^2)} < \eta < \frac{1}{2\sqrt{2}}$ uniformly over $\mathbf{U}' \in \mathcal{L}'$

$$\mathbb{P}\left(\nabla^{2}\widehat{f_{\Omega'}}(\mathbf{U}') \leq H_{\Omega'}(\mathbf{U}') - \eta \frac{\sqrt{2} \|\mathbf{U}'\|_{F}^{2}}{(1-\varepsilon)^{2}}\right)$$

$$\leq 2p^{2} \exp\left(-\frac{n\eta\sqrt{2}}{8\ln(32\sqrt{2}p^{2})}\right) \left(1 + \frac{7}{n^{2}\eta^{4}}\right).$$

Proof. Is provided in Appendix B.

This lemma shows that with probability depending on the parameters of the problem the sample average Hessian is not far from the expected Hessian at the same point. The next

result shows the expected Hessian at a point $\overline{\Omega}'$ close to the true parameter is lower bounded by a fraction of the expected Hessian at the true parameter value.

Lemma 4. For $\overline{\Omega'} = \mathbf{I} + \alpha \Delta \Omega'$, $\alpha \in [0,1]$, $\Delta \Omega' \in \mathcal{L}'$, $\varepsilon = \|\Delta \Omega'\|_2 \leq \frac{p}{6(p+2)} \cos^2 \phi_0$ the expected Hessian $H_{\overline{\Omega'}}(\Delta \Omega')$ is bounded from below by

$$H_{\overline{\Omega'}}(\Delta\Omega') \ge \frac{p}{2(p+2)}\cos^2\phi_0 \|\Delta\Omega'\|_F^2$$
.

Proof. Is provided in Appendix B.

Corollary 2. Under the conditions of Lemma 4 for τ satisfying $\frac{1+1/p^2}{\ln(32\sqrt{2}p^2)} \leq \tau \frac{p(1-\varepsilon)^2\cos^2\phi_0}{p+2} \leq 1$

$$\mathbb{P}\left(\frac{1}{2}\nabla^{2}\widehat{f_{\Omega'}}(\Delta\Omega') \leq (1-\tau)\frac{p}{4(p+2)}\cos^{2}\phi_{0} \|\Delta\Omega'\|_{F}^{2}\right)$$

$$\leq 2p^{2}\exp\left(-\frac{n\tau\cos^{2}\phi_{0}}{46\ln(7p)(1+\frac{2}{p})}\right)\left(1+\frac{2\cdot10^{3}(1+\frac{2}{p})^{4}}{n^{2}\tau^{4}\cos^{8}\phi_{0}}\right).$$

Proof. Is provided in Appendix B.

It now follows that with probability at least

$$1 - 2\exp\left(\frac{-nt^2}{2(1+1.7t)}\right) - 2p^2\exp\left(-\frac{n\tau\cos^2\phi_0}{46\ln(7p)(1+\frac{2}{p})}\right)\left(1 + \frac{2\cdot10^3(1+\frac{2}{p})^4}{n^2\tau^4\cos^8\phi_0}\right), (18)$$

$$\widehat{f_{\Omega'}} - \widehat{f_{\mathbf{I}}} \ge -tp \left\| \Delta \Omega' \right\|_F + (1 - \tau) \frac{p}{4(p+2)} \cos^2 \phi_0 \left\| \Delta \Omega' \right\|_F^2.$$

Demand positivity of the right-hand side to get

$$\|\Delta \Omega'\|_F > \frac{4t}{1-\tau} \frac{p+2}{\cos^2 \phi_0}.$$
 (19)

Set $\tau = \frac{3}{5}, t = \frac{\theta}{\sqrt{n}}$ and use (17) to get the statement of the theorem.

VII. NUMERICAL RESULTS

In this section we provide numerical simulations supporting our analysis. Figure 1 compares the behavior of Tyler's estimator with its 0.95 and 0.5-probability bounds for $p=50, \Theta_0=\mathbf{I}$. For a given n Tyler's estimator error bound was obtained by minimizing (19) with respect to t and τ under the probability constraint given by equating (18) to 0.95 or 0.5 respectively.

Figure 2 verifies the dependence of the performance on the dimension.

VIII. ACKNOWLEDGMENT

The authors would like to thank the associate editor and anonymous reviewers whose input greatly improved the paper.

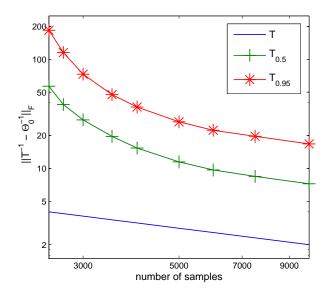


Fig. 1. Tyler's estimator performance bounds, p = 50, $\Theta_0 = \mathbf{I}$.

