The Annals of Statistics
2020, Vol. 48, No. 1, 86–110
https://doi.org/10.1214/18-AOS1793

ROBUST SPARSE COVARIANCE ESTIMATION BY THRESHOLDING TYLER'S M-ESTIMATOR

BY JOHN GOES[1,*], GILAD LERMAN[1,**] AND BOAZ NADLER[2]

[1]*School of Mathematics, University of Minnesota,* [*]*johngoes@umn.edu;* [**]*lerman@umn.edu*
[2]*Department of Computer Science and Applied Mathematics, Weizmann Institute of Science, boaz.nadler@weizmann.ac.il*

Estimating a high-dimensional sparse covariance matrix from a limited number of samples is a fundamental task in contemporary data analysis. Most proposals to date, however, are not robust to outliers or heavy tails. Toward bridging this gap, in this work we consider estimating a sparse shape matrix from n samples following a possibly heavy-tailed elliptical distribution. We propose estimators based on thresholding either Tyler's M-estimator or its regularized variant. We prove that in the joint limit as the dimension p and the sample size n tend to infinity with $p/n \to \gamma > 0$, our estimators are minimax rate optimal. Results on simulated data support our theoretical analysis.

1. Introduction. The covariance matrix $\mathbf{\Sigma}$ of a p-dimensional random variable X is a central object in statistical data analysis. Given n samples $\{\boldsymbol{x}_i\}_{i=1}^n$, an accurate estimate of $\mathbf{\Sigma}$ is needed in many tasks including PCA, clustering and discriminant analysis (Anderson (2003), Mardia, Kent and Bibby (1979)). The sample covariance matrix, which is the standard estimator for $\mathbf{\Sigma}$, is quite accurate when X is sub-Gaussian and $p \ll n$.

In several contemporary applications, however, the number of samples n and the dimension p are comparable, and the data may be heavy tailed. To accurately estimate the covariance matrix when n and p are comparable, additional assumptions, such as its approximate sparsity are typically made. Over the past decade, several sparse covariance matrix estimators were proposed and analyzed (Bickel and Levina (2008), Cai and Liu (2011), El Karoui (2008), Lam and Fan (2009), Rothman, Levina and Zhu (2009)). In addition, minimax rates for estimating high-dimensional sparse covariance matrices were established (Cai and Zhou (2012a, 2012b), Cai, Ren and Zhou (2016)).

With respect to heavy-tailed data, a popular model which we consider in this work is the elliptical distribution (Cambanis, Huang and Simons (1981), Fang, Kotz and Ng (1990), Frahm (2004), Kelker (1970)). An elliptical distribution is characterized by a $p \times p$ shape or scatter matrix $\boldsymbol{S}_p$, which is proportional to the population covariance matrix, when the latter exists. When the elliptical distribution is heavy tailed, the sample covariance is a poor estimate of the population covariance (Falk (2002)). Moreover, the elliptical distribution might be so heavy tailed as to not even have finite second moments, in which case its population covariance does not exist. Yet due to the structure of the elliptical distribution, even with heavy tails it is nonetheless possible to accurately estimate its shape matrix. This is useful in various applications, since the shape matrix preserves the directional properties of the distribution, such as its principal components.

Following Huber's pioneering work (Huber and Ronchetti (2009)), over the past decades several robust estimators of the covariance and shape matrix were proposed and theoretically studied; see Dümbgen, Nordhausen and Schuhmacher (2016), Dümbgen, Pauly and Schweizer (2015), Kent and Tyler (1991), Maronna (1976), Maronna and Yohai (2017) and

Received June 2017; revised November 2018.
MSC2010 subject classifications. Primary 62H12; secondary 62G35, 62G20.
Key words and phrases. Covariance matrix estimation, spectral norm, elliptical distribution, sparsity, Tyler's M-estimator, thresholding.

references therein. For elliptical distributions, Tyler (1987a) proposed a robust M-estimator for the scatter matrix $\boldsymbol{S}_p$ and an iterative scheme to compute it. Tyler's M-estimator has found widespread use in various applications involving heavy-tailed data. However, as it is defined only for $p < n$, in recent years several regularized variants, applicable also for $p > n$ were proposed and analyzed (Abramovich and Spencer (2007), Chen, Wiesel and Hero (2011), Ollila and Tyler (2014), Pascal, Chitour and Quek (2014), Sun, Babu and Palomar (2014), Wiesel (2012)). The spectral properties of Maronna's M-estimators and specifically Tyler's M-estimator and its regularized variants, in high dimensions as $n, p \to \infty$ with $p/n \to \gamma$ were studied by Dümbgen (1998), Couillet, Pascal and Silverstein (2014, 2015), Couillet, Kammoun and Pascal (2016), Couillet and McKay (2014), Zhang, Cheng and Singer (2016), among others. For a recent survey on Tyler's M-estimator and its variants, see Wiesel and Zhang (2014).

In this paper, we study the combination of heavy-tailed data with a "large p–large n" setting. As formulated in Section 2, we consider robust estimation of the shape matrix of an elliptical distribution, assuming it is approximately sparse. We address the following two challenges: (i) design a computationally efficient and statistically accurate estimator of the shape matrix $\boldsymbol{S}_p$, that is adaptive to its unknown sparsity parameters; (ii) provide theoretical guarantees on its accuracy, in the large p large n regime.

We make the following contributions. First, in Section 3 we propose simple and computationally efficient estimators for the sparse shape matrix of an elliptical distribution. These are based on thresholding either Tyler's M-estimator (TME) or its regularized variant. Second, we provide theoretical guarantees on their accuracy in the limit $n, p \to \infty$ with $p/n \to \gamma$. Theorems 1 and 2 show that the estimator $\hat{\boldsymbol{E}}$ based on thresholding either TME for $\gamma < 1$ or its regularized variant for any $\gamma \in (0, \infty)$, converges in spectral norm to a sparse shape matrix $\boldsymbol{S}_p$ at rate $\|\hat{\boldsymbol{E}} - \boldsymbol{S}_p\| = O_P((\log p/n)^{(1-q)/2})$, where q is the sparsity parameter of $\boldsymbol{S}_p$. Estimating a sparse shape matrix under a heavy-tailed elliptical distribution is thus possible with the same asymptotic error rate as estimating a sparse covariance matrix under sub-Gaussian distributions. Moreover, our estimators are rate optimal, as this rate coincides with the minimax rate for sparse covariance estimation with sub-Gaussian data (Cai and Zhou (2012a)).[1]

Our proofs follow the approach of Bickel and Levina (2008), with required modifications given that we analyze Tyler's M-estimators. Theorem 1, which concerns the TME and is thus valid for $p < n$, is proven in Section 5. The proof is relatively simple and heavily relies on Zhang, Cheng and Singer (2016), who studied the spectral properties of Tyler's M-estimator when $n, p \to \infty$. Theorem 2 considers the thresholded regularized TME, and is thus applicable also for $p > n$. As detailed in Section 6, its proof is far more involved, and combines a careful analysis of the regularized TME with several results in random matrix theory. Section 7 presents simulation results that support our theoretical analysis. With an eye toward practitioners, given that regularization is common also when $p < n$, we focus on the regularized TME. With Gaussian data, our thresholded TME estimator is as accurate as thresholding the sample covariance. In contrast, in the presence of heavy tails it is far more accurate. We also illustrate its potential utility in handling outliers. In addition, our estimator is quite fast to compute in practice, requiring only few seconds on a standard PC, say for $p = n = 1000$.

Our work is related to several recent papers, that also considered sparse shape or covariance matrix estimation with heavy-tailed data. Han, Lu and Liu (2014) considered a generalization of the elliptical distribution, denoted as the pair-elliptical distribution, with moderate

[1]Technically, the minimax rate was proven under the assumption that $p/n^{\beta} \to c$ with $\beta > 1$; see Remark 5 in Cai and Zhou (2012a). However, from personal communication with Professors Cai and Zhou, the same minimax rate should hold also when $\beta = 1$.

tails so the population covariance matrix exists. They proposed an estimator for the population covariance and derived finite sample approximation bounds, which depend on various properties of the distribution. For well-behaved elliptical distributions with an exactly sparse covariance matrix, their estimator is minimax rate optimal under the Frobenius norm. Soloveychik and Wiesel (2014) considered estimating a covariance matrix from a convex subset of all positive semidefinite matrices. They added a convex regularization term to the TME and solved the resulting optimization problem by a semidefinite program (SDP). They proved the existence of their estimator and its asymptotic consistency for fixed dimension p and $n \to \infty$. However, their SDP-based method is computationally demanding even for moderate values of n and p. Sun, Babu and Palomar (2016) considered a wider nonconvex class of matrices, and derived an SDP-based algorithm with lower time complexity.

Chen, Gao and Ren (2018) considered an elliptical distribution, corrupted by an ϵ-contamination model. They proposed several estimators for the shape matrix of the elliptical distribution, based on a generalization of Tukey's depth function. Under a notion of sparsity different from the one considered here, they proved their estimator is minimax rate optimal when $n, p \to \infty$ and $(\log p)/n \to 0$. However, from a practical perspective this depth function estimator has a significant limitation—it is intractable to compute. Balakrishnan et al. (2017) considered an ϵ-contamination model for a Gaussian distribution with sparse covariance matrix $\boldsymbol{\Sigma}$, such that $\|\boldsymbol{\Sigma} - \boldsymbol{I}\|_0 \leq s$ for a fixed $s \geq 0$. They proposed a polynomial-time algorithm to robustly estimate $\boldsymbol{\Sigma}$ under this model and established an upper bound on its error under Frobenius norm, assuming $n, p \to \infty$ and $(\log p)/n \to c \geq 0$. Our work in contrast provides a computationally efficient and rate optimal estimator for an approximately sparse shape matrix of a potentially heavy-tailed elliptical distribution in the high-dimensional setting $p, n \to \infty$ with $p/n \to \gamma$. Finally, Avella-Medina et al. (2018) developed rate optimal robust sparse covariance estimators for heavy-tailed distributions via a different approach than the one presented here, based on various robust pilot estimators. Further discussion and directions for future research appear in Section 8.

2. Problem setting. With precise definitions below, given n i.i.d. observations from an elliptical distribution, the problem we study is how to estimate its $p \times p$ *shape matrix* $\boldsymbol{S}_p$. Of particular interest to us is the high-dimensional regime, where both p, n are large and comparable. Following previous works, to be able to accurately estimate the shape matrix in this regime we assume that it is approximately sparse. For completeness, we first introduce some notation, briefly review the elliptical distribution and the class of approximately sparse shape matrices we consider.

Notation. We denote vectors by bold lowercase letters as in $\boldsymbol{v}$, and matrices by bold uppercase letters as in $\boldsymbol{A}$. For a vector $\boldsymbol{v} \in \mathbb{R}^n$, $\|\boldsymbol{v}\|$ is its Euclidean norm, $\|\boldsymbol{v}\|_\infty = \max_i |v_i|$, and $B_R(\boldsymbol{u}) = \{\boldsymbol{v} \in \mathbb{R}^n | \|\boldsymbol{v} - \boldsymbol{u}\|_\infty \leq R\}$. The unit sphere in $\mathbb{R}^p$ is denoted $\mathbb{S}^{p-1}$. The identity matrix is $\boldsymbol{I}$ and $\boldsymbol{0}$ and $\boldsymbol{1}$ are the vectors of zeros and ones respectively, with dimensions clear from the context. For a matrix $\boldsymbol{A} = (a_{ij})$, $\|\boldsymbol{A}\|$ denotes its spectral norm, $\|\boldsymbol{A}\|_F$ its Frobenius norm, $\|\boldsymbol{A}\|_{\max} = \max_{i,j} |a_{ij}|$ and $\|\boldsymbol{A}\|_\infty = \max_i \sum_j |a_{ij}|$. We denote the set of $p \times p$ symmetric positive semidefinite and definite matrices by S_+^p and S_{++}^p, respectively. We say that an event occurs with high probability (abbreviated w.h.p.), if its probability is at least $1 - C\exp(-cp)$ for constants $c, C > 0$ independent of p.

Elliptical distribution and its shape matrix. A random vector $\boldsymbol{x} \in \mathbb{R}^p$ follows an elliptical distribution with location vector $\boldsymbol{\mu}$ if it has the form

$$\boldsymbol{x} = \boldsymbol{\mu} + u\boldsymbol{S}_p^{\frac{1}{2}}\boldsymbol{\xi} = \boldsymbol{\mu} + u\boldsymbol{z}, \tag{1}$$

where $\boldsymbol{\xi}$ is drawn uniformly from $\mathbb{S}^{p-1}$, $\boldsymbol{S}_p \in S_{++}^p$ and u is an arbitrary random or deterministic nonzero scalar, independent of $\boldsymbol{\xi}$.

In equation (1), $\boldsymbol{S}_p$ is not unique, as it can be arbitrarily scaled with u absorbing the inverse scaling factor. Without loss of generality, we thus fix

$$\operatorname{tr}(\boldsymbol{S}_p) = p,$$

and refer to $\boldsymbol{S}_p$ as the *shape matrix*. This normalization is natural in the sense that the mean variance of the p coordinates of $\boldsymbol{z}$ is one. If the distribution is elliptical and the population covariance $\boldsymbol{\Sigma}$ exists, then $\boldsymbol{\Sigma} = c\boldsymbol{S}_p$ for some constant $c > 0$; see, for example, Soloveychik and Wiesel (2014).

An important property of the elliptical distribution is that if $\boldsymbol{x}_1, \boldsymbol{x}_2$ are independent random vectors from (1), then $\boldsymbol{x}_1 - \boldsymbol{x}_2$ has an elliptical distribution with the same shape matrix $\boldsymbol{S}_p$ but with a zero location vector $\boldsymbol{\mu} = \boldsymbol{0}$. When the goal is to estimate the shape matrix $\boldsymbol{S}_p$, this allows to remove the typically unknown location vector by a symmetrization principle (Dümbgen (1998)). Specifically, all $\boldsymbol{x}_i - \boldsymbol{x}_j$ are elliptically distributed with location vector $\boldsymbol{\mu} = \boldsymbol{0}$, and one may estimate the shape matrix using all of these pairwise differences (Dümbgen (1998), Sirkiä, Taskinen and Oja (2007)). As discussed by Nordhausen and Tyler (2015), such a procedure is beneficial also for nonelliptical distributions. The resulting $O(n^2)$ pairs are, however, dependent which may complicate the analysis of the resulting estimator. For simplicity, we shall thus assume to have initially observed $2n$ i.i.d. samples $\tilde{\boldsymbol{x}}_1, \ldots, \tilde{\boldsymbol{x}}_{2n}$ from model (1) and in what follows consider the n differences $\boldsymbol{x}_i = \tilde{\boldsymbol{x}}_{2i} - \tilde{\boldsymbol{x}}_{2i-1}$ for $i = 1, \ldots, n$ which form an i.i.d. sample from the elliptical distribution (1) with location vector $\boldsymbol{\mu} = \boldsymbol{0}$.

Approximate sparsity of the shape matrix. Following Bickel and Levina (2008), we consider the following class of row/column approximately sparse shape matrices with fixed parameters $0 \le q \le 1$, $M > 0$ and $s_p > 0$:

$$\mathcal{U}(q, s_p, M) = \left\{ \boldsymbol{A} \in S_{++}^p : a_{ii} \le M, \sum_{j=1}^{p} |a_{ij}|^q \le s_p, 1 \le i \le p \right\}.$$

Problem statement. Let $\{\boldsymbol{x}_i\}_{i=1}^n$ be n i.i.d. samples from the model (1) with location vector $\boldsymbol{\mu} = \boldsymbol{0}$ and a sparse shape matrix $\boldsymbol{S}_p \in \mathcal{U}(q, s_p, M)$. We consider the following two problems: (i) without explicit knowledge of q, s_p and M, design a computationally efficient and statistically accurate estimator of the shape matrix $\boldsymbol{S}_p$; (ii) provide theoretical guarantees on its accuracy, in the asymptotic limit as $p, n \to \infty$ with $p/n \to \gamma \in (0, \infty)$.

