Regularized Robust Estimation of Mean and Covariance Matrix Under Heavy-Tailed Distributions

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

***Abstract*—In this paper, the joint mean-covariance estimation problem is considered under the scenario that the number of samples is small relative to the problem dimension. The samples are assumed drawn independently from a heavy-tailed distribution of the elliptical family, which can model scenarios where the commonly adopted Gaussian assumption is violated either because of the data generating process or the contamination of outliers. Under the assumption that prior knowledge of the mean and covariance matrix is available, we propose a regularized estimator defined as the minimizer of a penalized loss function, which combines the prior information and the information provided by the samples. The loss function is chosen to be the negative log-likelihood function of the Cauchy distribution as a conservative representative of heavy-tailed distributions, and the penalty term is constructed with the prior being its global minimizer. The resulting regularized estimator shrinks the mean and the covariance matrix to the prior target. The existence and uniqueness of the estimator for finite samples are established under certain regularity conditions. Numerical algorithms are derived for the estimator based on the majorization-minimization framework with guaranteed convergence and simulation results demonstrate that the proposed estimator achieves better estimation accuracy compared to the benchmark estimators.**

***Index Terms*—Iterative shrinkage, majorization-minimization, regularization, robust estimation.**

I. Introduction

The robust estimation of the mean and covariance matrix has been a research topic in signal processing for decades and enjoyed a wide range of applications. Examples include direction of arrival estimation [2], anomaly detection in wireless sensor networks [3], space-time adaptive processing detection problem [4], impulsive noise attenuation in image processing [5], high-resolution frequency estimation [6], and portfolio optimization in finance [7]; see [8] for a general review and the references therein. An intuitive and common approach is to estimate the mean and covariance matrix by sample average, which coincides with the maximum likelihood estimator (MLE) under the assumption that the samples are independent and identically drawn from a Gaussian distribution. However, in many applications the samples follow a distribution with a heavier tail than the Gaussian distribution, either due to the intrinsic mechanism of the application (e.g., financial time series [9]) or the existence of outliers, such as faulty observations. In this case, the Gaussian assumption can result in a completely unreliable estimate.

A way to address the aforementioned issue is to assume instead that the underlying distribution is some heavy-tailed elliptical distribution, for example, the Student's t-distribution with a small degree of freedom parameter. In the seminal work [10], Tyler proposed an M-estimator that estimates the normalized covariance matrix for samples drawn from an elliptical distribution with a known mean. The estimator possesses two advantages against other robust M-estimators: it is the "most robust" one in a min-max sense, and it is distribution-free. Moreover, Tyler proved that under certain regularity conditions on the empirical distribution of the samples, the estimator exists and is unique. This results in a simple numerical algorithm that is guaranteed to converge to the unique solution. Tyler's estimator has been widely adopted and proved to work effectively in various signal processing related fields, e.g. [11]–[16]. Nevertheless, despite the above-mentioned merits, the following shortcomings narrow down its scope of application in reality. First, a relatively accurate estimation requires a large number of samples compared to the dimension of the problem. Second, it requires a prior estimate of the mean. Consequently, the estimation procedure is separated into two steps, and the shape of the distribution cannot be taken into account when estimating the mean.

Indeed, estimation problems in modern applications are often subject to a shortage of samples compared to the dimension of the parameters being estimated, e.g., bioinformatics, financial engineering and massive MIMO communication systems [3], [4], [17]–[22]. In this case, the information on the parameters to be estimated provided by the samples is not sufficient for statistical inference, and approaches that depend purely on the samples are most likely to yield unreliable estimates. To tackle this problem, shrinkage estimators are proposed that combine the sample information with *a priori* information. A method for designing shrinkage estimators is to penalize the original loss function with a penalty term that attains its minimum at the prior. Motivated by this idea, various versions of shrinkage Tyler's covariance estimator have been proposed, and have been demonstrated to work effectively when the number of samples is relatively small compared to the dimension of the problem [3], [4], [18], [20], [22], [23].

In practice, there are scenarios in which both the mean and the covariance are required to be estimated from the samples, from fundamental techniques such as data decorrelation and principal

Manuscript received July 14, 2014; revised November 25, 2014 and March 09, 2015; accepted March 11, 2015. Date of publication March 26, 2015; date of current version May 13, 2015. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Joseph Tabrikian. This work was supported by the Hong Kong RGC 617312 research grant. Preliminary results of this work were published in IEEE 8th Sensors and Array Multichannel Signal Processing Workshop (SAM), June 2014.

The authors are with the 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.2015.2417513

component analysis [24], to high-level applications such as portfolio optimization in financial engineering [25]. A two-step estimation can be done by plugging a previously estimated mean (e.g., sample mean or sample median) into the shrinkage Tyler's covariance estimator. However, estimating the mean without taking into account the correlation of each component may be inadequate, especially in the presence of outliers. For example, as mentioned in [8], a multivariate outlier may have none of its components outlying by just looking at a particular coordinate, and being treated as a normal sample consequently. Moreover, the estimation error of the mean propagates to the covariance estimation since the true mean is substituted by an erroneous estimate. A way to design a robust estimator that estimates the mean and covariance jointly and performs well in the small sample scenario remains unclear.

This paper focuses on the joint mean and covariance estimation problem assuming independent and identically distributed (i.i.d.) samples from a heavy-tailed elliptical distribution when the sample size is small relative to the problem dimension. A regularized mean-covariance estimator is proposed defined as the minimizer of a penalized or regularized loss function. The loss function is chosen to be the negative log-likelihood function of the Cauchy distribution for its heavy-tail property that is capable of modeling abnormal observations as well as the tractability of analysis (note that the samples are not assumed to be drawn from a Cauchy distribution). The proposed estimator shrinks the mean and covariance matrix towards a prior target. Theoretical results including the existence and uniqueness of the estimator are proved under certain regularity conditions on the samples. The conditions indicate that the shrinkage estimator overcomes the drawback of the Cauchy maximum likelihood estimator without shrinkage, as it exists even when the number of samples is smaller than the problem dimension. Different numerical algorithms are provided and compared for the proposed shrinkage estimator based on the majorization-minimization framework with provable convergence.

The paper is organized as follows: in Section II, we introduce the robust parameter estimation problem with samples drawn from a distribution of the elliptical family. In Section III, we propose a regularized robust estimator for the joint mean-covariance estimation problem with small sample size, and establish the conditions for the existence and uniqueness of the shrinkage estimator. Algorithms for the proposed estimator based on the majorization-minimization framework are derived in Section IV. Simulation studies on both the estimator performance and the algorithm convergence are conducted in Section V. 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 $\mathbf{\Sigma}$ 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 $\mathbb{S}_{++}^K$ in the usual sense is $\mathbb{S}_+^K \setminus \mathbb{S}_{++}^K$, which contains all rank deficient matrices in $\mathbb{S}_+^K$. In this paper, with a slight 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 $\mathbf{\Sigma}^k$ 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 "$\mathbf{\Sigma}$ converges" equivalently as "a sequence of matrices $\mathbf{\Sigma}^k$ converges" for notation simplicity without causing confusion.

II. Robust Estimation of Mean and Covariance Matrix

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

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

with the location and scatter parameter $(\boldsymbol{\mu}_0, \mathbf{R}_0)$ 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. We are interested in the problem of estimating the mean and trace normalized covariance matrix (if they exist), which are $\boldsymbol{\mu}_0$ and $\mathbf{R}_0/\mathrm{Tr}(\mathbf{R}_0)$, respectively. The reason for not estimating the covariance matrix, but the trace-normalized one, will be explained at the end of this section. In the rest of the paper the notation $P_N(\cdot)$ stands for the empirical distribution.

A. Tyler's Estimator for Covariance Estimation

In [10] Tyler proposed a distribution-free estimator that estimates the trace-normalized covariance matrix with a known $\boldsymbol{\mu}_0$. Without loss of generality, $\boldsymbol{\mu}_0$ is assumed zero. Specifically, the estimator $\hat{\mathbf{R}}$ is defined as the solution of the following fixed-point equation:

$$\frac{K}{N}\sum_{i=1}^{N} \frac{\mathbf{x}_i \mathbf{x}_i^T}{\mathbf{x}_i^T \mathbf{R}^{-1} \mathbf{x}_i} = \mathbf{R}. \tag{2}$$

The estimator $\hat{\mathbf{R}}$ can be interpreted as the MLE of $\mathbf{R}$ by fitting the normalized samples $\mathbf{s}_i = \frac{\mathbf{x}_i}{\|\mathbf{x}_i\|_2}$ to an angular central Gaussian distribution with pdf [26]–[28]

$$f(\mathbf{s}) = \frac{\Gamma\left(\frac{K}{2}\right)}{2\pi^{\frac{K}{2}}} \det(\mathbf{R})^{-\frac{1}{2}} \left(\mathbf{s}^T \mathbf{R}^{-1} \mathbf{s}\right)^{-K/2}. \tag{3}$$

Tyler established sufficient conditions for the existence of $\hat{\mathbf{R}}$ and proved that it is unique up to a positive scale factor [10], [26]. The estimator was shown to be strongly consistent and asymptotically normal. Tyler's estimator has two advantages against the others: it is the most robust estimator in the sense that its maximum asymptotic variance is less than the maximum asymptotic variance of any other consistent and uniformly asymptotically normal estimator, and it is distribution-free in the sense that its asymptotic standard deviation does not depend on the form of $g(\cdot)$. To numerically compute $\hat{\mathbf{R}}$, Tyler proposed the following iterative algorithm:

$$\begin{aligned} \tilde{\mathbf{R}}_{t+1} &= \frac{K}{N}\sum_{i=1}^{N} \frac{\mathbf{x}_i \mathbf{x}_i^T}{\mathbf{x}_i^T \mathbf{R}_t^{-1} \mathbf{x}_i} \\ \mathbf{R}_{t+1} &= \frac{\tilde{\mathbf{R}}_{t+1}}{\mathrm{Tr}(\tilde{\mathbf{R}}_{t+1})}, \end{aligned} \tag{4}$$

and also proved its convergence.

B. Cauchy MLE for Mean and Covariance Estimation

Tyler's estimator assumes a given $\boldsymbol{\mu}_0$, which can be substituted by *a priori* consistent estimator $\hat{\boldsymbol{\mu}}$. This leads to a two-step estimation procedure, where the shape of the distribution cannot be taken into account when estimating $\boldsymbol{\mu}_0$. A widely utilized robust M-estimator for the joint mean-covariance estimation problem is the MLE of the Student's t-distribution [29], [30]. This is defined as the minimizer of the negative log-likelihood function of a Student's t-distribution with degree of freedom ν (a particular case of (1)), which takes the following form:

$$L^{\nu}(\boldsymbol{\mu},\mathbf{R}) = \frac{N}{2}\log\det(\mathbf{R}) + \frac{K+\nu}{2}\sum_{i=1}^{N}\log\left(\nu+(\mathbf{x}_i-\boldsymbol{\mu})^T\mathbf{R}^{-1}(\mathbf{x}_i-\boldsymbol{\mu})\right).$$

The estimator satisfies the following system of fixed-point equations:

$$\begin{aligned}\frac{\nu+K}{N}\sum_{i=1}^{N}\frac{\mathbf{x}_i-\boldsymbol{\mu}}{\nu+(\mathbf{x}_i-\boldsymbol{\mu})^T\mathbf{R}^{-1}(\mathbf{x}_i-\boldsymbol{\mu})} &= \mathbf{0}\\ \frac{\nu+K}{N}\sum_{i=1}^{N}\frac{(\mathbf{x}_i-\boldsymbol{\mu})(\mathbf{x}_i-\boldsymbol{\mu})^T}{\nu+(\mathbf{x}_i-\boldsymbol{\mu})^T\mathbf{R}^{-1}(\mathbf{x}_i-\boldsymbol{\mu})} &= \mathbf{R}.\end{aligned} \tag{5}$$

The solution of (5) can be interpreted as a weighted sample average, for which the weight decreases as the sample gets farther away from the center, i.e., the weight is inversely proportional to $d_i^2=(\mathbf{x}_i-\boldsymbol{\mu})^T\mathbf{R}^{-1}(\mathbf{x}_i-\boldsymbol{\mu})$. This property indicates that the estimator is less sensitive to outliers than the sample average. The degree of down-weighting increases as ν decreases. The properties of the MLE of the Student's t-distribution are well-studied in the literature [31]–[34]. Under the condition that for any hyperplane H with $0\le\dim(H)\le K-1$, $P_N(H)<\frac{\dim(H)+\nu}{K+\nu}$, the solution $(\hat{\boldsymbol{\mu}},\hat{\mathbf{R}})$ to (5) exists, and it is unique when $\nu\ge 1$.