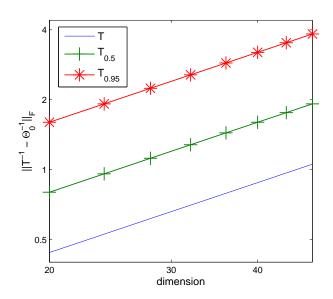


Fig. 2. Tyler's estimator performance bounds, $n=2500, \Theta_0(p)=\mathbf{I}_p$.

APPENDIX A MOMENTS OF RATIOS OF QUADRATIC FORMS

Lemma 5. Let $\mathbf{x}' \sim \mathcal{U}(\mathbf{I}), \mathbf{U}' \in \mathcal{S}(p), \mathbf{\Omega}' = \mathbf{I} + \Delta \mathbf{\Omega}' \succ 0$ and $\varepsilon = \|\Delta \mathbf{\Omega}'\|_2 < 1$, then the moments defined in (13) satisfy

$$\frac{R^{\nu}(\mathbf{U}', \mathbf{I}; \mathbf{I})}{(1+\varepsilon)^{\nu}} \le R^{\nu}(\mathbf{U}', \mathbf{\Omega}'; \mathbf{I}) \le \frac{R^{\nu}(\mathbf{U}', \mathbf{I}; \mathbf{I})}{(1-\varepsilon)^{\nu}}, \nu \in \mathbb{N}. \quad (20)$$

In addition

$$\begin{split} & \left| R^{\nu}(\Delta \mathbf{\Omega}', \mathbf{I} + \alpha \Delta \mathbf{\Omega}'; \mathbf{I}) - R^{\nu}(\Delta \mathbf{\Omega}', \mathbf{I}; \mathbf{I}) + \nu \alpha R^{\nu+1}(\Delta \mathbf{\Omega}', \mathbf{I}; \mathbf{I}) \right| \\ & \leq \frac{\nu(\nu+1)\alpha^2}{2(1-\alpha\varepsilon)^{\nu+2}} R^{\nu+2}(\Delta \mathbf{\Omega}', \mathbf{I}; \mathbf{I}). \end{split}$$

Proof. Consider

$$R^{\nu}(\mathbf{U}', \mathbf{\Omega}'; \mathbf{I}) = \mathbb{E}\left[\left(\frac{\mathbf{x}'^{T} \mathbf{U}' \mathbf{x}'}{\mathbf{x}'^{T} (\mathbf{I} + \Delta \mathbf{\Omega}') \mathbf{x}'}\right)^{\nu}\right], \quad (21)$$

since

$$1 + \varepsilon \ge 1 + \mathbf{x}'^T \Delta \mathbf{\Omega}' \mathbf{x}' \ge 1 - \varepsilon,$$

the variable under expectation in (21) can be bounded as

$$\left(\frac{\mathbf{x}'^T\mathbf{U}'\mathbf{x}'}{1+\varepsilon}\right)^{\nu} \leq \left(\mathbf{x}'^T\mathbf{U}'\mathbf{x}'\left(1+\mathbf{x}'^T\Delta\mathbf{\Omega}'\right)^{-1}\right)^{\nu} \leq \left(\frac{\mathbf{x}'^T\mathbf{U}'\mathbf{x}'}{1-\varepsilon}\right)^{\nu}.$$

By taking the expectations (20) follows. We continue

$$\begin{split} R^{\nu}(\Delta \mathbf{\Omega}', \mathbf{I} + \alpha \Delta \mathbf{\Omega}'; \mathbf{I}) &= \mathbb{E}\left[\left(\frac{\mathbf{x}'^T \Delta \mathbf{\Omega}' \mathbf{x}'}{\mathbf{x}'^T (\mathbf{I} + \alpha \Delta \mathbf{\Omega}') \mathbf{x}'}\right)^{\nu}\right] \\ &= \mathbb{E}\left[\left(\mathbf{x}'^T \Delta \mathbf{\Omega}' \mathbf{x}' \left(1 + \alpha \mathbf{x}'^T \Delta \mathbf{\Omega}' \mathbf{x}'\right)^{-1}\right)^{\nu}\right] \\ &= \mathbb{E}\left[\left(\mathbf{x}'^T \Delta \mathbf{\Omega}' \mathbf{x}'\right)^{\nu} \left(1 - \nu \alpha \mathbf{x}'^T \Delta \mathbf{\Omega}' \mathbf{x}' + g(\alpha \Delta \mathbf{\Omega}', \mathbf{x})\right)\right], \end{split}$$

where $g(\alpha \Delta \Omega', \mathbf{x}')$ is bounded by

$$|g(\alpha \Delta \Omega', \mathbf{x}')| \le \nu(\nu + 1)\alpha^2 \frac{(\mathbf{x}'^T \Delta \Omega' \mathbf{x}')^2}{2(1 - \alpha \varepsilon)^{\nu + 2}}.$$