3. Sparse shape matrix estimation.

If the elliptical distribution is sub-Gaussian, then thresholding the sample covariance matrix, proposed by Bickel and Levina (2008) and El Karoui (2008), yields an accurate estimate of $\boldsymbol{S}_p$ up to a multiplicative scaling. As illustrated in Section 7, however, in the presence of heavy tails, thresholding the sample covariance may give a poor estimate of the shape matrix.

To handle heavy tails, we propose the following approach: compute Tyler's M-estimator (TME) or its regularized variant, and threshold it. In Section 3.1, we review TME and its regularized variant. We prove that computing the latter is computationally efficient. Section 3.2 presents our proposed estimators. A theoretical analysis of their accuracy appears in Section 3.3.

3.1. *TME and its regularized variant*.

TME, proposed by Tyler (1987a) for elliptical distributions with a known location vector, which w.l.o.g. is assumed to be $\boldsymbol{0}$, is a $p \times p$ matrix $\hat{\boldsymbol{\Sigma}}$ which satisfies the nonlinear equation

$$\frac{p}{n} \sum_{i=1}^{n} \frac{\boldsymbol{x}_i \boldsymbol{x}_i^T}{\boldsymbol{x}_i^T \hat{\boldsymbol{\Sigma}}^{-1} \boldsymbol{x}_i} = \hat{\boldsymbol{\Sigma}}. \tag{2}$$

Here, samples $\boldsymbol{x}_i$ lying at the origin are ignored as they provide no information on the scatter matrix, and n is the number of samples not at the origin. As solutions to (2) can be multiplied by an arbitrary constant, Tyler (1987a) considered the normalization $\operatorname{tr}(\hat{\boldsymbol{\Sigma}}) = p$, and suggested to solve equation (2) by the following iterations, starting from an arbitrary $\hat{\boldsymbol{\Sigma}}_1 \in S_{++}^p$:

$$\hat{\boldsymbol{\Sigma}}_{k+1} = p \sum_{i=1}^{n} \frac{\boldsymbol{x}_i \boldsymbol{x}_i^T}{\boldsymbol{x}_i^T \hat{\boldsymbol{\Sigma}}_k^{-1} \boldsymbol{x}_i} \Big/ \operatorname{tr}\left(\sum_{i=1}^{n} \frac{\boldsymbol{x}_i \boldsymbol{x}_i^T}{\boldsymbol{x}_i^T \hat{\boldsymbol{\Sigma}}_k^{-1} \boldsymbol{x}_i} \right).$$

Kent and Tyler (1988), Theorems 1 and 2, showed that if any linear subspace in $\mathbb{R}^p$ of dimension $1 \le d \le p-1$ contains less than nd/p of the data samples, then there exists a unique solution to equation (2), and the above iterations converge to it. With n i.i.d. observations from an elliptical distribution, no samples lie at the origin and this condition holds with probability 1.

TME enjoys several important properties: First, it may equivalently be defined as the minimizer of the following cost function, over all positive definite matrices with the constraint $\operatorname{tr}(\boldsymbol{R}) = p$,

$$L(\boldsymbol{R}) = \frac{p}{n} \sum_{i=1}^{n} \log\left(\boldsymbol{x}_i^T \boldsymbol{R}^{-1} \boldsymbol{x}_i\right) + \log\left(\det(\boldsymbol{R})\right). \tag{3}$$

Equation (3) implies that $\hat{\boldsymbol{\Sigma}}$ is the maximum likelihood estimator of the shape matrix of the angular central Gaussian distribution (Tyler (1987b)) and of the generalized elliptical distribution (Frahm and Jaekel (2010)). Second, for data i.i.d. from a continuous elliptical distribution with p fixed, it is the "most robust" estimator of the shape matrix as $n \to \infty$ (Tyler (1987a), Remark 3.1). TME outperforms the sample covariance in a variety of applications, including finance (Frahm and Jaekel (2007)), anomaly detection in wireless sensor networks (Chen, Wiesel and Hero (2011)), antenna array processing (Ollila and Koivunen (2003)) and radar detection (Ollila and Tyler (2012)).

As the TME does not exist when $p > n$, several regularized variants have been proposed and analyzed (Abramovich and Spencer (2007), Chen, Wiesel and Hero (2011), Pascal, Chitour and Quek (2014), Sun, Babu and Palomar (2014), Wiesel (2012)). Even when $p \le n$, it is common to add small regularization to the TME. Following Sun, Babu and Palomar (2014), we consider the following regularized TME $\hat{\boldsymbol{\Sigma}}(\alpha)$, defined as the solution of

$$\hat{\boldsymbol{\Sigma}}(\alpha) = \frac{1}{1+\alpha} \frac{p}{n} \sum_{i=1}^{n} \frac{\boldsymbol{x}_i \boldsymbol{x}_i^T}{\boldsymbol{x}_i^T \hat{\boldsymbol{\Sigma}}(\alpha)^{-1} \boldsymbol{x}_i} + \frac{\alpha}{1+\alpha} \boldsymbol{I}, \tag{4}$$

where α is a regularization parameter. If $\alpha = 0$, equation (4) reverts to equation (2). While regularization toward general target matrices is possible (Wiesel (2012)), here for simplicity we consider only regularization toward the identity. In contrast to the TME of equation (2), where solutions can be multiplied by an arbitrary positive scalar, any solution to equation (4) satisfies $\operatorname{tr}(\hat{\boldsymbol{\Sigma}}(\alpha)^{-1}) = p$, for any value of α (Pascal, Chitour and Quek (2014), Proposition III.1).

Sun, Babu and Palomar (2014), Theorem 11 and Proposition 13, derived a sufficient and necessary condition for existence of a unique positive definite matrix which solves equation (4). Again, ignoring samples at the origin, the condition is that any linear subspace in $\mathbb{R}^p$ of dimension $1 \le d \le p-1$ contains less than $(1+\alpha)nd/p$ of the samples. Since $\alpha > 0$, this condition is weaker than for the original TME. In particular, with data i.i.d. from a continuous distribution, equation (4) has a unique solution for $\alpha > \max(0, p/n - 1)$; see also Pascal, Chitour and Quek (2014), Theorem III.1. With n i.i.d. samples from an elliptical distribution, these conditions hold with probability 1.

Sun, Babu and Palomar (2014), Proposition 18, further showed that starting from any positive definite initial guess, the following iterations

$$\hat{\mathbf{\Sigma}}_{k+1}(\alpha) = \frac{1}{1+\alpha}\frac{p}{n}\sum_{i=1}^{n}\frac{\boldsymbol{x}_i\boldsymbol{x}_i^T}{\boldsymbol{x}_i^T\hat{\mathbf{\Sigma}}_k(\alpha)^{-1}\boldsymbol{x}_i} + \frac{\alpha}{1+\alpha}\boldsymbol{I} \tag{5}$$

converge to the unique solution. Various properties of TME and its regularized variant, in the limit as $p, n \to \infty$ with $p/n \to \gamma$, were proven by Couillet, Kammoun and Pascal (2016), Couillet and McKay (2014), Dümbgen (1998), Zhang, Cheng and Singer (2016).

The following lemma, proven in the Appendix, shows that if α is sufficiently large and $\hat{\mathbf{\Sigma}}(\alpha)$ exists, then the iterations (5), starting with $\hat{\mathbf{\Sigma}}_1(\alpha) = \alpha\boldsymbol{I}/(1+\alpha)$, have a uniform linear convergence rate from the first iteration. As far as we know, this result is new and is of independent interest.

To state the lemma, let $e_k = \|\hat{\mathbf{\Sigma}}(\alpha) - \hat{\mathbf{\Sigma}}_k(\alpha)\|$ be the error after k iterations, $\tilde{\boldsymbol{X}}$ be the $p \times n$ matrix whose columns are $\{\boldsymbol{x}_i/\|\boldsymbol{x}_i\|\}_{i=1}^n$ and let

$$C(\tilde{\boldsymbol{X}}) = \frac{p}{n}\|\tilde{\boldsymbol{X}}\tilde{\boldsymbol{X}}^T\| = \frac{p}{n}\left\|\sum_{i=1}^{n}\frac{\boldsymbol{x}_i\boldsymbol{x}_i^T}{\|\boldsymbol{x}_i\|^2}\right\|.$$

Note that for a given data set, $C(\tilde{\boldsymbol{X}})$ is fixed and can be computed a priori.

LEMMA 1. *Let $\{\boldsymbol{x}_i\}_{i=1}^n$ be a data set in $\mathbb{R}^p$ with constant $C(\tilde{\boldsymbol{X}})$ and let $0 < R < 1$. Suppose that $\alpha > \max((3 + R^{-1})C(\tilde{\boldsymbol{X}}) - 1, 0)$ and let $\hat{\mathbf{\Sigma}}(\alpha)$ be a solution of (4). Then the iterations of equations (5), starting from $\hat{\mathbf{\Sigma}}_1(\alpha) = \frac{\alpha}{1+\alpha}\boldsymbol{I}$, uniformly and linearly converge to $\hat{\mathbf{\Sigma}}(\alpha)$ with the ratio R. That is,*

$$e_{k+1} \le Re_k \le R^k e_1 \qquad \text{for all } k \ge 1. \tag{6}$$

A straightforward calculation yields the bound $C(\tilde{\boldsymbol{X}}) \ge p/n$. Hence, the above assumptions on α imply that $\alpha > \max(0, p/n - 1)$, and consequently guarantee the existence and uniqueness of $\hat{\mathbf{\Sigma}}(\alpha)$ in our setting.

Lemma 1 implies that for sufficiently large α, calculating $\hat{\mathbf{\Sigma}}(\alpha)$ is computationally efficient. Specifically, for accuracy ϵ with convergence ratio R, at most $\lceil \log_{R^{-1}}(\epsilon^{-1})\rceil$ iterations are needed. If $n > p$, then at each iteration, the matrix inversion costs $O(p^3)$ operations and the other operations are $O(np^2)$. For $n < p$, one may first perform an SVD of the data and calculate $\hat{\mathbf{\Sigma}}(\alpha)$ restricted to the subspace $W = \mathrm{Span}(\boldsymbol{x}_i)$ whose dimension is at most n. This suffices, since for any $\boldsymbol{v} \perp W$, by definition $\hat{\mathbf{\Sigma}}(\alpha)\boldsymbol{v} = \alpha/(1+\alpha)\boldsymbol{v}$. Each iteration then costs at most $O(n^3)$. Therefore, the total cost of computing $\hat{\mathbf{\Sigma}}(\alpha)$ within accuracy ϵ is $O(\log(\epsilon^{-1})(n+p)\min(n,p)^2)$.

Our theoretical analysis below studies the regularized TME as $p, n \to \infty$ and $p/n \to \gamma \in (0, \infty)$, but with a fixed value of α. The next lemma shows that for data sampled from an elliptical distribution, with high probability $C(\tilde{\boldsymbol{X}})$ is bounded by a constant that depends on $\|\boldsymbol{S}_p\|$ and on the ratio p/n.

LEMMA 2. *Let $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ be i.i.d. from equation (1) with $\boldsymbol{\mu} = \boldsymbol{0}$ and shape matrix $\boldsymbol{S}_p$. Then, with probability $> 1 - \exp(-cp)$, where $c = c(\|\boldsymbol{S}_p\|) > 0$,*

$$C(\tilde{\boldsymbol{X}}) \le 2\|\boldsymbol{S}_p\|(1 + 2\sqrt{p/n})^2. \tag{7}$$

3.2. *TME-based thresholding estimators.* One possible approach to construct a sparse and robust estimator for the shape matrix is to add a suitable penalty to the original cost functional equation (3) of the TME. For various structural assumptions on the shape matrix, this was proposed by Soloveychik and Wiesel (2014) and by Sun, Babu and Palomar (2016).

With a sparsity inducing penalty, however, this approach in general leads to a nonconvex and potentially difficult to optimize objective. Instead, we opt for thresholding the (regularized) TME, which as discussed above, can be computed efficiently in practical polynomial time.

For a matrix $\boldsymbol{A} = (a_{ij})$ and threshold $t > 0$, define the entry-wise hard-thresholding operator by

$$\tau_t(\boldsymbol{A}) = \big(\mathbf{1}(|a_{ij}| > t)a_{ij}\big).$$

For $n > p$, where the TME $\hat{\boldsymbol{\Sigma}}$ exists and by definition has unit trace, our proposed estimator for the shape matrix $\boldsymbol{S}_p$ takes the form

$$\hat{\boldsymbol{S}}_p = \tau_t(\hat{\boldsymbol{\Sigma}}), \tag{8}$$

where the threshold $t = t(p, n)$ is specified below. Similarly, for general p, n, our estimator based on the regularized TME is

$$\hat{\boldsymbol{S}}_p = \tau_t\left(p\frac{\hat{\boldsymbol{\Sigma}}(\alpha) - \frac{\alpha}{1+\alpha}\boldsymbol{I}}{\operatorname{tr}(\hat{\boldsymbol{\Sigma}}(\alpha) - \frac{\alpha}{1+\alpha}\boldsymbol{I})}\right). \tag{9}$$

Note that both $\hat{\boldsymbol{\Sigma}}$ in equation (8) and the argument matrix prior to thresholding in equation (9) have rank at most $\min(n, p)$.

3.3. *Accuracy of the thresholded TME.* The following Theorems 1 and 2, proved in Sections 5 and 6, respectively, establish the asymptotic accuracy of equations (8) and (9) as estimates of the shape matrix $\boldsymbol{S}_p$.

THEOREM 1. *Consider a sequence $(n, p, \boldsymbol{S}_p)$ where $n \to \infty$, $p = p_n \to \infty$ with $p/n \to \gamma \in (0, 1)$, and $\boldsymbol{S}_p \in \mathcal{U}(q, s_p, M)$. For each triplet $(n, p, \boldsymbol{S}_p)$, let $\hat{\boldsymbol{\Sigma}}$ be the TME of n i.i.d. samples $\{\boldsymbol{x}_i\}_{i=1}^n \subset \mathbb{R}^p$ from the elliptical distribution (1). Then there exists a constant M' depending only on γ such that for any fixed $M'' > M'$, the thresholded TME of equation (8) with threshold $t_n = M''\sqrt{\log p/n}$, approaches $\boldsymbol{S}_p$ in spectral norm at a rate*

$$\|\tau_{t_n}(p\hat{\boldsymbol{\Sigma}}) - \boldsymbol{S}_p\| = \mathcal{O}_P\left(s_p \cdot \left(\frac{\log p}{n}\right)^{(1-q)/2}\right).$$

THEOREM 2. *Consider a sequence $(n, p, \boldsymbol{S}_p)$ as in Theorem 1, here with $p/n \to \gamma \in (0, \infty)$ and with the additional assumption that $\|\boldsymbol{S}_p\| \le s_{\max}$. For $\alpha > \max(0, \gamma - 1 + s_{\max}(1 + \sqrt{\gamma})^2)$, let $\hat{\boldsymbol{\Sigma}}(\alpha)$ be the regularized TME of n i.i.d. samples $\{\boldsymbol{x}_i\}_{i=1}^n \subset \mathbb{R}^p$ from the elliptical distribution (1). Then there exists a M' depending only on γ and α such that for any fixed $M'' > M'$, the estimator of equation (9) with $t_n = M''\sqrt{\log p/n}$, converges in spectral norm to $\boldsymbol{S}_p$ at rate*

$$\left\|\tau_{t_n}\left(p\frac{(\hat{\boldsymbol{\Sigma}}(\alpha) - \frac{\alpha}{1+\alpha}\boldsymbol{I})}{\operatorname{tr}(\hat{\boldsymbol{\Sigma}}(\alpha) - \frac{\alpha}{1+\alpha}\boldsymbol{I})}\right) - \boldsymbol{S}_p\right\| = \mathcal{O}_P\left(s_p\left(\frac{\log p}{n}\right)^{(1-q)/2}\right).$$

Several remarks regarding Theorem 2 are in place.