In this paper, we are going to adapt the M-estimator defined as the solution of (5) with $\nu=1$, which corresponds to the MLE of a Cauchy distribution, to high-dimensional estimation problem. For completeness, we first introduce some of its properties that serve as the basis of analysis in the high dimension regime.

Following the idea of [31] and [35], we construct the augmented samples $\bar{\mathbf{x}}_i=[\mathbf{x}_i;1]\in\mathbb{R}^{K+1}$ and define the following change of variable:

$$\boldsymbol{\Sigma}=\begin{bmatrix}\mathbf{R}+\boldsymbol{\mu}\boldsymbol{\mu}^T & \boldsymbol{\mu}\\ \boldsymbol{\mu}^T & 1\end{bmatrix}. \tag{6}$$

The negative log-likelihood function of the Cauchy distribution $L^1(\boldsymbol{\mu},\mathbf{R})$ ($L(\boldsymbol{\mu},\mathbf{R})$ hereafter) can be equivalently expressed as:

$$\begin{aligned}L(\boldsymbol{\mu},\mathbf{R}) &= L(\boldsymbol{\Sigma})\\ &= \frac{N}{2}\log\det(\boldsymbol{\Sigma})+\frac{K+1}{2}\sum_{i=1}^{N}\log\left(\bar{\mathbf{x}}_i^T\boldsymbol{\Sigma}^{-1}\bar{\mathbf{x}}_i\right),\end{aligned}$$

which can be virtually viewed as fitting $\bar{\mathbf{s}}_i=\frac{\bar{\mathbf{x}}_i}{\|\bar{\mathbf{x}}_i\|_2}$ to a $(K+1)$-dimensional angular central Gaussian distribution with the center being zero.

C. Asymptotics

It is known that if $(\hat{\boldsymbol{\mu}},\hat{\mathbf{R}})$ satisfies estimating equations (5) with $\{\mathbf{x}_i\}$ being i.i.d. samples from an elliptical distribution, then as $N\to+\infty$, the asymptotic values $(\boldsymbol{\mu}_\infty,\mathbf{R}_\infty)$ will converge in probability to the unique solution of the system of equations

$$\begin{aligned}(K+1)E_f\left\{\frac{\mathbf{x}-\boldsymbol{\mu}}{1+(\mathbf{x}-\boldsymbol{\mu})^T\mathbf{R}^{-1}(\mathbf{x}-\boldsymbol{\mu})}\right\} &= \mathbf{0}\\ (K+1)E_f\left\{\frac{(\mathbf{x}-\boldsymbol{\mu})(\mathbf{x}-\boldsymbol{\mu})^T}{1+(\mathbf{x}-\boldsymbol{\mu})^T\mathbf{R}^{-1}(\mathbf{x}-\boldsymbol{\mu})}\right\} &= \mathbf{R}\end{aligned} \tag{7}$$

with the relation $\boldsymbol{\mu}_\infty=\boldsymbol{\mu}_0$ and $\mathbf{R}_\infty=c\mathbf{R}_0$ for some scalar c that depends on $f(\cdot)$ [33].

From the above result, we conclude that if the mean of the underlying distribution $f(\mathbf{x})$ exists, we can estimate it by $\hat{\boldsymbol{\mu}}$. However, the covariance matrix can only be estimated up to a scale factor that depends on $f(\cdot)$, which is unknown. For this reason, we focus on the joint estimation of the mean and covariance matrix normalized by its trace, i.e., $\boldsymbol{\mu}_0$ and $\mathbf{R}_0/\mathrm{Tr}(\mathbf{R}_0)$.

III. Regularized Robust Estimator of Mean and Covariance Matrix

Assuming the samples are drawn independently from a continuous distribution $f(\mathbf{x})$, then the condition $N>K+1$ guarantees the existence of the Cauchy MLE with probability one [31], and a reliable estimation requires even more samples. However, in some applications the number of samples is relatively small compared to the number of parameters being estimated. In this case, the algorithm designed for the estimator may fail to converge. Motivated by the idea of [20], we regularize the Cauchy MLE by shrinking the estimator to a prior target $(\mathbf{t},\mathbf{T})$. The advantage of a shrinkage estimator is twofold: it provides a way to incorporate prior information into the estimator, and it helps stabilizing the estimator in the small sample situation.

We devise the following penalty function:

$$\begin{aligned}h(\boldsymbol{\mu},\mathbf{R}) = \alpha\left(K\log\left(\mathrm{Tr}(\mathbf{R}^{-1}\mathbf{T})\right)+\log\det(\mathbf{R})\right)\\ +\gamma\log\left(1+(\boldsymbol{\mu}-\mathbf{t})^T\mathbf{R}^{-1}(\boldsymbol{\mu}-\mathbf{t})\right)\end{aligned} \tag{8}$$

for some finite-valued nonnegative parameters α and γ. The following proposition shows that $(\mathbf{t},r\mathbf{T})$ minimizes $h(\boldsymbol{\mu},\mathbf{R})$ with $r>0$, therefore justifies that $h(\boldsymbol{\mu},\mathbf{R})$ is indeed a proper penalty function.

Proposition 1: The minimizer of (8) on the set $\mathbb{R}^K\times\mathbb{S}_{++}^K$ is given by $(\mathbf{t},r\mathbf{T})$ for $r>0$.

Proof: See Appendix A. ∎

The scale-invariant property of the minimizers of $h(\boldsymbol{\mu},\mathbf{R})$ is important due to the following reason. Since asymptotically $(\boldsymbol{\mu}_0,c\mathbf{R}_0)$ minimizes $L(\boldsymbol{\mu},\mathbf{R})$ with c depending on the unknown $f(\cdot)$, a way of setting a shrinkage target $\mathbf{T}$ for $\mathbf{R}_0$ (or $\mathbf{R}_0/\mathrm{Tr}(\mathbf{R}_0)$) is by adding a penalty term $h(\boldsymbol{\mu},\mathbf{R})$ that is minimized for $\mathbf{R}$ proportional to $\mathbf{T}$ (by passing the value of c).

The regularized estimation problem is stated below with a shrinkage target $(\mathbf{t}, \mathbf{T})$ for $(\boldsymbol{\mu}, \mathbf{R})$:

$$\begin{aligned}\underset{\boldsymbol{\mu},\mathbf{R}\succ\mathbf{0}}{\text{minimize}} \quad & \frac{(K+1)}{2}\sum_{i=1}^{N}\log\left(1+(\mathbf{x}_i-\boldsymbol{\mu})^T\mathbf{R}^{-1}(\mathbf{x}_i-\boldsymbol{\mu})\right) \\ & +\alpha\left(K\log\left(\mathrm{Tr}(\mathbf{R}^{-1}\mathbf{T})\right)+\log\det(\mathbf{R})\right) \\ & +\gamma\log\left(1+(\boldsymbol{\mu}-\mathbf{t})^T\mathbf{R}^{-1}(\boldsymbol{\mu}-\mathbf{t})\right)+\frac{N}{2}\log\det(\mathbf{R}).\end{aligned} \tag{9}$$

The shrinkage estimator $(\hat{\boldsymbol{\mu}}, \hat{\mathbf{R}})$ with $\hat{\mathbf{R}} \succ \mathbf{0}$, which is defined as the solution of problem (9), has to satisfy the following fixed-point equations:

$$\begin{aligned}\mathbf{R} &= \frac{K+1}{N+2\alpha}\sum_{i=1}^{N}\frac{(\mathbf{x}_i-\boldsymbol{\mu})(\mathbf{x}_i-\boldsymbol{\mu})^T}{1+(\mathbf{x}_i-\boldsymbol{\mu})^T\mathbf{R}^{-1}(\mathbf{x}_i-\boldsymbol{\mu})} \\ &\quad+\frac{2\gamma}{N+2\alpha}\frac{(\boldsymbol{\mu}-\mathbf{t})(\boldsymbol{\mu}-\mathbf{t})^T}{1+(\boldsymbol{\mu}-\mathbf{t})\mathbf{R}^{-1}(\boldsymbol{\mu}-\mathbf{t})}+\frac{2\alpha K}{N+2\alpha}\frac{\mathbf{T}}{\mathrm{Tr}(\mathbf{R}^{-1}\mathbf{T})} \\ \mathbf{0} &= (K+1)\sum_{i=1}^{N}\frac{(\mathbf{x}_i-\boldsymbol{\mu})}{1+(\mathbf{x}_i-\boldsymbol{\mu})^T\mathbf{R}^{-1}(\mathbf{x}_i-\boldsymbol{\mu})} \\ &\quad+2\gamma\frac{\mathbf{t}-\boldsymbol{\mu}}{1+(\boldsymbol{\mu}-\mathbf{t})^T\mathbf{R}^{-1}(\boldsymbol{\mu}-\mathbf{t})},\end{aligned} \tag{10}$$

which are derived by setting the gradient of the objective function, denoted by $L^{\mathrm{shrink}}(\boldsymbol{\mu}, \mathbf{R})$, to zero. Note that since α and γ are finite-valued, the effect of the penalty term on $(\hat{\boldsymbol{\mu}}, \hat{\mathbf{R}})$ vanishes as $N \to +\infty$. As a result, $(\hat{\boldsymbol{\mu}}, \hat{\mathbf{R}}/\mathrm{Tr}(\hat{\mathbf{R}})) \to (\boldsymbol{\mu}_0, \mathbf{R}_0/\mathrm{Tr}(\mathbf{R}_0))$ asymptotically.

By defining $\bar{\mathbf{x}}_i = [\mathbf{x}_i; 1]$, $\bar{\mathbf{t}} = [\mathbf{t}; 1]$, and using the reparametrization (6), problem (9) is equivalent to

$$\begin{aligned}\underset{\boldsymbol{\Sigma}\succ\mathbf{0}}{\text{minimize}} \quad & \left(\frac{N}{2}+\alpha\right)\log\det(\boldsymbol{\Sigma})+\frac{(K+1)}{2}\sum_{i=1}^{N}\log\left(\bar{\mathbf{x}}_i^T\boldsymbol{\Sigma}^{-1}\bar{\mathbf{x}}_i\right) \\ & +\alpha K\log\left(\mathrm{Tr}\left(\mathbf{S}^T\boldsymbol{\Sigma}^{-1}\mathbf{S}\mathbf{T}\right)\right)+\gamma\log\left(\bar{\mathbf{t}}^T\boldsymbol{\Sigma}^{-1}\bar{\mathbf{t}}\right) \\ \text{subject to} \quad & \boldsymbol{\Sigma}_{K+1,K+1}=1\end{aligned} \tag{11}$$

with $\mathbf{S}$ being a selection matrix defined as $\mathbf{S} = \begin{bmatrix}\mathbf{I}_K \\ \mathbf{0}_{1\times K}\end{bmatrix}$. Denote the objective function by $L^{\mathrm{shrink}}(\boldsymbol{\Sigma})$. A sufficient condition for the existence of the estimator $(\hat{\boldsymbol{\mu}}, \hat{\mathbf{R}})$ is stated in the following theorem.

Theorem 2: For the loss function

$$\begin{aligned}&L^{\mathrm{shrink}}(\boldsymbol{\mu}, \mathbf{R}) \\ &= \frac{(K+1)}{2}\sum_{i=1}^{N}\log\left(1+(\mathbf{x}_i-\boldsymbol{\mu})^T\mathbf{R}^{-1}(\mathbf{x}_i-\boldsymbol{\mu})\right) \\ &\quad+\alpha\left(K\log\left(\mathrm{Tr}(\mathbf{R}^{-1}\mathbf{T})\right)+\log\det(\mathbf{R})\right) \\ &\quad+\gamma\log\left(1+(\boldsymbol{\mu}-\mathbf{t})^T\mathbf{R}^{-1}(\boldsymbol{\mu}-\mathbf{t})\right)+\frac{N}{2}\log\det(\mathbf{R})\end{aligned} \tag{12}$$

defined on $\mathbb{R}^K \times \mathbb{S}_{++}^K$, under the assumption that $\mathbf{T}$ is full rank, a minimum of $L^{\mathrm{shrink}}(\boldsymbol{\mu}, \mathbf{R})$ exists under the following condition: for any hyperplane $H \subset \mathbb{R}^K$ with dimension $0 \le \dim(H) < K$, if H contains $\mathbf{t}$,

$$P_N(H) < \frac{(2\alpha+N)\dim(H)+N}{(K+1)N},$$

and if H does not contain $\mathbf{t}$,

$$P_N(H) < \frac{(2\alpha+N)\dim(H)+N+2\gamma}{(K+1)N}.$$

Proof: See Appendix B. ∎

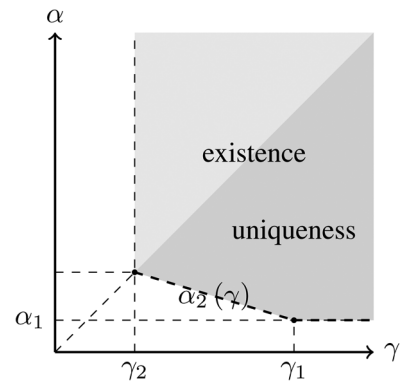


Fig. 1. Values that the regularization parameters α and γ can take for the existence and uniqueness of the shrinkage estimator.