We finally obtain

$$|R^{\nu}(\Delta \mathbf{\Omega}', \mathbf{I} + \alpha \Delta \mathbf{\Omega}'; \mathbf{I}) - R^{\nu}(\Delta \mathbf{\Omega}', \mathbf{I}; \mathbf{I}) + \nu \alpha R^{\nu+1}(\Delta \mathbf{\Omega}', \mathbf{I}; \mathbf{I})|$$

$$\leq \frac{\nu(\nu+1)\alpha^2}{2(1-\alpha\varepsilon)^{\nu+2}} R^{\nu+2}(\Delta \mathbf{\Omega}', \mathbf{I}; \mathbf{I}).$$

Given $\mathbf{x}' \sim \mathcal{U}(\mathbf{I})$, we can always represent it as $\mathbf{x}' = \frac{\mathbf{z}'}{\|\mathbf{z}'\|}$, where \mathbf{z}' is standard normally distributed $\mathbf{z}' \sim \mathcal{N}(0, \mathbf{I})$, to obtain

$$\frac{\mathbf{x}'^T \mathbf{U}' \mathbf{x}'}{\mathbf{x}'^T \mathbf{\Omega}' \mathbf{x}'} = \frac{\mathbf{z}'^T \mathbf{U}' \mathbf{z}'}{\mathbf{z}'^T \mathbf{\Omega}' \mathbf{z}'}$$

using this identity, we develop formulas for $R^{\nu}(\mathbf{U}', \mathbf{I}; \mathbf{I})$, [44]:

$$R^{1}(\mathbf{U}', \mathbf{I}; \mathbf{I}) = \frac{\operatorname{Tr}(\mathbf{U}')}{p},$$
(22)

$$R^{2}(\mathbf{U}', \mathbf{I}; \mathbf{I}) = \frac{\left(\operatorname{Tr}\left(\mathbf{U}'\right)\right)^{2} + 2\operatorname{Tr}\left(\left(\mathbf{U}'\right)^{2}\right)}{p(p+2)} = \frac{\left(\operatorname{Tr}\left(\mathbf{U}'\right)\right)^{2} + 2\left\|\mathbf{U}'\right\|_{F}^{2}}{p(p+2)},$$
(23)

$$R^{3}(\mathbf{U}', \mathbf{I}; \mathbf{I}) = \frac{\operatorname{Tr}(\mathbf{U}')^{3} + 6\operatorname{Tr}(\mathbf{U}')\operatorname{Tr}(\mathbf{U}'^{2}) + 8\operatorname{Tr}(\mathbf{U}'^{3})}{p(p+2)(p+4)}.$$
 (24)

A general moment bound can also be obtained as

$$R^{\nu}(\mathbf{U}', \mathbf{I}; \mathbf{I}) \le \frac{\left(\frac{\nu}{2}\right)! \|\mathbf{U}'\|_F^{\nu}}{\sqrt{p}^{\nu}}, \nu = 2, 4, \dots$$
 (25)

APPENDIX B CONCENTRATION BOUNDS

Lemma 6. (Vector Bernstein Inequality) [45] Let $\xi_1, \ldots, \xi_n \in \mathbb{R}^k$ be i.i.d zero-mean random vectors and suppose there exist $\sigma, L > 0$ such that

$$\mathbb{E} \|\boldsymbol{\xi}_1\|^r \le \frac{r!}{2} \sigma^2 L^{r-2}, r = 2, 3, \dots,$$

then for $t \geq 0$

$$\mathbb{P}\left(\left\|\widehat{\pmb{\xi}}\right\| \geq t\sigma\right) \leq 2\exp\left(\frac{-nt^2}{2(1+1.7t\frac{L}{\sigma})}\right).$$