REMARK 1. As noted by Bickel and Levina (2008), page 2580, if $\boldsymbol{S}_p \in \mathcal{U}(q, s_p, M)$ then $\|\boldsymbol{S}_p\| \leq M^{1-q} s_p$ which may grow with p. Since we analyze the regularized TME with a *fixed* value of α, we explicitly require that $\|\boldsymbol{S}_p\| \leq s_{\max}$ independent of p. If $\|\boldsymbol{S}_p\|$ grows to infinity with p, then the regularization α should also grow to infinity with p, such that $\alpha > c\|\boldsymbol{S}_p\|$ for some constant $c > 0$. While beyond the scope of this paper, we believe an analogue of Theorem 2 should hold in this case.

REMARK 2. The convergence rate in Theorems 1 and 2 coincides with the minimax rate for sparse covariance estimation with sub-Gaussian data, derived by Cai and Zhou (2012a). Since the Gaussian distribution is a particular case of an elliptical distribution, our estimators are thus minimax rate optimal. Furthermore, in light of Lemmas 1 and 2, computing the regularized TME and subsequently thresholding it, is computationally efficient.

REMARK 3. Several authors proposed the regularized Tyler's M-estimator as is (without modification), as an estimate of the shape matrix. In this case, choosing the value of the regularization constant is crucial (Chen, Wiesel and Hero (2011), Couillet and McKay (2014)). Setting the regularization parameter is also important to maximize the detection probability in various signal processing applications (Kammoun et al. (2018)). In contrast, as we subtract the regularization $\alpha/(1+\alpha)\boldsymbol{I}$ prior to thresholding, at least asymptotically, the precise value of α is unimportant, provided it is sufficiently large. This is also evident in the simulations in Section 7. From a practical perspective, we thus suggest to use a value of α as described in Lemma 1, with say $R = 1/2$, which is not only sufficient for existence but also guarantees fast convergence of the iterations that compute the regularized TME.

4. Preliminaries. In proving Theorems 1 and 2, we shall make frequent use of the following auxiliary lemmas. The first is a simple inequality. Let A, B be nonnegative random variables. Then for any $c > 0$ and $\lambda > 0$,

$$\Pr(AB > c) \leq \Pr(A > \lambda c) + \Pr(B > 1/\lambda). \tag{10}$$

Next is the following well-known result, which states that TME and regularized TME are unable to distinguish an elliptical distribution from a Gaussian one. Its proof (omitted) follows directly from the fact that (regularized) TME for data $\boldsymbol{x}_i$ is identical to that of data $t_i\boldsymbol{x}_i$, where t_i are arbitrary positive real valued numbers.

LEMMA 3. *TME or regularized TME with $\alpha > \max(0, p/n - 1)$ under an elliptical distribution with zero location vector and shape matrix $\boldsymbol{S}_p$ has the same distribution as under a zero mean Gaussian distribution with covariance $\boldsymbol{S}_p$.*

The following two results from random matrix theory, concerning the spectral norm of a Wishart matrix and concentration of quadratic forms will also be of use; see, for example, Davidson and Szarek (2001), Theorem 2.13 and Rudelson and Vershynin (2013), Theorem 1.1.

LEMMA 4. *Let $\{\boldsymbol{\xi}_i\}_{i=1}^n \subset \mathbb{R}^p$ be i.i.d. $N(0, \boldsymbol{I})$, and let $\boldsymbol{T}_n = \frac{1}{n}\sum_i \boldsymbol{\xi}_i\boldsymbol{\xi}_i^T$. Then, $\mathbb{E}[\|\boldsymbol{T}_n\|] \leq (1+\sqrt{p/n})^2$ and*

$$\Pr\big(\|\boldsymbol{T}_n\| > (1+\sqrt{p/n}+t)^2\big) \leq \exp(-nt^2/2).$$

LEMMA 5. *Let $\boldsymbol{A} \in \mathbb{R}^{p\times p}$ and $\boldsymbol{\xi} \sim N(\boldsymbol{0}, \boldsymbol{I})$. Then there exist absolute constants $c_1, c_2 > 0$ such that for all $\epsilon > 0$,*

$$\Pr\big(\big|\boldsymbol{\xi}^T\boldsymbol{A}\boldsymbol{\xi} - \operatorname{tr}(\boldsymbol{A})\big| > \epsilon\big) \leq 2\exp\bigg(-c_1 \min\bigg\{\frac{c_2^2\epsilon^2}{\|\boldsymbol{A}\|_F^2}, \frac{c_2\epsilon}{\|\boldsymbol{A}\|}\bigg\}\bigg).$$

Finally, the following auxiliary lemma, proved in the supplemental article (Goes, Lerman and Nadler (2020)), is a slight modification of a result by Bickel and Levina (2008), page 2583.

LEMMA 6. *Assume $\boldsymbol{B} \in \mathcal{U}(q, s_p, M)$. Let $\boldsymbol{A}$ be a matrix such that*

$$\|\boldsymbol{A} - \boldsymbol{B}\|_{\max} \le C_1\sqrt{\log p/n},$$

for some $C_1 > 0$. Suppose we threshold $\boldsymbol{A}$ at level $t = K\sqrt{\log p/n}$, with $K > C_1$. Then there exists a constant $C_2 = C_2(C_1, K, q) < \infty$ such that

$$\|\tau_t(\boldsymbol{A}) - \boldsymbol{B}\| \le C_2 s_p(\log p/n)^{(1-q)/2}.$$

5. Proof of Theorem 1. Let $\hat{\boldsymbol{S}}$ be the sample covariance of $\{\boldsymbol{x}_i\}_{i=1}^n$. In light of Lemma 3, we may assume that $\boldsymbol{x}_i$ are all i.i.d. $N(\boldsymbol{0}, \boldsymbol{S}_p)$. The proof proceeds in three steps: (i) reducing to a bound on $\|\hat{\boldsymbol{\Sigma}} - \hat{\boldsymbol{S}}\|_{\max}$; (ii) expressing $\hat{\boldsymbol{\Sigma}}$ as a weighted covariance matrix whose weights are all uniformly close to a constant, with high probability; and (iii) bounding $\|\hat{\boldsymbol{\Sigma}} - \hat{\boldsymbol{S}}\|_{\max}$.

5.1. *Step* 1: *Reduction from $\|\tau_{t_n}(\hat{\boldsymbol{\Sigma}}) - \boldsymbol{S}_p\|$ to $\|\hat{\boldsymbol{\Sigma}} - \hat{\boldsymbol{S}}\|_{\max}$.* By Lemma 6, it suffices to prove that $\|\hat{\boldsymbol{\Sigma}} - \boldsymbol{S}_p\|_{\max} = \mathcal{O}_P(\sqrt{\log p/n})$. By the triangle inequality,

$$\|\hat{\boldsymbol{\Sigma}} - \boldsymbol{S}_p\|_{\max} \le \|\hat{\boldsymbol{\Sigma}} - \hat{\boldsymbol{S}}\|_{\max} + \|\hat{\boldsymbol{S}} - \boldsymbol{S}_p\|_{\max}.$$

Since the proof of Theorem 1 of Bickel and Levina (2008) shows that

$$\|\hat{\boldsymbol{S}} - \boldsymbol{S}_p\|_{\max} = \mathcal{O}_P(\sqrt{\log p/n})$$

it thus suffices to show that

$$\|\hat{\boldsymbol{\Sigma}} - \hat{\boldsymbol{S}}\|_{\max} = \mathcal{O}_P(\sqrt{\log p/n}). \tag{11}$$

5.2. *Step* 2: *The weights of TME.* By Zhang, Cheng and Singer (2016), Lemma 2.1, TME can be written as a weighted covariance matrix,

$$\hat{\boldsymbol{\Sigma}} = p\sum_{i=1}^n w_i \boldsymbol{x}_i \boldsymbol{x}_i^T \Big/ \operatorname{tr}\left(\sum_{i=1}^n w_i \boldsymbol{x}_i \boldsymbol{x}_i^T\right),$$

where the weights w_i are the unique solution of

$$\arg\min_{w_i>0, \sum w_i = 1} -\sum_{i=1}^n \ln w_i + \frac{n}{p}\ln\det\left(\sum_{i=1}^n w_i \boldsymbol{x}_i \boldsymbol{x}_i^T\right). \tag{12}$$

This characterization is important because of the following result.

LEMMA 7. *Consider a sequence $(n, p, \boldsymbol{S}_p)$ where $n, p \to \infty$ with $p/n \to \gamma \in (0, 1)$, and $\boldsymbol{S}_p \in S_{++}^p$. For every triplet $(n, p, \boldsymbol{S}_p)$, let $\boldsymbol{x}_i \overset{i.i.d.}{\sim} N(\boldsymbol{0}, \boldsymbol{S}_p)$ and let $\{w_i\}_{i=1}^n$ be the corresponding weights of equation (12). Then there exist positive constants C, c and c' depending only on γ such that for any $0 < \epsilon < c'$, and sufficiently large n,*

$$\Pr\Big[\max_i |n w_i - 1| \ge \epsilon\Big] \le Cne^{-c\epsilon^2 n}. \tag{13}$$

The case $\boldsymbol{S}_p = \boldsymbol{I}$ was proved by Zhang, Cheng and Singer (2016), Lemma 2.2. Its generalization to an arbitrary $\boldsymbol{S}_p \in S_{++}^p$ is proved in the supplemental article (Goes, Lerman and Nadler (2020)).

5.3. *Step* 3: *Bounding* $\|\hat{\boldsymbol{\Sigma}} - \hat{\boldsymbol{S}}\|_{\max}$. The proof concludes by applying the following lemma which establishes equation (11). Its proof is in Appendix A.2.

LEMMA 8. *Let* $\hat{\boldsymbol{\Sigma}}$ *and* $\hat{\boldsymbol{S}}$ *be the TME and the sample covariance matrix of* $\boldsymbol{x}_1, \dots, \boldsymbol{x}_n$ *i.i.d. from* $N(\boldsymbol{0}, \boldsymbol{S}_p)$, *where* $\boldsymbol{S}_p \in \mathcal{U}(q, s_p, M)$ *with* $\operatorname{tr}(\boldsymbol{S}_p) = p$. *Assume that* $p, n \to \infty$, *with* $p/n \to \gamma \in (0, 1)$. *Then there exist positive constants* C, c *and* c' *that depend only on* γ, *such that for all* $\epsilon \in (0, c')$ *and* n *sufficiently large*

$$\Pr\bigl(\|\hat{\boldsymbol{\Sigma}} - \hat{\boldsymbol{S}}\|_{\max} \geq \epsilon\bigr) \leq Cne^{-c\epsilon^2 n}.$$

6. Proof of Theorem 2. We first introduce and prove a slightly modified version of Theorem 2. We then show how Theorem 2 follows from it. The modified theorem uses the following proposition, proved in Appendix A.3.

PROPOSITION 1. *Let* $\boldsymbol{y}, \boldsymbol{\xi}_1, \dots, \boldsymbol{\xi}_{n-1} \in \mathbb{R}^p$ *be i.i.d.* $N(\boldsymbol{0}, \boldsymbol{I})$ *and denote*

$$Q = Q(r) = \frac{1}{p}\boldsymbol{y}^T \left(\frac{1}{n} \sum_{j=1}^{n-1} \boldsymbol{\xi}_j \boldsymbol{\xi}_j^T + \alpha \frac{n}{p} \frac{1}{r} \boldsymbol{S}_p^{-1} \right)^{-1} \boldsymbol{y},$$

where $\boldsymbol{S}_p \in S_{++}^p$ *with* $\|\boldsymbol{S}_p\| \leq s_{\max}$. *Assume that* $\alpha > \max(0, p/n - 1 + s_{\max}(1 + \sqrt{p/n})^2)$, *and define*

$$r_{\min} = \frac{n}{p} \frac{\alpha}{1 + \alpha - p/n}, \qquad r_{\max} = \frac{n}{p} \frac{\alpha}{1 + \alpha - p/n - s_{\max}(1 + \sqrt{p/n})^2}.$$

Then there exists a unique $r = r(p, n, \alpha, \boldsymbol{S}_p) \in [r_{\min}, r_{\max}]$, *such that*

$$\mathbb{E}\bigl[Q(r)\bigr] = \frac{1}{1 + \alpha - p/n}, \tag{14}$$

where the expectation is over $\boldsymbol{y}$ *and* $\boldsymbol{\xi}_1, \dots, \boldsymbol{\xi}_{n-1}$.

6.1. *A reformulation of the main result.* We now introduce the modified theorem.

THEOREM 3. *Consider the setting of Theorem 2. Then there exists an* M' *depending only on* γ *and* α *such that for any fixed* $M'' > M'$, *the estimator* $\tau_{t_n}(\hat{\boldsymbol{\Sigma}}(\alpha) - \alpha \boldsymbol{I}/(1+\alpha))$ *with* $t_n = M''\sqrt{\frac{\log p}{n}}$, *converges in spectral norm to a multiple of* $\boldsymbol{S}_p$,

$$\left\| \tau_{t_n}\left(\hat{\boldsymbol{\Sigma}}(\alpha) - \frac{\alpha}{1+\alpha} \boldsymbol{I} \right) - \frac{p}{n} \frac{r}{1+\alpha} \boldsymbol{S}_p \right\| = \mathcal{O}_P\left(s_p \left(\frac{\log p}{n} \right)^{(1-q)/2} \right),$$

where the scalar $r = r(p, n, \alpha, \boldsymbol{S}_p)$ *is specified in Proposition 1.*

6.2. *Proof of Theorem 3.* By Lemma 3, we may assume $\boldsymbol{x}_i \overset{i.i.d.}{\sim} N(\boldsymbol{0}, \boldsymbol{S}_p)$. Following the argument in Section 5.1, combining Lemma 6 with the fact that by Proposition 1, $r < r_{\max}$, it suffices to show that

$$\left\| \left(\hat{\boldsymbol{\Sigma}}(\alpha) - \frac{\alpha}{1+\alpha} \boldsymbol{I} \right) - \frac{p}{n} \frac{r}{1+\alpha} \hat{\boldsymbol{S}} \right\|_{\max} = \mathcal{O}_P(\sqrt{\log p/n}). \tag{15}$$

To establish equation (15), we first express $\hat{\boldsymbol{\Sigma}}(\alpha)$ as the sum of $\frac{\alpha}{1+\alpha}\boldsymbol{I}$ and weighted $\boldsymbol{x}_i \boldsymbol{x}_i^T$ terms, where the weights are the root of some equation. Next we show that this root is concentrated near the vector $r\boldsymbol{1}/n$, with r specified in Proposition 1.