The existence condition requires the samples to be sufficiently spread out in the $\mathbb{R}^K$ space. The condition is more relaxed compared to that without regularization, which is $P_N(H) < \frac{\dim(H)+1}{K+1}$ as stated in [31]. Notice that setting α and γ to zero recovers the condition without regularization.

Corollary 3: Assume that the population distribution $f(\cdot)$ is continuous, then the estimator $(\hat{\boldsymbol{\mu}}, \hat{\mathbf{R}})$ exists for $N > 1$ if $\alpha, \gamma \ge 0$ and either of the following conditions is satisfied:

(i) if $\gamma > \gamma_1$, then $\alpha > \alpha_1$,

(ii) if $\gamma_2 < \gamma \le \gamma_1$, then $\alpha > \alpha_2(\gamma)$,

where

$$\begin{aligned}\alpha_1 &= \frac{1}{2}(K-N), \\ \alpha_2(\gamma) &= \frac{1}{2}\left(K+1-N-\frac{2\gamma+N-K-1}{N-1}\right),\end{aligned}$$

and $\gamma_1 = \frac{1}{2}K$, $\gamma_2 = \frac{1}{2}(K+1-N)$.

Proof: Notice that since the distribution is continuous, then with probability one, a d-dimensional hyperplane at most touches $d+1$ points. The condition in Theorem 2 can be simplified as follows:

$$\begin{cases}\frac{\min\{d,N\}}{N} < \frac{(2\alpha+N)d+N}{(K+1)N} \\ \frac{\min\{d+1,N\}}{N} < \frac{(2\alpha+N)d+2\gamma+N}{(K+1)N}\end{cases}, \forall\, 0 \le d \le K-1,$$

which is equivalent to

$$\begin{cases}(K+1-2\alpha-N)d < N, \\ \qquad \forall\, 0 \le d \le \min\{K-1, N\}, \\ (K+1-2\alpha-N)d < 2\gamma+N-K-1, \\ \qquad \forall\, 0 \le d \le \min\{K-1, N-1\}.\end{cases} \tag{13}$$

Since both α and γ are nonnegative, the conditions above is satisfied if $N \ge K+1$. Under the case that $N \le K$, some algebraic calculation reveals that (13) is equivalent to the statement of the corollary. ∎

The feasible region of (α, γ) is shown pictorially in Fig. 1. The condition in Corollary 3 implies the tradeoff between the regularization parameters α and γ. Specifically, since $\alpha_2(\gamma)$ is decreasing in γ, the condition indicates that when the confidence on the prior information of $\boldsymbol{\mu}$ gets weaker, which corresponds to a smaller value of γ, the regularization for $\mathbf{R}$ should be stronger.

For the special case that $\alpha = \gamma$, the condition reduces to $\alpha > \frac{1}{2}(K+1-N)$, or equivalently $N > K+1-2\alpha$. The lower bound on the number of samples is decreased by 2α as a result of regularization.

Now that we have established the existence condition for the shrinkage estimator $(\hat{\boldsymbol{\mu}}, \hat{\mathbf{R}})$, in the following theorem, we are going to show that the estimator is unique when $\gamma \geq \alpha$.

Theorem 4: Under the regularity condition in Theorem 2, $(\hat{\boldsymbol{\mu}}, \hat{\mathbf{R}})$ is unique when $\gamma \geq \alpha$.

Proof: See Appendix C. ∎

IV. Algorithms

In this section, we are going to derive algorithms for the regularized estimator based on the majorization-minimization (MM) framework. The concept of MM [36], [37] is briefly introduced below.

Consider the following optimization problem:

$$\begin{aligned} \underset{\mathbf{x}}{\text{minimize}} \quad & f(\mathbf{x}) \\ \text{subject to} \quad & \mathbf{x} \in \mathcal{X}, \end{aligned} \tag{14}$$

where $\mathbf{x} \in \mathbb{R}^n$, $f(\cdot)$ is assumed to be continuous in $\mathbf{x}$ and $\mathcal{X}$ is a convex set.

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

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

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 both $\mathbf{x}$ and $\mathbf{x}_t$. The update of $\mathbf{x}$ is given by

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

The limit points of sequence $\{\mathbf{x}_t\}$ are proved to be the stationary points of the original problem (14) [37].

The idea of majorizing $f(\mathbf{x})$ by a surrogate function can also be applied blockwise. Specifically, $\mathbf{x}$ is partitioned into m blocks as $\mathbf{x} = (\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(m)})$, where each n_i-dimensional block $\mathbf{x}^{(i)} \in \mathcal{X}_i$ and $\mathcal{X} = \prod_{i=1}^m \mathcal{X}_i$. At the $(t+1)$-th iteration, $\mathbf{x}^{(i)}$ is updated by solving the following problem:

$$\begin{aligned} \underset{\mathbf{x}^{(i)}}{\text{minimize}} \quad & g_i\left(\mathbf{x}^{(i)}|\mathbf{x}_t\right) \\ \text{subject to} \quad & \mathbf{x}^{(i)} \in \mathcal{X}_i \end{aligned} \tag{15}$$

with $i = (t \bmod m) + 1$ and the continuous surrogate function $g_i(\mathbf{x}^{(i)}|\mathbf{x}_t)$ satisfying the following properties:

$$\begin{aligned} f(\mathbf{x}_t) &= g_i\left(\mathbf{x}_t^{(i)}|\mathbf{x}_t\right), \\ f\left(\mathbf{x}_t^{(1)}, \ldots, \mathbf{x}^{(i)}, \ldots, \mathbf{x}_t^{(m)}\right) &\leq g_i\left(\mathbf{x}^{(i)}|\mathbf{x}_t\right) \forall \mathbf{x}^{(i)} \in \mathcal{X}_i, \\ f'\left(\mathbf{x}_t;\mathbf{d}_i^0\right) &= g_i'\left(\mathbf{x}_t^{(i)};\mathbf{d}_i|\mathbf{x}_t\right) \\ & \quad \forall \mathbf{x}_t^{(i)} + \mathbf{d}_i \in \mathcal{X}_i, \\ \mathbf{d}_i^0 &\triangleq (\mathbf{0};\ldots;\mathbf{d}_i;\ldots;\mathbf{0}). \end{aligned}$$

In short, at each iteration, the block MM applies the ordinary MM algorithm to one block while keeping the value of the other blocks fixed. The blocks are updated in cyclic order.

It is proved in [37] that under the following conditions:

(C1) the level set $\{\mathbf{x}|f(\mathbf{x}) \leq f(\mathbf{x}_0)\}$ with $\mathbf{x}_0$ being the initial point is compact;

(C2) each sub-problem (15) has a unique solution for any $\mathbf{x}_t \in \mathcal{X}$ for at least $(m-1)$ blocks;

(C3) at any stationary point $\mathbf{x}^\star$, $f'(\mathbf{x}^\star;\mathbf{d}) \geq 0$ for all $\mathbf{d} = (\mathbf{d}_1;\ldots;\mathbf{d}_m)$ with $f'(\mathbf{x}^\star;\mathbf{d}_i^0) \geq 0$,

the sequence $\{\mathbf{x}_t\}$ generated by block MM converges to the set of stationary points of (14).

In the rest of this section, for any continuous function $f(\mathbf{x})$, we define $f(\mathbf{y}) = +\infty$ when $\lim_{\mathbf{x}\to\mathbf{y}} f(\mathbf{x}) = +\infty$. With a slight abuse of notation, $(\boldsymbol{\mu}_0, \mathbf{R}_0)$ refers to the initial point of the algorithm in this section.

Recall that the optimization problem takes the form

$$\begin{aligned} \underset{\boldsymbol{\mu},\mathbf{R}\succcurlyeq\mathbf{0}}{\text{minimize}} \quad & \frac{N}{2}\log\det(\mathbf{R}) \\ & + \frac{(K+1)}{2}\sum_{i=1}^N \log\left(1+(\mathbf{x}_i-\boldsymbol{\mu})^T\mathbf{R}^{-1}(\mathbf{x}_i-\boldsymbol{\mu})\right) \\ & + \alpha\left(K\log\left(\mathrm{Tr}(\mathbf{R}^{-1}\mathbf{T})\right) + \log\det(\mathbf{R})\right) \\ & + \gamma\log\left(1+(\boldsymbol{\mu}-\mathbf{t})^T\mathbf{R}^{-1}(\boldsymbol{\mu}-\mathbf{t})\right). \end{aligned} \tag{16}$$

Before introducing the algorithms, we first state the following lemma, which is needed for proving the convergence of the algorithms.

Lemma 5: Given any initial point $(\boldsymbol{\mu}_0, \mathbf{R}_0)$ with $\mathbf{R}_0 \succ \mathbf{0}$, the level set $\mathcal{X}^0 = \{(\boldsymbol{\mu},\mathbf{R})|L^{\text{shrink}}(\boldsymbol{\mu},\mathbf{R}) \leq L^{shrink}(\boldsymbol{\mu}_0,\mathbf{R}_0)\}$ is compact.

Proof: Under the conditions stated in Theorem 2, a unique minimizer $(\hat{\boldsymbol{\mu}}, \hat{\mathbf{R}})$ of $L^{\text{shrink}}(\boldsymbol{\mu},\mathbf{R})$ exists. Observe that

$$\begin{aligned} \lambda_{\max}(\boldsymbol{\Sigma}) &= \sup_{\|\mathbf{y}\|=1} \mathbf{y}^T\boldsymbol{\Sigma}\mathbf{y} \\ &= \sup_{\|\tilde{\mathbf{y}}\|^2+y^2=1} [\,\tilde{\mathbf{y}}^T \quad y\,] \begin{bmatrix} \mathbf{R}+\boldsymbol{\mu}\boldsymbol{\mu}^T & \boldsymbol{\mu} \\ \boldsymbol{\mu}^T & 1 \end{bmatrix} \begin{bmatrix} \tilde{\mathbf{y}} \\ y \end{bmatrix} \\ &= \sup_{\|\tilde{\mathbf{y}}\|^2+y^2=1} \left\{\tilde{\mathbf{y}}^T\mathbf{R}\tilde{\mathbf{y}} + \left(\tilde{\mathbf{y}}^T\boldsymbol{\mu}+y\right)^2\right\} \\ &\geq \sup_{\|\tilde{\mathbf{y}}\|^2=1} \left\{\tilde{\mathbf{y}}^T(\mathbf{R}+\boldsymbol{\mu}\boldsymbol{\mu}^T)\tilde{\mathbf{y}}\right\}. \end{aligned}$$

Suppose $\lambda_{\max}(\mathbf{R}) \to +\infty$, then $\lambda_{\max}(\boldsymbol{\Sigma}) \to +\infty$. By Theorem 2, it is known that $L^{\text{shrink}}(\boldsymbol{\mu},\mathbf{R}) = L^{\text{shrink}}(\boldsymbol{\Sigma}) \to +\infty$ as $\lambda_{\max}(\boldsymbol{\Sigma}) \to +\infty$, which contradicts the fact that on $\mathcal{X}^0$, $L^{\text{shrink}}(\boldsymbol{\mu},\mathbf{R})$ is bounded above. Therefore $\mathbf{R}$ is bounded. Now suppose $\boldsymbol{\mu}$ is unbounded, set $\tilde{\mathbf{y}} = \boldsymbol{\mu}/\|\boldsymbol{\mu}\|$ and $y = 0$ we have $\lambda_{\max}(\boldsymbol{\Sigma}) \geq \|\boldsymbol{\mu}\|^2 \to +\infty$. Therefore $\mathcal{X}^0$ is bounded. The continuity of $L^{\text{shrink}}(\boldsymbol{\mu},\mathbf{R})$ implies $\mathcal{X}^0$ is closed. Hence $\mathcal{X}^0$ is compact. ∎

A. Majorization-Minimization

By the concavity of $\log(\cdot)$, at point $(\boldsymbol{\mu}_t, \mathbf{R}_t)$ function (12) is majorized by

$$\begin{aligned} & L(\boldsymbol{\mu},\mathbf{R}|\boldsymbol{\mu}_t,\mathbf{R}_t) \\ &= \frac{K+1}{2}\sum w_i(\boldsymbol{\mu}_t,\mathbf{R}_t)(\mathbf{x}_i-\boldsymbol{\mu})^T\mathbf{R}^{-1}(\mathbf{x}_i-\boldsymbol{\mu}) \\ &\quad + \gamma w_{\mathbf{t}}(\boldsymbol{\mu}_t,\mathbf{R}_t)(\mathbf{t}-\boldsymbol{\mu})^T\mathbf{R}^{-1}(\mathbf{t}-\boldsymbol{\mu}) \\ &\quad + \left(\frac{N}{2}+\alpha\right)\log\det(\mathbf{R}) + \alpha K\frac{\mathrm{Tr}(\mathbf{R}^{-1}\mathbf{T})}{\mathrm{Tr}\left(\mathbf{R}_t^{-1}\mathbf{T}\right)} + \text{const.} \end{aligned} \tag{17}$$

with weights

$$w_i(\boldsymbol{\mu},\mathbf{R}) = \frac{1}{1+(\mathbf{x}_i-\boldsymbol{\mu})^T\mathbf{R}^{-1}(\mathbf{x}_i-\boldsymbol{\mu})}$$
$$w_{\mathbf{t}}(\boldsymbol{\mu},\mathbf{R}) = \frac{1}{1+(\mathbf{t}-\boldsymbol{\mu})^T\mathbf{R}^{-1}(\mathbf{t}-\boldsymbol{\mu})}. \quad (18)$$

Setting the gradient of $L(\boldsymbol{\mu},\mathbf{R}|\boldsymbol{\mu}_t,\mathbf{R}_t)$ to zero leads to the update equations in Algorithm 1.