Lemma 7. (Matrix Bernstein Inequality) [46] Let $S_1, ..., S_n \in S(p)$ be i.i.d zero-mean random matrices and suppose there exist $\sigma, L > 0$ such that

$$\mathbb{E} \|\mathbf{S}_1\|_2^r \le \frac{r!}{2} \sigma^2 L^{r-2}, r = 2, 3, \dots,$$

then for $t > \frac{\sigma}{4L} \frac{1 + 1/p^2}{\ln \frac{64\sqrt{2}p^2L^2}{2}}$

$$\mathbb{P}\left(\lambda_{\max}(\widehat{\mathbf{S}}) \ge t\sigma\right)$$

$$\le 2p^2 \exp\left(-\frac{nt\sigma}{8L \ln \frac{64\sqrt{2}p^2L^2}{\sigma^2}}\right) \left(1 + \frac{6}{n^2t^2\sigma^2 \ln^2(1 + \frac{t}{\sigma})}\right).$$

Lemma 8. Let $\mathbf{x}'_i \sim \mathcal{U}(\mathbf{I}), i = 1, \dots, n$ then for $t \geq 0$

$$\mathbb{P}\left(\left\|\widehat{p\mathbf{x}'\mathbf{x}'^T} - \mathbf{I}\right\|_F \ge tp\right) \le 2\exp\left(\frac{-nt^2}{2(1+1.7t)}\right)$$

Proof. Define n centered random vectors

$$\boldsymbol{\xi}_i = \operatorname{vec}\left(\mathbf{x}_i'\mathbf{x}_i'^T - \frac{1}{p}\mathbf{I}\right) \in \mathbb{R}^{p^2}, i = 1, \dots, n,$$

and consider powers of their norms

$$\|\boldsymbol{\xi}_1\|^r = \left[\operatorname{Tr}\left(\left(\mathbf{x}_1'\mathbf{x}_1'^T - \frac{1}{p}\mathbf{I}\right)^2\right)\right]^{\frac{r}{2}}$$
$$= \left[\operatorname{Tr}\left(\left(1 - \frac{2}{p}\right)\mathbf{x}_1'\mathbf{x}_1'^T + \frac{1}{p^2}\mathbf{I}\right)\right]^{\frac{r}{2}} = \left(1 - \frac{1}{p}\right)^{\frac{r}{2}},$$

which are deterministic quantities. Set

$$\sigma = L = 1$$
,

and apply Lemma 6 to obtain

$$\mathbb{P}\left(\left\|\widehat{\pmb{\xi}}\right\| \geq t\right) \leq 2\exp\left(\frac{-nt^2}{2(1+1.7t)}\right).$$

Multiply $\hat{\boldsymbol{\xi}}$ by p to get the statement.

Proof of Lemma 2.

$$\left| \nabla \widehat{f}_{\mathbf{I}}(\mathbf{U}') \right| = \left| \operatorname{Tr} \left(\left(\widehat{p} \widehat{\mathbf{x}' \mathbf{x}'^T} - \mathbf{I} \right) \mathbf{U}' \right) \right|.$$

Apply Lemma 8 and the Cauchy-Schwartz inequality to get the statement. $\hfill\Box$

Proof of Lemma 3. For a linear operator L, its restriction \widetilde{L} to a linear subspace satisfies

$$\left\|\widetilde{L}\right\|_{2} \leq \left\|L\right\|_{2}$$

thus we can apply Lemma 7 to bound the deviation of $\nabla^2 \widehat{f_{\Omega'}}$ restricted to \mathcal{L}' from its expectation $H_{\Omega'}$ also restricted to \mathcal{L}' . Define n centered random quadratic forms

$$S_{i}(\mathbf{U}') = \operatorname{Tr}\left((\mathbf{\Omega}'^{-1}\mathbf{U}')^{2}\right) - p\left(\frac{\mathbf{x}_{i}'^{T}\mathbf{U}'\mathbf{x}_{i}'}{\mathbf{x}_{i}'^{T}\mathbf{\Omega}'\mathbf{x}_{i}'}\right)^{2} - H_{\mathbf{\Omega}'}(\mathbf{U}')$$

$$= p\left(R^{2}(\mathbf{U}, \mathbf{\Omega}'; \mathbf{I}) - \left(\frac{\mathbf{x}_{i}'^{T}\mathbf{U}'\mathbf{x}_{i}'}{\mathbf{x}_{i}'^{T}\mathbf{\Omega}'\mathbf{x}_{i}'}\right)^{2}\right), i = 1, \dots, n. \quad (26)$$