Following the definition of the regularized TME, we write $\hat{\boldsymbol{\Sigma}}(\alpha)$ as

$$\hat{\boldsymbol{\Sigma}}(\alpha) = \frac{1}{1+\alpha}\frac{p}{n}\sum_{i=1}^{n} w_i \boldsymbol{x}_i \boldsymbol{x}_i^T + \frac{\alpha}{1+\alpha}\boldsymbol{I}, \tag{16}$$

where the weight vector $\boldsymbol{w} = (w_1, \ldots, w_n)^T$ satisfies

$$w_i = \frac{1}{\boldsymbol{x}_i^T \hat{\boldsymbol{\Sigma}}(\alpha)^{-1} \boldsymbol{x}_i} = \frac{1}{\boldsymbol{x}_i^T \left(\frac{1}{1+\alpha}\frac{p}{n}\sum_{j=1}^{n} w_j \boldsymbol{x}_j \boldsymbol{x}_j^T + \frac{\alpha}{1+\alpha}\boldsymbol{I}\right)^{-1} \boldsymbol{x}_i}. \tag{17}$$

Next consider the function $g : \mathbb{R}^n \to \mathbb{R}^n$ whose n components are

$$g(\boldsymbol{v})_i = v_i - \frac{1}{\boldsymbol{x}_i^T \left(\frac{1}{1+\alpha}\frac{p}{n}\sum_{k=1}^{n} v_k \boldsymbol{x}_k \boldsymbol{x}_k^T + \frac{\alpha}{1+\alpha} n\boldsymbol{I}\right)^{-1} \boldsymbol{x}_i}. \tag{18}$$

Comparing equation (18) to equation (17), since $\hat{\boldsymbol{\Sigma}}(\alpha)$ is unique (Sun, Babu and Palomar (2014), Theorem 11), then the n nonlinear equations $g(\boldsymbol{v}) = \boldsymbol{0}$ have a unique solution, which is thus $n\boldsymbol{w}$. The next three lemmas state properties of g used to prove that as $p, n \to \infty$, with $p/n \to \gamma$, this root concentrates around $\boldsymbol{u} = r\boldsymbol{1}$, with r given in Proposition 1. Lemma 9 is proven in Appendix A.4 and the other two in the supplemental article (Goes, Lerman and Nadler (2020)). All three lemmas assume the setting of Theorem 3, and their generic constants depend only on γ, α and $s_{\max}$. Our analysis of the weights w_i follows the pioneering works of Couillet, Pascal and Silverstein (2014, 2015), who proved that the weights in Maronna's M-estimators converge to suitable constants, and Zhang, Cheng and Singer (2016), who derived concentration results for the weights of Tyler's M-estimator as $p, n \to \infty$ with $p/n \to \gamma < 1$.

LEMMA 9. *There exist $C, c > 0$ such that for any $\epsilon \in (0, 1)$*

$$\Pr\left(\|g(\boldsymbol{u})\|_\infty > \epsilon\right) < Cpe^{-cp\epsilon^2}.$$

LEMMA 10. *There exist $c', c_L, C, c > 0$ such that*

$$\Pr\left(\exists \boldsymbol{v} \in B_{c'}(\boldsymbol{u}), \|\nabla g(\boldsymbol{v}) - \nabla g(\boldsymbol{u})\|_{\max} > c_L \|\boldsymbol{v} - \boldsymbol{u}\|_\infty\right) < Cp^2 e^{-cp}.$$

LEMMA 11. *There exist $c_H, C, c > 0$ such that*

$$\Pr\left(\|(\nabla g(\boldsymbol{u}))^{-1}\|_\infty > c_H\right) < Cpe^{-cp}. \tag{19}$$

Lemmas 9 and 10 show that w.h.p. $g(\boldsymbol{u})$ is small and ∇g is Lipschitz near $\boldsymbol{u}$. These two properties are consistent with the root of g being close to $\boldsymbol{u}$. To rigorously prove this, following Zhang, Cheng and Singer (2016), we consider the function $f(\boldsymbol{v}) = (\nabla g(\boldsymbol{u}))^{-1} g(\boldsymbol{v})$. Lemma 11 shows that the matrix $(\nabla g(\boldsymbol{u}))^{-1}$ is w.h.p. not extremely large. Finally, the following lemma combines these properties of g to infer that its root is close to $\boldsymbol{u}$.

LEMMA 12. *Let $f : \mathbb{R}^n \to \mathbb{R}^n$, $\boldsymbol{u} \in \mathbb{R}^n$ and $C > 0$. Assume that*:

1. $\nabla f(\boldsymbol{u}) = \boldsymbol{I}$;
2. $\|\nabla f(\boldsymbol{v}) - \nabla f(\boldsymbol{u})\|_{\max} \le C\|\boldsymbol{v} - \boldsymbol{u}\|_\infty$ *for all* $\|\boldsymbol{v} - \boldsymbol{u}\|_\infty \le 3\|f(\boldsymbol{u})\|_\infty$;
3. $\|f(\boldsymbol{u})\|_\infty < \min\{1/(9C), 1/3\}$.

Then there exists a $\tilde{\boldsymbol{v}} \in \mathbb{R}^n$ such that $f(\tilde{\boldsymbol{v}}) = \boldsymbol{0}$ and $\|\tilde{\boldsymbol{v}} - \boldsymbol{u}\|_\infty < 3\|f(\boldsymbol{u})\|_\infty$.

Lemma 12 is slightly stronger than Lemma 3.1 of Zhang, Cheng and Singer (2016), as it has a weaker requirement that the Lipschitz condition in Lemma 12 holds in a smaller ball $\|\boldsymbol{v}-\boldsymbol{u}\|_\infty \le 3\|f(\boldsymbol{u})\|_\infty$, instead of the original requirement $\|\boldsymbol{v}-\boldsymbol{u}\|_\infty < 1$ in their Lemma 3.1. A careful inspection shows that their original proof is still valid under this weaker assumption.

To apply Lemma 12 to $f(\boldsymbol{v}) = (\nabla g(\boldsymbol{u}))^{-1} g(\boldsymbol{v})$, we verify that the three conditions of the lemma hold with high probability. The first condition is trivially satisfied. For the other two conditions, by Lemmas 9 and 11, w.h.p.

$$\|f(\boldsymbol{u})\|_\infty \le \|(\nabla g(\boldsymbol{u}))^{-1}\|_\infty \cdot \|g(\boldsymbol{u})\|_\infty \le c_H \epsilon.$$

Similarly, by Lemmas 10 and 11, for all $\|\boldsymbol{v}-\boldsymbol{u}\|_\infty \le c'$, w.h.p.

$$\|\nabla f(\boldsymbol{v}) - \nabla f(\boldsymbol{u})\|_{\max} \le \|(\nabla g(\boldsymbol{u}))^{-1}\|_\infty \cdot \|\nabla g(\boldsymbol{v}) - \nabla g(\boldsymbol{u})\|_{\max} \le c_H c_L \|\boldsymbol{v}-\boldsymbol{u}\|_\infty.$$

Since for sufficiently small ϵ, $c_H\epsilon < \min\{1/(9c_L c_H), 1/3\}$, both the second and third conditions of Lemma 12 are thus satisfied with constant $C = c_L c_H$.

To conclude, with probability at least $1 - Cp^2 e^{-cp\epsilon^2}$ all three conditions of Lemma 12 hold, so there exists $\tilde{\boldsymbol{v}} \in \mathbb{R}^n$ such that $f(\tilde{\boldsymbol{v}}) = \boldsymbol{0}$ and $\|\tilde{\boldsymbol{v}} - \boldsymbol{u}\|_\infty \le 3\|f(\boldsymbol{u})\|_\infty < 3c_H\epsilon$. Since $n\boldsymbol{w}$ is the unique root of $g(\boldsymbol{v})$ and also of $f(\boldsymbol{v})$,

$$\Pr\big(\|n\boldsymbol{w} - r\boldsymbol{1}\|_\infty > 3c_H\epsilon\big) < Cp^2 e^{-cp\epsilon^2}. \tag{20}$$

Next we use equation (20) to bound the LHS of equation (15). First, by equation (16),

$$\begin{aligned}
&\left\| \left(\hat{\boldsymbol{\Sigma}}(\alpha) - \frac{\alpha}{1+\alpha}\boldsymbol{I} \right) - \frac{1}{1+\alpha}\frac{p}{n} r\hat{\boldsymbol{S}} \right\|_{\max} \\
&\quad = \frac{1}{1+\alpha}\frac{p}{n} \left\| \sum_{i=1}^n w_i \boldsymbol{x}_i \boldsymbol{x}_i^T - r\frac{1}{n}\sum_{i=1}^n \boldsymbol{x}_i \boldsymbol{x}_i^T \right\|_{\max} \\
&\quad \le \frac{1}{1+\alpha}\frac{p}{n} \|n\boldsymbol{w} - r\boldsymbol{1}\|_\infty \left\| \frac{1}{n}\sum_{i=1}^n \boldsymbol{x}_i \boldsymbol{x}_i^T \right\|_{\max}.
\end{aligned}$$

Using this inequality and equation (10) with $\lambda = 1/[s_{\max}(1+2\sqrt{\gamma})^2]$,

$$\begin{aligned}
&\Pr\left(\left\| \left(\hat{\boldsymbol{\Sigma}}(\alpha) - \frac{\alpha}{1+\alpha}\boldsymbol{I} \right) - \frac{1}{1+\alpha}\frac{p}{n} r\hat{\boldsymbol{S}} \right\| > \epsilon \right) \\
&\quad \le \Pr\left(\frac{1}{1+\alpha}\frac{p}{n} \|n\boldsymbol{w} - r\boldsymbol{1}\|_\infty \|\hat{\boldsymbol{S}}\| > \epsilon \right) \\
&\quad \le \Pr\left(\|n\boldsymbol{w} - r\boldsymbol{1}\|_\infty > \epsilon \frac{n(1+\alpha)}{p s_{\max}(1+2\sqrt{\gamma})^2} \right) \\
&\qquad + \Pr\big(\|\hat{\boldsymbol{S}}\| > s_{\max}(1+2\sqrt{\gamma})^2 \big).
\end{aligned}$$

Since $\hat{\boldsymbol{S}} = \boldsymbol{S}_p^{1/2}(\frac{1}{n}\sum_i \boldsymbol{\xi}_i \boldsymbol{\xi}_i^T)\boldsymbol{S}_p^{1/2}$ with $\boldsymbol{\xi}_i \sim N(0, \boldsymbol{I})$, by Lemma 4 the second term is exponentially small in p. By equation (20), the first term is bounded by $C' p^2 e^{-cp\epsilon^2}$. Hence, equation (15) holds, which concludes the proof of Theorem 3.

6.3. *Concluding the proof of Theorem 2.* Similar to Theorems 1 and 3, to prove Theorem 2 it suffices to show that

$$\left\| p\left(\hat{\boldsymbol{\Sigma}}(\alpha) - \frac{\alpha}{1+\alpha}\boldsymbol{I} \right) \Big/ \operatorname{tr}\left(\hat{\boldsymbol{\Sigma}}(\alpha) - \frac{\alpha}{1+\alpha}\boldsymbol{I} \right) - \hat{\boldsymbol{S}} \right\|_{\max} = \mathcal{O}_P(\sqrt{\log p/n}). \tag{21}$$

Equation (15) combined with Proposition 1 imply that for $r > r_{\min} > 0$,

$$\left\| \frac{n}{p} \frac{1+\alpha}{r} \left(\hat{\boldsymbol{\Sigma}}(\alpha) - \frac{\alpha}{1+\alpha} \boldsymbol{I} \right) - \hat{\boldsymbol{S}} \right\|_{\max} = \mathcal{O}_P(\sqrt{\log p/n}). \tag{22}$$

Since $\operatorname{tr}(\hat{\boldsymbol{S}})$ is tightly concentrated around p, we may replace $\hat{\boldsymbol{S}}$ in equation (22) by $p\hat{\boldsymbol{S}}/\operatorname{tr}(\hat{\boldsymbol{S}})$. Equation (21) follows by the following lemma, proven in the supplemental article (Goes, Lerman and Nadler (2020)), combined with the fact that w.h.p. $\|\hat{\boldsymbol{S}}\|_{\max} \le 2\|\boldsymbol{S}_p\|_{\max} \le 2M$.

LEMMA 13. *Let $\boldsymbol{B} \in S_+^p$ with $\operatorname{tr}(\boldsymbol{B}) = p$ and $\|\boldsymbol{B}\|_{\max} \le b_{\max}$. Suppose that $\boldsymbol{A} \in S_+^p$ satisfies $\|\boldsymbol{A} - \boldsymbol{B}\|_{\max} < \epsilon \le 1/2$. Then*

$$\left\| \frac{p\boldsymbol{A}}{\operatorname{tr}(\boldsymbol{A})} - \boldsymbol{B} \right\|_{\max} \le 2(1 + b_{\max})\epsilon. \tag{23}$$

7. Numerical experiments. Focusing on the regularized TME, we present simulations that support our theoretical analysis. Section 7.1 compares the regularized TME, the sample covariance and their thresholded versions. Section 7.2 considers the sensitivity of the proposed estimator to α. Section 7.3 demonstrates a modified estimator to cope with outliers.

7.1. *Comparison of thresholded TME with covariance estimators.* We considered the following shape matrix, also used by Bickel and Levina (2008):

$$\boldsymbol{S}_p = (s_{ij}) = \left(0.7^{|i-j|}\right).$$

Note that excluding the diagonal all rows of this matrix have ℓ_1 norm bounded by $2/(1 - 0.7) - 2 = 14/3$. Hence, by the Gershgorin disk theorem, for any p, this matrix has a finite spectral norm, $\|\boldsymbol{S}_p\| \le s_{\max} = 1 + 14/3 = 17/3$. This is in accordance with our assumptions in Theorem 2.

We generated data from a Gaussian scale mixture, which is a particular case of equation (1). Here, u and $\boldsymbol{\xi}$ are independent, with $\boldsymbol{\xi} \sim N(\boldsymbol{0}, \boldsymbol{I})$. We considered three different choices for the random variables u_i: (i) $u_i = 1$, so $\{\boldsymbol{x}_i\}_{i=1}^n$ are i.i.d. $N(\boldsymbol{0}, \boldsymbol{S}_p)$; (ii) $u_i \sim \text{Laplace}(0, 1)$, a heavy-tailed distribution with finite moments; and (iii) $u_i \sim \text{Cauchy}(0, 1)$, so the distribution does not even have a well-defined mean or covariance.

We computed four estimators for the shape matrix: (i) SampCov: the sample covariance scaled to have trace p, $p\hat{\boldsymbol{S}}/\operatorname{tr}(\hat{\boldsymbol{S}})$; (ii) th-SampCov: the thresholded version of SampCov, $\tau_t(p\hat{\boldsymbol{S}}/\operatorname{tr}(\hat{\boldsymbol{S}}))$; (iii) RegTME: the regularized TME, normalized to have trace p,

$$\frac{p(\boldsymbol{\Sigma}(\alpha) - \frac{\alpha}{1+\alpha}\boldsymbol{I})}{\operatorname{tr}(\boldsymbol{\Sigma}(\alpha) - \frac{\alpha}{1+\alpha}\boldsymbol{I})};$$

and (iv) th-RegTME: the thresholded version of RegTME in equation (9). We choose $\alpha = 10$, and threshold at level $t = \sqrt{(\log p)/n}$. Our stopping rule for (5) is $\|p\hat{\boldsymbol{\Sigma}}_{k+1}/\operatorname{tr}(\hat{\boldsymbol{\Sigma}}_{k+1}) - p\hat{\boldsymbol{\Sigma}}_k/\operatorname{tr}(\hat{\boldsymbol{\Sigma}}_k)\|_F < 10^{-12}$, or $k = 1400$ iterations.

