Regularized Tyler's Scatter Estimator: Existence, Uniqueness, and Algorithms

Ying Sun, *Student Member, IEEE*, Prabhu Babu, and Daniel P. Palomar, *Fellow, IEEE*

***Abstract*—This paper considers the regularized Tyler's scatter estimator for elliptical distributions, which has received considerable attention recently. Various types of shrinkage Tyler's estimators have been proposed in the literature and proved work effectively in the "large p small n" scenario. Nevertheless, the existence and uniqueness properties of the estimators are not thoroughly studied, and in certain cases the algorithms may fail to converge. In this work, we provide a general result that analyzes the sufficient condition for the existence of a family of shrinkage Tyler's estimators, which quantitatively shows that regularization indeed reduces the number of required samples for estimation and the convergence of the algorithms for the estimators. For two specific shrinkage Tyler's estimators, we also proved that the condition is necessary and the estimator is unique. Finally, we show that the two estimators are actually equivalent. Numerical algorithms are also derived based on the majorization-minimization framework, under which the convergence is analyzed systematically.**

***Index Terms*—Tyler's scatter estimator, shrinkage estimator, existence, uniqueness, majorization-minimization.**

I. Introduction

COVARIANCE estimation has been a long existing problem in various signal processing related fields, including multiantenna communication systems, social networks, bioinformatics, as well as financial engineering. A well known and easy to implement estimator is the sample covariance matrix. Under the assumption of clean samples, the estimator is consistent by the Law of Large Numbers. However, the performance of the sample covariance matrix is vulnerable to data corrupted by noise and outliers, which is often the case in real-world applications.

As a remedy, robust estimators are proposed aimed at limiting the influence of erroneous observations so as to achieve better performance in non-Gaussian scenarios [1], [2]. Recently, Tyler's scatter estimator [3] has received considerable attention both theoretically and practically in signal processing related fields, e.g., [4]–[8] to name a few, see [9] for a comprehensive overview. Tyler's estimator estimates the normalized scatter matrix (equivalently the normalized covariance matrix if the covariance exists) assuming that the underlying distribution is elliptically symmetric. The estimator is shown to enjoy the following advantages against the others: it is distribution-free in the sense that its asymptotic variance does not depend on the parametric form of the underlying distribution, and it is also the most robust estimator in a min-max sense.

In addition to non-Gaussian observations, another problem we face in practice is the "large p small n" problem, which refers to high dimensional statistical inference with insufficient number of samples. It is obvious that the sample covariance matrix is singular when the number of samples is smaller than the dimension, and Tyler's estimator has the same drawback. In order to handle this problem, [10] borrowed the diagonal loading idea [11] and proposed a regularized Tyler's estimator that shrinks towards identity. A rigorous proof for the existence, uniqueness, and convergence properties is provided in [12], where a systematic way of choosing the regularization parameter is also proposed. However, the estimator is criticized for not being derived from a meaningful cost function. To overcome this issue, a new scale-invariant shrinkage Tyler's estimator, defined as a minimizer of a penalized cost function, was recently proposed in [13]. By showing that the objective function is geodesically convex, Wiesel proved that any algorithm that converges to the local minimum of the objective function is actually the global minimum. Numerical algorithms are provided for the estimator and simulation results demonstrate the estimator is robust and effective in the sample deficient scenario. Despite the good properties, the existence and uniqueness properties of the estimator remains unclear.

In this paper, we study the shrinkage Tyler's estimator and try to answer the unsolved problems mentioned above. First, we give a proof that states the sufficient condition for the existence of shrinkage Tyler's estimator with penalized cost function taking a general form. Second, we propose a Kullback-Leibler divergence (KL divergence) penalized cost function that results in a shrinkage Tyler's estimator similar to the heuristic diagonal loading one considered in [10], [12]. We then move to these two specific estimators and show that under the condition $P_N(S) < \frac{(1+\alpha_0)\dim(S)}{K}$ and the shrinkage target matrix being positive definite, the estimators exist, where N is the number of samples, K is the dimension of the samples and α_0 controls the amount of penalty added to the cost function, $P_N(S)$ stands for the proportion of samples contained in a proper subspace S. In addition, we prove it is also a necessary condition, provided that $\alpha_0 > 0$. Although derived from different cost functions, and also with different estimation equation, we prove that the two

Manuscript received January 26, 2014; revised May 20, 2014 and July 29, 2014; accepted July 30, 2014. Date of publication August 18, 2014; date of current version August 27, 2014. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Pascal Larzabal. This work was supported by the Hong Kong Research Grant RGC 617312.

The authors are with the Department of Electronic and Computer Engineering, Hong Kong University of Science and Technology (HKUST), Clear Water Bay, Hong Kong (e-mail: ysunac@ust.hk; eeprabhubabu@ust.hk; palomar@ust.hk).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TSP.2014.2348944

shrinkage estimators are actually equivalent. Under the assumption that the underlying distribution is continuous, the condition simplifies to $N > \frac{K}{1+\alpha_0}$. Compared with the existence condition for Tyler's estimator, which is $P_N(S) < \frac{\dim(S)}{K}$, or $N > K$ under continuity assumption, this result clearly demonstrates that regularization can relax the requirement on the number of samples, hence shows its capability of handling large dimension estimation problems. Algorithms for the shrinkage estimators based on majorization-minimization framework are provided, where the convergence can be analyzed systematically.

It is worth mentioning the work [14], where the same condition $N > \frac{K}{1+\alpha_0}$ is also independently derived for the KL penalty based shrinkage estimator that shrinks the covariance matrix to identity in the complex field, assuming the samples are linearly independent. [14] eliminates the additional trace normalization step in [12] by showing that the trace of the inverse of the estimator is equal to K. Different from that approach, our work gives an interpretation of the estimator as the minimizer of a KL divergence penalized cost function. Starting from the cost function, we establish the existence condition with a different proof from [14]. In addition, we extend the result (in the real field) since the condition $P_N(S) < \frac{(1+\alpha_0)\dim(S)}{K}$ implies $N > \frac{K}{1+\alpha_0}$ if the samples are linearly independent, and we consider a general positive definite shrinkage target matrix as in [13].

The paper is organized as follows: In Section II, we briefly review Tyler's estimator for samples drawn from the elliptical family. In Section III, the two types of shrinkage estimators, i.e., one proposed in [13] and another derived based on KL divergence are considered, and a rigorous proof for the existence and uniqueness of the estimators is provided. Algorithms based on majorization-minimization are presented in Section IV. Numerical examples follow in Section V, and we conclude in Section VI.

Notation: $\mathbb{R}^n$ stands for n-dimensional real-valued vector space, $\|\cdot\|_2$ stands for vector Frobenius norm. $\mathbb{S}_+^K$ stands for symmetric positive semidefinite $K \times K$ matrices, which is a closed cone in $\mathbb{R}^{K\times K}$, $\mathbb{S}_{++}^K$ denotes symmetric positive definite $K \times K$ matrices. $\lambda_{\max}$ and $\lambda_{\min}$ stand for the largest and smallest eigenvalue of a matrix $\boldsymbol{\Sigma}$ respectively. $\det(\cdot)$ and $\mathrm{Tr}(\cdot)$ stand for matrix determinant and trace respectively. $\|\cdot\|_F$ is the matrix Frobenius norm.

The boundary of the open set $\mathbb{S}_{++}^K$ is conventionally defined as $\mathbb{S}_+^K \backslash \mathbb{S}_{++}^K$, which contains all rank deficient matrices in $\mathbb{S}_+^K$. With a slightly abuse of notation, we also include matrices with all eigenvalues $\lambda \to +\infty$ into the boundary of $\mathbb{S}_{++}^K$. Therefore a sequence of matrices $\boldsymbol{\Sigma}^k$ converges to the boundary of $\mathbb{S}_{++}^K$ iff $\lambda_{\max}^k \to +\infty$ or $\lambda_{\min}^k \to 0$. In the rest of the paper, we will use the statement "$\boldsymbol{\Sigma}$ converges" equivalently as "a sequence of matrices $\boldsymbol{\Sigma}^k$ converges" for notation simplicity.

II. Robust Covariance Matrix Estimation

In this paper, we assume a number N of K-dimensional samples $\{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$ are drawn from an elliptical population distribution with probability density function (pdf) of the form

$$f(\mathbf{x}) = \det(\boldsymbol{\Sigma}_0)^{-\frac{1}{2}} g\left((\mathbf{x}-\boldsymbol{\mu}_0)^T \boldsymbol{\Sigma}_0^{-1}(\mathbf{x}-\boldsymbol{\mu}_0)\right) \quad (1)$$

with location and scatter parameter $(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0)$ in $\mathbb{R}^K \times \mathbb{S}_{++}^K$. The nonnegative function $g(\cdot)$, which is called the density generator, determines the shape of the pdf. In most of the popularly used distributions, e.g., the Gaussian and the Student's t-distribution, $g(\cdot)$ is a decreasing function and determines the decay of the tails of the distribution. Given $\boldsymbol{\mu}_0$, our problem of interest is to estimate the covariance matrix. We can always center the pdf by defining $\tilde{\mathbf{x}} = \mathbf{x} - \boldsymbol{\mu}_0$, hence without loss of generality in the rest of the paper we assume $\boldsymbol{\mu}_0 = \mathbf{0}$. We use the notation P_N and $f(\cdot)$ for the empirical and the population distributions, respectively. It is known that the covariance matrix of an elliptical distribution takes the form $c_g\boldsymbol{\Sigma}_0$ with c_g being a constant that depends on $g(\cdot)$ [1], hence it is unlikely to have a good covariance estimator without prior knowledge of g. In this paper, instead of trying to find the parametric form of g and get an estimator of $c_g\boldsymbol{\Sigma}_0$, we are interested in estimating the normalized covariance matrix $\frac{\boldsymbol{\Sigma}_0}{\mathrm{Tr}(\boldsymbol{\Sigma}_0)}$.

The commonly used sample covariance matrix, which also happens to be the maximum likelihood estimator for the normal distribution, estimates $c_g\boldsymbol{\Sigma}_0$ asymptotically, however it is sensitive to outliers. This motivates the research for estimators robust to outliers in the data and, in fact, many researchers in the statistics literature have addressed this problem by proposing various robust covariance estimators like M-estimators [15], S-estimators [16], MVE [17], and MCD [18] to name a few, see [1], [2] for a complete overview. For example, in [15], Maronna analyzed the properties of the M-estimators, which are given as the solution $\boldsymbol{\Sigma}$ to the equation

$$\boldsymbol{\Sigma} = \frac{1}{N}\sum_{i=1}^{N} u\left(\mathbf{x}_i^T \boldsymbol{\Sigma}^{-1}\mathbf{x}_i\right)\mathbf{x}_i\mathbf{x}_i^T, \quad (2)$$

where the choice of function $u(\cdot)$ determines a whole family of different estimators. Under some technical conditions on $u(s)$ (i.e., $u(s) \geq 0$ for $s > 0$ and nonincreasing, and $su(s)$ is strictly increasing), Maronna proved that there exists a unique $\boldsymbol{\Sigma}$ that solves (2), and gave an iterative algorithm to arrive at that solution. He also established its consistency and robustness. A number of well known estimators take the form (2) and in [15] Maronna gave two examples, with one being the maximum likelihood estimator for multivariate Student's t-distribution, and the other being the Huber's estimator [19]. Both of them are popular for handling heavy tails and outliers in the data.

For all the robust covariance estimators, there is a tradeoff between their efficiency, which measures the variance (estimation accuracy) of the estimator, and robustness, which quantifies the sensitivity of the estimator to outliers. As these two quantities are opposed in nature, a considerable effort has to be put in designing estimators that achieve the right balance between these two quantities. In [3], Tyler dealt with this problem by proposing an estimator that is distribution-free and the "most robust" estimator in min-max sense. Tyler's estimator of $\boldsymbol{\Sigma}$ is given as the solution of the following equation

$$\boldsymbol{\Sigma} = \frac{K}{N}\sum_{i=1}^{N}\frac{\mathbf{x}_i\mathbf{x}_i^T}{\mathbf{x}_i^T\boldsymbol{\Sigma}^{-1}\mathbf{x}_i}, \quad (3)$$

where the results of [15] cannot be applied since $su(s) = K$ is not strictly increasing. Tyler established the conditions for the existence of a solution to the fixed-point equation (3), as well as the fact that the estimator is unique up to a positive scale factor, in the sense that $\mathbf{\Sigma}$ solves (3) if and only if $c\mathbf{\Sigma}$ solves (3) for some positive scalar c. The estimator was shown to be strongly consistent and asymptotically normal with its asymptotic standard deviation independent of g.

Tyler's fixed-point (3) can be alternatively interpreted as follows. Consider the normalized samples defined as $\mathbf{s} = \frac{\mathbf{x}}{\|\mathbf{x}\|_2}$, it is known that the probability distribution of $\mathbf{s}$ takes the form [20]–[22]

$$f(\mathbf{s}) = \frac{\Gamma\left(\frac{K}{2}\right)}{2\pi^{K/2}} \det(\mathbf{\Sigma})^{-\frac{1}{2}} (\mathbf{s}^T \mathbf{\Sigma}^{-1} \mathbf{s})^{-K/2}. \quad (4)$$

Given N samples from the normalized distribution $\{\mathbf{s}_i\}$, the maximum likelihood estimator of $\mathbf{\Sigma}$ can be obtained by minimizing the negative log-likelihood function

$$L(\mathbf{\Sigma}) = \sum_{i=1}^{N} \frac{K}{2} \log\left(\mathbf{s}_i^T \mathbf{\Sigma}^{-1} \mathbf{s}_i\right) + \frac{N}{2} \log\det(\mathbf{\Sigma}), \quad (5)$$

which is equivalent to minimizing

$$L^{\mathrm{Tyler}}(\mathbf{\Sigma}) = \sum_{i=1}^{N} \frac{K}{2} \log\left(\mathbf{x}_i^T \mathbf{\Sigma}^{-1} \mathbf{x}_i\right) + \frac{N}{2} \log\det(\mathbf{\Sigma}). \quad (6)$$

If a minimum $\hat{\mathbf{\Sigma}} \succ \mathbf{0}$ of the function $L^{\mathrm{Tyler}}(\mathbf{\Sigma})$ exists, it needs to satisfy the stationary equation given in (3), which was originally derived by Tyler in [3]. In [3], [21], the authors provided the condition for existence of a nonsingular solution to (3) based on the following reasoning. Notice that $\hat{\mathbf{\Sigma}}$ must be nonsingular, and the function $L^{\mathrm{Tyler}}(\mathbf{\Sigma})$ is unbounded above on the boundary of positive definite matrices, this implies the existence of a minimum. Based on these observations, Kent and Tyler established the existence conditions by showing $L^{\mathrm{Tyler}}(\mathbf{\Sigma}) \to +\infty$ on the boundary. Specifically, under the conditions that: (i) no $\mathbf{x}_i$ lies on the origin, and (ii) for any proper subspace $S \subseteq \mathbb{R}^K$, $P_N(S) < \frac{\dim(S)}{K}$, where $P_N(S) \triangleq \frac{\sum_{i=1}^{N} 1_{\{\mathbf{x}_i \in S\}}}{N}$ stands for the proportion of samples in S, then a nonsingular minimum of the problem (6) exists, which is equivalent to (3) having a solution. In words, the above mentioned conditions require the number of samples to be sufficiently large, and the samples should be spread out in the whole space.