Bound the moments of $S_1(\mathbf{U}')$

$$\begin{aligned} &|\mathbb{E}[S_{1}(\mathbf{U}')^{r}]| = \left| p^{r} \sum_{j=0}^{r} (-1)^{j} {r \choose j} \left[R^{2}(\mathbf{U}', \mathbf{\Omega}'; \mathbf{I}) \right]^{r-j} R^{2j}(\mathbf{U}', \mathbf{\Omega}'; \mathbf{I}) \right| \\ &\leq p^{r} \max_{j} \frac{r!}{(r-j)! j!} \left(\frac{\|\mathbf{U}'\|_{F}^{2}}{(1-\varepsilon)^{2p}} \right)^{r-j} \frac{j! \|\mathbf{U}'\|_{F}^{2j}}{(1-\varepsilon)^{2j} p^{j}} \leq \frac{r!}{2} \frac{2}{(1-\varepsilon)^{2r}} \|\mathbf{U}'\|_{F}^{2r}, \end{aligned}$$

which allows us to set

$$\frac{\sigma}{\sqrt{2}} = L = \frac{1}{(1 - \varepsilon)^2}.$$

Use Lemma 7 to get for $\frac{1}{2\sqrt{2}}\frac{1+1/p^2}{\ln(32\sqrt{2}p^2)}<\eta<\frac{1}{2\sqrt{2}}$

$$\begin{split} & \mathbb{P}\left(\lambda_{\min}\left(\nabla^{2}\widehat{f_{\Omega'}}|_{\mathcal{L}'}\right) \leq \lambda_{\min}(H_{\Omega'}|_{\mathcal{L}'}) - \eta\sigma\right) \\ & = \mathbb{P}\left(\lambda_{\min}\left(\widehat{S}|_{\mathcal{L}'}\right) \leq -\eta\sigma\right) = \mathbb{P}\left(\lambda_{\max}\left(-\widehat{S}|_{\mathcal{L}'}\right) \geq \frac{\sqrt{2}\eta}{(1-\varepsilon)^{2}}\right) \\ & \leq 2p^{2}\exp\left(-\frac{n\eta\sqrt{2}}{8\ln(32\sqrt{2}p^{2})}\right)\left(1 + \frac{3(1-\varepsilon)^{4}}{n^{2}\eta^{2}\ln^{2}\left(1 + \frac{\eta(1-\varepsilon)^{2}}{\sqrt{2}}\right)}\right) \\ & \leq 2p^{2}\exp\left(-\frac{n\eta\sqrt{2}}{8\ln(32\sqrt{2}p^{2})}\right)\left(1 + \frac{7}{n^{2}\eta^{4}}\right). \end{split}$$

Proof of Lemma 4. From (14) we have

$$H_{\overline{\mathbf{\Omega}'}}(\Delta \mathbf{\Omega}') = \operatorname{Tr}\left(\left(\overline{\mathbf{\Omega}'}^{-1} \Delta \mathbf{\Omega}'\right)^{2}\right) - pR^{2}(\Delta \mathbf{\Omega}', \overline{\mathbf{\Omega}'}; \mathbf{I}).$$

Using Lemma 5 we obtain

$$H_{\overline{\Omega'}}(\Delta\Omega') = \operatorname{Tr}\left(\left[\left(\mathbf{I} + \alpha\Delta\Omega'\right)^{-1}\Delta\Omega'\right]^{2}\right) - pR^{2}(\Delta\Omega', \mathbf{I} + \alpha\Delta\Omega'; \mathbf{I})$$

$$= \operatorname{Tr}\left(\Delta\Omega'^{2} - 2\alpha\Delta\Omega'^{3} + h(\alpha\Delta\Omega')\Delta\Omega'^{2}\right)$$

$$- p\left[R^{2}(\Delta\Omega', \mathbf{I}; \mathbf{I}) - 2\alpha R^{3}(\Delta\Omega', \mathbf{I}; \mathbf{I}) + l(\alpha\Delta\Omega')\right].$$

Use the Lagrange remainder form for $h(\alpha \Delta \Omega')$ and apply Lemma 5 to bound $l(\alpha \Delta \Omega')$ and get

$$|h(\alpha \Delta \mathbf{\Omega}')| \leq \frac{3(\alpha \varepsilon)^2}{(1 - \alpha \varepsilon)^4}, \quad |l(\alpha \Delta \mathbf{\Omega}')| \leq \frac{3\alpha^2}{(1 - \alpha \varepsilon)^4} R^4(\Delta \mathbf{\Omega}', \mathbf{I}; \mathbf{I}).$$