We measured the accuracy of an estimator $\hat{\boldsymbol{S}}_p$ by the logarithm of its averaged relative error (LRE). That is, for 100 different realizations, we independently generated n i.i.d. samples in $\mathbb{R}^p$, and each time estimated $(\hat{\boldsymbol{S}}_p)_i$, where $i = 1, \ldots, 100$. The LRE was then computed as follows:

$$\text{LRE} = \log\left(\frac{1}{100} \sum_{i=1}^{100} \frac{\|(\hat{\boldsymbol{S}}_p)_i - \boldsymbol{S}_p\|}{\|\boldsymbol{S}_p\|} \right).$$

We considered sample sizes $n \in [100, 1000]$ and the following three ratios $p/n \in \{0.5, 1, 2\}$. Figure 1 shows the LRE of the four estimators. As expected theoretically, for $u_i \equiv 1$ thresholding the sample covariance or the regularized TME yield similar errors. In contrast, for

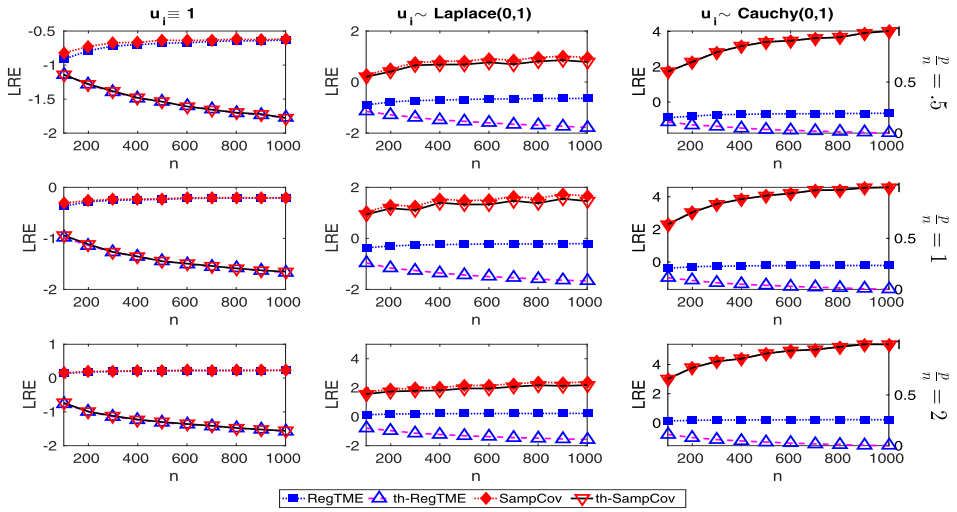

FIG. 1. *Comparison of the LRE of the four estimators with data i.i.d. from a Gaussian scale mixture. The rows correspond to* $p/n = 0.5, 1, 2$. *The columns correspond to* $u_i \equiv 1$, $u_i \overset{i.i.d.}{\sim}$ Laplace$(0, 1)$ *and* $u_i \overset{i.i.d.}{\sim}$ Cauchy$(0, 1)$.

heavy-tailed data the thresholded sample covariance performs poorly, whereas the thresholded regularized TME is still an accurate estimate of $\boldsymbol{S}_p$. Note that since the regularized TME is invariant to the scaling u_i, the resulting errors of the regularized TME and its thresholded version (the blue squares and triangles) are the same for all three distributions of u_i.

7.2. *Sensitivity of regularized TME to choice of* α. Next we study how the error and runtime of th-RegTME depend on the regularization parameter α. We consider the Gaussian model with covariance $\boldsymbol{S}_p$, and explore the behavior of th-RegTME for the following values of α: 0.2, 0.4, 0.6, 0.8, 1, 2, 3, $\ldots$, 20 and the following three cases: $(p, n) = (800, 400)$, $(p, n) = (800, 200)$ and $(p, n) = (400, 200)$. Even though the regularized TME does not exist if $\alpha < \max(0, p/n - 1)$, our algorithm, with a stopping criterion based on scaled matrices converged for all considered values of α. For a similar property upon scaling scatter matrices, see Chen, Wiesel and Hero (2011). The left panel of Figure 2 shows the LRE of th-RegTME as a function of α. The maximal LRE occurs at $p/n - 1$ and larger values of α yield slightly smaller errors, which are nearly identical for all large values of α. This is in accordance with Theorem 2, which states that asymptotically all large values of α yield the same error rate. The right panel of Figure 2 displays the logarithm of the runtime of th-RegTME as a function of α, showing a sharp increase in runtime as $p/n - 1$ approaches α.

Next we explore the behavior of th-RegTME for $p = 480$, $\alpha = 1, 2, 3, 4$ and $n = 60, 64, 68, \ldots, 300$. The left panel of Figure 3 shows the error of th-RegTME as a func-

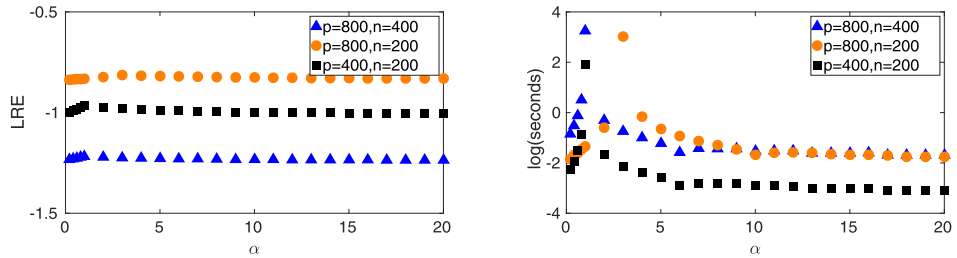

FIG. 2. *LRE and log-runtime of th-RegTME on elliptical data for different choices of* α *and three choices of* p *and* n.

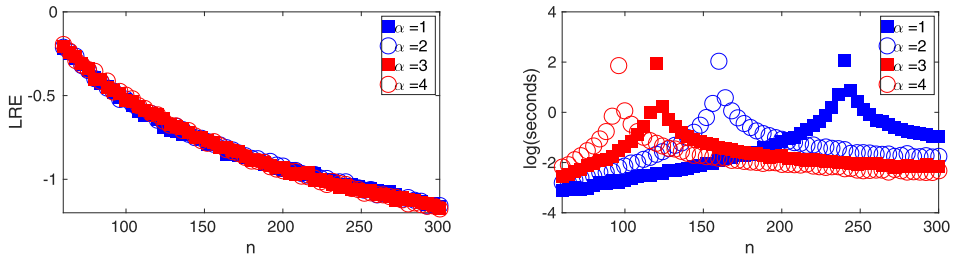

FIG. 3. *LRE and log-runtime of th-RegTME versus number of samples n, at $p = 480$ and $\alpha = 1, 2, 3, 4$.*

tion of n. Again, in accordance with theory, α has little effect on the accuracy. Of particular interest is the runtime, seen in the right panel of Figure 3. Here, we see a sharp increase in runtime as $p/n - 1$ approaches α. For $n \geq \frac{p}{\alpha+1}$, the runtime decreases as α increases.

These experiments indicate that one may generally prefer larger α, particularly for faster runtime. We propose to choose a value of α close to the bound in Lemma 1 for some $R \in (0, 1)$, which guarantees fast convergence.

7.3. *Regularized TME in the presence of outliers.* We conclude the numerical section with an illustrative example of the ability of the regularized TME to detect outliers, and upon their removal and thresholding, to provide a robust and accurate estimate of a sparse shape matrix. For a related rigorous study on the ability of Maronna's M-estimator to detect outliers, see Morales-Jimenez, Couillet and McKay (2015).

To this end, we consider the following ϵ-contamination mixture model: $(1 - \epsilon)n$ of the observed data, the inliers, follow an elliptical distribution with the same sparse shape matrix $\boldsymbol{S}_p$ as above. The remaining ϵn of the samples, the outliers, follow an elliptical distribution with shape matrix $\boldsymbol{U}(p\boldsymbol{D}/\operatorname{tr}(\boldsymbol{D}))\boldsymbol{U}'$, where $\boldsymbol{U}$ is a unitary matrix, uniformly distributed with Haar measure, and $\boldsymbol{D}$ is a diagonal matrix. In our first experiment, the diagonal entries d_{ii} are all i.i.d. uniformly distributed over $[1, 5]$, so the outliers are rather diffuse. In our second experiment $d_{11} = p, d_{22} = p/2$ and all other $d_{ii} = 1$, so the outliers are nearly on a 2-d randomly rotated subspace.

Given n samples from this ϵ-contamination model, and without knowledge of ϵ, the task is to accurately estimate the shape matrix $\boldsymbol{S}_p$. Since both the inliers and outliers have potentially heavy-tailed distributions, it might not be possible to detect the outliers by simple schemes, such as those based on the norm of a sample or the number of its neighbors in a given radius. However, recall that by our theoretical analysis, in the absence of outliers ($\epsilon = 0$), the corresponding weights w_i in the regularized TME are all approximately equal. For $\epsilon \ll 1$, with all samples normalized to have unit norm, we thus expect the inliers to still all have similar weights, and the outliers to have quite different weights, hopefully smaller though not necessarily so. With further details in Appendix A.5, our proposed procedure for robustness to outliers is to estimate the mean and standard deviation of the inliers' weights. Then exclude all samples whose weights are outside, say, the mean plus or minus two standard deviations, recompute the regularized TME on the remaining samples and threshold it.

Figure 4 illustrates the robustness of this procedure to outliers in two different settings. From left to right, for $\epsilon = 0.2$ and $\epsilon = 0.4$, it shows the weights of the n normalized samples $\boldsymbol{x}_i/\|\boldsymbol{x}_i\|$, sorted so the first ϵn of them are the outliers. The blue horizontal line is a robust estimate of the mean weight of the inliers, and the two red lines are this estimated mean plus and minus two standard deviations. The top row corresponds to the first outlier model with $d_{ii} \sim U[1, 5]$. The second row corresponds to our second outlier model with $D = \operatorname{diag}(p, p/2, 1, \ldots, 1)$. Note that this outlier shape matrix has a spectral norm $O(p)$, which does not satisfy our requirement that $\|\boldsymbol{D}\| \leq s_{\max}$. As indeed observed empirically, the

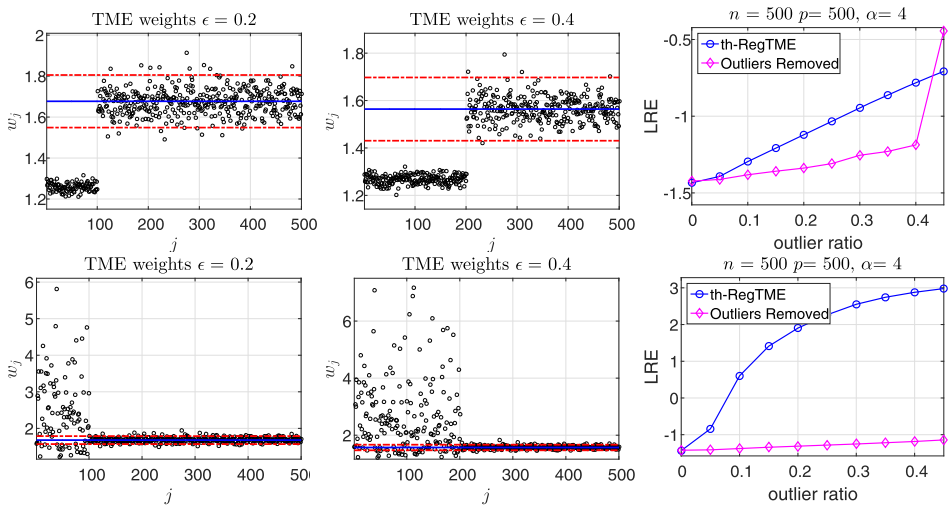

FIG. 4. *The TME weights for* $\epsilon = 0.2, 0.4$ *and the log relative error (LRE) of thresholding the regularized TME, before and after outlier removal, versus* ϵ. *Top row* $D_{ii} \sim U[1, 5]$. *Bottom row* $D = \mathrm{diag}(p, p/2, 1, \ldots, 1)$.

weights of the outliers do not so tightly concentrate around some value. Yet, our outlier exclusion procedure still succeeds to exclude most of these outliers. The error of the thresholded TME with outliers removed, compared to that of thresholding the original TME is shown in the right column of Figure 4.

This simple example illustrates the potential ability of TME to screen outliers in high-dimensional settings, at least for small contamination levels. A detailed study of this issue is an interesting topic for future research.

8. Summary and discussion. In this paper, we proposed simple estimators for the shape matrix of possibly heavy-tailed elliptical distributions, assuming the shape matrix is approximately sparse. We further analyzed their error, showing that under the spectral norm they are minimax rate optimal in a high-dimensional setting with $p/n \to \gamma$.

There are several directions for future research. One direction is to study whether our proposed approach, of thresholding a regularized TME provides accurate estimates of a sparse covariance matrix for other heavy-tailed distributions beyond the elliptical distribution. Another direction is to extend our results to the case $p = n^\beta$, with $\beta > 1$. Our current analysis assumed the regularization parameter α of TME is *fixed*, whereas if $p = n^\beta$ with $\beta > 1$, just to ensure its existence would require $\alpha \to \infty$. Handling this case thus requires extending our analysis to allow α to grow with n and p.

A question of practical interest is how to set the threshold parameter in a data-driven fashion. Bickel and Levina (2008), Section 3, proposed a cross validation procedure to set the threshold. Rigorously proving that this provides a good estimate in the case of (regularized) TME is an interesting topic for future research.

While our work focused on approximate sparsity of the shape matrix, robust inference under other common assumptions can also be studied. For example, one might assume that the first few leading eigenvectors of $\mathbf{\Sigma}$ are sparse, also known as sparse-PCA, or that $\mathbf{\Sigma}$ is the combination of a low rank and a sparse matrix. In particular, a robust sparse-PCA estimator may be constructed by applying a sparse-PCA procedure to Tyler's M-estimator.

Finally, another direction for future work is to develop a computationally efficient algorithm for sparse covariance estimation in the presence of a small fraction of arbitrary outliers. This setting was considered in Chen, Gao and Ren (2018), but without a computationally

tractable estimator. Our promising preliminary results in Section 7.3 suggest to study whether the regularized TME offers such robustness, and under which outlier models.

APPENDIX A: PROOFS

A.1. Complexity of calculating the regularized TME.