To arrive at the estimator satisfying (3), Tyler proposed the following iterative algorithm:

$$\begin{aligned} \tilde{\mathbf{\Sigma}}_{t+1} &= \frac{K}{N} \sum_{i=1}^{N} \frac{\mathbf{x}_i \mathbf{x}_i^T}{\mathbf{x}_i^T \mathbf{\Sigma}_t^{-1} \mathbf{x}_i} \\ \mathbf{\Sigma}_{t+1} &= \frac{\tilde{\mathbf{\Sigma}}_{t+1}}{\mathrm{Tr}(\tilde{\mathbf{\Sigma}}_{t+1})} \end{aligned} \quad (7)$$

that converges to the unique (up to a positive scale factor) solution of (3).

The robust property of Tyler's estimator can be understood intuitively as follows: by normalizing the samples, i.e., $\mathbf{s} = \frac{\mathbf{x}}{\|\mathbf{x}\|_2}$, the magnitude of an outlier is more unlikely to make the estimator break down. In other words, the estimator is not sensitive to the magnitude of samples, only their direction can affect the performance.

III. Regularized Covariance Matrix Estimation

The regularity conditions for the existence of Tyler's estimator leads to a condition on the number of samples that $N \geq K + 1$ [21], [23]. In some practical applications the number of samples is not sufficient, Tyler's iteration (7) may not converge. In these scenarios, a more sensible approach is to shrink Tyler's estimator to some known a priori estimate of $\mathbf{\Sigma}$. In the literature of robust estimators, there exists two different shrinkage based approaches for Tyler's estimator.

In the first approach, the authors in [10], [12] proposed the following estimator:

$$\begin{aligned} \tilde{\mathbf{\Sigma}}_{t+1} &= \frac{1}{1+\alpha_0} \frac{K}{N} \sum_{i=1}^{N} \frac{\mathbf{x}_i \mathbf{x}_i^T}{\mathbf{x}_i^T \mathbf{\Sigma}_t^{-1} \mathbf{x}_i} + \frac{\alpha_0}{1+\alpha_0} \mathbf{I} \\ \mathbf{\Sigma}_{t+1} &= \frac{\tilde{\mathbf{\Sigma}}_{t+1}}{\mathrm{Tr}(\tilde{\mathbf{\Sigma}}_{t+1})}, \end{aligned} \quad (8)$$

which is a slightly modified version of the original Tyler's iteration in (7), with the modification being including an identity matrix in the first step of the iteration that aims at shrinking the estimator towards the identity matrix. This resembles the idea of regularizing an estimator via diagonal loading [11], [24]. In [12], Chen *et al.* proved the uniqueness of the estimator obtained by the iteration (8) based on concave Perron-Frobenius theory, and gave a method to choose the regularization weight α_0. Although this estimator is widely used and performs well in practice, it is still considered to be heuristic as it does not have an interpretation based on minimizing a cost function.

As a second approach, in [13], the author took a different route and derived a new shrinkage-based Tyler's estimator that has a clear interpretation based on minimizing the penalized negative log-likelihood function

$$L^{\mathrm{Wiesel}}(\mathbf{\Sigma}) = \frac{2}{N} L^{\mathrm{Tyler}}(\mathbf{\Sigma}) + \alpha_0 h^{\mathrm{target}}(\mathbf{\Sigma}), \quad (9)$$

where $h^{\mathrm{target}}(\mathbf{\Sigma}) = K \log(\mathrm{Tr}(\mathbf{\Sigma}^{-1}\mathbf{T})) + \log\det(\mathbf{\Sigma})$ is a function with minimum at the desired target matrix $\mathbf{T}$, hence it will shrink the solution of (9) towards the target. By showing the cost function $L^{\mathrm{Wiesel}}(\mathbf{\Sigma})$ is geodesically convex, the author proved that any local minimum over the set of positive definite matrices is a global minimum [13]. He then derived an iterative algorithm based on majorization-minimization that monotonically decreases the cost function at each iteration:

$$\mathbf{\Sigma}_{t+1} = \frac{1}{1+\alpha_0} \frac{K}{N} \sum_{i=1}^{N} \frac{\mathbf{x}_i \mathbf{x}_i^T}{\mathbf{x}_i^T \mathbf{\Sigma}_t^{-1} \mathbf{x}_i} + \frac{\alpha_0}{1+\alpha_0} \frac{K\mathbf{T}}{\mathrm{Tr}\left(\mathbf{\Sigma}_t^{-1}\mathbf{T}\right)}. \quad (10)$$

Even though the author in [13] showed that the cost function is convex in geodesic space, the existence and uniqueness of the global minimizer remains unknown. Moreover, it is mentioned in [13] that for some values of α_0 the cost function becomes unbounded below and the iterations do not converge.

In this section, we address the following points: (i) we give the missing interpretation based on minimizing a cost function for the estimator in (8), and we also prove its existence and uniqueness; (ii) we prove the iteration in (10) with an additional trace normalization step converges to a unique point and also establish the conditions on the regularization parameter α_0 to ensure the existence of the solution. For both cases, the cost function takes the form of penalized negative log-likelihood function with different penalizing functions. Our methodology for the proofs hinges on techniques used by Tyler in [21], [23].

We start with a proof of existence for a minimizer of a general penalized negative log-likelihood function in the following theorem, the proof of existence of the two aforementioned cases $L^{\text{Tyler}}(\boldsymbol{\Sigma})$ and $L^{\text{Wiesel}}(\boldsymbol{\Sigma})$ are just special cases of the general result.

The idea of proving the existence is to establish the regularity conditions under which the cost function takes value $+\infty$ on the boundary of the set $\mathbb{S}_{++}^K$, a minimum then exists by the continuity of the cost function. The main result is established in Theorem 3, and the following lemma is needed.

Lemma 1: For any continuous function $f(\cdot)$ defined on the set $\mathbb{S}_{++}$, there exists a $\hat{\boldsymbol{\Sigma}} \succ \mathbf{0}$ such that $f(\hat{\boldsymbol{\Sigma}}) \leq f(\boldsymbol{\Sigma}), \forall \boldsymbol{\Sigma} \succ \mathbf{0}$ if $f(\boldsymbol{\Sigma}) \to +\infty$ on the boundary of the set $\mathbb{S}_{++}$.

Definition 2: For any continuous function $f(s)$ defined on $s > 0$, define the quantities

$$a_f = \sup\left\{a | s^{a/2} \exp\left(-f\left(s\right)\right) \to 0 \text{ as } s \to +\infty\right\} \tag{11}$$

and

$$a_f^{'} = \inf\left\{a | s^{a/2} \exp\left(-f\left(s\right)\right) \to 0 \text{ as } s \to 0\right\}. \tag{12}$$

In this paper we are particularly interested in the functions $f(s) = c\log s$ and $f(s) = cs$ with some positive scalar $c < +\infty$. For $f(s) = c\log s$, $a_f = a_f^{'} = 2c$ and, for $f(s) = cs$, $a_f = +\infty$, $a_f^{'} = 0$. We restrict our attention to the case $a_f \geq 0$.

Consider the penalized cost function takes the general form $\tilde{L}(\boldsymbol{\Sigma}) = L^{\rho}(\boldsymbol{\Sigma}) + h(\boldsymbol{\Sigma})$ with original cost function

$$L^{\rho}(\boldsymbol{\Sigma}) = \frac{N}{2}\log\det(\boldsymbol{\Sigma}) + \sum\nolimits_{i=1}^{N}\rho\left(\mathbf{x}_i^T\boldsymbol{\Sigma}^{-1}\mathbf{x}_i\right) \tag{13}$$

where $\rho(\cdot)$ is a continuous function, and the penalty term

$$h(\boldsymbol{\Sigma}) = \alpha\log\det(\boldsymbol{\Sigma}) + \sum_{l=1}^{L}\alpha_l h_l\left(\mathrm{Tr}\left(\mathbf{A}_l^T\boldsymbol{\Sigma}^{-1}\mathbf{A}_l\right)\right) \tag{14}$$

where $\mathrm{Tr}(\mathbf{A}_l^T\boldsymbol{\Sigma}^{-1}\mathbf{A}_l)$ measures the difference between $\boldsymbol{\Sigma}$ and the positive semidefinite matrix $\mathbf{A}_l\mathbf{A}_l^T$. $h_l(\cdot)$ is, in general, an increasing function that increases the penalty as $\boldsymbol{\Sigma}$ deviates from $\mathbf{A}_l\mathbf{A}_l^T$, which is considered to be the prior target that we wish to shrink $\boldsymbol{\Sigma}$ to.

We first give an intuitive argument on the condition that ensures the existence of the estimator. Since the estimator $\hat{\boldsymbol{\Sigma}}$ is defined as the minimizer of the penalized loss function, it exists if $\tilde{L}(\boldsymbol{\Sigma}) \to +\infty$ on the boundary of $\mathbb{S}_{++}^K$ by Lemma 1, and clearly $\hat{\boldsymbol{\Sigma}}$ is nonsingular. We infer $\boldsymbol{\Sigma}$ by the samples $\{\mathbf{x}_i\}$, if the samples are concentrated on some subspace, naturally we "guess" the distribution is degenerate, i.e., $\hat{\boldsymbol{\Sigma}}$ is singular. Therefore, the samples are required to be sufficiently spread out in the whole space so that the inference leads to a nonsingular $\hat{\boldsymbol{\Sigma}}$. Under the case when we have a prior information that $\boldsymbol{\Sigma}$ should be close to the matrix $\mathbf{A}_l\mathbf{A}_l^T$, to ensure $\hat{\boldsymbol{\Sigma}}$ being nonsingular we need to distribute more $\mathbf{x}_i$'s in the null space of $\mathbf{A}_l\mathbf{A}_l^T$ and hence less in the range of $\mathbf{A}_l\mathbf{A}_l^T$. To formalize this intuition, we give the following theorem.

Theorem 3: For cost function

$$\begin{aligned}\tilde{L}(\boldsymbol{\Sigma}) = {} & \frac{N}{2}\log\det\left(\boldsymbol{\Sigma}\right) + \sum\nolimits_{i=1}^{N}\rho\left(\mathbf{x}_i^T\boldsymbol{\Sigma}^{-1}\mathbf{x}_i\right)\\ & + \left(\alpha\log\det\left(\boldsymbol{\Sigma}\right) + \sum\nolimits_{l=1}^{L}\alpha_l h_l\left(\mathrm{Tr}\left(\mathbf{A}_l^T\boldsymbol{\Sigma}^{-1}\mathbf{A}_l\right)\right)\right)\end{aligned} \tag{15}$$

defined on positive definite matrices $\boldsymbol{\Sigma} \succ \mathbf{0}$ with $\rho(\cdot)$ and $h(\cdot)$ being continuous functions, define a_ρ and $a_\rho^{'}$ for ρ, a_l and $a_l^{'}$ for $\alpha_l h_l$'s according to (11) and (12), then $\tilde{L}(\boldsymbol{\Sigma}) \to +\infty$ on the boundary of the set $\mathbb{S}_{++}^K$ if the following conditions are satisfied:

i) no $\mathbf{x}_i$ lies on the origin;
ii) for any proper subspace S

$$P_N(S) < \min\left\{1 - \frac{(N+2\alpha)\left(K - \dim(S)\right) - \sum_{l\in\upsilon} a_l}{a_\rho N}, \frac{(N+2\alpha)\dim(S) - \sum_{l\in\omega} a_l^{'}}{a_\rho^{'} N}\right\}$$

where sets ω and υ are defined as $\omega = \{l | \mathbf{A}_l \subseteq S\}$, $\upsilon = \{l | \mathbf{A}_l \nsubseteq S\}$;

iii) $(-\frac{N}{2} - \alpha)K + \frac{a_\rho}{2}N + \frac{1}{2}\sum_l a_l^{'} < 0$ and $\frac{a_\rho}{2}N - (\frac{N}{2} + \alpha)K + \frac{1}{2}\sum_l a_l > 0$.

Proof: See Appendix A. ■

Remark 4: Condition (i) avoids the scenario when $\mathbf{x}_i^T\boldsymbol{\Sigma}^{-1}\mathbf{x}_i$ takes value 0 and $\rho(s)$ is undefined at $s = 0$, for example $\rho(s) = \log(s)$ for the log-likelihood function. The first part in condition (ii), $P_N(S) < 1 - \frac{(N+2\alpha)(K-\dim(S)) - \sum_{l\in\upsilon} a_l}{a_\rho N}$, ensures $\tilde{L}(\boldsymbol{\Sigma}) \to +\infty$ under the case that some but not all eigenvalues λ_j of $\boldsymbol{\Sigma}$ tend to zero, and the second part in condition (ii), $P_N(S) < \frac{(N+2\alpha)\dim(S) - \sum_{l\in\omega} a_l^{'}}{a_\rho^{'} N}$, ensures $\tilde{L}(\boldsymbol{\Sigma}) \to +\infty$ under the case that some but not all eigenvalues λ_j of $\boldsymbol{\Sigma}$ tend to positive infinity. Together they force $\tilde{L}(\boldsymbol{\Sigma}) \to +\infty$ when $\frac{\lambda_{\max}}{\lambda_{\min}} \to 0$. The first part of condition (iii) ensures $\tilde{L}(\boldsymbol{\Sigma}) \to +\infty$ when all $\lambda \to +\infty$ and the second part ensures $\tilde{L}(\boldsymbol{\Sigma}) \to +\infty$ when all $\lambda \to 0$.

Corollary 5: Assuming the population distribution $f(\cdot)$ is continuous, and the matrices $\mathbf{A}_l$ are full rank, condition (ii) in Theorem 3 simplifies to:

$$\begin{cases}\sum_l a_l - (N+2\alpha)(K-d) > a_\rho(d-N)\\ \alpha > \frac{a_\rho^{'} - N}{2}\end{cases}, \quad \forall 1 \leq d \leq K-1.$$

Proof: The conclusion follows easily from the following two facts: given that the population distribution $f(\cdot)$ is continuous, and no $\mathbf{x}_i$ lies on the origin, any $1 \leq d < K$ sample points define a proper subspace S with $\dim(S) = d$ with probability one; and since $\mathbf{A}_l$'s are full rank, the set $\omega = \emptyset$. ■

Under the regularity conditions provided in Theorem 3, Lemma 1 implies a minimizer $\hat{\boldsymbol{\Sigma}}$ of $\tilde{L}(\boldsymbol{\Sigma})$ exists and is positive definite, therefore it needs to satisfy the condition $\frac{\partial\tilde{L}(\boldsymbol{\Sigma})}{\partial\boldsymbol{\Sigma}} = \mathbf{0}$.

We then show how Theorem 3 works for Tyler's estimator defined as the nonsingular minimizer of (6). Notice that the loss function $L^{\text{Tyler}}(\boldsymbol{\Sigma})$ is scale-invariant, we have $L^{\text{Tyler}}(c\boldsymbol{\Sigma}_0) =$

$L^{\text{Tyler}}(\mathbf{\Sigma}_0) = \text{constant}$ for any positive definite $\mathbf{\Sigma}_0$. This implies that there are cases when $\mathbf{\Sigma}$ goes to the boundary of $\mathbb{S}^K_{++}$ and $L^{\text{Tyler}}(\mathbf{\Sigma})$ will not go to positive infinity. Due to this reason, condition (iii) is violated in Theorem 3. To handle the scaling issue, we introduce a trace constraint $\text{Tr}(\mathbf{\Sigma}) = 1$.