Use formulas (23), (24) to get

$$H_{\overline{\Omega'}}(\Delta\Omega') = \|\Delta\Omega'\|_F^2 - 2\alpha \operatorname{Tr} \left(\Delta\Omega'^3\right) - \frac{\left(\operatorname{Tr} \left(\Delta\Omega'\right)\right)^2 + 2\|\Delta\Omega'\|_F^2}{p+2}$$

$$+ 2\alpha \frac{\left(\operatorname{Tr} \left(\Delta\Omega'\right)\right)^3 + 6\operatorname{Tr} \left(\Delta\Omega'\right)\|\Delta\Omega'\|_F^2 + 8\operatorname{Tr} \left(\Delta\Omega'^3\right)}{(p+2)(p+4)}$$

$$+ r(\alpha\Delta\Omega')\|\Delta\Omega'\|_F^2$$

$$= H_{\Omega_0}(\Delta\Omega) + r(\alpha\Delta\Omega')\|\Delta\Omega'\|_F^2$$

$$- 2\alpha \left(\frac{(p^2 + 6p)\operatorname{Tr} \left(\Delta\Omega'^3\right) - \left(\operatorname{Tr} \left(\Delta\Omega'\right)\right)^3 - 6\operatorname{Tr} \left(\Delta\Omega'\right)\|\Delta\Omega'\|_F^2}{(p+2)(p+4)}\right)$$

$$\geq \left(\frac{p}{n+2}\cos^2\phi - 2\alpha \left\|\Delta\mathbf{\Omega}'\right\|_2 + r(\alpha\Delta\mathbf{\Omega}')\right) \left\|\Delta\mathbf{\Omega}'\right\|_F^2,$$

where due to the condition $\|\Delta\Omega'\|_2 = \varepsilon \leq \frac{1}{6}$,

$$|r(\alpha \Delta \mathbf{\Omega}')| \leq \frac{3(\alpha \varepsilon)^2}{(1 - \alpha \varepsilon)^4} + p \frac{3\alpha^2}{(1 - \alpha \varepsilon)^4} \frac{2 \|\Delta \mathbf{\Omega}'\|_F^2}{p^2}$$
$$\leq \frac{(\alpha \varepsilon)^2 (3 + \frac{6}{p} p \varepsilon^2)}{(1 - \alpha \varepsilon)^4} \leq \frac{\varepsilon^2 (3 + 6\varepsilon^2)}{(1 - \varepsilon)^4} \leq \frac{1}{6}.$$

Here we have applied the bound from (25) to $R^4(\Delta\Omega', \mathbf{I}; \mathbf{I})$. Finally,

$$2\alpha \|\Delta \mathbf{\Omega}'\|_{2} \leq 2\varepsilon \leq \frac{1}{3}$$

$$H_{\overline{\mathbf{\Omega}'}}(\Delta \mathbf{\Omega}') \geq \frac{p}{2(p+2)} \cos^{2} \phi_{0} \|\Delta \mathbf{\Omega}'\|_{F}^{2}.$$

Proof of Corollary 2. Set

$$\frac{1}{2\sqrt{2}} \frac{1 + 1/p^2}{\ln(32\sqrt{2}p^2)} \le \eta = \tau \frac{(1-\varepsilon)^2}{\sqrt{2}} \frac{p}{2(p+2)} \cos^2 \phi_0 \le \frac{1}{2\sqrt{2}},$$

recall that $\varepsilon \leq \frac{p}{6(p+2)}\cos^2\phi_0 \leq \frac{1}{6}$ and use Lemma 3 to get

$$\begin{split} & \mathbb{P}\left(\nabla^{2}\widehat{f_{\overline{\Omega'}}}(\mathbf{U'}) \leq (1-\tau)\frac{p}{2(p+2)}\cos^{2}\phi_{0} \|\mathbf{U'}\|_{F}^{2}\right) \\ & \leq 2p^{2}\exp\left(-\frac{n\eta\sqrt{2}}{8\ln(32\sqrt{2}p^{2})}\right)\left(1+\frac{7}{n^{2}\eta^{4}}\right) \\ & \leq 2p^{2}\exp\left(-\frac{n\tau\cos^{2}\phi_{0}}{46\ln(7p)(1+\frac{2}{p})}\right)\left(1+\frac{2\cdot10^{3}(1+\frac{2}{p})^{4}}{n^{2}\tau^{4}\cos^{8}\phi_{0}}\right). \end{split}$$