Algorithm 1: Majorization-Minimization

1) Initialize $\mathbf{R}_0$ as an arbitrary positive definite matrix, and $\boldsymbol{\mu}_0$ as an arbitrary vector.

2) Iterate

$$\boldsymbol{\mu}_{t+1} = \frac{(K+1)\sum_{i=1}^N w_i(\boldsymbol{\mu}_t,\mathbf{R}_t)\mathbf{x}_i + 2\gamma w_{\mathbf{t}}(\boldsymbol{\mu}_t,\mathbf{R}_t)\mathbf{t}}{(K+1)\sum_{i=1}^N w_i(\boldsymbol{\mu}_t,\mathbf{R}_t) + 2\gamma w_{\mathbf{t}}(\boldsymbol{\mu}_t,\mathbf{R}_t)}$$
$$\mathbf{R}_{t+1} = \frac{K+1}{N+2\alpha}\sum_{i=1}^N w_i(\boldsymbol{\mu}_t,\mathbf{R}_t)(\mathbf{x}_i-\boldsymbol{\mu}_{t+1})(\mathbf{x}_i-\boldsymbol{\mu}_{t+1})^T + \frac{2\gamma}{N+2\alpha}w_{\mathbf{t}}(\boldsymbol{\mu}_t,\mathbf{R}_t)(\mathbf{t}-\boldsymbol{\mu}_{t+1})(\mathbf{t}-\boldsymbol{\mu}_{t+1})^T + \frac{2\alpha K}{N+2\alpha}\frac{\mathbf{T}}{\mathrm{Tr}\left(\mathbf{R}_t^{-1}\mathbf{T}\right)} \quad (19)$$

with $w_i(\boldsymbol{\mu},\mathbf{R})$ and $w_{\mathbf{t}}(\boldsymbol{\mu},\mathbf{R})$ given in (18) until convergence.

Lemma 6: The pair $(\boldsymbol{\mu}_{t+1},\mathbf{R}_{t+1})$ given by (19) uniquely minimizes the surrogate function (17).

Proof: Since the surrogate function $L(\boldsymbol{\mu},\mathbf{R}|\boldsymbol{\mu}_t,\mathbf{R}_t)$ upper bounds the cost function $L^{\mathrm{shrink}}(\boldsymbol{\mu},\mathbf{R})$ globally, the minimum of $L(\boldsymbol{\mu},\mathbf{R}|\boldsymbol{\mu}_t,\mathbf{R}_t)$ exists with $\mathbf{R}\succ\mathbf{0}$ and $\boldsymbol{\mu}$ being finite. Observe that $L(\boldsymbol{\mu},\mathbf{R}|\boldsymbol{\mu}_t,\mathbf{R}_t)$ is the negative log-likelihood function of a Gaussian distribution and has a unique stationary point $(\boldsymbol{\mu}_{t+1},\mathbf{R}_{t+1})$ on $\mathbb{R}^K\times\mathbb{S}_+^K$, it has to be the minimum. ∎

Proposition 7: The sequence $\{(\boldsymbol{\mu}_t,\mathbf{R}_t)\}$ generated by Algorithm 1 converges to

(i) the set of stationary points of problem (16) if $\alpha>\gamma$;
(ii) the global minimizer of problem (16) if $\alpha\le\gamma$.

Proof: By Lemma 5 we have that the initial level set $\mathcal{X}^0$ is compact. Furthermore, by Lemma 6, $(\boldsymbol{\mu}_{t+1},\mathbf{R}_{t+1})$ uniquely minimizes the surrogate function $L(\boldsymbol{\mu},\mathbf{R}|\boldsymbol{\mu}_t,\mathbf{R}_t)$. Hence the sequence $\{(\boldsymbol{\mu}_t,\mathbf{R}_t)\}$ converges to the set of stationary points of $L^{\mathrm{shrink}}(\boldsymbol{\mu},\mathbf{R})$ [37]. It has been proved in Theorem 4 that the stationary point of problem (16) is unique, and it is the global minimum when $\alpha\le\gamma$. Therefore in this case $\{(\boldsymbol{\mu}_t,\mathbf{R}_t)\}$ converges to the global minimizer of problem (16). ∎

B. Block Majorization-Minimization

Instead of upperbounding the whole function at point $(\boldsymbol{\mu}_t,\mathbf{R}_t)$, majorization can also be applied blockwise. Specifically, an upperbound for $L^{\mathrm{shrink}}(\boldsymbol{\mu}_t,\mathbf{R})$ can be obtained as:

$$L(\mathbf{R}|\boldsymbol{\mu}_t,\mathbf{R}_t) = \frac{K+1}{2}\sum_{i=1}^N w_i(\boldsymbol{\mu}_t,\mathbf{R}_t)(\mathbf{x}_i-\boldsymbol{\mu}_t)^T\mathbf{R}^{-1}(\mathbf{x}_i-\boldsymbol{\mu}_t) + \alpha K\frac{\mathrm{Tr}(\mathbf{R}^{-1}\mathbf{T})}{\mathrm{Tr}\left(\mathbf{R}_t^{-1}\mathbf{T}\right)} + \gamma w_{\mathbf{t}}(\boldsymbol{\mu}_t,\mathbf{R}_t)(\mathbf{t}-\boldsymbol{\mu}_t)^T\mathbf{R}^{-1}(\mathbf{t}-\boldsymbol{\mu}_t) + \left(\frac{N}{2}+\alpha\right)\log\det(\mathbf{R}) + \mathrm{const.} \quad (20)$$

with the value of $\boldsymbol{\mu}$ fixed as $\boldsymbol{\mu}_t$, which leads to the update equation (22) in Algorithm 2 for $\mathbf{R}$.

Then we fix the value of $\mathbf{R}$ as $\mathbf{R}_{t+1}$ and get an upperbound for $L^{\mathrm{shrink}}(\boldsymbol{\mu},\mathbf{R}_{t+1})$ as follows:

$$L(\boldsymbol{\mu}|\boldsymbol{\mu}_t,\mathbf{R}_{t+1}) = \frac{K+1}{2}\sum_{i=1}^N w_i(\boldsymbol{\mu}_t,\mathbf{R}_{t+1})(\mathbf{x}_i-\boldsymbol{\mu}_t)^T\mathbf{R}_{t+1}^{-1}(\mathbf{x}_i-\boldsymbol{\mu}_t) + \gamma w_{\mathbf{t}}(\boldsymbol{\mu}_t,\mathbf{R}_{t+1})(\mathbf{t}-\boldsymbol{\mu}_t)^T\mathbf{R}_{t+1}^{-1}(\mathbf{t}-\boldsymbol{\mu}_t) + \mathrm{const.} \quad (21)$$

which leads to the update (23) in Algorithm 2 for $\boldsymbol{\mu}$.

Notice that the update order can be reversed, i.e., first fix $\mathbf{R}_t$ and get $\boldsymbol{\mu}_{t+1}$, then fix $\boldsymbol{\mu}_{t+1}$ and get $\mathbf{R}_{t+1}$. This leads to similar iterations to Algorithm 2.

Algorithm 2: Block Majorization-Minimization

1) Initialize $\mathbf{R}_0$ as an arbitrary positive definite matrix, and $\boldsymbol{\mu}_0$ as an arbitrary vector.

2) Iterate

$$\mathbf{R}_{t+1} = \frac{K+1}{N+2\alpha}\sum_{i=1}^N w_i(\boldsymbol{\mu}_t,\mathbf{R}_t)(\mathbf{x}_i-\boldsymbol{\mu}_t)(\mathbf{x}_i-\boldsymbol{\mu}_t)^T + \frac{2\gamma}{N+2\alpha}w_{\mathbf{t}}(\boldsymbol{\mu}_t,\mathbf{R}_t)(\mathbf{t}-\boldsymbol{\mu}_{\mathbf{t}})(\mathbf{t}-\boldsymbol{\mu}_{\mathbf{t}})^T + \frac{2\alpha K}{N+2\alpha}\frac{\mathbf{T}}{\mathrm{Tr}\left(\mathbf{R}_t^{-1}\mathbf{T}\right)} \quad (22)$$
$$\boldsymbol{\mu}_{t+1} = \frac{(K+1)\sum_{i=1}^N w_i(\boldsymbol{\mu}_t,\mathbf{R}_{t+1})\mathbf{x}_i + 2\gamma w_{\mathbf{t}}(\boldsymbol{\mu}_t,\mathbf{R}_{t+1})\mathbf{t}}{(K+1)\sum_{i=1}^N w_i(\boldsymbol{\mu}_t,\mathbf{R}_{t+1}) + 2\gamma w_{\mathbf{t}}(\boldsymbol{\mu}_t,\mathbf{R}_{t+1})} \quad (23)$$

with $w_i(\boldsymbol{\mu},\mathbf{R})$ and $w_{\mathbf{t}}(\boldsymbol{\mu},\mathbf{R})$ given in (18) until convergence.

Proposition 8: The sequences $\{(\boldsymbol{\mu}_t,\mathbf{R}_t)\}$ generated by Algorithm 2 converge to

(i) the set of stationary points of problem (16) if $\alpha>\gamma$;
(ii) the global minimizer of problem (16) if $\alpha\le\gamma$.

Proof: We verify the sufficient conditions (**C1**), (**C2**) and (**C3**) for block MM established in [37]. First, the level set $\mathcal{X}^0$ is compact by Lemma 5. Second, with fixed $\boldsymbol{\mu}_t$, the surrogate function $L(\mathbf{R}|\boldsymbol{\mu}_t,\mathbf{R}_t)$ upperbounds $L^{\mathrm{shrink}}(\boldsymbol{\mu}_t,\mathbf{R})$, therefore $L(\mathbf{R}|\boldsymbol{\mu}_t,\mathbf{R}_t)\to+\infty$ when $\mathbf{R}$ goes to the boundary of $\mathbb{S}_{++}^K$. This implies that $\mathbf{R}_{t+1}$ given by (22), which satisfies the stationary condition of $L(\mathbf{R}|\boldsymbol{\mu}_t,\mathbf{R}_t)$, is the unique minimizer of $L(\mathbf{R}|\boldsymbol{\mu}_t,\mathbf{R}_t)$. Similarly we can prove that $\boldsymbol{\mu}_{t+1}$ given by (23) is the unique minimizer of $L(\boldsymbol{\mu}|\boldsymbol{\mu}_t,\mathbf{R}_{t+1})$. Condition (**C3**) is satisfied naturally. Therefore $\{(\boldsymbol{\mu}_t,\mathbf{R}_t)\}$ converges to the set of stationary points of problem (16), which is unique and the global minimum when $\alpha\le\gamma$. ∎

C. Special Case for $\alpha=\gamma$

In this subsection we provide an algorithm for the special case $\alpha=\gamma$, which is simpler than the previously described ones. It has been proved in Theorem 4 that when $\alpha=\gamma$, the objective function $L^{\mathrm{shrink}}(\boldsymbol{\mu},\mathbf{R})$ is scale-invariant and has a unique

minimizer on $\mathbb{R}^K \times \mathbb{S}_+^K$ up to a positive scale factor. The $\boldsymbol{\Sigma}$ reparametrization (6) yields the following equivalent optimization problem:

$$\begin{aligned}\underset{\boldsymbol{\Sigma}\succcurlyeq\mathbf{0}}{\text{minimize}} \quad & \left(\frac{N}{2}+\alpha\right)\log\det(\boldsymbol{\Sigma})+\frac{K+1}{2}\sum_{i=1}^{N}\log\left(\bar{\mathbf{x}}_i^T\boldsymbol{\Sigma}^{-1}\bar{\mathbf{x}}_i\right)\\ & +\alpha K\log\left(\mathrm{Tr}\left(\mathbf{S}^T\boldsymbol{\Sigma}^{-1}\mathbf{S}\mathbf{T}\right)\right)+\alpha\log\left(\bar{\mathbf{t}}^T\boldsymbol{\Sigma}^{-1}\bar{\mathbf{t}}\right)\\ \text{subject to} \quad & \boldsymbol{\Sigma}_{K+1,K+1}=1. \end{aligned} \tag{24}$$

For this special case, in addition to Algorithm 1 and 2, we propose a more efficient one as described in Algorithm 3. The convergence follows the same reasoning as Proposition 17 in [22].