For the Tyler's problem of minimizing (6), we seek the condition that ensures $L^{\text{Tyler}}(\mathbf{\Sigma}) \to +\infty$ when $\mathbf{\Sigma}$ goes to the boundary of the set $\{\mathbf{\Sigma} | \mathbf{\Sigma} \succ \mathbf{0},\ \text{Tr}(\mathbf{\Sigma}) = 1\}$ relative to $\{\mathbf{\Sigma} | \mathbf{\Sigma} \succcurlyeq \mathbf{0},\ \text{Tr}(\mathbf{\Sigma}) = 1\}$. The condition implies that there is a unique minimizer $\hat{\mathbf{\Sigma}}$ that minimizes $L^{\text{Tyler}}(\mathbf{\Sigma})$ over the set $\{\mathbf{\Sigma} | \mathbf{\Sigma} \succ \mathbf{0},\ \text{Tr}(\mathbf{\Sigma}) = 1\}$, and by it is equivalent to the existence of a unique (up to a positive scale factor) minimizer $\mathbf{\Sigma}^\star$ that minimizes $L^{\text{Tyler}}(\mathbf{\Sigma})$ over the set $\mathbb{S}^K_{++}$ since $L^{\text{Tyler}}(\mathbf{\Sigma})$ is scale-invariant.

The constraint $\text{Tr}(\mathbf{\Sigma}) = 1$ excludes the case that any of $\lambda_j \to +\infty$ and the case all $\lambda_j \to 0$, hence we only need to let $L^{\text{Tyler}}(\mathbf{\Sigma}) \to +\infty$ under the case that some but not all $\lambda_j \to 0$, which corresponds to the condition $P_N(S) < 1 - \frac{(N+2\alpha)(K-\dim(S)) - \sum_{l\in v} a_l}{a_\rho N}$ in Theorem 3. For Tyler's cost function $L^{\text{Tyler}}(\mathbf{\Sigma})$, we have $\rho(s) = \frac{K}{2}\log s$ and $\alpha = 0$, $a_\rho = a'_\rho = K$, therefore Theorem 3 leads to the condition on the samples: $P_N(S) < \frac{\dim(S)}{K}$, or $N \geq K+1$ if the population distribution $f(\cdot)$ is continuous, which reduces to the condition given in [21].

A. Regularization via Wiesel's Penalty

In [13], Wiesel proposed a regularization penalty $h(\mathbf{\Sigma})$ that results in a shrinkage estimator. Specifically, the penalty terms that encourage shrinkage towards an identity matrix and more generally towards an arbitrary prior matrix $\mathbf{T}$ are defined as follows:

$$\begin{aligned} h^{\text{identity}}(\mathbf{\Sigma}) &= K\log(\text{Tr}(\mathbf{\Sigma}^{-1})) + \log\det(\mathbf{\Sigma}) \\ h^{\text{target}}(\mathbf{\Sigma}) &= K\log(\text{Tr}(\mathbf{\Sigma}^{-1}\mathbf{T})) + \log\det(\mathbf{\Sigma}). \end{aligned} \tag{16}$$

As can be seen the penalty terms are scale-invariant. Wiesel justified the choice of the above mentioned penalty functions by showing that the minimizer is some scaled multiple of $\mathbf{I}$ (or $\mathbf{T}$). Thus adding this penalty terms to the Tyler's cost function would yield estimators that are shrunk towards $\mathbf{I}$ (or $\mathbf{T}$). In the rest of this subsection we consider the general case h^{target} only, where the penalty term shrinks $\mathbf{\Sigma}$ to scalar multiples of $\mathbf{T}$, and we make the assumption that $\mathbf{T}$ is positive definite, which is reasonable since $\mathbf{\Sigma}$ must be a positive definite matrix. The cost function is restated below for convenience

$$\begin{aligned} L^{\text{Wiesel}}(\mathbf{\Sigma}) = \log\det(\mathbf{\Sigma}) &+ \frac{K}{N}\sum_{i=1}^N \log\left(\mathbf{x}_i^T\mathbf{\Sigma}^{-1}\mathbf{x}_i\right) \\ &+ \alpha_0(K\log(\text{Tr}(\mathbf{\Sigma}^{-1}\mathbf{T})) + \log\det(\mathbf{\Sigma})). \end{aligned} \tag{17}$$

Minimizing $L^{\text{Wiesel}}(\mathbf{\Sigma})$ gives the fixed-point condition

$$\mathbf{\Sigma} = \frac{1}{1+\alpha_0}\frac{K}{N}\sum_{i=1}^N \frac{\mathbf{x}_i\mathbf{x}_i^T}{\mathbf{x}_i^T\mathbf{\Sigma}^{-1}\mathbf{x}_i} + \frac{\alpha_0}{1+\alpha_0}\frac{K\mathbf{T}}{\text{Tr}\left(\mathbf{\Sigma}^{-1}\mathbf{T}\right)}. \tag{18}$$

Recall that in the absence of regularization (i.e., $\alpha_0 = 0$), a solution to the fixed-point equation exists under the condition $P_N(S) < \frac{\dim(S)}{N}$. With the regularization, however, it is not clear. We start giving a result for the uniqueness and then come back to the existence.

Theorem 6: If (18) has a solution, then it is unique up to a positive scale factor.

Proof: It is easy to see if $\mathbf{\Sigma}$ solves (18), $c\mathbf{\Sigma}$ is also a solution for $c > 0$. Without loss of generality assume $\mathbf{\Sigma} = \mathbf{I}$ is a solution, otherwise define $\tilde{\mathbf{x}}_i = \mathbf{\Sigma}^{-\frac{1}{2}}\mathbf{x}_i$ and $\tilde{\mathbf{T}} = \mathbf{\Sigma}^{-\frac{1}{2}}\mathbf{T}\mathbf{\Sigma}^{-\frac{1}{2}}$, and that there exists another solution $\mathbf{\Sigma}_1$. Denote the eigenvalues of $\mathbf{\Sigma}_1$ as $\lambda_1 \geq \cdots \geq \lambda_K$ with at least one strictly inequality, then under the condition that $\mathbf{T}$ is positive definite

$$\begin{aligned} \mathbf{\Sigma}_1 &= \frac{1}{1+\alpha_0}\frac{K}{N}\sum_{i=1}^N \frac{\mathbf{x}_i\mathbf{x}_i^T}{\mathbf{x}_i^T\mathbf{\Sigma}_1^{-1}\mathbf{x}_i} + \frac{\alpha_0}{1+\alpha_0}\frac{K\mathbf{T}}{\text{Tr}\left(\mathbf{\Sigma}_1^{-1}\mathbf{T}\right)} \\ &\prec \frac{1}{1+\alpha_0}\frac{K}{N}\sum_{i=1}^N \frac{\mathbf{x}_i\mathbf{x}_i^T}{\lambda_1^{-1}\mathbf{x}_i^T\mathbf{x}_i} + \frac{\alpha_0}{1+\alpha_0}\frac{K\mathbf{T}}{\text{Tr}\left(\lambda_1^{-1}\mathbf{T}\right)} \\ &= \lambda_1\mathbf{I}, \end{aligned}$$

where the inequality follows from the fact that $\text{Tr}(\mathbf{S}\mathbf{\Sigma}_1^{-1}) > \text{Tr}(\lambda_1^{-1}\mathbf{S})$ for any positive definite matrix $\mathbf{S}$ and the last equality follows from the assumption that $\mathbf{I}$ is a solution to (18). We have the contradiction $\lambda_1 < \lambda_1$, hence all the eigenvalues of $\mathbf{\Sigma}_1$ should be equal, i.e., $\mathbf{\Sigma}_1 = \lambda\mathbf{I}$. ■

Before establishing the existence condition, we give an example when the solution to (18) does not exist for illustration.

Example 7: Consider the case when all $\mathbf{x}_i$'s are aligned in one direction. Eigendecompose $\mathbf{\Sigma} = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^T$ and choose $\mathbf{u}_1$ to be aligned with the $\mathbf{x}_i$'s, let $\lambda_1 \to +\infty$ while others $0 < c \leq \lambda < +\infty$. Ignoring the constant terms, the boundedness of $L^{\text{Wiesel}}(\mathbf{\Sigma})$ is equivalent to the boundedness of $(1 + \alpha_0 - K)\log\lambda_1$, hence it is unbounded below if $\alpha_0 < K - 1$.

The example shows that $L^{\text{Wiesel}}(\mathbf{\Sigma})$ can be unbounded below, implying that (18) has no solution if the data are too concentrated and α_0 is small. The following theorems gives the exact tradeoff between data dispersion and the choice of α_0.

Theorem 8: A unique solution to (18) exists (up to a positive scale factor) if the following conditions are satisfied:

i) no $\mathbf{x}_i$ lies on the origin;
ii) for any proper subspace $S \subseteq \mathbb{R}^K$, $P_N(S) < \frac{(1+\alpha_0)\dim(S)}{K}$,

and they are the global minima of the loss function (17).

Proof: We start by rewriting the function including a scale factor $\frac{N}{2}$ w.r.t. (17) for convenience:

$$\begin{aligned} L^{\text{Wiesel}}(\mathbf{\Sigma}) = \frac{N}{2}\log\det(\mathbf{\Sigma}) &+ \frac{K}{2}\sum_{i=1}^N \log\left(\mathbf{x}_i^T\mathbf{\Sigma}^{-1}\mathbf{x}_i\right) \\ &+ \frac{\alpha_0 N}{2}\left(K\log\left(\text{Tr}\left(\mathbf{\Sigma}^{-1}\mathbf{T}\right)\right) + \log\det\left(\mathbf{\Sigma}\right)\right). \end{aligned} \tag{19}$$

Invoke Theorem 3 with $\rho(s) = \frac{K}{2}\log(s)$, $h_1(s) = K\log(s)$, $\alpha = \alpha_1 = \frac{\alpha_0 N}{2}$ and $\mathbf{A}_1 = \mathbf{T}^{\frac{1}{2}}$, hence $a_\rho = a'_\rho = K$ and $a_1 = a'_1 = 2\alpha K$. By the same reasoning as for the Tyler's loss function, the condition $P_N(S) < 1 - \frac{(N+2\alpha)(K-\dim(S)) - \sum_{l\in v} a_l}{a_\rho N}$, which is $P_N(S) < \frac{(1+\alpha_0)\dim(S)}{K}$ since $\mathbf{T}$ is full rank, ensures the existence of a unique solution to (18) under the constraint $\mathbf{\Sigma} \succ \mathbf{0}$ and $\text{Tr}(\mathbf{\Sigma}) = 1$. Hence a unique (up to a positive scale factor) solution to (18) exists on the set of $\mathbb{S}^K_{++}$ by the scale-invariant property of $L^{\text{Wiesel}}(\mathbf{\Sigma})$. ■

To make the existence condition easy to check, we use Corollary 5. Theorem 8 then simplifies to $\alpha_0 > \frac{K}{N} - 1$ or, equivalently $N > \frac{K}{1+\alpha_0}$, from which we can see that compared to the condition without regularization, shrinkage allows less number of samples, and the minimum number depends on α_0.

At last, we show that the condition $P_N(S) < \frac{(1+\alpha_0)\dim(S)}{K}$ is also necessary in the following proposition.

Proposition 9: If (18) admits a solution on $\mathbb{S}_{++}^K$, then for any proper subspace $S \subseteq \mathbb{R}^K$, $P_N(S) < \frac{(1+\alpha_0)\dim(S)}{K}$, provided that $\mathbf{T}$ is positive definite and $\alpha_0 > 0$.

Proof: For a proper subspace S, define $\mathbf{P}$ as the orthogonal projection matrix associated to S, i.e., $\mathbf{P}\mathbf{x} = \mathbf{x}$, $\forall \mathbf{x} \in S$. Assume the solution is $\mathbf{I}$. Multiplying both sides of (18) by matrix $\mathbf{I} - \mathbf{P}$ and taking the trace we have

$$\begin{aligned} &K - \dim(S) \\ &= \frac{1}{1+\alpha_0}\frac{K}{N}\sum_{i=1}^{N}\frac{\mathbf{x}_i^T\left(\mathbf{I}-\mathbf{P}\right)\mathbf{x}_i}{\mathbf{x}_i^T\mathbf{x}_i} + \frac{\alpha_0 K}{1+\alpha_0}\frac{\mathrm{Tr}(\mathbf{T}-\mathbf{T}\mathbf{P})}{\mathrm{Tr}(\mathbf{T})}. \end{aligned}$$

If $\mathbf{x}_i \in S$, then $\mathbf{x}_i^T(\mathbf{I}-\mathbf{P})\mathbf{x}_i = 0$, and $\mathbf{x}_i^T(\mathbf{I}-\mathbf{P})\mathbf{x}_i \le \mathbf{x}_i^T\mathbf{x}_i$ if $\mathbf{x}_i \notin S$. Moreover, $\mathrm{Tr}(\mathbf{T}\mathbf{P}) > 0$ since $\mathbf{T}$ is positive definite. This therefore implies

$$K - \dim(S) < \frac{1}{1+\alpha_0}\frac{K}{N}(N - NP_N(S)) + \frac{\alpha_0 K}{1+\alpha_0}.$$

Rearranging the terms yields $P_N(S) < \frac{(1+\alpha_0)\dim(S)}{K}$. ■

B. Regularization via Kullback-Leibler Divergence Penalty

An ideal penalty term should increase as $\mathbf{\Sigma}$ deviates from the prior target $\mathbf{T}$. Wiesel's penalty function discussed in the last subsection satisfies this property and, in this subsection, we propose another penalty that has this property. The penalty that we choose is the KL divergence between $\mathcal{N}_\Sigma(\mathbf{0}, \mathbf{\Sigma})$ and $\mathcal{N}_T(\mathbf{0}, \mathbf{T})$, i.e., two zero-mean Gaussians with covariance matrices $\mathbf{\Sigma}$ and $\mathbf{T}$, respectively. The formula for the KL divergence is as follows [25], [26]

$$D_{KL}(\mathcal{N}_T \| \mathcal{N}_\Sigma) = \frac{1}{2}\left(\mathrm{Tr}\left(\mathbf{\Sigma}^{-1}\mathbf{T}\right) - K - \log\left(\frac{\det\left(\mathbf{T}\right)}{\det(\mathbf{\Sigma})}\right)\right).$$

Ignoring the constant terms results in the following loss function:

$$\begin{aligned} L^{\mathrm{KL}}(\mathbf{\Sigma}) =& \log\det(\mathbf{\Sigma}) + \frac{K}{N}\sum_{i=1}^{N}\log\left(\mathbf{x}_i^T\mathbf{\Sigma}^{-1}\mathbf{x}_i\right) \\ &+ \alpha_0(\mathrm{Tr}(\mathbf{\Sigma}^{-1}\mathbf{T}) + \log\det(\mathbf{\Sigma})) \end{aligned} \tag{20}$$

with the following fixed-point condition:

$$\mathbf{\Sigma} = \frac{1}{1+\alpha_0}\frac{K}{N}\sum_{i=1}^{N}\frac{\mathbf{x}_i\mathbf{x}_i^T}{\mathbf{x}_i^T\mathbf{\Sigma}^{-1}\mathbf{x}_i} + \frac{\alpha_0}{1+\alpha_0}\mathbf{T}. \tag{21}$$

Unlike the penalty function discussed in the last subsection, KL divergence penalty encourages shrinkage towards $\mathbf{T}$ without scaling ambiguity. This can be easily seen, as the minimizer of the KL divergence penalty is just $\mathbf{T}$. Notice that (21) is similar to the diagonal loading in (8), but without the heuristic normalizing step.

Theorem 10: If (21) has a solution, then it is unique.