REFERENCES

- [1] O. Ledoit and M. Wolf, "Improved estimation of the covariance matrix of stock returns with an application to portfolio selection," *Journal of empirical finance*, vol. 10, no. 5, pp. 603–621, 2003.
- [2] P. J. Bickel and E. Levina, "Regularized estimation of large covariance matrices," *The Annals of Statistics*, pp. 199–227, 2008.
- [3] A. J. Rothman, P. J. Bickel, E. Levina, and J. Zhu, "Sparse permutation invariant covariance estimation," *Electronic Journal of Statistics*, vol. 2, pp. 494–515, 2008.
- [4] D. E. Tyler, "A distribution-free M-estimator of multivariate scatter," *The Annals of Statistics*, vol. 15, no. 1, pp. 234–251, 1987.
- [5] V. A. Marcenko and L. A. Pastur, "Distribution of eigenvalues for some sets of random matrices," *Sbornik: Mathematics*, vol. 1, no. 4, pp. 457–483, 1967.
- [6] C. A. Tracy and H. Widom, "On orthogonal and symplectic matrix ensembles," *Communications in Mathematical Physics*, vol. 177, no. 3, pp. 727–754, 1996.
- [7] N. El Karoui, "Tracy-Widom limit for the largest eigenvalue of a large class of complex sample covariance matrices," *The Annals of Probability*, vol. 35, no. 2, pp. 663–714, 2007.
- [8] Z. Bai and J.-F. Yao, "Limit theorems for sample eigenvalues in a generalized spiked population model," *arXiv* preprint arXiv:0806.1141, 2008.
- [9] J. Baik and J. W. Silverstein, "Eigenvalues of large sample covariance matrices of spiked population models," *Journal of Multivariate Analysis*, vol. 97, no. 6, pp. 1382–1408, 2006.
- [10] L. S. Cardoso, M. Debbah, P. Bianchi, and J. Najim, "Cooperative spectrum sensing using random matrix theory,"

- in 3rd International Symposium on Wireless Pervasive Computing. IEEE, 2008, pp. 334–338.
- [11] R. Couillet and W. Hachem, "Fluctuations of spiked random matrix models and failure diagnosis in sensor networks," *IEEE Transactions on Information Theory*, vol. 59, no. 1, pp. 509–525, 2012.
- [12] X. Mestre and M. A. Lagunas, "Modified subspace algorithms for DoA estimation with large arrays," *IEEE Transactions on Signal Processing*, vol. 56, no. 2, pp. 598–614, 2008.
- [13] R. Schmidt, "Multiple emitter location and signal parameter estimation," *IEEE Transactions on Antennas and Propagation*, vol. 34, no. 3, pp. 276–280, 1986.
- [14] R. Couillet, J. W. Silverstein, Z. Bai, and M. Debbah, "Eigen-inference for energy estimation of multiple sources," *IEEE Transactions on Information Theory*, vol. 57, no. 4, pp. 2420–2439, 2011.
- [15] P. Vallet, P. Loubaton, and X. Mestre, "Improved subspace estimation for multivariate observations of high dimension: the deterministic signals case," *IEEE Transactions on Information Theory*, vol. 58, no. 2, pp. 1043–1068, 2012.
- [16] R. Vershynin, "Introduction to the non-asymptotic analysis of random matrices," *Compressed sensing: theory and applications, Edited by Y. Eldar and G. Kutyniok, Cambridge University Press*, 2012.
- [17] —, "How close is the sample covariance matrix to the actual covariance matrix?" *Journal of Theoretical Probability*, vol. 25, no. 3, pp. 655–686, 2012.
- [18] N. Srivastava and R. Vershynin, "Covariance estimation for distributions with $2+\varepsilon$ moments," *arXiv* preprint *arXiv*:1106.2775, 2011.
- [19] P. Ravikumar, M. J. Wainwright, G. Raskutti, and B. Yu, "High-dimensional covariance estimation by minimizing l₁-penalized log-determinant divergence," *Electronic Journal of Statistics*, vol. 5, pp. 935–980, 2011.
- [20] P. J. Huber, "Robust estimation of a location parameter," *The Annals of Mathematical Statistics*, vol. 35, no. 1, pp. 73–101, 1964.
- [21] R. A. Maronna, "Robust M-estimators of multivariate location and scatter," *The annals of statistics*, pp. 51–67, 1976.
- [22] G. Frahm, "Generalized elliptical distributions: theory and applications," *Universität zu Köln*, 2004.
- [23] Y. Chen, A. Wiesel, and A. O. Hero, "Robust shrinkage estimation of high-dimensional covariance matrices," *IEEE Transactions on Signal Processing*, vol. 59, no. 9, pp. 4097–4107, 2011.
- [24] E. Ollila and V. Koivunen, "Robust antenna array processing using M-estimators of pseudo-covariance," 14th IEEE Proceedings on Personal, Indoor and Mobile Radio Communications, vol. 3, pp. 2659–2663, 2003.
- [25] Y. I. Abramovich, N. K. Spencer, and M. D. Turley, "Time-varying autoregressive (TVAR) models for multiple radar observations," *IEEE Transactions on Signal Processing*, vol. 55, no. 4, pp. 1298–1311, 2007.
- [26] E. Ollila, D. Tyler, V. Koivunen, and H. Poor, "Complex elliptically symmetric distributions: survey, new results