Algorithm 3: Algorithm for $\alpha = \gamma$

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

2) Iterate

$$\tilde{\boldsymbol{\Sigma}}_{t+1}=\frac{K+1}{N+2\alpha}\sum_{i=1}^{N}\frac{\bar{\mathbf{x}}_i\bar{\mathbf{x}}_i^T}{\bar{\mathbf{x}}_i^T\boldsymbol{\Sigma}_t^{-1}\bar{\mathbf{x}}_i}+\frac{2\alpha}{N+2\alpha}\left(\frac{K\mathbf{S}\mathbf{T}\mathbf{S}^T}{\mathrm{Tr}\left(\mathbf{S}^T\boldsymbol{\Sigma}_t^{-1}\mathbf{S}\mathbf{T}\right)}+\frac{\bar{\mathbf{t}}\bar{\mathbf{t}}^T}{\bar{\mathbf{t}}^T\boldsymbol{\Sigma}_t^{-1}\bar{\mathbf{t}}}\right)$$

$$\boldsymbol{\Sigma}_{t+1}=\frac{\tilde{\boldsymbol{\Sigma}}_{t+1}}{(\tilde{\boldsymbol{\Sigma}}_{t+1})_{K+1,K+1}}$$

until convergence.

D. Accelerated Majorization-Minimization

Recall it is proved in Theorem 4 that the stationary condition (10) can be embedded in (30), which is equivalent to the equation system (31)–(33).

Since ζ can be solved as $\zeta = \frac{N+2\gamma}{N+2\alpha}$, we can substitute the value of ζ into (32) and eliminate (33). This leads to the iterations described in Algorithm 4. Compared to Algorithm 1, the update equation of $\boldsymbol{\mu}_{t+1}$ in Algorithm 4 is unchanged, while that of $\mathbf{R}_{t+1}$ is multiplied by a factor of β_t. In the case that $\alpha = \gamma$, Algorithm 4 and Algorithm 3 are identical.

Algorithm 4: Accelerated Majorization-Minimization

1) Initialize $\mathbf{R}_0$ as an arbitrary positive definite matrix, and $\boldsymbol{\mu}_0$ as an arbitrary vector.

2) Iterate

$$\boldsymbol{\mu}_{t+1}=\frac{(K+1)\sum_{i=1}^{N}w_i(\boldsymbol{\mu}_t,\mathbf{R}_t)\mathbf{x}_i+2\gamma w_{\mathbf{t}}(\boldsymbol{\mu}_t,\mathbf{R}_t)\mathbf{t}}{(K+1)\sum_{i=1}^{N}w_i(\boldsymbol{\mu}_t,\mathbf{R}_t)+2\gamma w_{\mathbf{t}}(\boldsymbol{\mu}_t,\mathbf{R}_t)}$$

$$\begin{aligned}\mathbf{R}_{t+1}=\beta_t\Bigg\{&\frac{K+1}{N+2\alpha}\sum_{i=1}^{N}w_i(\boldsymbol{\mu}_t,\mathbf{R}_t)(\mathbf{x}_i-\boldsymbol{\mu}_{t+1})(\mathbf{x}_i-\boldsymbol{\mu}_{t+1})^T\\ &+\frac{2\gamma}{N+2\alpha}w_{\mathbf{t}}(\boldsymbol{\mu}_t,\mathbf{R}_t)(\mathbf{t}-\boldsymbol{\mu}_{t+1})(\mathbf{t}-\boldsymbol{\mu}_{t+1})^T\\ &+\frac{2\alpha K}{N+2\alpha}\frac{\mathbf{T}}{\mathrm{Tr}\left(\mathbf{R}_t^{-1}\mathbf{T}\right)}\Bigg\}\end{aligned} \tag{25}$$

with $w_i(\boldsymbol{\mu},\mathbf{R})$ and $w_{\mathbf{t}}(\boldsymbol{\mu},\mathbf{R})$ given in (18) and

$$\beta_t=\frac{N+2\gamma}{(K+1)\sum_{i=1}^{N}w_i(\boldsymbol{\mu}_t,\mathbf{R}_t)+2\gamma w_{\mathbf{t}}(\boldsymbol{\mu}_t,\mathbf{R}_t)},$$

until convergence.

V. Numerical Results

In this section, we conduct a simulation study of the performance of the proposed shrinkage estimator and the convergence of the numerical algorithms presented in Section IV. The estimation error is measured by the symmetrized KL divergence

$$\begin{aligned}\mathrm{err}(\hat{\boldsymbol{\mu}},\hat{\mathbf{R}})=\frac{1}{2}E\Big\{&D_{KL}\left(\mathcal{N}(\hat{\boldsymbol{\mu}},\hat{\mathbf{R}})\|\mathcal{N}(\boldsymbol{\mu}_0,\mathbf{R}_0)\right)\\ &+D_{KL}\left(\mathcal{N}(\boldsymbol{\mu}_0,\mathbf{R}_0)\,\Big\|\,\mathcal{N}(\hat{\boldsymbol{\mu}},\hat{\mathbf{R}})\right)\Big\},\end{aligned}$$

where all covariance matrices are normalized by their traces. The expected error of the estimator is approximated by averaging 200 independent simulations with randomly generated data sets following the same underlying distribution. In all the following simulations, if not specified, the scatter parameter $\mathbf{R}_0(\beta)$ is set to be a Toeplitz matrix of the form

$$(\mathbf{R}_0)_{ij}=\beta^{|i-j|}$$

with $K = 100$, $\beta = 0.8$, and the distribution center $\boldsymbol{\mu}_0$ is fixed to be $\mathbf{1}$.

The first simulation compares the estimation error of the sample average estimator, the Cauchy MLE and the MLE with samples drawn from a Student's t-distribution with degree of freedom parameter ν, denoted as $t_\nu(\boldsymbol{\mu}_0,\mathbf{R}_0)$. In this simulation, ν varies from 1 to 100 and the number of samples N is set to be 120. The NMSE of $\hat{\boldsymbol{\mu}}$ and $\hat{\mathbf{R}}$ plotted in Fig. 2 illustrates that even in the case that the tail of the underlying distribution is not as heavy as the Cauchy distribution, which corresponds to a large value of ν, the estimation error incurred by fitting the samples to a Cauchy distribution is small compared to the MLE that assumes perfect knowledge of ν. For this reason, the negative log-likelihood function of Cauchy distribution is an acceptable choice as the loss function of estimating the parameters $(\boldsymbol{\mu},\mathbf{R})$ of an elliptical distribution.

The second simulation shows the performance of the proposed shrinkage estimator with a small sample size. The estimators that we consider are the sample average, the Cauchy MLE, the plug-in estimator (first estimate the location by sample mean or sample median then estimate covariance by shrinkage Tyler's estimator [22] with the estimated mean), and the proposed shrinkage Cauchy MLE. As for the tuning parameter of the shrinkage estimators, we define $\rho(\alpha) = \frac{N}{N+2\alpha}$ and search for the $\rho^\star$ on the grid $\{0.1, 0.2. \ldots, 1\}$ that yields the shrinkage estimator with the smallest estimation error as proposed in [20], so as to eliminate the effect of parameter tuning that is important but not the focus of this paper. The shrinkage target $\mathbf{T}$ is set to be $\mathbf{I}/K$ motivated by the idea of diagonal loading [17] and $\mathbf{t}$ is set to be the sample mean and the sample median. Notice that in both cases the shrinkage target does not depend on any prior knowledge of the true parameter. Fig. 3 shows that for Gaussian distributed samples, heavy-tailed samples $\mathbf{x}_i \sim t_3(\boldsymbol{\mu}_0,\mathbf{R}_0)$ and elliptically distributed samples $\mathbf{x}_i \sim \boldsymbol{\mu}+\sqrt{\tau}\mathbf{u}$, where $\tau \sim \chi^2$ and $\mathbf{u} \sim \mathcal{N}(\mathbf{0},\mathbf{R})$, the proposed shrinkage estimator with the shrinkage target being the sample mean achieves the smallest estimation error. The proposed shrinkage estimator is robust to the class of outliers that are distributed far away from the "good" samples. A plot in Fig. 3(d) shows the estimation error versus the percentage of outliers with normal samples

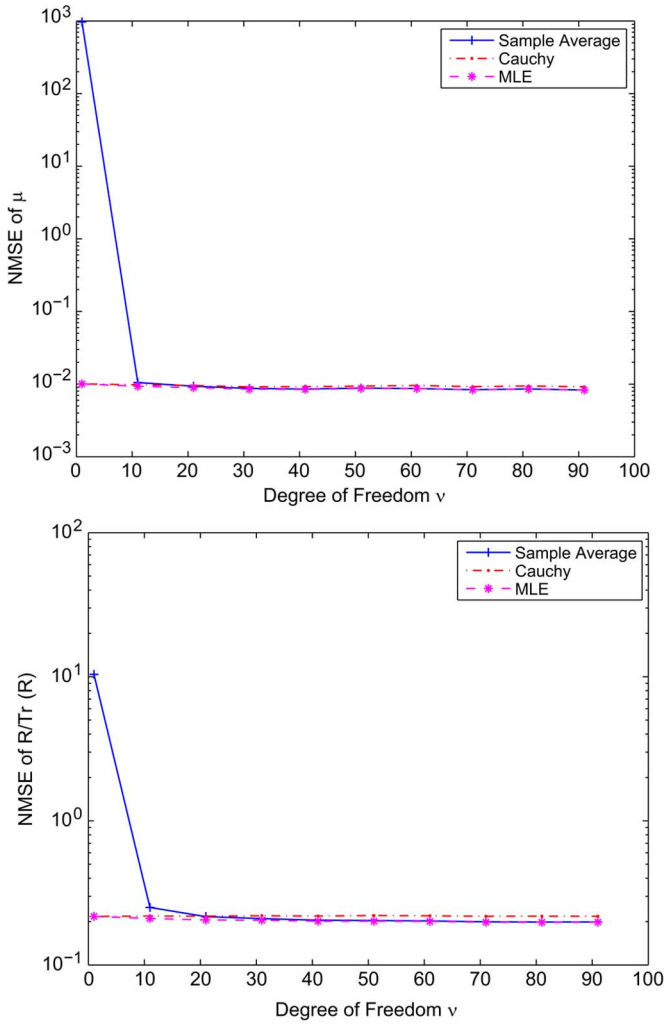


Fig. 2. NMSE of $\hat{\boldsymbol{\mu}}$ and $\hat{\mathbf{R}}$ with $N = 120$ 100-dimensional samples drawn from a Student's t-distribution with degree of freedom parameter ν.

$\mathbf{x}_i \sim \mathcal{N}(\boldsymbol{\mu}_0, \mathbf{R}_0)$ and outliers $\mathbf{x}_{\text{outlier}} \sim \boldsymbol{\mu}_0 + r\mathbf{s}$, where $\mathbf{s}$ is uniformly distributed on a $(K-1)$-dimensional sphere and r is uniformly distributed on the interval $[2l, 2l+1]$, l is set to be $l \triangleq \max\{\|\mathbf{x}_i\|_2\}$. The total number of samples is $N = 120$.