Proof: Without loss of generality, we assume $\mathbf{\Sigma} = \mathbf{I}$ solves (21). Assume there is another matrix $\mathbf{\Sigma}_1$ that solves (21), denote the largest eigenvalue of $\mathbf{\Sigma}_1$ as λ_1 and suppose $\lambda_1 > 1$. We then have the following contradiction:

$$\begin{aligned} \mathbf{\Sigma}_1 &\preccurlyeq \frac{1}{1+\alpha_0}\frac{K}{N}\sum_{i=1}^{N}\frac{\mathbf{x}_i\mathbf{x}_i^T}{\lambda_1^{-1}\mathbf{x}_i^T\mathbf{x}_i} + \frac{\alpha_0}{1+\alpha_0}\mathbf{T} \\ &< \frac{\lambda_1}{1+\alpha_0}\frac{K}{N}\sum_{i=1}^{N}\frac{\mathbf{x}_i\mathbf{x}_i^T}{\mathbf{x}_i^T\mathbf{x}_i} + \frac{\alpha_0\lambda_1}{1+\alpha_0}\mathbf{T} = \lambda_1\mathbf{I}, \end{aligned}$$

which gives contradiction $\lambda_1 < \lambda_1$, hence $\lambda_1 \le 1$. Similarly, suppose the smallest eigenvalue of $\mathbf{\Sigma}_1$ satisfies $\lambda_K < 1$. We then have

$$\begin{aligned} \mathbf{\Sigma}_1 &\succcurlyeq \frac{1}{1+\alpha_0}\frac{K}{N}\sum_{i=1}^{N}\frac{\mathbf{x}_i\mathbf{x}_i^T}{\lambda_K^{-1}\mathbf{x}_i^T\mathbf{x}_i} + \frac{\alpha_0}{1+\alpha_0}\mathbf{T} \\ &> \frac{\lambda_K}{1+\alpha_0}\frac{K}{N}\sum_{i=1}^{N}\frac{\mathbf{x}_i\mathbf{x}_i^T}{\mathbf{x}_i^T\mathbf{x}_i} + \frac{\alpha_0\lambda_K}{1+\alpha_0}\mathbf{T} = \lambda_K\mathbf{I}, \end{aligned}$$

which is a contradiction and hence $\lambda_K \ge 1$, from which $\mathbf{\Sigma}_1 = \mathbf{I}$ follows. ■

Theorem 11: A unique solution to (21) exists, if

i) no $\mathbf{x}_i$ lies on the origin;

ii) $P_N(S) < \frac{(1+\alpha_0)\dim(S)}{K}$,

and it is the global minimum of loss function (20).

Proof: Equivalently, we can define

$$\begin{aligned} L^{\mathrm{KL}}(\mathbf{\Sigma}) =& \frac{N}{2}\log\det(\mathbf{\Sigma}) + \frac{K}{2}\sum_{i=1}^{N}\log\left(\mathbf{x}_i^T\mathbf{\Sigma}^{-1}\mathbf{x}_i\right) \\ &+ \frac{\alpha_0 N}{2}(\mathrm{Tr}(\mathbf{\Sigma}^{-1}\mathbf{T}) + \log\det(\mathbf{\Sigma})). \end{aligned} \tag{22}$$

Invoke Theorem 3 with $\rho(s) = \frac{K}{2}\log(s)$, $h_1(s) = s$, $\alpha = \alpha_1 = \frac{\alpha_0 N}{2}$ and $\mathbf{A}_1 = \mathbf{T}^{\frac{1}{2}}$, hence $a_\rho = a_\rho^{'} = K$, $a_1 = +\infty$, $a_1^{'} = 0$. Since $\mathbf{T}$ is full rank and $a_1 = +\infty$, condition (ii) reduces to $P_N(S) < \frac{(1+\alpha_0)\dim(S)}{K}$. Condition (iii) is satisfied, hence an interior minimum exists. Furthermore, it is the unique minimum, hence it is global. ■

Remark 12: The only difference between the regularized estimator discussed in this subsection and the heuristic estimator in (8) is the extra normalizing step in (8). With the trace normalization, [12] proved that the iteration (8) converges to a unique solution without any assumption of the data. However, the iteration implied by (21), which is based on minimizing a negative log-likelihood function penalized via the KL divergence function, requires some regularity conditions to be satisfied (cf. Theorem 11). According to Corollary 5, the condition simplifies to $\alpha_0 > \frac{K}{N} - 1$ if the population distribution is continuous.

Proposition 13: If (21) admits a solution on $\mathbb{S}_{++}^K$, then for any proper subspace $S \subseteq \mathbb{R}^K$, $P_N(S) < \frac{(1+\alpha_0)\dim(S)}{K}$, provided that $\mathbf{T}$ is positive definite and $\alpha_0 > 0$.

Proof: Multiply both sides of (21) by $\mathbf{T}^{-\frac{1}{2}}$ and define $\tilde{\mathbf{\Sigma}} = \mathbf{T}^{-\frac{1}{2}}\mathbf{\Sigma}\mathbf{T}^{-\frac{1}{2}}$, $\tilde{\mathbf{x}}_i = \mathbf{T}^{-\frac{1}{2}}\mathbf{x}_i$ yields

$$\tilde{\mathbf{\Sigma}} = \frac{1}{1+\alpha_0}\frac{K}{N}\sum_{i=1}^{N}\frac{\tilde{\mathbf{x}}_i\tilde{\mathbf{x}}_i^T}{\tilde{\mathbf{x}}_i^T\tilde{\mathbf{\Sigma}}^{-1}\tilde{\mathbf{x}}_i} + \frac{\alpha_0}{1+\alpha_0}\mathbf{I}. \tag{23}$$

The rest of the proof follows the same reasoning as Proposition 9. ■

Finally, we show in the following proposition that the Wiesel's shrinkage estimator defined as solution to (18) and KL shrinkage estimator defined as solution to (21) are equivalent.

Proposition 14: The solution to fixed point (21) solves (18) and, conversely, any solution of (18) solves (21) with a proper scale factor.

Proof: If α_0 is zero, the statement is trivial. We consider the case $\alpha_0 \neq 0$. Following the argument of previous proposition we arrive at (23). It has been shown in [14] that the unique solution $\tilde{\mathbf{\Sigma}}$ to (23) satisfies $\mathrm{Tr}(\tilde{\mathbf{\Sigma}}^{-1}) = K$ given $\alpha_0 > 0$, hence $\mathrm{Tr}(\mathbf{\Sigma}^{-1}\mathbf{T}) = K$. Substitute it into (18) yields exactly (21) with solution $\mathbf{\Sigma}$, which indicates $\mathbf{\Sigma}$ solves (18). The second part of the proposition follows from the fact that Wiesel's fixed-point (18) has a unique solution up to a positive scale factor. ■

IV. Algorithms

Before going the specific algorithms, we first briefly introduce the concept of majorization-minimization [27], [28]. Consider the following optimization problem

$$\begin{array}{ll} \underset{\mathbf{x}}{\text{minimize}} & f(\mathbf{x}) \\ \text{subject to} & \mathbf{x} \in \mathcal{X} \end{array} \tag{24}$$

where $f(\cdot)$ is assumed to be a continuous function, not necessarily convex, and $\mathcal{X}$ is a closed convex set.

At a given point $\mathbf{x}_t$, the majorization-minimization algorithm finds a surrogate function $g(\mathbf{x}|\mathbf{x}_t)$ that satisfies the following properties:

$$\begin{aligned} f(\mathbf{x}_t) &= g(\mathbf{x}_t|\mathbf{x}_t) \\ f(\mathbf{x}) &\leq g(\mathbf{x}|\mathbf{x}_t),\ \forall \mathbf{x} \in \mathcal{X} \\ f'(\mathbf{x}_t;\mathbf{d}) &= g'(\mathbf{x}_t;\mathbf{d}|\mathbf{x}_t),\ \forall \mathbf{x}_t + \mathbf{d} \in \mathcal{X} \end{aligned} \tag{25}$$

with $f'(\mathbf{x};\mathbf{d})$ stands for directional derivative. The surrogate function $g(\mathbf{x}|\mathbf{x}_t)$ is assumed to be continuous in $\mathbf{x}$ and $\mathbf{x}_t$[1].

The majorization-minimization algorithm updates $\mathbf{x}$ as

$$\mathbf{x}_{t+1} = \arg\min_{\mathbf{x}\in\mathcal{X}} g(\mathbf{x}|\mathbf{x}_t).$$

It is proved that every limit point of the sequence $\{\mathbf{x}_t\}$ is a stationary point of problem (24), and under the assumption that the level set $\mathcal{X}^0 = \{\mathbf{x}|f(\mathbf{x}) \leq f(\mathbf{x}_0)\}$ is compact, the distance between $\{\mathbf{x}_t\}$ and the set of stationary points reduces to zero in the limit [28].

In the rest of this section, for any continuous differentiable function $f(\mathbf{y})$, we define $f(\mathbf{y}) = +\infty$ when $\lim_{\mathbf{x}\to\mathbf{y}} f(\mathbf{x}) = +\infty$.

A. Regularization via Wiesel's Penalty

In [13], Wiesel derived Tyler's iteration (7) but without the trace normalization step, from the majorization-minimization perspective, with surrogate function $g(\mathbf{\Sigma}|\mathbf{\Sigma}_t)$ for (6) defined as

$$g(\mathbf{\Sigma}|\mathbf{\Sigma}_t) = \frac{N}{2}\log\det(\mathbf{\Sigma}) + \sum_{i=1}^{N}\frac{K}{2}\frac{\mathbf{x}_i^T\mathbf{\Sigma}^{-1}\mathbf{x}_i}{\mathbf{x}_i^T\mathbf{\Sigma}_t^{-1}\mathbf{x}_i} + \text{const.} \tag{26}$$

[1]Notice that if both $f(\mathbf{x})$ and $g(\mathbf{x}|\mathbf{x}_t)$ are continuously differentiable, then the first two conditions in (25) imply the third.

A positive definite stationary point of $g(\mathbf{\Sigma}|\mathbf{\Sigma}_t)$ satisfies the first equation of (7). By the same technique, to solve the problem

$$\begin{array}{ll} \underset{\mathbf{\Sigma}}{\text{minimize}} & \log\det(\mathbf{\Sigma}) + \frac{K}{N}\sum_{i=1}^{N}\log\left(\mathbf{x}_i^T\mathbf{\Sigma}^{-1}\mathbf{x}_i\right) \\ & \quad + \alpha_0(K\log(\mathrm{Tr}(\mathbf{\Sigma}^{-1}\mathbf{T})) + \log\det(\mathbf{\Sigma})) \\ \text{subject to} & \mathbf{\Sigma} \succcurlyeq \mathbf{0}. \end{array} \tag{27}$$

Wiesel derived the iteration (10) by majorizing (17) with function

$$\begin{aligned} &(1+\alpha_0)\log\det(\mathbf{\Sigma}) + \frac{K}{N}\sum_{i=1}^{N}\frac{\mathbf{x}_i^T\mathbf{\Sigma}^{-1}\mathbf{x}_i}{\mathbf{x}_i^T\mathbf{\Sigma}_t^{-1}\mathbf{x}_i} \\ &\qquad + \frac{\alpha_0 K}{\mathrm{Tr}\left(\mathbf{\Sigma}_t^{-1}\mathbf{T}\right)}\mathrm{Tr}\left(\mathbf{\Sigma}^{-1}\mathbf{T}\right) + \text{const.} \end{aligned} \tag{28}$$

It is worth pointing out that if we do the change of variable $\boldsymbol{\psi} = \mathbf{\Sigma}^{-1}$ in $L^{\text{Wiesel}}(\mathbf{\Sigma})$ and linearize the term $\log(\mathbf{x}_i^T\mathbf{\Sigma}^{-1}\mathbf{x}_i)$, this also leads to the same iteration (10).

In the rest of this subsection, we prove the convergence of the iteration (10) proposed by Wiesel, but with an additional trace normalization step, i.e., our modified iteration takes the form:

$$\begin{aligned} \tilde{\mathbf{\Sigma}}_{t+1} &= \frac{1}{1+\alpha_0}\frac{K}{N}\sum_{i=1}^{N}\frac{\mathbf{x}_i\mathbf{x}_i^T}{\mathbf{x}_i^T\mathbf{\Sigma}_t^{-1}\mathbf{x}_i} + \frac{\alpha_0}{1+\alpha_0}\frac{K\mathbf{T}}{\mathrm{Tr}\left(\mathbf{\Sigma}_t^{-1}\mathbf{T}\right)} \\ \mathbf{\Sigma}_{t+1} &= \frac{\tilde{\mathbf{\Sigma}}_{t+1}}{\mathrm{Tr}(\tilde{\mathbf{\Sigma}}_{t+1})}. \end{aligned} \tag{29}$$

Denote the set $\mathcal{S} = \{\mathbf{\Sigma}|\mathrm{Tr}(\mathbf{\Sigma}) = 1,\ \mathbf{\Sigma} \succcurlyeq \mathbf{0}\}$.

Algorithm 1: Wiesel's shrinkage estimator

---

1) Initialize $\mathbf{\Sigma}_0$ as an arbitrary positive definite matrix.
2) Iterate

$$\begin{aligned} \tilde{\mathbf{\Sigma}}_{t+1} &= \frac{1}{1+\alpha_0}\frac{K}{N}\sum_{i=1}^{N}\frac{\mathbf{x}_i\mathbf{x}_i^T}{\mathbf{x}_i^T\mathbf{\Sigma}_t^{-1}\mathbf{x}_i} + \frac{\alpha_0}{1+\alpha_0}\frac{K\mathbf{T}}{\mathrm{Tr}\left(\mathbf{\Sigma}_t^{-1}\mathbf{T}\right)} \\ \mathbf{\Sigma}_{t+1} &= \frac{\tilde{\mathbf{\Sigma}}_{t+1}}{\mathrm{Tr}(\tilde{\mathbf{\Sigma}}_{t+1})} \end{aligned}$$

until convergence.

Lemma 15: The set $\mathcal{X}^0 = \{\mathbf{\Sigma}|L^{\text{Wiesel}}(\mathbf{\Sigma}) \leq L^{\text{Wiesel}}(\mathbf{\Sigma}_0)\} \cap \mathcal{S}$ is a compact set.

Proof: $\mathrm{Tr}(\mathbf{\Sigma}) = 1$ implies the set $\mathcal{X}^0$ is bounded. The set is closed follows easily from the fact that $L^{\text{Wiesel}}(\mathbf{\Sigma}) \to +\infty$ when $\mathbf{\Sigma}$ tends to be singular. ■

Lemma 16: The $\tilde{\mathbf{\Sigma}}_{t+1}$ given in (29) is the unique minimizer of surrogate function (28).