PROOF OF LEMMA 1. We arbitrarily fix a solution $\hat{\boldsymbol{\Sigma}}(\alpha)$ of (4). Since $\hat{\boldsymbol{\Sigma}}(\alpha)$ is invariant to scaling of the data, we assume that $\|\boldsymbol{x}_i\| = 1$, $1 \le i \le n$. We first analyze the quantity $e_1 = \|\hat{\boldsymbol{\Sigma}}(\alpha) - \hat{\boldsymbol{\Sigma}}_1(\alpha)\|$. To this end, let $\lambda_{\max} = \|\hat{\boldsymbol{\Sigma}}(\alpha)\|$. Taking the spectral norm in equation (4), together with the fact that $\frac{1}{\boldsymbol{x}^T \hat{\boldsymbol{\Sigma}}(\alpha)^{-1}\boldsymbol{x}} \le \lambda_{\max}$ for any vector $\boldsymbol{x}$ with $\|\boldsymbol{x}\| = 1$,

$$\lambda_{\max} \le \frac{1}{1+\alpha}\left\| \frac{p}{n}\sum_{i=1}^{n} \frac{\boldsymbol{x}_i \boldsymbol{x}_i^T}{\boldsymbol{x}_i^T \Sigma^{-1} \boldsymbol{x}_i} \right\| + \frac{\alpha}{1+\alpha} \le \frac{1}{1+\alpha}\lambda_{\max} C(\tilde{\boldsymbol{X}}) + \frac{\alpha}{1+\alpha}.$$

Equivalently, for $1 + \alpha > C(\tilde{\boldsymbol{X}})$,

$$\lambda_{\max} \le \frac{\alpha}{1+\alpha - C(\tilde{\boldsymbol{X}})}.$$

Combining this inequality with the fact that by equation (4) $\hat{\boldsymbol{\Sigma}}(\alpha) - \frac{\alpha}{1+\alpha}\boldsymbol{I} \in S_+^p$,

$$e_1 = \left\| \hat{\boldsymbol{\Sigma}}(\alpha) - \frac{\alpha}{1+\alpha}\boldsymbol{I} \right\| = \lambda_{\max} - \frac{\alpha}{1+\alpha} \le \frac{\alpha}{1+\alpha}\frac{C(\tilde{\boldsymbol{X}})}{1+\alpha - C(\tilde{\boldsymbol{X}})}. \tag{24}$$

Next we analyze the error e_k. We denote $\boldsymbol{E}_k = \hat{\boldsymbol{\Sigma}}(\alpha) - \hat{\boldsymbol{\Sigma}}_k(\alpha)$ and write

$$\hat{\boldsymbol{\Sigma}}_k(\alpha) = \hat{\boldsymbol{\Sigma}}(\alpha) - \boldsymbol{E}_k = \hat{\boldsymbol{\Sigma}}(\alpha)^{1/2}\big(\boldsymbol{I} - \hat{\boldsymbol{\Sigma}}(\alpha)^{-1/2}\boldsymbol{E}_k\hat{\boldsymbol{\Sigma}}(\alpha)^{-1/2}\big)\hat{\boldsymbol{\Sigma}}(\alpha)^{1/2}.$$

Since $\hat{\boldsymbol{\Sigma}}(\alpha)$ and $\hat{\boldsymbol{\Sigma}}_k(\alpha)$ are invertible, so is $\boldsymbol{I} - \hat{\boldsymbol{\Sigma}}(\alpha)^{-1/2}\boldsymbol{E}_k\hat{\boldsymbol{\Sigma}}(\alpha)^{-1/2}$. Let $\boldsymbol{B}_k = \boldsymbol{I} - (\boldsymbol{I} - \hat{\boldsymbol{\Sigma}}(\alpha)^{-1/2}\boldsymbol{E}_k\hat{\boldsymbol{\Sigma}}(\alpha)^{-1/2})^{-1}$ and $\boldsymbol{R}_k = \hat{\boldsymbol{\Sigma}}(\alpha)^{-1/2}\boldsymbol{B}_k\hat{\boldsymbol{\Sigma}}(\alpha)^{-1/2}$. Then

$$\hat{\boldsymbol{\Sigma}}_k(\alpha)^{-1} = \hat{\boldsymbol{\Sigma}}(\alpha)^{-1/2}(\boldsymbol{I} - \boldsymbol{B}_k)\hat{\boldsymbol{\Sigma}}(\alpha)^{-1/2} = \hat{\boldsymbol{\Sigma}}(\alpha)^{-1} - \boldsymbol{R}_k.$$

Subtracting equation (5) from equation (4) gives

$$\begin{aligned}\boldsymbol{E}_{k+1} &= \frac{1}{1+\alpha}\frac{p}{n}\sum_i \boldsymbol{x}_i\boldsymbol{x}_i^T\left(\frac{1}{\boldsymbol{x}_i^T\hat{\boldsymbol{\Sigma}}(\alpha)^{-1}\boldsymbol{x}_i} - \frac{1}{\boldsymbol{x}_i^T\hat{\boldsymbol{\Sigma}}(\alpha)^{-1}\boldsymbol{x}_i - \boldsymbol{x}_i^T\boldsymbol{R}_k\boldsymbol{x}_i}\right)\\ &= \frac{1}{1+\alpha}\frac{p}{n}\sum_i \frac{\boldsymbol{x}_i\boldsymbol{x}_i^T}{\boldsymbol{x}_i^T\hat{\boldsymbol{\Sigma}}(\alpha)^{-1}\boldsymbol{x}_i}\left(1 - \frac{1}{1-\delta_{ki}}\right),\end{aligned}$$

where $\delta_{ki} = \boldsymbol{x}_i^T\boldsymbol{R}_k\boldsymbol{x}_i / \boldsymbol{x}_i^T\hat{\boldsymbol{\Sigma}}(\alpha)^{-1}\boldsymbol{x}_i$.

Let $D_k = \max_{1\le i\le n}|\delta_{ki}/(1-\delta_{ki})|$. Since all terms $\boldsymbol{x}_i\boldsymbol{x}_i^T/\boldsymbol{x}_i^T\hat{\boldsymbol{\Sigma}}(\alpha)^{-1}\boldsymbol{x}_i$ are positive semidefinite, the above equation implies that

$$\begin{aligned}\|\boldsymbol{E}_{k+1}\| &\le D_k\left\|\frac{1}{1+\alpha}\frac{p}{n}\sum_i \frac{\boldsymbol{x}_i\boldsymbol{x}_i^T}{\boldsymbol{x}_i^T\hat{\boldsymbol{\Sigma}}(\alpha)^{-1}\boldsymbol{x}_i}\right\|\\ &= D_k\left\|\hat{\boldsymbol{\Sigma}}(\alpha) - \frac{\alpha}{1+\alpha}\boldsymbol{I}\right\| = D_k e_1.\end{aligned} \tag{25}$$

Equation (24) gives a bound on e_1. We now bound D_k. Since $\hat{\boldsymbol{\Sigma}}(\alpha) \ge \frac{\alpha}{1+\alpha}\boldsymbol{I}$,

$$\|\hat{\boldsymbol{\Sigma}}(\alpha)^{-1/2}\boldsymbol{E}_k\hat{\boldsymbol{\Sigma}}(\alpha)^{-1/2}\| \le \|\hat{\boldsymbol{\Sigma}}(\alpha)^{-1}\|e_k \le \frac{1+\alpha}{\alpha}e_k.$$

Assume this quantity is strictly smaller than one, then

$$\|\boldsymbol{B}_k\| = \|\boldsymbol{I} - (\boldsymbol{I} - \hat{\boldsymbol{\Sigma}}(\alpha)^{-1/2}\boldsymbol{E}_k\hat{\boldsymbol{\Sigma}}(\alpha)^{-1/2})^{-1}\| \le \frac{1+\alpha}{\alpha}\frac{e_k}{1-\frac{1+\alpha}{\alpha}e_k}. \tag{26}$$

Finally, given the relation between $\boldsymbol{R}_k$ and $\boldsymbol{B}_k$,

$$|\delta_{ki}| = \frac{|\boldsymbol{x}_i^T\boldsymbol{R}_k\boldsymbol{x}_i|}{\boldsymbol{x}_i^T\hat{\boldsymbol{\Sigma}}(\alpha)^{-1}\boldsymbol{x}_i} = \frac{|(\hat{\boldsymbol{\Sigma}}(\alpha)^{-1/2}\boldsymbol{x}_i)^T\boldsymbol{B}_k(\hat{\boldsymbol{\Sigma}}(\alpha)^{-1/2}\boldsymbol{x}_i)|}{\|\hat{\boldsymbol{\Sigma}}(\alpha)^{-1/2}\boldsymbol{x}_i\|^2} \le \|\boldsymbol{B}_k\|.$$

Thus, assuming that the bound on $\|\boldsymbol{B}_k\|$ in equation (26) is less than one,

$$D_k = \max_i \frac{|\delta_{ki}|}{1-\delta_{ki}} \le \frac{\|\boldsymbol{B}_k\|}{1-\|\boldsymbol{B}_k\|} \le \frac{1+\alpha}{\alpha}e_k \cdot \frac{1}{1-2\frac{1+\alpha}{\alpha}e_k}. \tag{27}$$

Inserting (27) and (24) into (25) yields that

$$\frac{e_{k+1}}{e_k} \le \frac{C(\tilde{\boldsymbol{X}})}{1+\alpha-C(\tilde{\boldsymbol{X}})}\frac{1}{1-2\frac{1+\alpha}{\alpha}e_k}. \tag{28}$$

For the proof to hold, the bound in (26) needs to be less than one, namely that $\frac{1+\alpha}{\alpha}e_k < 0.5$. For $0 < R < 1$ and $1+\alpha > (3+R^{-1})C(\tilde{\boldsymbol{X}})$, equation (24) implies that $\frac{1+\alpha}{\alpha}e_1 < \frac{1}{2+R^{-1}}$ and combining this with equation (28) results in the estimate $e_2/e_1 < R$. Since $R < 1$, induction implies that for $k > 1$, $\frac{1+\alpha}{\alpha}e_k < \frac{1}{2+R^{-1}} < 0.5$, as required, and so equation (6) holds. Since this convergence holds with any solution of (4), this solution thus has to be unique. □

PROOF OF LEMMA 2. Since the regularized TME is invariant to scaling, we may assume that all $u_i \sim \chi_p^2$, and express $\boldsymbol{x}_i = \boldsymbol{S}_p^{\frac{1}{2}}\boldsymbol{\xi}_i$, where $\boldsymbol{\xi}_i \sim N(\boldsymbol{0}, \boldsymbol{I})$. Let $\boldsymbol{U}\boldsymbol{D}\boldsymbol{U}^T$ be the eigendecomposition of $\boldsymbol{S}_p$. Redefining $\boldsymbol{\xi} = \boldsymbol{U}\boldsymbol{\xi}$, then $\|\boldsymbol{x}_i\|^2 = \boldsymbol{\xi}_i^T\boldsymbol{D}\boldsymbol{\xi}_i$ and

$$C(\tilde{\boldsymbol{X}}) = \left\| \boldsymbol{S}_p^{\frac{1}{2}}\left(\frac{1}{n}\sum_{i=1}^n \frac{\boldsymbol{\xi}_i\boldsymbol{\xi}_i^T}{\frac{1}{p}\boldsymbol{\xi}_i^T\boldsymbol{D}\boldsymbol{\xi}_i}\right)\boldsymbol{S}_p^{\frac{1}{2}} \right\|.$$

Combining Lemma 5 with a union bound yields

$$\Pr\left(\max_i \left|\frac{1}{p}\boldsymbol{\xi}_i^T\boldsymbol{D}\boldsymbol{\xi}_i - \frac{1}{p}\operatorname{tr}(\boldsymbol{D})\right| > \epsilon\right) < 2n\exp\left(-c_1\min\left\{\frac{c_2^2p^2\epsilon^2}{\|\boldsymbol{D}\|_F^2}, \frac{c_2p\epsilon}{\|\boldsymbol{D}\|}\right\}\right).$$

Since $\|\boldsymbol{D}\| = \|\boldsymbol{S}_p\|$ and $\|\boldsymbol{D}\|_F^2 \le p\|\boldsymbol{S}_p\|^2$, for any fixed ϵ the above probability is exponentially small in p. Taking say $\epsilon = 1/2$ and recalling that $\operatorname{tr}(\boldsymbol{D}) = p$, gives that with high probability,

$$C(\tilde{\boldsymbol{X}}) \le 2\|\boldsymbol{S}_p\| \cdot \left\|\frac{1}{n}\sum_{i=1}^n \boldsymbol{\xi}_i\boldsymbol{\xi}_i^T\right\|.$$

Equation (7) follows since by Lemma 4, w.h.p. $\|\frac{1}{n}\sum_{i=1}^n \boldsymbol{\xi}_i\boldsymbol{\xi}_i^T\| \le (1+2\sqrt{p/n})^2$. □

A.2. Proof of Lemma 8. To prove the lemma, we shall use the following auxiliary result, whose proof appears in the supplemental article (Goes, Lerman and Nadler (2020)).

LEMMA 14. *Assume the setting of Lemma 8. There exist constants C, c and $c' < 1$ depending on γ such that $\forall\epsilon \in (0, c')$ and n sufficiently large,*

$$\Pr\left(\left|\frac{p}{T_w} - 1\right| > \epsilon\right) \le Cne^{-cn\epsilon^2}. \tag{29}$$

PROOF OF LEMMA 8. By definition,

$$\|\hat{\boldsymbol{\Sigma}} - \hat{\boldsymbol{S}}\|_{\max} = \left\| \sum_{i=1}^{n} \left(\frac{p w_i}{T_w} - \frac{1}{n} \right) \boldsymbol{x}_i \boldsymbol{x}_i^T \right\|_{\max}$$

$$\leq \left\| \frac{np\boldsymbol{w}}{T_w} - \mathbf{1} \right\|_{\infty} \cdot \left\| \frac{1}{n} \sum_{i=1}^{n} \boldsymbol{x}_i \boldsymbol{x}_i^T \right\|_{\max}. \tag{30}$$

Since $\boldsymbol{x}_i \sim N(0, \boldsymbol{S}_p)$ with $\boldsymbol{S}_p \in \mathcal{U}(q, s_p, M)$, then w.h.p., $\|\frac{1}{n}\sum \boldsymbol{x}_i \boldsymbol{x}_i^T\|_{\max} \leq 2M$. As for the first term on the RHS of equation (30), by the triangle inequality,

$$\left\| \frac{np\boldsymbol{w}}{T_w} - \mathbf{1} \right\|_{\infty} = \left\| \frac{np\boldsymbol{w}}{T_w} - n\boldsymbol{w} + n\boldsymbol{w} - \mathbf{1} \right\|_{\infty} \leq \|n\boldsymbol{w}\|_{\infty} \left| \frac{p}{T_w} - 1 \right| + \|n\boldsymbol{w} - \mathbf{1}\|_{\infty}.$$

Hence,

$$\Pr\left(\left\| \frac{np\boldsymbol{w}}{T_w} - \mathbf{1} \right\|_{\infty} > \epsilon \right) \leq \Pr\left(\|n\boldsymbol{w}\|_{\infty} \left| \frac{p}{T_w} - 1 \right| > \epsilon/2 \right) + \Pr(\|n\boldsymbol{w} - \mathbf{1}\|_{\infty} > \epsilon/2).$$

Lemma 7 provides an exponential bound on the second term. For the first term, applying equation (10) with $\lambda = 2$ gives

$$\Pr\left(\|n\boldsymbol{w}\|_{\infty} \left| \frac{p}{T_w} - 1 \right| > \epsilon/2 \right) \leq \Pr(\|n\boldsymbol{w}\|_{\infty} > 2) + \Pr\left(\left| \frac{p}{T_w} - 1 \right| > \epsilon/4 \right)$$

$$\leq \Pr(\|n\boldsymbol{w} - \mathbf{1}\|_{\infty} > 1) + \Pr\left(\left| \frac{p}{T_w} - 1 \right| > \epsilon/4 \right).$$

By Lemmas 7 and 14, these two probabilities are exponentially small. □

A.3. Proof of Proposition 1. To prove the existence of a unique $r^* = r^*(p, n, \alpha, \boldsymbol{S}_p)$ which satisfies equation (14), we first show that $\mathbb{E}[Q(r)]$ is strictly monotone increasing in r and then use the intermediate value theorem.

To simplify notation, let $\boldsymbol{T} = \frac{1}{n}\sum_{j=1}^{n-1} \boldsymbol{\xi}_j \boldsymbol{\xi}_j^T$ and $\beta = \beta(r) = \frac{n\alpha}{pr}$. Then

$$\mathbb{E}[Q(r)] = \mathbb{E}_{\boldsymbol{\xi}_i}[\mathbb{E}_{\boldsymbol{y}}[Q(r)]] = \mathbb{E}\left[\frac{1}{p} \operatorname{tr}((\boldsymbol{T} + \beta \boldsymbol{S}_p^{-1})^{-1}) \right], \tag{31}$$

where the expectation is now only over the random variables $\boldsymbol{\xi}_i$.