The third simulation analyzes the sensitivity of the shrinkage estimator to the shrinkage target $(\mathbf{t}, \mathbf{T})$. The underlying distribution is chosen to be $t_3(\boldsymbol{\mu}_0, \mathbf{R}_0)$. The estimation error of the shrinkage estimator with $\mathbf{t} = t\mathbf{1}$, $t \in \{0.1, 0.3, \ldots, 0.9\}$, and $\mathbf{T}$ being a Toeplitz matrix $\mathbf{R}_0(\beta)$, $\beta \in \{0.1, 0.3, 0.5, 0.7\}$ is listed in Table I for $N = 80$ and 150 respectively. The table indicates that the estimation error decreases as $(\mathbf{t}, \mathbf{T})$ gets closer to the true parameter $(\boldsymbol{\mu}_0, \mathbf{R}_0)$. For $N = 150$, the estimation error of the MLE is 260.84, which turns out to be much larger than the maximum error for the $N = 150$ case in Table I. The reason is that although $\mathbf{t} = 0.1 \times \mathbf{1}$ is far away from $\boldsymbol{\mu}_0$, the regularization parameter γ is small so that $\boldsymbol{\mu}$ is estimated majorly based on the samples, and $\mathbf{T}$ is close to the identity matrix $\mathbf{I}$ when $\beta = 0.1$, which still helps in improving the estimation accuracy by shrinking the eigenvalues of $\hat{\mathbf{R}}$ towards the center in the small sample regime. To conclude, one can expect that a more informative prior ($(\mathbf{t}, \mathbf{T})$ is closer to $(\boldsymbol{\mu}_0, \mathbf{R}_0)$) leading to a more accurate estimator $(\hat{\boldsymbol{\mu}}, \hat{\mathbf{R}})$. Even if the prior on the parameters is completely wrong, the shrinkage estimator performs

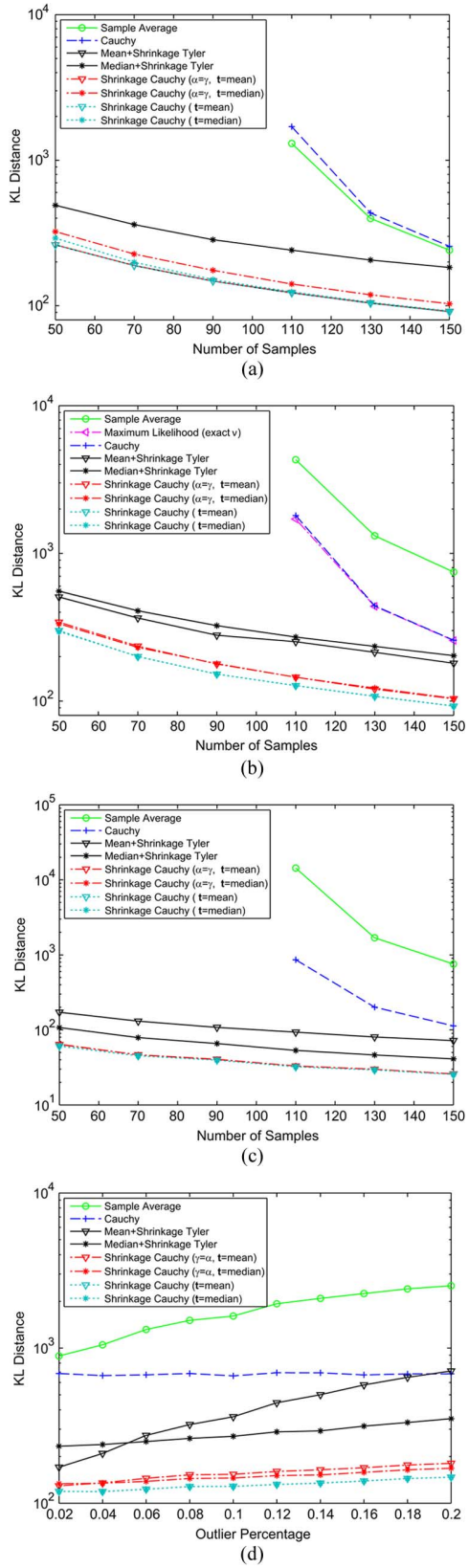


Fig. 3. Performance comparison for different estimators. (a) Gaussian distributed samples (b) Student's t-distributed samples ($\nu = 3$) (c) Elliptically distributed samples ($\mathbf{x}_i \sim \sqrt{\tau}\mathbf{u}$) (d) Gaussian distributed samples with outlier contamination.

TABLE I
SENSITIVITY ANALYSIS: AVERAGED ESTIMATION ERROR OF THE PROPOSED SHRINKAGE ESTIMATOR FOR DIFFERENT VALUES OF $(\mathbf{t}, \mathbf{T})$

	$\beta = 0.1$		$\beta = 0.3$		$\beta = 0.5$		$\beta = 0.7$	
	$N = 80$	$N = 150$	$N = 80$	$N = 150$	$N = 80$	$N = 150$	$N = 80$	$N = 150$
$t = 0.1$	153.47	86.01	138.48	80.56	125.42	73.76	110.48	63.24
$t = 0.3$	148.87	84.25	133.98	79.00	121.18	72.22	105.22	61.58
$t = 0.5$	138.67	80.36	124.42	75.46	112.06	68.72	94.33	57.92
$t = 0.7$	111.14	69.47	100.57	64.90	87.86	58.54	68.90	47.63
$t = 0.9$	52.78	37.01	45.20	32.47	35.42	27.85	20.09	17.66

TABLE II
AVERAGE NUMBER OF ITERATIONS REQUIRED FOR ALGORITHMS, I.E., MM, BLOCK MM AND ACCELERATED MM, TO CONVERGE

	MM			Block MM (first update $\mathbf{R}$ then $\boldsymbol{\mu}$)			Block MM (first update $\boldsymbol{\mu}$ then $\mathbf{R}$)			Accelerated MM		
	$K = 30$	$K = 50$	$K = 100$	$K = 30$	$K = 50$	$K = 100$	$K = 30$	$K = 50$	$K = 100$	$K = 30$	$K = 50$	$K = 100$
$\rho = 0.2$	1015.07	1596.16	2944.42	1021.06	1604.055	2955.325	1028.85	1622.74	3007.825	55.215	56.215	57.93
$\rho = 0.5$	419.505	666.975	1250.175	421.045	668.91	1252.775	426.615	680.485	1281.28	23.995	23.94	24.06
$\rho = 0.8$	309.73	484.775	907.175	311.96	487.66	911.045	321.175	507.835	963.06	40.51	37.4	34.83

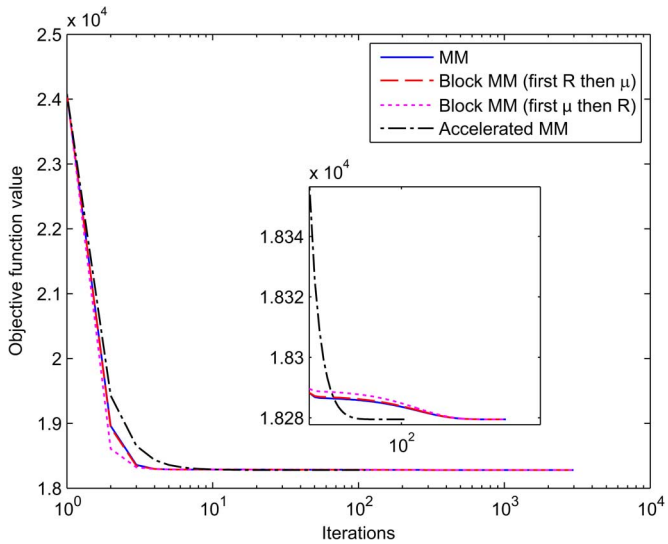


Fig. 4. Convergence comparison for algorithms in Section IV.

no worse than the Cauchy MLE provided that the tuning parameters α and γ are properly selected.

Fig. 4 and Table II demonstrate the convergence of the algorithms provided in Section IV. The convergence criterion is set to be $\|\mathbf{R}_t - \mathbf{R}_{t+1}\|_F < 10^{-5}$ and $\|\boldsymbol{\mu}_t - \boldsymbol{\mu}_{t+1}\|_F < 10^{-5}$. Fig. 4 plots the evolution curve of the objective value versus the number of iterations. The parameters are set to be $N = 80$, $K = 100$, $\mathbf{t} = 0.9 \times \mathbf{1}$, $\mathbf{T} = \mathbf{I}/K$, $\gamma = 360$ and $\alpha = 160$. While the computational cost per iteration of all the algorithms is roughly the same, it can be seen that the accelerated algorithm requires far fewer iterations than the MM algorithms (MM and block MM). Table II lists the average number of iterations that each of the algorithms require to converge as α and K changes. In this simulation, α and γ are set to be equal, consequently Algorithm 3 and Algorithm 4 turn out to be the same. The ratio N/K is fixed to be 1.2. The data indicates that the accelerated MM algorithm not only converges much faster, but also is not very sensitive to the problem dimension. For these reasons, the accelerated MM is recommended for practical applications.

Finally, we test the performance of the proposed shrinkage estimator on a real financial data set. We choose weekly close prices p_t from Jan 1, 2010 to June 8, 2014, 230 weeks in total, of $K = 40$ stocks selected from the S&P 500 index components provided by Yahoo Finance. The samples are constructed as $r_t = \log p_t - \log p_{t-1}$, i.e., the weekly log-returns. The process r_t is assumed stationary. The vector $\mathbf{r}_t$ is constructed by stacking the log-returns of all K stocks. We compare estimator performance in the minimum variance portfolio set up, that is, we allocate the portfolio weights to minimize the overall variance. The problem can be formulated mathematically as

$$\begin{aligned} \underset{\mathbf{w}}{\text{minimize}} \quad & \mathbf{w}^T \boldsymbol{\Sigma} \mathbf{w} \\ \text{subject to} \quad & \mathbf{1}^T \mathbf{w} = 1 \end{aligned} \tag{26}$$

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

To estimate $\boldsymbol{\Sigma}$, we use a rolling window approach with window size N. In particular, for the nonshrinkage estimators, at week t we use the log-returns of the previous N weeks to estimate the normalized covariance $\boldsymbol{\Sigma}$ and find the optimal portfolio allocation $\mathbf{w}_t^\star$ according to problem (26). For the shrinkage estimators, we further divide the N weeks returns into two parts with the first $N^{\text{train}} = N - N^{\text{val}}$ weeks for estimating $\boldsymbol{\Sigma}$ for different regularization parameters (α, γ) and find the allocation strategy $\mathbf{w}$, and the remaining N^{val} weeks as validation data for selecting the best $(\alpha^\star, \gamma^\star)$, which is the one that yields the smallest empirical portfolio variance $\text{Var}(\{\mathbf{w}^T \mathbf{r}_i\}_{i=t-N^{\text{val}},\ldots,t-1})$. The normalized covariance $\boldsymbol{\Sigma}$ is then estimated with the overall N weeks log-return and tuning parameter $(\alpha^\star, \gamma^\star)$.

Then we use $\mathbf{w}_t^\star$ to invest for N^{test} weeks and collect the returns $v_t = (\mathbf{w}_t^\star)^T \mathbf{r}_t$. Every N^{test} weeks we rebalance the portfolio based on this procedure. Fig. 5 compares the variance (risk) of the portfolio constructed based on different estimators.

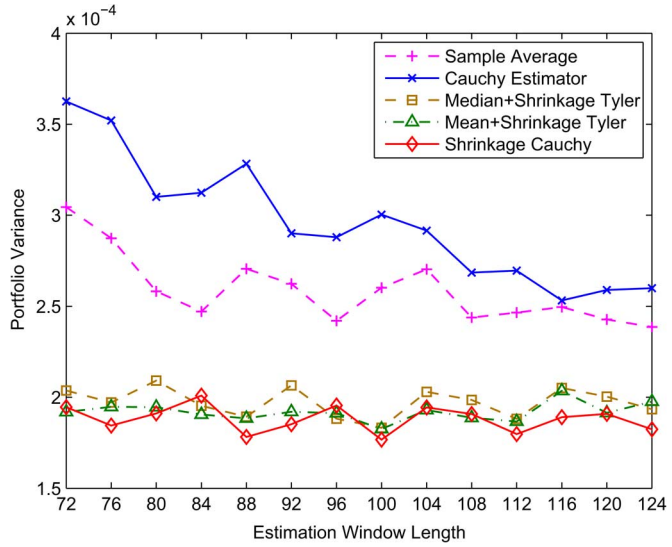


Fig. 5. Risk (variance) comparison of portfolio constructed based on different covariance estimators.