Proof: For surrogate function (28), its value goes to positive infinity when $\frac{\lambda_{\max}(\mathbf{\Sigma})}{\lambda_{\min}(\mathbf{\Sigma})} \to +\infty$, since it majorizes $L^{\text{Wiesel}}(\mathbf{\Sigma})$ and $L^{\text{Wiesel}}(\mathbf{\Sigma}) \to +\infty$ in this case. Now consider the case when $\frac{\lambda_{\max}}{\lambda_{\min}} = O(1)$. Define $\bar{\mathbf{\Sigma}} = \frac{\mathbf{\Sigma}}{\mathbf{\Sigma}_{K,K}}$, then function (28) can be rewritten as

$$\begin{aligned} &(1+\alpha_0)(\log\det(\bar{\mathbf{\Sigma}}) + K\log\lambda_{\min}) \\ &\quad + \frac{K}{N}\sum_{i=1}^{N}\lambda_{\min}^{-1}\frac{\mathbf{x}_i^T\bar{\mathbf{\Sigma}}^{-1}\mathbf{x}_i}{\mathbf{x}_i^T\mathbf{\Sigma}_t^{-1}\mathbf{x}_i} + \frac{\alpha_0 K}{\mathrm{Tr}\left(\mathbf{\Sigma}_t^{-1}\mathbf{T}\right)}\mathrm{Tr}\left(\lambda_{\min}^{-1}\bar{\mathbf{\Sigma}}^{-1}\mathbf{T}\right) \\ &\quad + \text{const.} \end{aligned} \tag{30}$$

The terms $\log\det(\bar{\boldsymbol{\Sigma}})$, $\frac{\mathbf{x}_i^T\bar{\boldsymbol{\Sigma}}^{-1}\mathbf{x}_i}{\mathbf{x}_i^T\boldsymbol{\Sigma}_t^{-1}\mathbf{x}_i}$ and $\mathrm{Tr}(\bar{\boldsymbol{\Sigma}}^{-1}\mathbf{T})$ are all constants bounded away from both 0 and $+\infty$. It is easy to see that when $\lambda_{\min}\to 0$ or $\lambda_{\min}\to+\infty$, (30) goes to $+\infty$. Therefore we conclude that the value of Wiesel's surrogate function (28) goes to $+\infty$ when $\boldsymbol{\Sigma}$ approaches the boundary of $\mathbb{S}_+^K$. The fact that $\tilde{\boldsymbol{\Sigma}}_{t+1}$ given in (29) is the unique solution to the stationary equation implies that it is the unique minimizer of (28) on the set $\mathbb{S}_+^K$. ■

Proposition 17: The sequence $\{\boldsymbol{\Sigma}_t\}$ generated by Algorithm 1 converges to the global minimizer of problem (27).

Proof: It is proved in Theorem 8 that under the conditions provided in Theorem 8, the minimizer $\hat{\boldsymbol{\Sigma}}$ of problem

$$\begin{array}{ll}\underset{\boldsymbol{\Sigma}\succ\mathbf{0}}{\text{minimize}} & \log\det(\boldsymbol{\Sigma})+\frac{K}{N}\sum_{i=1}^N\log\left(\mathbf{x}_i^T\boldsymbol{\Sigma}^{-1}\mathbf{x}_i\right)\\ & +\alpha_0(K\log(\mathrm{Tr}(\boldsymbol{\Sigma}^{-1}\mathbf{T}))+\log\det(\boldsymbol{\Sigma}))\\ \text{subject to} & \mathrm{Tr}(\boldsymbol{\Sigma})=1\end{array}\quad(31)$$

exists and is unique, furthermore, it solves problem (27). It is also proved that the objective function $L^{\mathrm{Wiesel}}(\boldsymbol{\Sigma})\to+\infty$ on the boundary of the set $\mathcal{S}$. We now show that the sequence $\{\boldsymbol{\Sigma}_t\}$ converges to unique minimizer of (31).

Denote the surrogate function in general as $g(\boldsymbol{\Sigma}|\boldsymbol{\Sigma}_t)$, by Lemma 16 we therefore have the following inequality

$$\begin{aligned}L^{\mathrm{Wiesel}}(\boldsymbol{\Sigma}_t)&=g(\boldsymbol{\Sigma}_t|\boldsymbol{\Sigma}_t)\ge g(\tilde{\boldsymbol{\Sigma}}_{t+1}|\boldsymbol{\Sigma}_t)\\&\ge L^{\mathrm{Wiesel}}(\tilde{\boldsymbol{\Sigma}}_{t+1})=L^{\mathrm{Wiesel}}(\boldsymbol{\Sigma}_{t+1}),\end{aligned}$$

which means $\{L^{\mathrm{Wiesel}}(\boldsymbol{\Sigma}_t)\}$ is a non-increasing sequence.

Assume that there exists converging subsequence $\boldsymbol{\Sigma}_{t_j}\to\boldsymbol{\Sigma}_\infty$, then

$$\begin{aligned}g(\boldsymbol{\Sigma}|\boldsymbol{\Sigma}_{t_j})&\ge g\left(\tilde{\boldsymbol{\Sigma}}_{t_j+1}|\boldsymbol{\Sigma}_{t_j}\right)\ge L^{\mathrm{Wiesel}}\left(\tilde{\boldsymbol{\Sigma}}_{t_j+1}\right)\\&=L^{\mathrm{Wiesel}}\left(\boldsymbol{\Sigma}_{t_j+1}\right)\ge L^{\mathrm{Wiesel}}\left(\boldsymbol{\Sigma}_{t_{j+1}}\right)=g\left(\boldsymbol{\Sigma}_{t_{j+1}}|\boldsymbol{\Sigma}_{t_{j+1}}\right),\\&\qquad\forall\boldsymbol{\Sigma}\succ\mathbf{0}.\end{aligned}$$

Letting $j\to+\infty$ results in

$$g(\boldsymbol{\Sigma}|\boldsymbol{\Sigma}_\infty)\ge g(\boldsymbol{\Sigma}_\infty|\boldsymbol{\Sigma}_\infty),\ \forall\boldsymbol{\Sigma}\succ\mathbf{0},$$

which implies that the directional derivative $L^{\mathrm{Wiesel}\prime}(\boldsymbol{\Sigma}_\infty;\boldsymbol{\Delta})\ge\mathbf{0},\forall\boldsymbol{\Sigma}_\infty+\boldsymbol{\Delta}\succ\mathbf{0}$. The limit $\boldsymbol{\Sigma}_\infty$ is nonsingular since if $\boldsymbol{\Sigma}_\infty$ is singular $L^{\mathrm{Wiesel}}(\boldsymbol{\Sigma}_\infty)=+\infty$, but $L^{\mathrm{Wiesel}}(\boldsymbol{\Sigma}_\infty)\le L^{\mathrm{Wiesel}}(\boldsymbol{\Sigma}_0)<+\infty$ given that $\boldsymbol{\Sigma}_0\succ\mathbf{0}$, which is a contradiction. Since $\boldsymbol{\Sigma}_\infty\succ\mathbf{0}$ and the function is continuously differentiable, we have $\frac{\partial L^{\mathrm{Wiesel}}(\boldsymbol{\Sigma})}{\partial\boldsymbol{\Sigma}}|_{\boldsymbol{\Sigma}_\infty}=\mathbf{0}$. Since $\mathrm{Tr}(\boldsymbol{\Sigma}_\infty)=1$, $\boldsymbol{\Sigma}_\infty=\hat{\boldsymbol{\Sigma}}$.

The set $\mathcal{X}^0=\{\boldsymbol{\Sigma}|L(\boldsymbol{\Sigma})\le L(\boldsymbol{\Sigma}_0)\}\cap\mathcal{S}$ is a compact set, and $\{\boldsymbol{\Sigma}_t\}$ lies in this set, hence $\{\boldsymbol{\Sigma}_t\}$ converges to $\hat{\boldsymbol{\Sigma}}$. ■

B. Regularization via Kullback-Leibler Penalty

Following the same approach, for the KL divergence penalty problem:

$$\begin{array}{ll}\underset{\boldsymbol{\Sigma}}{\text{minimize}} & \log\det(\boldsymbol{\Sigma})+\frac{K}{N}\sum_{i=1}^N\log\left(\mathbf{x}_i^T\boldsymbol{\Sigma}^{-1}\mathbf{x}_i\right)\\ & +\alpha_0(\mathrm{Tr}(\boldsymbol{\Sigma}^{-1}\mathbf{T})+\log\det(\boldsymbol{\Sigma}))\\ \text{subject to} & \boldsymbol{\Sigma}\succcurlyeq\mathbf{0}.\end{array}\quad(32)$$

We can majorize $L^{\mathrm{KL}}(\boldsymbol{\Sigma})$ at $\boldsymbol{\Sigma}_t$ by function

$$(1+\alpha_0)\log\det(\boldsymbol{\Sigma})+\frac{K}{N}\sum_{i=1}^N\frac{\mathbf{x}_i^T\boldsymbol{\Sigma}^{-1}\mathbf{x}_i}{\mathbf{x}_i^T\boldsymbol{\Sigma}_t^{-1}\mathbf{x}_i}+\alpha_0\mathrm{Tr}(\boldsymbol{\Sigma}^{-1}\mathbf{T}),\quad(33)$$

the stationary condition leads to the iteration

$$\boldsymbol{\Sigma}_{t+1}=\frac{1}{1+\alpha_0}\frac{K}{N}\sum_{i=1}^N\frac{\mathbf{x}_i\mathbf{x}_i^T}{\mathbf{x}_i^T\boldsymbol{\Sigma}_t^{-1}\mathbf{x}_i}+\frac{\alpha_0}{1+\alpha_0}\mathbf{T}.\quad(34)$$

Algorithm 2 summarizes the procedure for KL shrinkage estimator.

Algorithm 2: KL divergence penalized shrinkage estimator

1) Initialize $\boldsymbol{\Sigma}_0$ as an arbitrary positive definite matrix.
2) Iterate

$$\boldsymbol{\Sigma}_{t+1}=\frac{1}{1+\alpha_0}\frac{K}{N}\sum_{i=1}^N\frac{\mathbf{x}_i\mathbf{x}_i^T}{\mathbf{x}_i^T\boldsymbol{\Sigma}_t^{-1}\mathbf{x}_i}+\frac{\alpha_0}{1+\alpha_0}\mathbf{T}$$

until convergence.

Proposition 18: The sequence $\{\boldsymbol{\Sigma}_t\}$ generated by Algorithm 2 converges to the global minimizer of problem (32).

Proof: We verify the assumptions required for the convergence of algorithm [28], namely (25) and the compactness of initial level set $\mathcal{X}^0=\{\boldsymbol{\Sigma}|L^{\mathrm{KL}}(\boldsymbol{\Sigma})\le L^{\mathrm{KL}}(\boldsymbol{\Sigma}_0),\ \boldsymbol{\Sigma}\succ\mathbf{0}\}$.

The first condition in (25) is satisfied by construction. To verify the second condition, we see that the gradient of the surrogate function $g(\boldsymbol{\Sigma}|\boldsymbol{\Sigma}_t)$ has a unique zero. Since $g(\boldsymbol{\Sigma}|\boldsymbol{\Sigma}_t)$ is a global upperbound for $L^{\mathrm{KL}}(\boldsymbol{\Sigma})$, $g(\boldsymbol{\Sigma}|\boldsymbol{\Sigma}_t)\to+\infty$ as $\boldsymbol{\Sigma}$ goes to the boundary of $\mathbb{S}_+^K$. By the continuity of $g(\boldsymbol{\Sigma}|\boldsymbol{\Sigma}_t)$, a minimizer $\boldsymbol{\Sigma}^\star\succ\mathbf{0}$ exists and has to satisfy $\frac{\partial g}{\partial\boldsymbol{\Sigma}}=\mathbf{0}$. Therefore the unique zero has to be the global minimum, i.e., $\boldsymbol{\Sigma}_{t+1}=\arg\min_{\boldsymbol{\Sigma}\succcurlyeq\mathbf{0}}g(\boldsymbol{\Sigma}|\boldsymbol{\Sigma}_t)$. The last condition is satisfied since $L^{\mathrm{KL}}(\boldsymbol{\Sigma})$ and $g(\boldsymbol{\Sigma}|\boldsymbol{\Sigma}_t)$ are continuously differentiable on $\mathbb{S}_{++}^K$.

It is proved in Theorems 10 and 11 that on set $\mathbb{S}_{++}^K$, $L^{\mathrm{KL}}(\boldsymbol{\Sigma})$ has a unique stationary point and it is the global minimum. Furthermore, the conditions in Theorem 11 ensures $L^{\mathrm{KL}}(\boldsymbol{\Sigma})\to+\infty$ when $\boldsymbol{\Sigma}$ goes to the boundary of $\mathbb{S}_+^K$. The initial set $\mathcal{X}^0=\{\boldsymbol{\Sigma}|L^{\mathrm{KL}}(\boldsymbol{\Sigma})\le L^{\mathrm{KL}}(\boldsymbol{\Sigma}_0),\ \boldsymbol{\Sigma}\succ\mathbf{0}\}$ is compact follows easily.

Therefore the sequence $\{\boldsymbol{\Sigma}_t\}$ converges to the set of stationary points, hence the global minimum of problem (32). ■

C. Estimation With Structure Constraints

In this subsection, we briefly discuss the covariance estimation problem with structure constraints. In general, the uniqueness of the estimator cannot be guaranteed. However, algorithms can still be derived based on majorization-minimization when the constraint set $\mathcal{C}$ is convex. In this case, we can majorize the objective functions $L^{\mathrm{Wiesel}}(\boldsymbol{\Sigma})$ and $L^{\mathrm{KL}}(\boldsymbol{\Sigma})$ by

$$\begin{aligned}&g^{\mathrm{Wiesel}}(\boldsymbol{\Sigma}|\boldsymbol{\Sigma}_t)\\&=(1+\alpha_0)\mathrm{Tr}\left(\boldsymbol{\Sigma}_t^{-1}\boldsymbol{\Sigma}\right)+\frac{K}{N}\sum_{i=1}^N\frac{\mathbf{x}_i^T\boldsymbol{\Sigma}^{-1}\mathbf{x}_i}{\mathbf{x}_i^T\boldsymbol{\Sigma}_t^{-1}\mathbf{x}_i}+\frac{\alpha_0K\mathrm{Tr}\left(\boldsymbol{\Sigma}^{-1}\mathbf{T}\right)}{\mathrm{Tr}\left(\boldsymbol{\Sigma}_t^{-1}\mathbf{T}\right)}\end{aligned}$$

and

$$g^{\mathrm{KL}}(\boldsymbol{\Sigma}|\boldsymbol{\Sigma}_t) = (1+\alpha_0)\mathrm{Tr}\left(\boldsymbol{\Sigma}_t^{-1}\boldsymbol{\Sigma}\right) + \frac{K}{N}\sum_{i=1}^{N}\frac{\mathbf{x}_i^T\boldsymbol{\Sigma}^{-1}\mathbf{x}_i}{\mathbf{x}_i^T\boldsymbol{\Sigma}_t^{-1}\mathbf{x}_i} + \alpha_0\mathrm{Tr}(\boldsymbol{\Sigma}^{-1}\mathbf{T})$$

respectively, ignoring the constant term. Without any additional constraint, setting the gradient of $g(\cdot)$ to zero yields update

$$\boldsymbol{\Sigma}_{t+1} = \boldsymbol{\Sigma}_t^{\frac{1}{2}}\left(\boldsymbol{\Sigma}_t^{-\frac{1}{2}}\mathbf{M}_t\boldsymbol{\Sigma}_t^{-\frac{1}{2}}\right)^{1/2}\boldsymbol{\Sigma}_t^{\frac{1}{2}},$$

where

$$\mathbf{M}_t^{\mathrm{Wiesel}} = \frac{1}{1+\alpha_0}\frac{K}{N}\sum_{i=1}^{N}\frac{\mathbf{x}_i\mathbf{x}_i^T}{\mathbf{x}_i^T\boldsymbol{\Sigma}_t^{-1}\mathbf{x}_i} + \frac{\alpha_0}{1+\alpha_0}\frac{K\mathbf{T}}{\mathrm{Tr}\left(\boldsymbol{\Sigma}_t^{-1}\mathbf{T}\right)}$$

and

$$\mathbf{M}_t^{\mathrm{KL}} = \frac{1}{1+\alpha_0}\frac{K}{N}\sum_{i=1}^{N}\frac{\mathbf{x}_i\mathbf{x}_i^T}{\mathbf{x}_i^T\boldsymbol{\Sigma}_t^{-1}\mathbf{x}_i} + \frac{\alpha_0}{1+\alpha_0}\mathbf{T}.$$

Notice that $\mathbf{M}_t$ is exactly the update we derived by only majorizing the $\log\mathrm{Tr}(\cdot)$ terms in the previous subsection, and $\boldsymbol{\Sigma}_{t+1}$ is the geometric mean between matrices $\boldsymbol{\Sigma}_t$ and $\mathbf{M}_t$ [29]. Intuitively $\boldsymbol{\Sigma}_{t+1}$ can be viewed as a smoothed update of $\boldsymbol{\Sigma}_t$.