First, we show that for any $\boldsymbol{S}_p \in \mathcal{S}_p^{++}$, $\mathbb{E}[Q(r)]$ is strictly monotone increasing in r. Indeed, differentiating with respect to r and using the identity $\operatorname{tr}(\boldsymbol{AB}) = \operatorname{tr}(\boldsymbol{BA})$

$$\frac{d}{dr}\mathbb{E}[Q(r)] = \frac{n\alpha}{pr^2}\mathbb{E}\left[\frac{1}{p} \operatorname{tr}((\boldsymbol{T} + \beta \boldsymbol{S}_p^{-1})^{-2} \boldsymbol{S}_p^{-1}) \right].$$

Using Bhatia (1997), Problem III.6.14 and Jensen's inequality,

$$\frac{d}{dr}\mathbb{E}[Q(r)] \geq \frac{n\alpha}{pr^2} \lambda_{\min}(\boldsymbol{S}_p^{-1}) \mathbb{E}\left[\frac{1}{p} \operatorname{tr}((\boldsymbol{T} + \beta \boldsymbol{S}_p^{-1})^{-2}) \right]$$

$$\geq \frac{n\alpha}{pr^2} \frac{1}{s_{\max}} \mathbb{E}[1/\lambda_1(\boldsymbol{T} + \beta \boldsymbol{S}_p^{-1})^2]$$

$$\geq \frac{n\alpha}{pr^2} \frac{1}{s_{\max}} \frac{1}{\mathbb{E}[\lambda_1(\boldsymbol{T} + \beta \boldsymbol{S}_p^{-1})]^2}.$$

Clearly, $\lambda_1(\boldsymbol{T} + \beta \boldsymbol{S}_p^{-1}) \leq \lambda_1(\boldsymbol{T}) + \beta/\lambda_{\min}(\boldsymbol{S}_p)$. Furthermore, upon averaging over the random variables $\boldsymbol{\xi}_i$, by Lemma 4, $\mathbb{E}[\lambda_1(\boldsymbol{T})] \leq (1 + \sqrt{p/n})^2$. Therefore, for any fixed $\boldsymbol{S}_p$, the

derivative of $\mathbb{E}[Q(r)]$ is strictly positive for any $r > 0$. Hence if there exists a solution to equation (14), then it must be unique.

Next we show that this solution must satisfy $r \geq r_{\min}$. By definition,

$$\begin{aligned}\mathbb{E}\big[Q(r)\big] &= \frac{1}{p}\sum_{j=1}^{n}\frac{1}{\lambda_j(\boldsymbol{T}+\beta\boldsymbol{S}_p^{-1})}\\ &\leq \frac{1}{p}\sum_{j=1}^{n}\frac{1}{\lambda_j(\beta\boldsymbol{S}_p^{-1})} = \frac{1}{\beta}\frac{1}{p}\sum_{j=1}^{p}\lambda_j(\boldsymbol{S}_p) = \frac{n\alpha}{pr}.\end{aligned} \tag{32}$$

Combining (32) with (14) implies that

$$r^*(p,n,\alpha,\boldsymbol{S}_p) \geq \frac{n}{p}\frac{\alpha}{1+\alpha-p/n} = r_{\min}.$$

Finally, we bound r from above. To this end, using Jensen's inequality

$$\begin{aligned}\mathbb{E}\big[Q(r)\big] &= \frac{1}{p}\mathbb{E}\bigg[\sum_j\frac{1}{\lambda_j(\boldsymbol{T}+\beta\boldsymbol{S}_p^{-1})}\bigg] \geq \frac{1}{p}\mathbb{E}\bigg[\sum_j\frac{1}{\beta/\lambda_j(\boldsymbol{S}_p)+\|\boldsymbol{T}\|}\bigg]\\ &\geq \mathbb{E}\bigg[\frac{1}{\beta+s_{\max}\|\boldsymbol{T}\|}\bigg] \geq \frac{1}{\beta+s_{\max}\mathbb{E}[\|\boldsymbol{T}\|]}.\end{aligned}$$

By Lemma 4, $\mathbb{E}[\|\boldsymbol{T}\|] \leq (1+\sqrt{p/n})^2$. Hence, for $\alpha > p/n - 1 + s_{\max}(1+\sqrt{p/n})^2$, the solution to equation (14) satisfies

$$r^* \leq r_{\max} = \frac{n}{p}\frac{\alpha}{1+\alpha-p/n-s_{\max}(1+\sqrt{\gamma})^2}.$$

A.4. Proof of Lemma 9. Let $\hat{\boldsymbol{S}} = \frac{1}{n}\sum_{k=1}^{n}\boldsymbol{x}_k\boldsymbol{x}_k^T$ and $\hat{\boldsymbol{T}} = \frac{1}{n}\sum_{k=1}^{n}\boldsymbol{\xi}_k\boldsymbol{\xi}_k^T$, where $\boldsymbol{x}_i = \boldsymbol{S}_p^{\frac{1}{2}}\boldsymbol{\xi}_i$ and $\boldsymbol{\xi}_i \overset{i.i.d.}{\sim} N(\boldsymbol{0},\boldsymbol{I})$. Then equation (18) may be written as

$$\begin{aligned}\frac{1}{r}g(\boldsymbol{u})_i &= 1-\frac{1/(1+\alpha)}{\frac{1}{p}\boldsymbol{x}_i^T(\hat{\boldsymbol{S}}+\beta\boldsymbol{I})^{-1}\boldsymbol{x}_i} = 1-\frac{1/(1+\alpha)}{\frac{1}{p}\boldsymbol{\xi}_i^T(\boldsymbol{S}_p^{-\frac{1}{2}}\hat{\boldsymbol{S}}\boldsymbol{S}_p^{-\frac{1}{2}}+\beta\boldsymbol{S}_p^{-1})^{-1}\boldsymbol{\xi}_i}\\ &= 1-\frac{1}{1+\alpha}\frac{1}{\frac{1}{p}\boldsymbol{\xi}_i^T\boldsymbol{E}\boldsymbol{\xi}_i},\end{aligned} \tag{33}$$

where $\boldsymbol{E} = (\hat{\boldsymbol{T}}+\beta\boldsymbol{S}_p^{-1})^{-1}$ and $\beta = \alpha\frac{n}{p}\frac{1}{r}$. The quadratic form $\frac{1}{p}\boldsymbol{\xi}_i^T\boldsymbol{E}\boldsymbol{\xi}_i$ is difficult to analyze directly because $\boldsymbol{E}$ depends on $\boldsymbol{\xi}_i$. To disentangle this dependency, let $\hat{\boldsymbol{T}}_{-i} = \frac{1}{n}\sum_{k\neq i}\boldsymbol{\xi}_k\boldsymbol{\xi}_k^T$, and $\boldsymbol{E}_{-i} = (\hat{\boldsymbol{T}}_{-i}+\beta\boldsymbol{S}_p^{-1})^{-1}$. As $\boldsymbol{E}^{-1}$ and $\boldsymbol{E}_{-i}^{-1}$ differ by a rank-one matrix $\frac{1}{n}\boldsymbol{\xi}_i\boldsymbol{\xi}_i^T$, by the Sherman–Morrison formula,

$$\boldsymbol{E} = \boldsymbol{E}_{-i} - \frac{1}{n}\frac{\boldsymbol{E}_{-i}\boldsymbol{\xi}_i\boldsymbol{\xi}_i^T\boldsymbol{E}_{-i}}{1+\frac{1}{n}\boldsymbol{\xi}_i^T\boldsymbol{E}_{-i}\boldsymbol{\xi}_i}.$$

Therefore, denoting by Q_i the quadratic form

$$Q_i(r) \equiv Q_i = \frac{1}{p}\boldsymbol{\xi}_i^T\boldsymbol{E}_{-i}\boldsymbol{\xi}_i, \tag{34}$$

it follows that

$$\frac{1}{p}\boldsymbol{\xi}_i^T\boldsymbol{E}\boldsymbol{\xi}_i = Q_i - \frac{\frac{p}{n}Q_i^2}{1+\frac{p}{n}Q_i} = \frac{Q_i}{1+\frac{p}{n}Q_i}. \tag{35}$$

Plugging this expression into equation (33) gives

$$\frac{1}{r}g(\boldsymbol{u})_i = \frac{Q_i(1+\alpha-\frac{p}{n})-1}{(1+\alpha)Q_i}. \tag{36}$$

Next, to establish a concentration bound for $g(\boldsymbol{u})_i/r$, we study the concentration of Q_i. Since $\boldsymbol{\xi}_i \sim N(\mathbf{0}, \boldsymbol{I})$ and is independent of $\boldsymbol{E}_{-i}$,

$$\mathbb{E}Q_i = \mathbb{E}\,\mathrm{tr}(\boldsymbol{E}_{-i})/p.$$

We first show that Q_i concentrates tightly around $\mathrm{tr}(\boldsymbol{E}_{-i})/p$ in view of concentration of quadratic forms. We then show that $\mathrm{tr}(\boldsymbol{E}_{-i})$ concentrates tightly around its mean using results about the concentration of certain functions of the eigenvalues of random matrices.

Applying Lemma 5 with $\boldsymbol{\xi} = \boldsymbol{\xi}_i$ and viewing the matrix $\boldsymbol{E}_{-i}$ as fixed,

$$\Pr\left(\left|Q_i - \frac{1}{p}\mathrm{tr}(\boldsymbol{E}_{-i})\right| > \epsilon\right) \le 2\exp\left(-c_1 \min\left\{\frac{c_2^2 p^2\epsilon^2}{\|\boldsymbol{E}_{-i}\|_F^2}, \frac{c_2 p\epsilon}{\|\boldsymbol{E}_{-i}\|}\right\}\right),$$

where the above probability is only w.r.t. $\boldsymbol{\xi}_i$. Next, given that $\boldsymbol{E}_{-i} = (\boldsymbol{T}_{-i} + \beta\boldsymbol{S}_p^{-1})^{-1}$, then $\|\boldsymbol{E}_{-i}\| \le \frac{s_{\max}}{\beta}$ and $\|\boldsymbol{E}_{-i}\|_F^2 \le p s_{\max}^2/\beta^2$. Thus,

$$\Pr\left(\left|Q_i - \frac{1}{p}\mathrm{tr}(\boldsymbol{E}_{-i})\right| > \epsilon\right) \le C\exp(-cp\epsilon^2), \tag{37}$$

where now the probability is over all of the $\boldsymbol{\xi}_k$'s.

It remains to obtain a concentration inequality for $\mathrm{tr}(\boldsymbol{E}_{-i})/p$. To this end, consider the following $p \times (n-1+p)$ matrix:

$$\boldsymbol{Y} = \begin{pmatrix}\boldsymbol{\xi}_1 & \cdots & \boldsymbol{\xi}_{i-1} & \boldsymbol{\xi}_{i+1} & \cdots & \boldsymbol{\xi}_n & \sqrt{n\beta}\boldsymbol{S}_p^{-1/2}\end{pmatrix}.$$

By definition, all entries of $\boldsymbol{Y}$ are independent, the first $p \times (n-1)$ are standard Gaussian random variables and the rest deterministic. Then, by Guionnet and Zeitouni (2000), Corollary 1.8b,[2] for any function $h : \mathbb{R} \to \mathbb{R}$ such that $h(x^2)$ is Lipschitz with constant L, and for any $\delta > 0$,

$$\Pr\left(\frac{1}{K}\left|\mathrm{tr}\,h\left(\frac{\boldsymbol{Y}\boldsymbol{Y}^T}{K}\right) - \mathbb{E}\,\mathrm{tr}\left(h\left(\frac{\boldsymbol{Y}\boldsymbol{Y}^T}{K}\right)\right)\right| > \delta\right) \le 2\exp\left(-\frac{\delta^2 K^2}{2L^2}\right), \tag{38}$$

where $K = 2p + n - 1$ and for a symmetric matrix $\boldsymbol{A}$ with eigenvalues λ_j, $\mathrm{tr}\,h(\boldsymbol{A}) = \sum_j h(\lambda_j)$.

Since $\boldsymbol{Y}\boldsymbol{Y}^T = n(\hat{\boldsymbol{T}}_{-i} + \beta\boldsymbol{S}_p^{-1}) = n\boldsymbol{E}_{-i}^{-1}$, consider the function

$$h(x) = \frac{n}{p}\cdot\frac{1}{x}$$

for which $\frac{1}{2p+n-1}\mathrm{tr}\,h(\boldsymbol{Y}\boldsymbol{Y}^T/(2p+n-1)) = \mathrm{tr}(\boldsymbol{E}_{-i})/p$. Next note that for sufficiently large n and sufficiently small ϵ,

$$\lambda_{\min}\left(\frac{\boldsymbol{Y}\boldsymbol{Y}^T}{2p+n-1}\right) = \frac{n}{2p+n-1}\lambda_{\min}(\hat{\boldsymbol{T}}_{-i} + \beta\boldsymbol{S}_p^{-1}) \ge \frac{1}{2\gamma+1+\epsilon}\frac{\beta}{s_{\max}} = x_0.$$

[2]There is a typo in the original paper. In the notation of their Corollary 1.8, $\boldsymbol{Z}$ should be replaced with $\boldsymbol{Z}/(M+N)$.

We thus apply the function h only in the interval $x \geq x_0$. The Lipschitz constant of $h(x^2)$ for n sufficiently large is bounded by

$$L \leq \left| \frac{d}{dx} h(x^2) \Big|_{x=x_0} \right| \leq 16 \frac{(\gamma + 0.5 + \epsilon)^3}{\gamma - \epsilon} \left(\frac{s_{\max}}{\beta} \right)^3 \leq 16 \frac{(\gamma + 1)^3}{\gamma} \left(\frac{s_{\max}}{\beta} \right)^3.$$

Hence, applying (38), there exists a positive constant c that depends on γ, α, r and $s_{\max}$ such that

$$\Pr\left(\frac{1}{p} |\operatorname{tr}(\boldsymbol{E}_{-i}) - \mathbb{E}\operatorname{tr}(\boldsymbol{E}_{-i})| > \delta \right) \leq 2 \exp(-cp^2\delta^2). \tag{39}$$

Next, by the triangle inequality,

$$\begin{aligned} \Pr\left(\left| Q_i - \frac{\mathbb{E}\operatorname{tr}(\boldsymbol{E}_{-i})}{p} \right| > \epsilon \right) \leq{}& \Pr\left(\left| Q_i - \frac{\operatorname{tr}(\boldsymbol{E}_{-i})}{p} \right| > \frac{\epsilon}{2} \right) \\ &+ \Pr\left(\left| \frac{\operatorname{tr}(\boldsymbol{E}_{-i})}{p} - \frac{\mathbb{E}\operatorname{tr}(\boldsymbol{E}_{-i})}{p} \right| > \frac{\epsilon}{2} \right). \end{aligned}$$

Combining the above equation with equations (37) and (39), implies that at the value of r specified in Proposition 1, for which $\mathbb{E}[\operatorname{tr}(\boldsymbol{E}_{-i})/p] = \frac{1}{1+\alpha-\frac{p}{n}}$,

$$\Pr\left(\left| Q_i - \frac{1}{1 + \alpha - \frac{p}{n}} \right| > \epsilon \right) < Ce^{-cp\epsilon^2}. \tag{40}$$

We are finally ready to establish a concentration result for $\frac{1}{r} g(\boldsymbol{u})$. Combining equation (36) and a union bound over all p coordinates of g,

$$\begin{aligned} \Pr\left(\left\| \frac{1}{r} g(\boldsymbol{u}) \right\|_\infty > \epsilon \right) &\leq p \Pr\left(\left| \frac{1}{r} g(\boldsymbol{u})_i \right| > \epsilon \right) \\ &\leq p \Pr\left(\left| \frac{Q_i(1 + \alpha - p/n) - 1}{(1 + \alpha) Q_i} \right| > \epsilon \right). \end{aligned}$$

Applying equation (10) with $\lambda = 1$ to the equation above gives

$$\Pr\left(\left\| \frac{1}{r} g(\boldsymbol{u}) \right\|_\infty > \epsilon \right) < p \Pr\left(\left| Q_i \left(1 + \alpha - \frac{p}{n} \right) - 1 \right| > \epsilon \right) + p \Pr((1 + \alpha) Q_i < 1).$$

By equation (40), the first term on the RHS is exponentially small in p. As for the second term, since $(1 + \alpha)^{-1} < (1 + \alpha - p/n)^{-1}$, then again by equation (40), $\Pr(Q_i < 1/(1 + \alpha))$ is also exponentially small in p. The lemma thus follows from the boundedness of r from above, as established in Proposition 1.