The parameters are set to be $N^{\text{val}} = 24$ and $N^{\text{test}} = 12$, the shrinkage targets are $\mathbf{t} = \mathbf{0}$ and $\mathbf{T} = \mathbf{I}/K$, and the step size of the searching grid for (α, γ) is set to be 0.2. The size of the rolling window N is varied from 72 to 124. We can see that in most cases the proposed shrinkage Cauchy MLE yields the lowest risk.

VI. Conclusion

In this paper, we have considered the robust mean-covariance estimation with samples drawn from an elliptical distribution that is capable of modeling heavy tails and outliers. In particular, we have proposed a robust shrinkage estimator by adding a penalty term to the Cauchy likelihood function, and established the existence and uniqueness result of the shrinkage estimator under certain regularity conditions. Efficient numerical algorithms have been provided based on the majorization-minimization framework with provable convergence and simulation results have shown that the proposed estimator works considerably better in the small sample scenario with the presence of erroneous observations.

Appendix

A. Proof for Proposition 1

We analyze the following nested optimization problem:

$$\min_{\mathbf{R}\succ\mathbf{0}} \min_{\boldsymbol{\mu}} \alpha \left(K \log\left(\operatorname{Tr}(\mathbf{R}^{-1}\mathbf{T})\right) + \log\det(\mathbf{R})\right) + \gamma \log\left(1 + (\boldsymbol{\mu}-\mathbf{t})^T\mathbf{R}^{-1}(\boldsymbol{\mu}-\mathbf{t})\right).$$

For any fixed value of $\mathbf{R} \succ \mathbf{0}$, it is easy to see that the optimal $\boldsymbol{\mu}$ is $\boldsymbol{\mu}^\star = \mathbf{t}$. Substituting the optimal $\boldsymbol{\mu}$ into the problem we have

$$\underset{\mathbf{R}\succ\mathbf{0}}{\text{minimize}} \; K \log\left(\operatorname{Tr}(\mathbf{R}^{-1}\mathbf{T})\right) + \log\det(\mathbf{R}).$$

Setting the gradient to zero yields the stationary condition

$$\mathbf{R} = \frac{K\mathbf{T}}{\operatorname{Tr}(\mathbf{R}^{-1}\mathbf{T})}$$

with solution $\mathbf{R} = r\mathbf{T}$ for any $r > 0$. We now show that the stationary points $\mathbf{R} = r\mathbf{T}$ are actually the minimizers.

We claim that the objective function $h(\mathbf{t}, \mathbf{R}) = K \log(\operatorname{Tr}(\mathbf{R}^{-1}\mathbf{T})) + \log\det(\mathbf{R})$ goes to positive infinity when $\mathbf{R}$ tends to a singular matrix. To see this, first notice that $h(\mathbf{t}, \mathbf{R})$ is scale-invariant, i.e., $h(\mathbf{t}, \mathbf{R}) = h(\mathbf{t}, r\mathbf{R})$, therefore we add the constraint $\operatorname{Tr}(\mathbf{R}) = 1$ to remove the scale ambiguity. Eigendecompose $\mathbf{R}$ as $\mathbf{R} = \mathbf{U}\boldsymbol{\Lambda}\mathbf{U}^T$, and define $\tilde{\mathbf{T}} = \mathbf{U}^T\mathbf{T}\mathbf{U}$. The diagonal components of $\tilde{\mathbf{T}}$ are all positive since $\mathbf{T} \succ \mathbf{0}$. Therefore $h(\mathbf{t}, \mathbf{R}) = K \log(\sum_{i=1}^K \lambda_i^{-1}\tilde{t}_{ii}) + \sum_{i=1}^K \log\lambda_i$ with λ_i being the i-th diagonal component of $\boldsymbol{\Lambda}$ and $\tilde{t}_{ii}$ being the i-th diagonal component of $\tilde{\mathbf{T}}$. Without loss of generality we can assume the ordering $\lambda_1 \geq \cdots \geq \lambda_K$. Consider the case $\lambda_j \to 0$ for some $1 < j \leq K$, then we have

$$\begin{aligned} h(\mathbf{t}, \mathbf{R}) &\geq K \log\left(\sum_{i=j}^K \lambda_i^{-1}\tilde{t}_{ii}\right) + \sum_{i=j}^K \log\lambda_i + \text{const.} \\ &\geq \frac{K\sum_{i=j}^K \log\lambda_i^{-1}}{K-j+1} + \sum_{i=j}^K \log\lambda_i + \text{const.} \to +\infty. \end{aligned}$$

Therefore, a minimizer of $h(\mathbf{t}, \mathbf{R})$ exists on $\mathbb{S}_{++}^K$ and has to satisfy the stationary condition. The scale-invariant property of $h(\mathbf{t}, \mathbf{R})$ implies $r\mathbf{T}$ must be a global minima.

B. Proof for Theorem 2

Notice the fact that $L^{\text{shrink}}(\boldsymbol{\mu}, \mathbf{R}) \to +\infty$ on the boundary of the feasible set $\mathbb{R}^K \times \mathbb{S}_{++}^K$ implies the minimum exists, we therefore seek for the condition that guarantees $L^{\text{shrink}}(\boldsymbol{\mu}, \mathbf{R}) \to +\infty$ on the boundary.

Define $\bar{\mathbf{x}}_i = [\mathbf{x}_i; 1]$, $\bar{\mathbf{t}} = [\mathbf{t}; 1]$ and matrix

$$\boldsymbol{\Sigma} = \begin{bmatrix} \mathbf{R} + \boldsymbol{\mu}\boldsymbol{\mu}^T & \boldsymbol{\mu} \\ \boldsymbol{\mu}^T & 1 \end{bmatrix},$$

we have the following identities

$$\begin{aligned} \bar{\mathbf{x}}_i^T\boldsymbol{\Sigma}^{-1}\bar{\mathbf{x}}_i &= 1 + (\mathbf{x}_i - \boldsymbol{\mu})^T\mathbf{R}^{-1}(\mathbf{x}_i - \boldsymbol{\mu}) \\ \bar{\mathbf{t}}^T\boldsymbol{\Sigma}^{-1}\bar{\mathbf{t}} &= 1 + (\mathbf{t} - \boldsymbol{\mu})^T\mathbf{R}^{-1}(\mathbf{t} - \boldsymbol{\mu}) \\ \mathbf{S}^T\boldsymbol{\Sigma}^{-1}\mathbf{S} &= \mathbf{R}^{-1} \end{aligned} \tag{27}$$

with $\mathbf{S} = \left[\begin{smallmatrix} \mathbf{I}_K \\ \mathbf{0}_{1\times K} \end{smallmatrix}\right]$. The loss function $L^{\text{shrink}}(\boldsymbol{\mu}, \mathbf{R})$ can be equivalently written in $\boldsymbol{\Sigma}$ as

$$\begin{aligned} &L^{\text{shrink}}(\boldsymbol{\mu}, \mathbf{R}) = L^{\text{shrink}}(\boldsymbol{\Sigma}) \\ &= \left(\alpha + \frac{N}{2}\right)\log\det(\boldsymbol{\Sigma}) + \frac{K+1}{2}\sum_i \log\left(\bar{\mathbf{x}}_i^T\boldsymbol{\Sigma}^{-1}\bar{\mathbf{x}}_i\right) \\ &\quad + \gamma\log\left(\bar{\mathbf{t}}^T\boldsymbol{\Sigma}^{-1}\bar{\mathbf{t}}\right) + \alpha K\log\left(\operatorname{Tr}\left(\mathbf{S}^T\boldsymbol{\Sigma}^{-1}\mathbf{S}\mathbf{T}\right)\right). \end{aligned}$$

Define the feasible set of $\boldsymbol{\Sigma}$ as $\mathcal{S} = \{\boldsymbol{\Sigma} | \boldsymbol{\Sigma} \succ \mathbf{0},\ \boldsymbol{\Sigma}_{K+1,K+1} = 1\}$. In the rest of the proof, we are going to find the condition that ensures $L^{\text{shrink}}(\boldsymbol{\Sigma}) \to +\infty$ as $(\lambda_{\max}(\boldsymbol{\Sigma})/\lambda_{\min}(\boldsymbol{\Sigma})) \to +\infty$ for all $\boldsymbol{\Sigma} \in \mathcal{S}$, which implies that a minimum exists on $\mathcal{S}$. Denote the eigenvalues of $\boldsymbol{\Sigma}$ as $\lambda_1 \geq \cdots \geq \lambda_{K+1}$, on the set $\mathcal{S}$ we have $\lambda_1 \geq 1$ and $\lambda_{K+1} \leq 1$.

The quantities $\bar{\mathbf{x}}_i^T\boldsymbol{\Sigma}^{-1}\bar{\mathbf{x}}_i$ and $\bar{\mathbf{t}}^T\boldsymbol{\Sigma}^{-1}\bar{\mathbf{t}}$ are all greater than or equal to 1 by the identities (27), therefore the corre-

sponding terms in $L^{\text{shrink}}(\mathbf{\Sigma})$ are nonnegative. For the term $\text{Tr}(\mathbf{S}^T\mathbf{\Sigma}^{-1}\mathbf{S}\mathbf{T})$ we have

$$\text{Tr}(\mathbf{S}^T\mathbf{\Sigma}^{-1}\mathbf{S}\mathbf{T}) = \text{Tr}\left(\left[\mathbf{T}^{\frac{1}{2}} \quad \mathbf{0} \right] \mathbf{\Sigma}^{-1} \begin{bmatrix} \mathbf{T}^{\frac{1}{2}} \\ \mathbf{0} \end{bmatrix} \right).$$

Eigendecompose $\mathbf{\Sigma}$ as $\mathbf{\Sigma} = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^T$ with $\mathbf{\Lambda} \triangleq \text{diag}(\lambda_1, \ldots, \lambda_{K+1})$, and denote the eigenvector corresponding to λ_j as $\mathbf{u}_j$, we can express $\text{Tr}(\mathbf{S}^T\mathbf{\Sigma}^{-1}\mathbf{S}\mathbf{T})$ as $\text{Tr}(\mathbf{S}^T\mathbf{\Sigma}^{-1}\mathbf{S}\mathbf{T}) = \sum_j \lambda_j^{-1}\|\tilde{\mathbf{t}}_j\|^2$, where $\tilde{\mathbf{t}}_j = [\mathbf{T}^{1/2} \quad \mathbf{0}]\,\mathbf{u}_j$.

Now define the function

$$\begin{aligned} G(\mathbf{\Sigma}) &= \exp\left(-L^{\text{shrink}}(\mathbf{\Sigma})\right) \\ &= \det(\mathbf{\Sigma})^{-\frac{N}{2}-\alpha} \prod_i \left(\bar{\mathbf{x}}_i^T\mathbf{\Sigma}^{-1}\bar{\mathbf{x}}_i\right)^{-\frac{K+1}{2}} \cdot \\ &\quad (\bar{\mathbf{t}}^T\mathbf{\Sigma}^{-1}\bar{\mathbf{t}})^{-\gamma} \left(\sum_{j=1}^{K+1} \lambda_j^{-1}\|\tilde{\mathbf{t}}_j\|^2\right)^{-\alpha K}. \end{aligned}$$

The function $L^{\text{shrink}}(\mathbf{\Sigma}) \to +\infty$ if and only if $G(\mathbf{\Sigma}) \to 0$.

Denote the subspace spanned by $\{\mathbf{u}_1, \ldots, \mathbf{u}_j\}$ as S_j and define $D_j = S_j \setminus S_{j-1} = \{\mathbf{x} \in \mathbb{R}^{K+1} | \mathbf{x} \in S_j, \mathbf{x} \notin S_{j-1}\}$ with $S_0 = \{0\}$ and $D_0 = \{0\}$. The D_j's partition the whole $\mathbb{R}^{K+1}$ space. Notice that $P_N\{S_0\} = 0$ since the last element of the augmented sample $\bar{\mathbf{x}}_i$ is 1, therefore $\sum_{j=1}^m P_N(D_j) = P_N(S_m)$ and $\sum_{j=m}^{K+1} P_N(D_j) = 1 - P_N(S_{m-1})$.

Partition the samples $\bar{\mathbf{x}}_i$ according to D_j's and define

$$G_j = \lambda_j^{-\frac{N}{2}-\alpha} \prod_{\bar{\mathbf{x}}_i \in D_j} \left(\bar{\mathbf{x}}_i^T\mathbf{\Sigma}^{-1}\bar{\mathbf{x}}_i\right)^{-\frac{K+1}{2}}$$

therefore the objective function can be written as $G(\mathbf{\Sigma}) = \prod_{j=1}^{K+1} G_j \cdot (\bar{\mathbf{t}}^T\mathbf{\Sigma}^{-1}\bar{\mathbf{t}})^{-\gamma}(\sum_{j=1}^{K+1}\lambda_j^{-1}\|\tilde{\mathbf{t}}_j\|^2)^{-\alpha K}$.