However, when constrained, a closed-form solution for $\boldsymbol{\Sigma}_{t+1}$ cannot be obtained in general. The surrogate function $g(\cdot)$ is convex since $\mathrm{Tr}(\boldsymbol{\Sigma}_t^{-1}\boldsymbol{\Sigma})$ is linear and $\mathrm{Tr}(\boldsymbol{\Sigma}^{-1}\mathbf{T})$ is convex, $\boldsymbol{\Sigma}_{t+1} = \arg\min_{\boldsymbol{\Sigma}\in\mathcal{C}} g(\boldsymbol{\Sigma}|\boldsymbol{\Sigma}_t)$ can be found numerically if $\mathcal{C}$ is convex. We consider two such examples.

1) Covariance Matrix With Toeplitz Structure: Toeplitz structure arises frequently in various signal processing related fields. For example, in time series analysis, the autocovariance matrix of a stationary process is Toeplitz. Imposing the Toeplitz structure on $\boldsymbol{\Sigma}$ we need to solve

$$\begin{aligned} \underset{\boldsymbol{\Sigma}}{\text{minimize}} \quad & g(\boldsymbol{\Sigma}|\boldsymbol{\Sigma}_t) \\ \text{subject to} \quad & \boldsymbol{\Sigma} \succcurlyeq \mathbf{0} \\ & \boldsymbol{\Sigma}_{ij} = \boldsymbol{\Sigma}_{i+1,j+1},\ \forall i,j = 1,\ldots,K-1 \end{aligned}$$

for each iteration. The additional constraint is linear.

2) Linear Additive Structure: Suppose $\boldsymbol{\Sigma}$ can be decomposed as $\boldsymbol{\Sigma} = \mathbf{S} + \mathrm{diag}(\sigma_1,\ldots,\sigma_K)$, where $\mathbf{S} \succcurlyeq \mathbf{0}$ is signal covariance and $\mathrm{diag}(\sigma_1,\ldots,\sigma_K)$ with $\sigma_i \in \mathcal{I}_i$ is noise covariance restricted to some interval. Then, at each iteration we solve

$$\begin{aligned} \underset{\boldsymbol{\Sigma},\mathbf{S},\{\sigma_i\}}{\text{minimize}} \quad & g(\boldsymbol{\Sigma}|\boldsymbol{\Sigma}_t) \\ \text{subject to} \quad & \boldsymbol{\Sigma} \succcurlyeq \mathbf{0} \\ & \mathbf{S} \succcurlyeq \mathbf{0} \\ & \boldsymbol{\Sigma} = \mathbf{S} + \mathrm{diag}(\sigma_1,\ldots,\sigma_K) \\ & \sigma_i \in \mathcal{I}_i. \end{aligned}$$

The additional constraint is convex.

D. Parameter Tuning

A crucial issue in regularized covariance estimator is to choose the penalty parameter α_0. We have shown that if the population distribution is continuous, for both Wiesel's penalty and KL divergence penalty, we require $\alpha_0 > \frac{K}{N} - 1$ to guarantee the existence of the regularized estimator.

There is a rich literature discussing the rules of parameter tuning developed for specific estimators. A standard way is to select α_0 by cross-validation. A method based on random matrix theory has also been investigated in a recent paper [30].

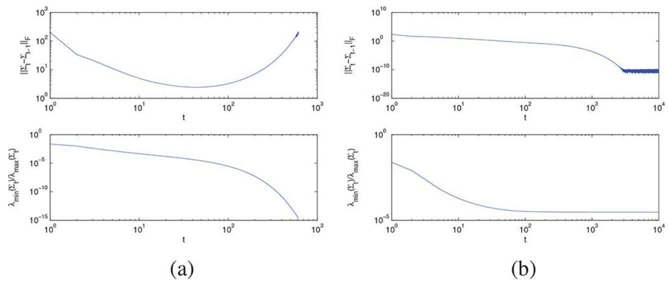


Fig. 1. Algorithm convergence of Wiesel's shrinkage estimator: (a) when the existence conditions are not satisfied with $\alpha_0 = 0.24$, and (b) when the existence conditions are satisfied with $\alpha_0 = 0.26$.

V. Numerical Results

In all of the simulations, the estimator performance is evaluated according to the criteria in [13], namely, the normalized mean-square error

$$\mathrm{NMSE} = \frac{E\left(\left\|\hat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}^{\mathrm{true}}\right\|_F^2\right)}{\left\|\boldsymbol{\Sigma}^{\mathrm{true}}\right\|_F^2}$$

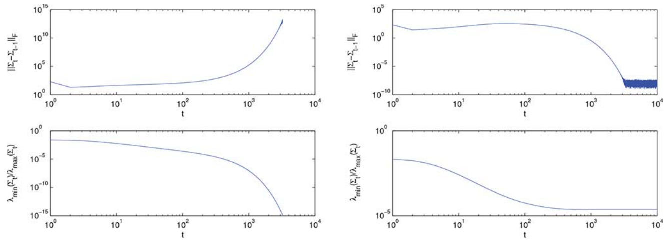
where all matrices $\boldsymbol{\Sigma}$ are all normalized by their trace. The expected value is approximated by 100 times Monte-Carlo simulations.

The first two simulations aims at illustrating the existence conditions for both Wiesel's shrinkage estimator and KL shrinkage estimator. We choose $N = 8$ and $K = 10$ with the samples drawn a Gaussian distribution $\mathcal{N}(\mathbf{0},\boldsymbol{\Sigma}_0)$, where $\boldsymbol{\Sigma}_0$ is a randomly generated positive definite covariance matrix. The shrinkage target $\mathbf{T}$ is also an arbitrary positive definite matrix. According to the result in Section III, $\alpha_0 > \frac{K}{N} - 1$, i.e., $\alpha_0 > 0.25$, is the necessary and sufficient condition for the existence of a positive definite estimator. We simulate two scenarios with $\alpha_0 = 0.24$ and 0.26. Fig. 1 plots $\|\boldsymbol{\Sigma}_t - \boldsymbol{\Sigma}_{t-1}\|_F$ and the inverse of the condition number, namely $\frac{\lambda_{\min}(\boldsymbol{\Sigma}_t)}{\lambda_{\max}(\boldsymbol{\Sigma}_t)}$, as a function of the number of iterations in log-scale for Wiesel's shrinkage estimator and with $\alpha_0 = 0.24$ (left) and $\alpha_0 = 0.26$ (right) respectively. Fig. 1 shows that for Wiesel's shrinkage estimator, when $\alpha_0 = 0.24$ $\boldsymbol{\Sigma}_t$ diverges, and when $\alpha_0 = 0.26$ $\boldsymbol{\Sigma}_t$ converges to a nonsingular limit. Fig. 2 shows similar situation happens for KL shrinkage estimator.

For the rest of the simulations, the shrinkage parameter α_0 is selected by grid search. That is, we define $\rho = \frac{1}{1+\alpha_0}$ and enumerate ρ uniformly on interval (0,1], and select the ρ (equivalently α_0) that gives the smallest error.

Fig. 3 demonstrates the performance of shrinkage Tyler's estimator in the sample deficient case. The tuning parameter is selected to be the one that yields the smallest NMSE for each estimator as proposed in [13]. We choose the example

$$\boldsymbol{\Sigma}(\beta)_{ij} = \beta^{|i-j|}$$



(a) (b)

Fig. 2. Algorithm convergence of KL shrinkage estimator: (a) when the existence conditions are not satisfied with $\alpha_0 = 0.24$, and (b) when the existence conditions are satisfied with $\alpha_0 = 0.26$.

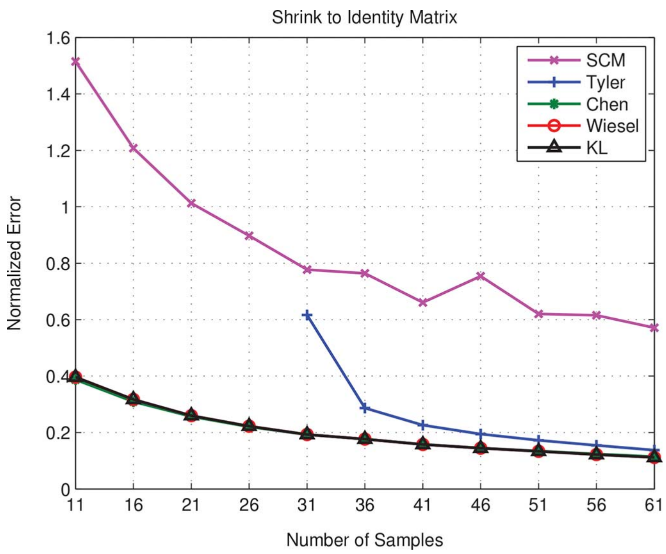


Fig. 3. Illustration of the benefit of shrinkage estimators with $K = 30$ and shrinkage target matrix $\mathbf{I}$.

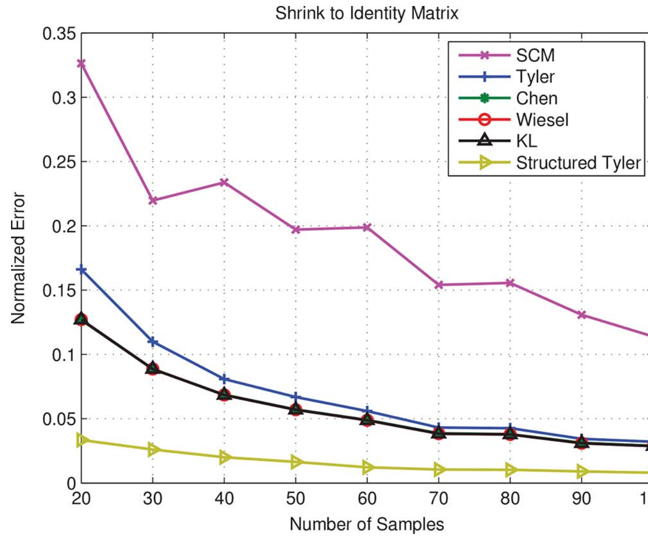


Fig. 4. Illustration of the benefit of shrinkage estimators with $K = 10$ and shrinkage target matrix $\mathbf{I}$.

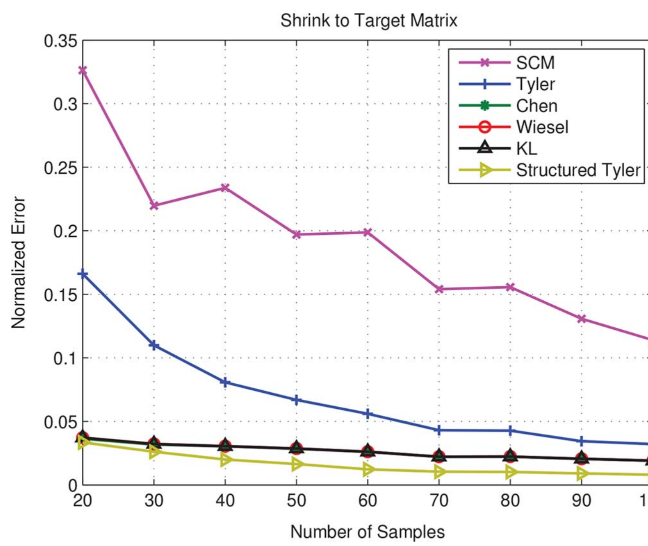


Fig. 5. Illustration of the benefit of shrinkage estimators with $K = 10$ and a knowledge-aided shrinkage target matrix $\mathbf{T}$.

with $K = 30$. In this simulation, the underlying distribution is chosen to be a Student's t-distribution with parameters $\boldsymbol{\mu}_0 = \mathbf{0}$, $\boldsymbol{\Sigma}_0 = \boldsymbol{\Sigma}(0.8)$, and $\nu = 3$, and the shrinkage target is set to be an identity matrix. The number of samples N starts from 11 to 61. The curve corresponding to Tyler's estimator starts at $N = 31$ since the condition for Tyler's estimator to exist is $N > K$, i.e., $N > 30$ in this case. The figure illustrates that both Tyler's estimator and shrinkage Tyler's estimator outperform the sample covariance matrix when all of them exist, shrinkage estimators exist even when $N \leq K$ and, moreover, achieve the best performance in all cases.

Figs. 4 and 5 compare the performance of different shrinkage Tyler's estimators following roughly one of the simulation set-up in [13] for a fair comparison. The samples are drawn from a Student's t-distribution with parameters $\boldsymbol{\mu}_0 = \mathbf{0}$, $\boldsymbol{\Sigma}_0 = \boldsymbol{\Sigma}(0.8)$ and $\nu = 3$. The number of samples N varies from 20 to 100. Fig. 4 shows the estimation error when setting $\mathbf{T} = \mathbf{I}$ and Fig. 5 shows that when setting $\mathbf{T} = \boldsymbol{\Sigma}(0.7)$, the searching step size of ρ is set to be 0.01. The result indicates that estimation accuracy is increased due to shrinkage when the number of sample is not enough. Wiesel's shrinkage estimator and KL shrinkage estimator yield the same NMSE. Interestingly, Chen's shrinkage estimator gives roughly the same NMSE, although with a different shrinkage parameter α_0. Chen's and KL shrinkage estimator thus find their advantage in practice since an easier way of choosing α_0 rather than cross-validation has been investigated in the literature [12], [30], a detailed comparison of them from random matrix theory perspective has also been provided in [30].

In both of the simulations, we include Tyler's estimator with a Toeplitz structure constraint as introduced in the previous section. The figures show that the structure constrained estimator achieves relatively better performance than all other estimators both when shrinking to $\mathbf{I}$ and shrinking to $\mathbf{T}$. Although structure constraint can be imposed on shrinkage estimators to achieve potentially even smaller estimation error, we leave out this simulation due to the heavy computational cost introduced both by a lack of a closed-form solution per iteration (a SDP need to be solved numerically) and grid searching for the best regularization parameter. The problem of accelerating the algorithm and investigating the effect of imposing structure constraint on shrinkage estimator are left for future work.

Finally, the performance of Tyler's estimator is tested on a real financial data set. We choose daily close prices p_t from Jan 1, 2008 to July 31, 2011, 720 days in total, of $K = 45$ stocks from the Hang Seng Index provided by Yahoo Finance. The samples are constructed as $r_t = \log p_t - \log p_{t-1}$, i.e., the daily log-returns. The process r_t is assumed to be stationary. The vector $\mathbf{r}_t$ is constructed by stacking the log-returns of all K stocks. $\mathbf{r}_t$ that is close to $\mathbf{0}$ (all elements are less than 10^{-6}) is discarded. We compare the performance of different covariance estimators in the minimum variance portfolio set up, that is, we allocate the portfolio weights to minimize the overall risk. The problem can be formulated formally as

$$\begin{array}{ll} \underset{\mathbf{w}}{\text{minimize}} & \mathbf{w}^T \boldsymbol{\Sigma} \mathbf{w} \\ \text{subject to} & \mathbf{1}^T \mathbf{w} = 1 \end{array} \tag{35}$$

with $\boldsymbol{\Sigma}$ being the covariance matrix of $\mathbf{r}_t$. Clearly the scaling of $\boldsymbol{\Sigma}$ does not affect the solution to this problem.