A.5. TME with outliers. Consider an ϵ-contamination model, where $(1 - \epsilon)n$ samples follow an elliptical distribution with shape matrix $\boldsymbol{S}_{\text{in}}$, and the remaining ϵn follow an elliptical distribution with shape matrix $\boldsymbol{S}_{\text{out}}$. We conjecture that under suitable assumptions, the inlier and outlier weights of the TME concentrate around two values, w_{in} and w_{out}, respectively.

For our procedure to select the inliers, we further assume that the inlier weights are approximately Gaussian distributed around w_{in} with an unknown standard deviation σ_{in}. To estimate w_{in} and σ_{in} we compute a nonparametric density estimate $\hat{f}(w)$ of all n weights (using MATLAB's `ksdensity` procedure). Then $w_{\text{in}} = \arg\max_w \hat{f}(w)$ is the weight with highest estimated density. Next, for some r we find the largest interval $[w_L, w_R]$ around w_{in} so that for $w \in [w_L, w_R]$ we have $\hat{f}(w) \geq r \max \hat{f}(w) = r \hat{f}(w_{\text{in}})$. Then, given our assumption that the weights are Gaussian distributed, $\sigma_{\text{in}} = \frac{1}{2}(w_R - w_L)/\sqrt{-2\log(r)}$. In our simulations, we used $r = 0.7$, which worked well across all different contamination levels.

Clearly, one may obtain improved estimates of these quantities, as well as the unknown ϵ, by fitting a mixture of two Gaussians to the weights. However, for our illustrative example, we opted for the above simpler procedure.

Acknowledgments. We thank the three anonymous referees for multiple suggestions that greatly improved the manuscript. We also thank Teng Zhang and Ofer Zeitouni for useful discussions, and Tony Cai, Harrison Zhou, Elizaveta Levina and Peter Bickel for correspondence regarding their papers. We also thank the IMA and the schools of the authors for supporting collaborative visits. B.N. is the incumbent William Petschek professorial chair of mathematics.

This work was supported by NSF Grants DMS-14-18386, DMS-18-21266 and GRFP-00039202, UMN Doctoral Dissertation Fellowship and the Feinberg Foundation Visiting Faculty Program Fellowship of the Weizmann Institute of Science.

SUPPLEMENTARY MATERIAL

Supplement to "Robust sparse covariance estimation by thresholding Tyler's M-estimator" (DOI: 10.1214/18-AOS1793SUPP; .pdf). This supplement contains additional theoretical results and proofs omitted from main article.

REFERENCES

ABRAMOVICH, Y. I. and SPENCER, N. K. (2007). Diagonally loaded normalised sample matrix inversion (LNSMI) for outlier-resistant adaptive filtering. In *Proceedings of the IEEE* 32*nd Intl. Conf. on Acoustics, Speech, and Signal Proc.* (*ICASSP*) 1105–1108.

ANDERSON, T. W. (2003). *An Introduction to Multivariate Statistical Analysis*, 3rd ed. *Wiley Series in Probability and Statistics*. Wiley, Hoboken, NJ. MR1990662

AVELLA-MEDINA, M., BATTEY, H. S., FAN, J. and LI, Q. (2018). Robust estimation of high-dimensional covariance and precision matrices. *Biometrika* **105** 271–284. MR3804402 https://doi.org/10.1093/biomet/asy011

BALAKRISHNAN, S., DU, S. S., LI, J. and SINGH, A. (2017). Computationally efficient robust sparse estimation in high dimensions. In *Conference on Learning Theory* 169–212.

BHATIA, R. (1997). *Matrix Analysis*. *Graduate Texts in Mathematics* **169**. Springer, New York. MR1477662 https://doi.org/10.1007/978-1-4612-0653-8

BICKEL, P. J. and LEVINA, E. (2008). Covariance regularization by thresholding. *Ann. Statist.* **36** 2577–2604. MR2485008 https://doi.org/10.1214/08-AOS600

CAI, T. and LIU, W. (2011). Adaptive thresholding for sparse covariance matrix estimation. *J. Amer. Statist. Assoc.* **106** 672–684. MR2847949 https://doi.org/10.1198/jasa.2011.tm10560

CAI, T. T., REN, Z. and ZHOU, H. H. (2016). Estimating structured high-dimensional covariance and precision matrices: Optimal rates and adaptive estimation. *Electron. J. Stat.* **10** 1–59. MR3466172 https://doi.org/10.1214/15-EJS1081

CAI, T. T. and ZHOU, H. H. (2012a). Optimal rates of convergence for sparse covariance matrix estimation. *Ann. Statist.* **40** 2389–2420. MR3097607 https://doi.org/10.1214/12-AOS998

CAI, T. T. and ZHOU, H. H. (2012b). Minimax estimation of large covariance matrices under ℓ_1-norm. *Statist. Sinica* **22** 1319–1349. MR3027084

CAMBANIS, S., HUANG, S. and SIMONS, G. (1981). On the theory of elliptically contoured distributions. *J. Multivariate Anal.* **11** 368–385. MR0629795 https://doi.org/10.1016/0047-259X(81)90082-8

CHEN, M., GAO, C. and REN, Z. (2018). Robust covariance and scatter matrix estimation under Huber's contamination model. *Ann. Statist.* **46** 1932–1960. MR3845006 https://doi.org/10.1214/17-AOS1607

CHEN, Y., WIESEL, A. and HERO, A. O. III (2011). Robust shrinkage estimation of high-dimensional covariance matrices. *IEEE Trans. Signal Process.* **59** 4097–4107. MR2865971 https://doi.org/10.1109/TSP.2011.2138698

COUILLET, R., KAMMOUN, A. and PASCAL, F. (2016). Second order statistics of robust estimators of scatter. Application to GLRT detection for elliptical signals. *J. Multivariate Anal.* **143** 249–274. MR3431431 https://doi.org/10.1016/j.jmva.2015.08.021

COUILLET, R. and MCKAY, M. (2014). Large dimensional analysis and optimization of robust shrinkage covariance matrix estimators. *J. Multivariate Anal.* **131** 99–120. MR3252638 https://doi.org/10.1016/j.jmva.2014.06.018

COUILLET, R., PASCAL, F. and SILVERSTEIN, J. W. (2014). Robust estimates of covariance matrices in the large dimensional regime. *IEEE Trans. Inform. Theory* **60** 7269–7278. MR3273053 https://doi.org/10.1109/TIT.2014.2354045

COUILLET, R., PASCAL, F. and SILVERSTEIN, J. W. (2015). The random matrix regime of Maronna's M-estimator with elliptically distributed samples. *J. Multivariate Anal.* **139** 56–78. MR3349480 https://doi.org/10.1016/j.jmva.2015.02.020

DAVIDSON, K. R. and SZAREK, S. J. (2001). Local operator theory, random matrices and Banach spaces. In *Handbook of the Geometry of Banach Spaces, Vol. I* 317–366. North-Holland, Amsterdam. MR1863696 https://doi.org/10.1016/S1874-5849(01)80010-3

DÜMBGEN, L. (1998). On Tyler's M-functional of scatter in high dimension. *Ann. Inst. Statist. Math.* **50** 471–491. MR1664575 https://doi.org/10.1023/A:1003573311481

DÜMBGEN, L., NORDHAUSEN, K. and SCHUHMACHER, H. (2016). New algorithms for M-estimation of multivariate scatter and location. *J. Multivariate Anal.* **144** 200–217. MR3434950 https://doi.org/10.1016/j.jmva.2015.11.009

DÜMBGEN, L., PAULY, M. and SCHWEIZER, T. (2015). M-functionals of multivariate scatter. *Stat. Surv.* **9** 32–105. MR3324616 https://doi.org/10.1214/15-SS109

EL KAROUI, N. (2008). Operator norm consistent estimation of large-dimensional sparse covariance matrices. *Ann. Statist.* **36** 2717–2756. MR2485011 https://doi.org/10.1214/07-AOS559

FALK, M. (2002). The sample covariance is not efficient for elliptical distributions. *J. Multivariate Anal.* **80** 358–377. MR1889781 https://doi.org/10.1006/jmva.2000.1983

FANG, K. T., KOTZ, S. and NG, K. W. (1990). *Symmetric Multivariate and Related Distributions. Monographs on Statistics and Applied Probability* **36**. CRC Press, London. MR1071174 https://doi.org/10.1007/978-1-4899-2937-2

FRAHM, G. (2004). Generalized elliptical distributions: Theory and applications. Ph.D. thesis, Univ. Köln.

FRAHM, G. and JAEKEL, U. (2007). Tyler's M-estimator, random matrix theory, and generalized elliptical distributions with applications to finance. Discussion Papers in Econometrics and Statistics, No. 2/07, Institute of Econometrics and Statistics, Univ. Cologne.

FRAHM, G. and JAEKEL, U. (2010). A generalization of Tyler's M-estimators to the case of incomplete data. *Comput. Statist. Data Anal.* **54** 374–393. MR2756433 https://doi.org/10.1016/j.csda.2009.08.019

GOES, J., LERMAN, G. and NADLER, B. (2020). Supplement to "Robust sparse covariance estimation by thresholding Tyler's M-estimator." https://doi.org/10.1214/18-AOS1793SUPP.

GUIONNET, A. and ZEITOUNI, O. (2000). Concentration of the spectral measure for large matrices. *Electron. Commun. Probab.* **5** 119–136. MR1781846 https://doi.org/10.1214/ECP.v5-1026

HAN, F., LU, J. and LIU, H. (2014). Robust scatter matrix estimation for high dimensional distributions with heavy tails. Technical Report, Princeton Univ., Princeton, NJ.

HUBER, P. J. and RONCHETTI, E. M. (2009). *Robust Statistics*, 2nd ed. *Wiley Series in Probability and Statistics*. Wiley, Hoboken, NJ. MR2488795 https://doi.org/10.1002/9780470434697

KAMMOUN, A., COUILLET, R., PASCAL, F. and ALOUINI, M.-S. (2018). Optimal design of the adaptive normalized matched filter detector using regularized Tyler estimators. *IEEE Trans. Aerosp. Electron. Syst.* **54** 755–769. MR3456682 https://doi.org/10.1109/TSP.2015.2494858

KELKER, D. (1970). Distribution theory of spherical distributions and a location-scale parameter generalization. *Sankhyā Ser. A* **32** 419–438. MR0287628

KENT, J. T. and TYLER, D. E. (1988). Maximum likelihood estimation for the wrapped Cauchy distribution. *J. Appl. Stat.* **15** 247–254.

KENT, J. T. and TYLER, D. E. (1991). Redescending M-estimates of multivariate location and scatter. *Ann. Statist.* **19** 2102–2119. MR1135166 https://doi.org/10.1214/aos/1176348388

LAM, C. and FAN, J. (2009). Sparsistency and rates of convergence in large covariance matrix estimation. *Ann. Statist.* **37** 4254–4278. MR2572459 https://doi.org/10.1214/09-AOS720

MARDIA, K. V., KENT, J. T. and BIBBY, J. M. (1979). *Multivariate Analysis*. Academic Press, London. MR0560319

MARONNA, R. A. (1976). Robust M-estimators of multivariate location and scatter. *Ann. Statist.* **4** 51–67. MR0388656

MARONNA, R. A. and YOHAI, V. J. (2017). Robust and efficient estimation of multivariate scatter and location. *Comput. Statist. Data Anal.* **109** 64–75. MR3603641 https://doi.org/10.1016/j.csda.2016.11.006

MORALES-JIMENEZ, D., COUILLET, R. and MCKAY, M. R. (2015). Large dimensional analysis of robust M-estimators of covariance with outliers. *IEEE Trans. Signal Process.* **63** 5784–5797. MR3405886 https://doi.org/10.1109/TSP.2015.2460225

NORDHAUSEN, K. and TYLER, D. E. (2015). A cautionary note on robust covariance plug-in methods. *Biometrika* **102** 573–588. MR3394276 https://doi.org/10.1093/biomet/asv022

OLLILA, E. and KOIVUNEN, V. (2003). Robust antenna array processing using M-estimators of pseudo-covariance. In *Proceedings of the IEEE* 14*th International Symposium on Personal*, *Indoor and Mobile Radio Communications* (*PIMRC*) 2659–2663.

OLLILA, E. and TYLER, D. E. (2012). Distribution-free detection under complex elliptically symmetric clutter distribution. In *IEEE* 7*th Sensor Array and Multichannel Signal Processing Workshop*, (*SAM*) 413–416.

OLLILA, E. and TYLER, D. E. (2014). Regularized M-estimators of scatter matrix. *IEEE Trans. Signal Process.* **62** 6059–6070. MR3281544 https://doi.org/10.1109/TSP.2014.2360826

PASCAL, F., CHITOUR, Y. and QUEK, Y. (2014). Generalized robust shrinkage estimator and its application to STAP detection problem. *IEEE Trans. Signal Process.* **62** 5640–5651. MR3273520 https://doi.org/10.1109/TSP.2014.2355779

ROTHMAN, A. J., LEVINA, E. and ZHU, J. (2009). Generalized thresholding of large covariance matrices. *J. Amer. Statist. Assoc.* **104** 177–186. MR2504372 https://doi.org/10.1198/jasa.2009.0101

RUDELSON, M. and VERSHYNIN, R. (2013). Hanson–Wright inequality and sub-Gaussian concentration. *Electron. Commun. Probab.* **18** no. 82, 9. MR3125258 https://doi.org/10.1214/ECP.v18-2865

SIRKIÄ, S., TASKINEN, S. and OJA, H. (2007). Symmetrised M-estimators of multivariate scatter. *J. Multivariate Anal.* **98** 1611–1629. MR2370110 https://doi.org/10.1016/j.jmva.2007.06.005

SOLOVEYCHIK, I. and WIESEL, A. (2014). Tyler's covariance matrix estimator in elliptical models with convex structure. *IEEE Trans. Signal Process.* **62** 5251–5259. MR3268109 https://doi.org/10.1109/TSP.2014.2348951

SUN, Y., BABU, P. and PALOMAR, D. P. (2014). Regularized Tyler's scatter estimator: Existence, uniqueness, and algorithms. *IEEE Trans. Signal Process.* **62** 5143–5156. MR3260359 https://doi.org/10.1109/TSP.2014.2348944

SUN, Y., BABU, P. and PALOMAR, D. P. (2016). Robust estimation of structured covariance matrix for heavy-tailed elliptical distributions. *IEEE Trans. Signal Process.* **64** 3576–3590. MR3515702 https://doi.org/10.1109/TSP.2016.2546222

TYLER, D. E. (1987a). A distribution-free M-estimator of multivariate scatter. *Ann. Statist.* **15** 234–251. MR0885734 https://doi.org/10.1214/aos/1176350263

TYLER, D. E. (1987b). Statistical analysis for the angular central Gaussian distribution on the sphere. *Biometrika* **74** 579–589. MR0909362 https://doi.org/10.1093/biomet/74.3.579

WIESEL, A. (2012). Unified framework to regularized covariance estimation in scaled Gaussian models. *IEEE Trans. Signal Process.* **60** 29–38. MR2932100 https://doi.org/10.1109/TSP.2011.2170685

WIESEL, A. and ZHANG, T. (2014). Structured robust covariance estimation. *Found. Trends Signal Process.* **8** 127–216. MR3438278 https://doi.org/10.1561/2000000053

ZHANG, T., CHENG, X. and SINGER, A. (2016). Marčenko–Pastur law for Tyler's M-estimator. *J. Multivariate Anal.* **149** 114–123. MR3507318 https://doi.org/10.1016/j.jmva.2016.03.010