- and applications," *IEEE Transactions on Signal Processing*, vol. 60, no. 11, pp. 5597–5625, 2012.
- [27] F. Bandiera, O. Besson, and G. Ricci, "Knowledge-aided covariance matrix estimation and adaptive detection in compound-Gaussian noise," *IEEE Transactions on Signal Processing*, vol. 58, no. 10, pp. 5391–5396, 2010.
- [28] F. Pascal, Y. Chitour, J. P. Ovarlez, P. Forster, and P. Larzabal, "Covariance structure maximum-likelihood estimates in compound Gaussian noise: Existence and algorithm analysis," *IEEE Transactions on Signal Pro*cessing, vol. 56, no. 1, pp. 34–48, 2008.
- [29] O. Besson and Y. Abramovich, "On the Fisher information matrix for multivariate elliptically contoured distributions," *Signal Processing Letters, IEEE*, vol. 20, no. 11, pp. 1130–1133, 2013.
- [30] M. Greco and F. Gini, "Cramér-Rao lower bounds on covariance matrix estimation for complex elliptically symmetric distributions," *IEEE Transactions on Signal Processing*, vol. 61, no. 24, pp. 6401–6409, 2013.
- [31] L. Duembgen, "The asymptotic behavior of Tylers Mestimator of scatter in high dimension," *Ann. Inst. Statist. Math*, vol. 50, pp. 471–491, 1997.
- [32] G. Frahm and K. Glombek, "Semicircle law of Tylers M-estimator for scatter," *Statistics & Probability Letters*, vol. 82, no. 5, pp. 959–964, 2012.
- [33] T. Zhang, X. Cheng, and A. Singer, "Marchenko-Pastur Law for Tyler's and Maronna's M-estimators," *arXiv* preprint arXiv:1401.3424, 2014.
- [34] R. Couillet and M. R. McKay, "Large dimensional analysis and optimization of robust shrinkage covariance matrix estimators," *arXiv preprint arXiv:1401.4083*, 2014.
- [35] R. Couillet, F. Pascal, and J. W. Silverstein, "The random matrix regime of Maronna's M-estimator with elliptically distributed samples," *arXiv preprint arXiv:1311.7034*, 2013.
- [36] —, "Robust M-estimation for array processing: a random matrix approach," *IEEE Transactions on Information Theory, Submitted for publication, Available:* http://arxiv. org/abs/1204.5320, 2012.
- [37] D. E. Tyler, "Statistical analysis for the angular central Gaussian distribution on the sphere," *Biometrika*, vol. 74, no. 3, pp. 579–589, 1987.
- [38] G. Frahm and U. Jaekel, "Tyler's M-estimator, random matrix theory, and Generalized Elliptical distributions with applications to finance," Tech. Rep., 2007.
- [39] A. Wiesel, "Geodesic convexity and covariance estimation," *IEEE Transactions on Signal Processing*, vol. 60, no. 12, pp. 6182–6189, 2012.
- [40] —, "Unified framework to regularized covariance estimation in scaled gaussian models," *IEEE Transactions on Signal Processing*, vol. 60, no. 1, pp. 29–38, 2012.
- [41] T. Zhang, A. Wiesel, and M. S. Greco, "Multivariate generalized gaussian distribution: Convexity and graphical models," *IEEE Transactions on Signal Processing*, vol. 61, no. 16, pp. 4141–4148, 2013.
- [42] J. K. Bradley and C. Guestrin, "Sample complexity of composite likelihood," *International Conference on Artificial Intelligence and Statistics*, pp. 136–160, 2012.

- [43] O. Ledoit and M. Wolf, "Some hypothesis tests for the covariance matrix when the dimension is large compared to the sample size," *Annals of Statistics*, pp. 1081–1102, 2002.
- [44] J. G. De Gooijer, "Exact moments of the sample autocorrelations from series generated by general ARIMA processes of order (p,d,q),d=0 or 1," *Journal of Econometrics*, vol. 14, no. 3, pp. 365–379, 1980.
- [45] V. Yurinskii, "Exponential inequalities for sums of random vectors," *Journal of multivariate analysis*, vol. 6, no. 4, pp. 473–499, 1976.
- [46] S. Minsker, "On some extensions of Bernstein's inequality for self-adjoint operators," *arXiv* preprint *arXiv*:1112.5448, 2011.