The simulation takes the following procedure. At day t, we use the previous N $\mathbf{r}_t$'s as samples to estimate the covariance matrix. The samples are split into two parts as $N = N^{\text{train}} + N^{\text{val}}$. For nonshrinkage estimators, at day t, we take the $\mathbf{r}_i$'s with $i \in [t - N^{\text{train}} - N^{\text{val}}, t-1]$ as samples to estimate the normalized covariance matrix $\boldsymbol{\Sigma}$. For a particular shrinkage estimator, the target matrix is set to be $\mathbf{I}$ and the tuning parameter ρ is chosen as follows: for each value of $\rho \in \{0.01, 0.02, \ldots, 1\}$, we calculate the shrinkage estimator $\boldsymbol{\Sigma}^\rho$ with samples $\mathbf{r}_i$, $i \in [t - N^{\text{train}} - N^{\text{val}}, t - N^{\text{val}} - 1]$ and the corresponding $\mathbf{w}^\rho$ by solving (35). We then take the $\mathbf{r}_i$'s with $i \in [t - N^{\text{val}}, t-1]$ as validation data and evaluate the variance of portfolio series $\{(\mathbf{w}^\rho)^T \mathbf{r}_i\}$ in this period, the best $\rho^\star$ is chosen to be the one that yields the smallest variance. Finally the shrinkage estimator is obtained using samples $\mathbf{r}_i$ with $i \in [t - N^{\text{train}} - N^{\text{val}}, t-1]$ and tuning parameter $\rho^\star$. With the allocation strategy $\mathbf{w}$ for each of the estimators as the solution to (35), we construct portfolio for the next N^{test} days and collect the returns. The procedure is repeated every N^{test} days till the end and the variance of the portfolio constructed based on different estimators is calculated.

In the simulation, we choose $N^{\text{val}} = N^{\text{test}} = 10$ and vary N^{train} from 70 to 100. Fig. 6 compares the variance (risk) of portfolio constructed based on different estimators, with one additional baseline portfolio constructed by equal investment in each asset. From the figure we can see shrinkage estimators achieves relatively better performance than the nonshrinkage ones.

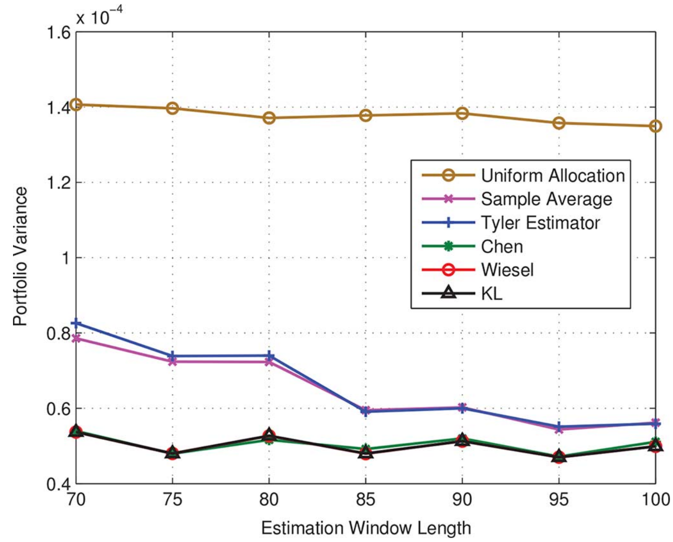


Fig. 6. Comparison of portfolio risk constructed based on different covariance estimators.

VI. Conclusion

In this work, we have given a rigorous proof for the existence and uniqueness of the regularized Tyler's estimator proposed in [13], and justified the heuristic diagonal loading shrinkage estimator in [10] with the KL divergence. Under the condition that samples are reasonably spread out, i.e., $P_N(S) < \frac{(1+\alpha_0)\dim(S)}{K}$, or $N > \frac{K}{1+\alpha_0}$ if the underlying distribution is continuous, the estimators have been shown to exist and be unique (up to a positive scale factor for Wiesel's estimator). Algorithms based on the majorization-minimization framework have also been provided with guaranteed convergence. Finally we have discussed structure constrained estimation and have shown via simulation that imposing such constraint helps improving estimation accuracy.

Appendix

A. Proof for Theorem 3

For the loss function

$$\tilde{L}(\boldsymbol{\Sigma}) = \frac{N}{2} \log\det(\boldsymbol{\Sigma}) + \sum\nolimits_{i=1}^{N} \rho\left(\mathbf{x}_i^T \boldsymbol{\Sigma}^{-1} \mathbf{x}_i\right) + \left(\alpha \log\det(\boldsymbol{\Sigma}) + \sum\nolimits_{l=1}^{L} \alpha_l h_l\left(\mathrm{Tr}\left(\mathbf{A}_l^T \boldsymbol{\Sigma}^{-1} \mathbf{A}_l\right)\right)\right)$$

where the regularization term is written in general as $\alpha \log\det(\boldsymbol{\Sigma}) + \sum_{l=1}^{L} \alpha_l h_l(\mathrm{Tr}(\mathbf{A}_l^T \boldsymbol{\Sigma}^{-1} \mathbf{A}_l))$. Define a_ρ, $a_\rho^{'}$ for $\rho(s)$ and a_l, $a_l^{'}$ for $\alpha_l h_l$ as in Definition 2.

Define function

$$G(\boldsymbol{\Sigma}) = \exp\{-\tilde{L}(\boldsymbol{\Sigma})\} = \det(\boldsymbol{\Sigma})^{-\frac{N}{2}-\alpha} \prod_i g_\rho\left(\mathbf{x}_i^T \boldsymbol{\Sigma}^{-1} \mathbf{x}_i\right) \cdot \prod g_l \left(\sum_j \lambda_j^{-1} \|\tilde{\mathbf{a}}_{lj}\|^2\right)$$

where $\tilde{\mathbf{a}}_{lj}$ is defined as the jth row of $\tilde{\mathbf{A}}_l = \mathbf{U}^T \mathbf{A}_l$ with $\mathbf{U}$ being the unitary matrix such that $\mathbf{U}\boldsymbol{\Lambda}\mathbf{U}^T = \boldsymbol{\Sigma}$, $\boldsymbol{\Lambda} = \mathrm{diag}(\lambda_1, \ldots, \lambda_K)$, and $g_\rho(s) = \exp\{-\rho(s)\}$, $g_l(s) = \exp\{-\alpha_l h_l(s)\}$. The eigenvalues λ_j is arranged in descending order, i.e., $\lambda_1 \geq \cdots \geq \lambda_K$, and denote the inverse of λ as φ, hence $\varphi_1 \leq \cdots \leq \varphi_K$.

Denote the eigenvectors corresponding to λ_j as $\mathbf{u}_j$, the subspace spanned by $\{\mathbf{u}_1, \ldots, \mathbf{u}_j\}$ as S_j and $D_j = S_j \backslash S_{j-1} = \{\mathbf{x} \in \mathbb{R}^K | \mathbf{x} \in S_j, \mathbf{x} \notin S_{j-1}\}$ with $S_0 = \{0\}$ and $D_0 = \{0\}$. By definition, D_j, $j = 0, \ldots, K$ partition the whole $\mathbb{R}^K$

space. Notice that $P_N\{S_0\} = 0$ by the assumption that no $\mathbf{x}_i$ lies on the origin, we have $\sum_{j=1}^{m} P_N(D_j) = P_N(S_m)$ and $\sum_{j=m}^{K} P_N(D_j) = 1 - P_N(S_{m-1})$.

Partition the samples $\mathbf{x}_i$ according to D_j, $j=0$ is excluded hereafter, define function

$$G_j = \begin{cases} \lambda_j^{-\frac{N}{2}-\alpha} \prod_{\mathbf{x}_i \in D_j} g_\rho\left(\mathbf{x}_i^T \boldsymbol{\Sigma}^{-1}\mathbf{x}_i\right), \text{ if } \exists \mathbf{x}_i \in D_j \\ \lambda_j^{-\frac{N}{2}-\alpha}, \text{ if no } \mathbf{x}_i \in D_j \end{cases}$$

and we have $G(\boldsymbol{\Sigma}) = \prod_{j=1}^{K} G_j(\boldsymbol{\Sigma}) \prod g_l(\sum_j \lambda_j^{-1}\|\tilde{\mathbf{a}}_{lj}\|^2)$.

For the $\mathbf{A}_l$'s, denote $\mathbf{A}_l = [\mathbf{a}_{l1}, \mathbf{a}_{l2}, \ldots, \mathbf{a}_{lp}]$. For each $\mathbf{a}_l$, there exists some D_j such that $\mathbf{a}_l \in D_j$, since the D_j's partition the whole space. Define q_l to be the maximum index of D_j that the $\mathbf{a}_l$'s belongs to. Therefore we have $\|\tilde{\mathbf{a}}_{lq_l}\| \neq 0$ and $\|\tilde{\mathbf{a}}_{lj}\| = 0$ for $j > q_l$.

We analyze the behavior of $G(\boldsymbol{\Sigma})$ at the boundary of its feasible set $\mathbb{S}_{++}^K$, by Lemma 1, we only need to ensure $G(\boldsymbol{\Sigma}) \to 0$, then there exists $\tilde{L}(\hat{\boldsymbol{\Sigma}}) \leq \tilde{L}(\boldsymbol{\Sigma})$, $\forall \boldsymbol{\Sigma} \succ \mathbf{0}$, and $\hat{\boldsymbol{\Sigma}} \succ \mathbf{0}$.

Consider the general case that some of the λ_j's go to zero, some remains bounded away from both 0 and positive infinity, and the rests tend to positive infinity. Formally, define two integers r and s that $1 \leq r \leq s \leq K$, such that $\lambda_j \to +\infty$ for $j \in [1, r]$, λ_j is bounded for $j \in (r, s]$ and $\lambda_j \to 0$ for $j \in (s, K]$. Denote some arbitrary small positive quantity by ϵ.

First we analyze the terms G_j with $\lambda_j \to 0$. Consider the samples $\mathbf{x}_i \in D_h$ for some $h \in (s, K]$, then $\mathbf{x}_i^T \boldsymbol{\Sigma}^{-1}\mathbf{x}_i = \sum_{j=1}^{h} \lambda_j^{-1}\|\mathbf{u}_j^T\mathbf{x}_i\|^2 \geq \lambda_h^{-1}\|\mathbf{u}_h^T\mathbf{x}_i\|^2$, which is $+\infty > (\mathbf{x}_i^T\boldsymbol{\Sigma}^{-1}\mathbf{x}_i)\lambda_h > 0$. Since $\lambda_h \to 0$, we have $\mathbf{x}_i^T\boldsymbol{\Sigma}^{-1}\mathbf{x}_i \to +\infty$. By definition

$$\begin{aligned} &\lim_{\lambda_h \to 0} g_\rho\left(\mathbf{x}_i^T\boldsymbol{\Sigma}^{-1}\mathbf{x}_i\right)\left(\mathbf{x}_i^T\boldsymbol{\Sigma}^{-1}\mathbf{x}_i\right)^{(a_\rho-\epsilon)/2} \\ &= \lim_{\lambda_h \to 0}\left\{g_\rho\left(\mathbf{x}_i^T\boldsymbol{\Sigma}^{-1}\mathbf{x}_i\right)\lambda_h^{-(a_\rho-\epsilon)/2}\right\} \\ &\quad \cdot \left\{\left(\left(\mathbf{x}_i^T\boldsymbol{\Sigma}^{-1}\mathbf{x}_i\right)\lambda_h\right)^{(a_\rho-\epsilon)/2}\right\} \\ &= 0 \end{aligned}$$

which implies $\lim_{\lambda_h \to 0}\{g_\rho(\mathbf{x}_i^T\boldsymbol{\Sigma}^{-1}\mathbf{x}_i)\lambda_h^{-(a_\rho-\epsilon)/2}\} = 0$, i.e., $g_\rho(\mathbf{x}_i^T\boldsymbol{\Sigma}^{-1}\mathbf{x}_i) = o(\lambda_h^{\frac{a_\rho-\epsilon}{2}})$. Therefore, if $\mathbf{x}_i \in D_j$, $g_\rho(\mathbf{x}_i^T\boldsymbol{\Sigma}^{-1}\mathbf{x}_i) = o(\lambda_j^{\frac{a_\rho-\epsilon}{2}})$. For each G_j we have

$$G_j = o\left(\lambda_j^{\frac{a_\rho-\epsilon}{2}\cdot NP_N(D_K) - \frac{N}{2} - \alpha - \epsilon}\right) \; \forall j \geq s+1.$$

In the second step, we analyze the terms G_j with $\lambda_j \to +\infty$. Consider the samples $\mathbf{x}_i \in D_h$ for some $h \in [1, r]$, we have shown that $0 < (\mathbf{x}_i^T\boldsymbol{\Sigma}^{-1}\mathbf{x}_i)\lambda_h < +\infty$. Since $\lambda_h \to +\infty$, $(\mathbf{x}_i^T\boldsymbol{\Sigma}^{-1}\mathbf{x}_i) \to 0$. Given that $a_\rho' > -\infty$, $g_\rho(\mathbf{x}_i^T\boldsymbol{\Sigma}^{-1}\mathbf{x}_i) = o(\varphi_h^{-\frac{a_\rho'+\epsilon}{2}})$ by

$$\lim_{\varphi_h \to 0} g_\rho(\mathbf{x}_i^T\boldsymbol{\Sigma}^{-1}\mathbf{x}_i)\left(\mathbf{x}_i^T\boldsymbol{\Sigma}^{-1}\mathbf{x}_i\right)^{\left(a_\rho'+\epsilon\right)/2} = 0.$$

Therefore for each G_j we have

$$G_j = o\left(\varphi_j^{\frac{N}{2}+\alpha-\frac{a_\rho'+\epsilon}{2}NP_N(D_j)-\epsilon}\right) \; \forall j \leq r.$$

For the G_j with λ_j being some constant, it is easy to see that $g_\rho(\mathbf{x}_i^T\boldsymbol{\Sigma}^{-1}\mathbf{x}_i) = O(1)$, which does not affect the order of $G(\boldsymbol{\Sigma})$.

Now we have characterized the g_ρ's, we move to the g_l's. Since $\|\tilde{\mathbf{a}}_{lq_l}\| \neq 0$ and $\|\tilde{\mathbf{a}}_{lj}\| = 0$ for $j > q_l$, by the same reasoning above, $g_l = o(\varphi_{q_l}^{-\frac{a_l'+\epsilon}{2}})$ if $q_l \leq r$ and $g_l = o(\lambda_{q_l}^{\frac{a_l-\epsilon}{2}})$ if $q_l \geq s+1$. Therefore

$$\begin{aligned} G(\boldsymbol{\Sigma}) &= \prod_{j=1}^{K} G_j(\boldsymbol{\Sigma}) \prod g_l\left(\sum_j \lambda_j^{-1}\|\tilde{\mathbf{a}}_{lj}\|^2\right) \\ &= \prod_{j=1}^{r} o\left(\varphi_j^{\frac{N}{2}+\alpha-\frac{a_\rho'+\epsilon}{2}NP_N(D_j)-\epsilon}\right) \\ &\quad \cdot \prod_{j=s+1}^{K} o\left(\lambda_j^{\frac{a_\rho-\epsilon}{2}\cdot NP_N(D_j)-\frac{N}{2}-\alpha-\epsilon}\right) \\ &\quad \cdot \prod_{\{l|q_l \geq s+1\}} o\left(\lambda_{q_l}^{\frac{a_l-\epsilon}{2}}\right) \prod_{\{l|q_l \leq r\}} o\left(\varphi_{q_l}^{-\frac{a_l'+\epsilon}{2}}\right) \end{aligned}$$

with $\prod_{\{l|q_l \geq s+1\}}$ defined to be 1 if the set $\{l|q_l \geq s+1\}$ is empty, and the same for $\prod_{\{l|q_l \leq r\}}$.

We make the following assumption:

$$\begin{aligned} &\left(\frac{N}{2}+\alpha-\epsilon\right)m - \frac{a_\rho'+\epsilon}{2}N\sum\nolimits_{j=1}^{m} P_N(D_j) - \sum\nolimits_{q_l \leq m}\frac{a_l'+\epsilon}{2} \\ &\quad \geq 0, \; \forall 1 \leq m \leq r, \\ &\quad \frac{a_\rho}{2}\cdot N\sum\nolimits_{j=m}^{K} P_N(D_j) - \left(\frac{N}{2}+\alpha+\epsilon\right)(K-m+1) \\ &\quad + \sum\nolimits_{q_l \geq m}\frac{a_l-\epsilon}{2} \geq 0, \; \forall K \geq m \geq s+1 \end{aligned} \tag{36}$$

by the order $\lambda_1 \geq \cdots \geq \lambda_K$ hence $\varphi_1 \leq \cdots \leq \varphi_K$, and base on the fact that

$$\begin{aligned} o\left(\lambda_1^{\alpha_1}\right)o\left(\lambda_2^{\alpha_2}\right) &= o\left(\lambda_1^{\alpha_1+\alpha_2}\right) \; if \; \alpha_2 \geq 0 \\ o\left(\varphi_1^{\alpha_1}\right)o\left(\varphi_2^{\alpha_2}\right) &= o\left(\varphi_2^{\alpha_1+\alpha_2}\right) \; if \; \alpha_1 \geq 0 \end{aligned}$$

the order of $G(\boldsymbol{\Sigma})$ is

$$\begin{aligned} G(\boldsymbol{\Sigma}) &= o\left(\varphi_r^{\left(\frac{N}{2}+\alpha-\epsilon\right)r - \frac{a_\rho'+\epsilon}{2}N\sum_{j=1}^{r}P_N(D_j) - \sum_{q_l \leq r}\frac{a_l'+\epsilon}{2}}\right) \\ &\cdot o\left(\lambda_{s+1}^{\frac{a_\rho-\epsilon}{2}\cdot N\sum_{j=s+1}^{K}P_N(D_j) - \left(\frac{N}{2}+\alpha+\epsilon\right)(K-s) + \sum_{q_l \geq s+1}\frac{a_l-\epsilon}{2}}\right) \end{aligned}$$

and it goes to zero.

Now we simplify the assumption (36). Since $\sum_{j=1}^{m} P_N(D_j) = P_N(S_m)$ and $\sum_{j=m}^{K} P_N(D_j) = 1 - P_N(S_{m-1})$, and r, s can take any value that satisfies $1 \leq r \leq s < K$, we end up with the following condition:

$$\begin{aligned} &\left(\frac{N}{2}+\alpha-\epsilon\right)d - \frac{a_\rho'+\epsilon}{2}NP_N(S_d) - \sum_{q_l \leq d}\frac{a_l'+\epsilon}{2} \geq 0, \\ &\frac{a_\rho-\epsilon}{2}\cdot N\left(1-P_N\left(S_d\right)\right) - \left(\frac{N}{2}+\alpha+\epsilon\right)(K-d) \\ &\quad + \sum_{q_l \geq d+1}\frac{a_l-\epsilon}{2} \geq 0, \end{aligned}$$

for all $1 \leq d \leq K-1$.

Define sets $\omega = \{l | q_l \leq d\}$ and $\upsilon = \{l | q_l > d\}$, consider when $l \in \omega$, which means $q_l \leq d$, by the definition of q_l, is equivalent to $\mathrm{range}(\mathbf{A}_l) \subseteq S_d$, similarly for $l \in \upsilon$, which means $q_l > d$, is equivalent to $\mathrm{range}(\mathbf{A}_l) \nsubseteq S_d$.

The condition should be valid for any $\mathbf{U}$ and $1 \leq d \leq K-1$, tidy up the expression and let $\epsilon \to 0$ results in: for any proper subspace S

$$P_N(S) < \min\left\{1 - \frac{(N+2\alpha)(K - \dim(S)) - \sum_{l\in\upsilon} a_l}{a_\rho N}, \frac{(N+2\alpha)\dim(S) - \sum_{l\in\omega} a_l^{'}}{a_\rho^{'} N}\right\}$$

where sets ω and υ are defined as $\omega = \{l | \mathrm{range}(\mathbf{A}_l) \subseteq S\}$, $\upsilon = \{l | \mathrm{range}(\mathbf{A}_l) \nsubseteq S\}$.

For the case $r = 0$, $1 \leq s < K$, which means no $\lambda \to +\infty$ and some, not all $\lambda \to 0$, following the same reasoning gives condition

$$P_N(S) < 1 - \frac{(N+2\alpha)(K - \dim(S)) - \sum_{l\in\upsilon} a_l}{a_\rho N}$$

and for $s = K$, $1 \leq r < K$, which means no $\lambda \to 0$ and some, not all $\lambda \to +\infty$, gives condition

$$P_N(S) < \frac{(N+2\alpha)\dim(S) - \sum_{l\in\omega} a_l^{'}}{a_\rho^{'} N}.$$

Notice that the above two conditions are included in the first one.

And finally under the scenario that all $\lambda \to +\infty$, it's easy to see $G(\mathbf{\Sigma}) = o(\varphi_K^{(\frac{N}{2}+\alpha-\epsilon)K - \frac{a_\rho^{'}}{2}N - \frac{1}{2}\sum_l (a_l^{'}+\epsilon)})$ goes to zero if $(-\frac{N}{2} - \alpha)K + \frac{a_\rho}{2}N + \frac{1}{2}\sum_l a_l^{'} < 0$, and under the case that all $\lambda \to 0$, $G(\mathbf{\Sigma}) = o(\lambda_1^{\frac{a_\rho}{2}\cdot N - (\frac{N}{2}+\alpha+\epsilon)K + \frac{1}{2}\sum_l (a_l - \epsilon)})$ goes to zero if $\frac{a_\rho}{2} \cdot N - (\frac{N}{2} + \alpha)K + \frac{1}{2}\sum_l a_l > 0$.

REFERENCES

[1] R. Maronna, D. Martin, and V. Yohai, *Robust Statistics: Theory and Methods, Ser. Wiley Series in Probability and Statistics*. New York, NY, USA: Wiley, 2006.

[2] P. Huber, *Robust Statistics, Ser. Wiley Series in Probability and Statistics-Applied Probability and Statistics Section Series*. New York, NY, USA: Wiley, 2004.

[3] D. E. Tyler, "A distribution-free M-estimator of multivariate scatter," *Ann. Statist.*, vol. 15, no. 1, pp. 234–251, 03, 1987.

[4] E. Ollila, H. Oja, and V. Koivunen, "Complex-valued ICA based on a pair of generalized covariance matrices," *Comput. Statist. Data Anal.*, vol. 52, no. 7, pp. 3789–3805, 2008.

[5] E. Ollila and V. Koivunen, "Influence function and asymptotic efficiency of scatter matrix based array processors: Case MVDR beamformer," *IEEE Trans. Signal Process.*, vol. 57, no. 1, pp. 247–259, Jan. 2009.

[6] E. Ollila and D. Tyler, "Distribution-free detection under complex elliptically symmetric clutter distribution," in *Proc. IEEE 7th Sensor Array Multichannel Signal Process. Workshop (SAM)*, Hoboken, NJ, USA, Jun. 2012, pp. 413–416.

[7] F. Pascal, Y. Chitour, J. Ovarlez, P. Forster, and P. Larzabal, "Covariance structure maximum-likelihood estimates in compound gaussian noise: Existence and algorithm analysis," *IEEE Trans. Signal Process.*, vol. 56, no. 1, pp. 34–48, Jan. 2008.

[8] A. Wiesel, "Geodesic convexity and covariance estimation," *IEEE Trans. Signal Process.*, vol. 60, no. 12, pp. 6182–6189, Dec. 2012.

[9] E. Ollila, D. Tyler, V. Koivunen, and H. Poor, "Complex elliptically symmetric distributions: Survey, new results and applications," *IEEE Trans. Signal Process.*, vol. 60, no. 11, pp. 5597–5625, Nov. 2012.

[10] Y. Abramovich and N. Spencer, "Diagonally loaded normalised sample matrix inversion (LNSMI) for outlier-resistant adaptive filtering," in *Proc. IEEE Int Conf. Acoust., Speech, Signal Process. (ICCASP)*, Honolulu, HI, USA, Apr. 2007, vol. 3, pp. III-1105–III-1108.

[11] O. Ledoit and M. Wolf, "A well-conditioned estimator for large-dimensional covariance matrices," *J. Multivar. Anal.*, vol. 88, no. 2, pp. 365–411, 2004.

[12] Y. Chen, A. Wiesel, and A. Hero, "Robust shrinkage estimation of highdimensional covariance matrices," *IEEE Trans. Signal Process.*, vol. 59, no. 9, pp. 4097–4107, Sep. 2011.

[13] A. Wiesel, "Unified framework to regularized covariance estimation in scaled Gaussian models," *IEEE Trans. Signal Process.*, vol. 60, no. 1, pp. 29–38, Jan. 2012.

[14] F. Pascal, Y. Chitour, and Y. Quek, "Generalized robust shrinkage estimator and its application to STAP detection problem," 2013 [Online]. Available: http://arxiv.org/abs/1311.6567

[15] R. A. Maronna, "Robust M-estimators of multivariate location and scatter," *Ann. Statist.*, vol. 4, no. 1, pp. 51–67, 1976.

[16] P. L. Davies, "Asymptotic behaviour of S-estimates of multivariate location parameters and dispersion matrices," *Ann. Statist.*, vol. 15, no. 3, pp. 1269–1292, 1987.

[17] S. Van Aelst and P. Rousseeuw, "Minimum volume ellipsoid," *Wiley Interdiscip. Rev. Comput. Statist.*, vol. 1, no. 1, pp. 71–82, 2009.

[18] R. W. Butler, P. L. Davies, and M. Jhun, "Asymptotics for the minimum covariance determinant estimator," *Ann. Statist.*, vol. 21, no. 3, pp. 1385–1400, 09, 1993.

[19] P. J. Huber, "Robust estimation of a location parameter," *Ann. Math. Statist.*, vol. 35, no. 1, pp. 73–101, 03, 1964.

[20] D. E. Tyler, "Statistical analysis for the angular central Gaussian distribution on the sphere," *Biometrika*, vol. 74, no. 3, pp. 579–589, 1987.

[21] J. T. Kent and D. E. Tyler, "Maximum likelihood estimation for the wrapped cauchy distribution," *J. Appl. Statist.*, vol. 15, no. 2, pp. 247–254, 1988.

[22] G. Frahm, "Generalized Elliptical Distributions," Ph.D. dissertation, Economic and Social Statist. Dept., Univ. of Cologne, Cologne, Germany, 2004.

[23] J. T. Kent and D. E. Tyler, "Redescending M-estimates of multivariate location and scatter," *Ann. Statist.*, vol. 19, no. 4, pp. 2102–2119, 12, 1991.

[24] M. Zhang, F. Rubio, D. Palomar, and X. Mestre, "Finite-sample linear filter optimization in wireless communications and financial systems," *IEEE Trans. Signal Process.*, vol. 61, no. 20, pp. 5014–5025, Oct. 15, 2013.

[25] S. Verdu, "Spectral efficiency in the wideband regime," *IEEE Trans. Inf. Theory*, vol. 48, no. 6, pp. 1319–1343, Jun. 2002.

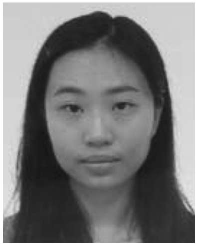
[26] T. M. Cover and J. A. Thomas, *Elements of Information Theory*. New York, NY, USA: Wiley , 2012.

[27] D. R. Hunter and K. Lange, "A tutorial on MM algorithms," *Amer. Statist.*, vol. 58, no. 1, pp. 30–37, 2004.

[28] M. Razaviyayn, M. Hong, and Z. Luo, "A unified convergence analysis of block successive minimization methods for nonsmooth optimization," *SIAM J. Optim.*, vol. 23, no. 2, pp. 1126–1153, 2013.

[29] T. Zhang, "A majorization-minimization algorithm for the Karcher mean of positive definite matrices," 2013 [Online]. Available: http://arxiv.org/abs/1312.4654

[30] R. Couillet and M. R. McKay, "Large dimensional analysis and optimization of robust shrinkage covariance matrix estimators," 2014 [Online]. Available: http://arxiv.org/abs/1401.4083



Ying Sun (S'14) received the B.Sc. degree in electronic information from the Huazhong University of Science and Technology, Wuhan, China, in 2011. She is currently pursuing the Ph.D. degree in electronic and computer engineering at The Hong Kong University of Science and Technology.

Her research interests include statistical signal processing, optimization algorithms and machine learning.

Prabhu Babu received the Ph.D. degree in electrical engineering from the Uppsala University, Sweden, in 2012.

He is currently a Research Associate with The Hong Kong University of Science and Technology.

Daniel P. Palomar (S'99–M'03–SM'08–F'12) received the Electrical Engineering and Ph.D. degrees from the Technical University of Catalonia (UPC), Barcelona, Spain, in 1998 and 2003, respectively.

He is a Professor in the Department of Electronic and Computer Engineering at the Hong Kong University of Science and Technology (HKUST), Hong Kong, which he joined in 2006. Since 2013 he is a Fellow of the Institute for Advance Study (IAS) at HKUST. He had previously held several research appointments, namely, at King's College London (KCL), London, UK; Technical University of Catalonia (UPC), Barcelona; Stanford University, Stanford, CA; Telecommunications Technological Center of Catalonia (CTTC), Barcelona; Royal Institute of Technology (KTH), Stockholm, Sweden; University of Rome "La Sapienza", Rome, Italy; and Princeton University, Princeton, NJ. His current research interests include applications of convex optimization theory, game theory, and variational inequality theory to financial systems and communication systems.

Dr. Palomar is an IEEE Fellow, a recipient of a 2004/06 Fulbright Research Fellowship, the 2004 Young Author Best Paper Award by the IEEE Signal Processing Society, the 2002/03 best Ph.D. prize in information technologies and communications by the Technical University of Catalonia (UPC), the 2002/03 Rosina Ribalta first prize for the Best Doctoral Thesis in information technologies and communications by the Epson Foundation, and the 2004 prize for the best Doctoral Thesis in Advanced Mobile Communications by the Vodafone Foundation and COIT. He serves as an Associate Editor of the IEEE TRANSACTIONS ON INFORMATION THEORY, and has been an Associate Editor of the IEEE TRANSACTIONS ON SIGNAL PROCESSING, a Guest Editor of the IEEE SIGNAL PROCESSING MAGAZINE 2010 Special Issue on "Convex Optimization for Signal Processing," the IEEE JOURNAL ON SELECTED AREAS IN COMMUNICATIONS 2008 Special Issue on "Game Theory in Communication Systems," and the IEEE JOURNAL ON SELECTED AREAS IN COMMUNICATIONS 2007 Special Issue on "Optimization of MIMO Transceivers for Realistic Communication Networks."