If $\bar{\mathbf{x}}_i \in D_h$, we have the following fact:

$$\bar{\mathbf{x}}_i^T\mathbf{\Sigma}^{-1}\bar{\mathbf{x}}_i = \sum_{j=1}^{h} \lambda_j^{-1} \left\|\bar{\mathbf{x}}_i^T\mathbf{u}_j\right\|^2 \geq \lambda_h^{-1} \left\|\bar{\mathbf{x}}_i^T\mathbf{u}_h\right\|^2 > 0.$$

Define two integers r and s with $0 \leq r \leq K, 1 \leq s \leq K+1$, and $r \leq s$, such that $\lambda_h \to +\infty$ for $h \in [1, r]$, λ_h is bounded for $h \in (r, s]$, and $\lambda_h \to 0$ for $h \in (s, K+1]$.

First consider the G_h's with $h \in [1, r]$. If $D_h \neq \emptyset$, then $\bar{\mathbf{x}}_i^T\mathbf{\Sigma}^{-1}\bar{\mathbf{x}}_i = \sum_{j=1}^h \lambda_j^{-1}\|\bar{\mathbf{x}}_i^T\mathbf{u}_j\|^2$. Since for all the λ_j's with $j \leq r$ all goes to infinity, the quantity $\bar{\mathbf{x}}_i^T\mathbf{\Sigma}^{-1}\bar{\mathbf{x}}_i$ must go to zero. This contradicts the fact that $\bar{\mathbf{x}}_i^T\mathbf{\Sigma}^{-1}\bar{\mathbf{x}}_i \geq 1$. Therefore no $\bar{\mathbf{x}}_i$ lies in D_h for $h \in [1, r]$, and $G_h = O(\lambda_h^{-(N/2)-\alpha})$.

Next, for the G_h's with $h \in (r, s]$, clearly G_h is bounded away from both 0 and $+\infty$, thus does not have any effect on the order of $G(\mathbf{\Sigma})$. Finally, consider the G_h's with $h \in (s, K]$, since $\lambda_h \to 0$, we have $\bar{\mathbf{x}}_i^T\mathbf{\Sigma}^{-1}\bar{\mathbf{x}}_i \to +\infty$. Since

$$+\infty > \lim_{\lambda_h \to 0} \left(\bar{\mathbf{x}}_i^T\mathbf{\Sigma}^{-1}\bar{\mathbf{x}}_i\right)\lambda_h > 0,$$

we have $\bar{\mathbf{x}}_i^T\mathbf{\Sigma}^{-1}\bar{\mathbf{x}}_i = O(\lambda_h^{-1})$, and $G_h = O(\lambda_h^{-(N/2)-\alpha+(K+1)/2NP_N(D_h)})$.

Having finished the terms associated with $\mathbf{x}_i$, we then analyze the terms contributed by regularization.

For the term $(\bar{\mathbf{t}}^T\mathbf{\Sigma}^{-1}\bar{\mathbf{t}})^{-\gamma}$, since the D_j's partition the whole space, there must exist some D_h that $\bar{\mathbf{t}} \in D_h$. By the previous analysis, we have $h > r$. If $h \leq s$ then $0 < (\bar{\mathbf{t}}^T\mathbf{\Sigma}^{-1}\bar{\mathbf{t}})^{-\gamma} < +\infty$, which is $O(1)$, and if $h \in (s, K]$, $(\bar{\mathbf{t}}^T\mathbf{\Sigma}^{-1}\bar{\mathbf{t}})^{-\gamma} = O(\lambda_h^{\gamma})$. In short, $(\bar{\mathbf{t}}^T\mathbf{\Sigma}^{-1}\bar{\mathbf{t}})^{-\gamma} = O(\lambda_h^{\gamma 1_{\{h>s\}}})$.

Finally we analyze the last term $(\sum_{j=1}^{K+1}\lambda_j^{-1}\|\tilde{\mathbf{t}}_j\|^2)^{-\alpha K}$. Recall the definition $\tilde{\mathbf{t}}_j = [\mathbf{T}^{\frac{1}{2}}\ \mathbf{0}]\mathbf{u}_j$, and denote each column of $\mathbf{T}^{\frac{1}{2}}$ as $\mathbf{t}_i \in \mathbb{R}^K$. Then for each vector $[\mathbf{t}_i; 0] \in \mathbb{R}^{K+1}$ there must exist some D_j to which it belongs. Denote the largest index of such D_j as q, we have $\|\tilde{\mathbf{t}}_q\| \neq 0$ and $\|\tilde{\mathbf{t}}_j\| = 0$ for all $j > q$. Repeating the previous reasoning we conclude that if $q \leq r$ or $q > s$, $(\sum_{j=1}^{K+1}\lambda_j^{-1}\|\tilde{\mathbf{t}}_j\|^2)^{-\alpha K} = O(\lambda_q^{\alpha K})$ and if $q \in (r, s]$, it is some constant. In short, $(\sum_{j=1}^{K+1}\lambda_j^{-1}\|\tilde{\mathbf{t}}_j\|^2)^{-\alpha K} = O(\lambda_q^{\alpha K 1_{\{q\leq r\}\cup\{q>s\}}})$.

Combining the three terms above, denote the partition that $\bar{\mathbf{t}}$ belongs to as D_h, i.e., $\bar{\mathbf{t}} \in D_h$, we have

$$\begin{aligned} G(\mathbf{\Sigma}) &= \prod_{j=1}^{K+1} G_j \cdot (\bar{\mathbf{t}}^T\mathbf{\Sigma}^{-1}\bar{\mathbf{t}})^{-\gamma} \left(\sum_{j=1}^{K+1}\lambda_j^{-1}\|\tilde{\mathbf{t}}_j\|^2\right)^{-\alpha K} \\ &= \prod_{j=1}^{r} O\left(\lambda_j^{-\frac{N}{2}-\alpha}\right) \prod_{j=s+1}^{K+1} O\left(\lambda_j^{-\frac{N}{2}-\alpha+\frac{K+1}{2}NP_N(D_j)}\right) \cdot \\ &\quad O\left(\lambda_h^{\gamma 1_{\{h>s\}}}\right) O\left(\lambda_q^{\alpha K 1_{\{q\leq r\}\cup\{q>s\}}}\right) \end{aligned}$$

with $\prod_{j=a}^b \triangleq 1$, if $a > b$. By the ordering $\lambda_1 \geq \cdots \geq \lambda_{K+1}$, to guarantee $G \to 0$ we impose the conditions

$$\left(-\frac{N}{2} - \alpha\right) m + \alpha K 1_{\{q\leq m\}} < 0, \forall 1 \leq m \leq r \tag{28}$$

and

$$\begin{aligned} &\left(-\frac{N}{2} - \alpha\right)(K+2-m) + \frac{K+1}{2}N\sum_{j=m}^{K+1} P_N(D_j) \\ &+\gamma 1_{\{m\leq h\}} + \alpha K 1_{\{m\leq q\}} > 0, \forall K+1 \geq m \geq s+1, \end{aligned} \tag{29}$$

where the first one forces the terms in the product corresponding to $\lambda \to +\infty$ to go to zero, and the second one forces the terms in the product corresponding to $\lambda \to 0$ to go to zero. Before simplifying the conditions, we first establish the following lemma.

Lemma 9: $q \leq m$ if and only if $S_m \supseteq \mathbb{R}^K$. If $S_m = \mathbb{R}^K$, the corresponding $\mathbf{\Sigma}$ takes the form $\mathbf{\Sigma} = \begin{bmatrix} \mathbf{R} & \mathbf{0} \\ \mathbf{0} & 1 \end{bmatrix}$, which implies $\boldsymbol{\mu} = \mathbf{0}$ and $\lambda_{K+1} = 1$.

Proof: Recall the definition of q, which is the largest index j of D_j that $[\mathbf{t}_i; 0]$ belongs to. Therefore $q \leq m$ if and only if $[\mathbf{t}_i; 0] \in S_m$ for all i. Under the assumption that $\mathbf{T}$ is full rank, we have $S_m \supseteq \mathbb{R}^K$. If $S_m = \mathbb{R}^K$, since S_m is a K-dimensional subspace spanned by $[\mathbf{u}_1, \ldots, \mathbf{u}_K]$, which are the eignenvectors of $\mathbf{\Sigma}$, we have $\mathbf{\Sigma} = \sum_{j=1}^K \lambda_j\mathbf{u}_j\mathbf{u}_j^T + \lambda_{K+1}\mathbf{e}_{K+1}\mathbf{e}_{K+1}^T$, where $\mathbf{e}_{K+1} \triangleq [\mathbf{0}_{K\times 1}; 1]$. Clearly $\mathbf{\Sigma}$ must take the form $\mathbf{\Sigma} = \begin{bmatrix} \mathbf{R} & \mathbf{0} \\ \mathbf{0} & 1 \end{bmatrix}$, which implies that $\boldsymbol{\mu} = \mathbf{0}$ and $\lambda_{K+1} = 1$. ∎

Consider the condition (28). r can take an arbitrary integer value from 0 to K. When $m \leq K-1$, or $m = K$ and $S_m \neq \mathbb{R}^K$,

the condition is satisfied automatically since $(-\frac{N}{2}-\alpha)m$ is always negative and $q > m$. When $S_m = \mathbb{R}^K$, $(-\frac{N}{2}-\alpha)K + \alpha K < 0$. We therefore have proved that condition (28) is satisfied for any proper subspace of $\mathbb{R}^{K+1}$.

Now we move to condition (29). Since $\sum_{j=m}^{K+1} P_N(D_j) = 1 - P_N(S_{m-1})$ and condition (29) should be valid for all $s \in [1, K+1]$. Substituting $d = m-1$ yields the condition that for any proper subspace S_d with $d \in [1, K]$

$$\left(-\frac{N}{2}-\alpha\right)(K+1-d) + \frac{K+1}{2}N\left(1 - P_N(S_d)\right) + \gamma 1_{\{d+1\le h\}} + \alpha K 1_{\{d+1\le q\}} > 0$$

We have $h \le d$ if $\bar{\mathbf{t}} \in S_d$ and $q \le d$ if $S_d = \mathbb{R}^K$ by the previous argument, therefore condition (29) is equivalent to: for any proper subspace $S \subseteq \mathbb{R}^{K+1}$, if $\bar{\mathbf{t}} \notin S$ and $S \neq \mathbb{R}^K$, we need

$$P_N(S) < \frac{\dim(S)(N+2\alpha) + 2\gamma - 2\alpha}{(K+1)N}$$

and if $\bar{\mathbf{t}} \in S$, which implies $S \neq \mathbb{R}^K$ since the last entry of $\bar{\mathbf{t}}$ is 1, we need

$$P_N(S) < \frac{\dim(S)(N+2\alpha) - 2\alpha}{(K+1)N}$$

and finally if $S = \mathbb{R}^K$, which implies $\bar{\mathbf{t}} \notin S$, which corresponds to the situation when $\lambda_K \to +\infty$ and $\lambda_{K+1} = 1$ according to Lemma 9, we actually do not have condition (29) since no $\lambda \to 0$.

Notice that $\bar{\mathbf{x}}$ and $\bar{\mathbf{t}}$ belong to the same subspace S if and only if $\mathbf{x}$ and $\mathbf{t}$ belong to the same hyperplane, and $\bar{\mathbf{x}} \in S$ if and only if $\mathbf{x} \in H$ with $\dim(H) = \dim(S) - 1$. Finally we arrive at the condition on samples: for any hyperplane $H \subset \mathbb{R}^K$ with dimension $0 \le \dim(H) < K$, if H contains $\mathbf{t}$,

$$P_N(H) < \frac{(2\alpha+N)\dim(H) + N}{(K+1)N}$$

if H does not contain $\mathbf{t}$,

$$P_N(H) < \frac{(2\alpha+N)\dim(H) + N + 2\gamma}{(K+1)N}.$$

C. Proof for Theorem 4

Define matrix

$$\tilde{\mathbf{\Sigma}} = \begin{bmatrix} \zeta^{-1}\mathbf{R} + \boldsymbol{\mu}\boldsymbol{\mu}^T & \boldsymbol{\mu} \\ \boldsymbol{\mu}^T & 1 \end{bmatrix}$$

and assume that the following equality is satisfied:

$$\tilde{\mathbf{\Sigma}} = \frac{K+1}{N+2\alpha}\sum_{i=1}^{N}\frac{\bar{\mathbf{x}}_i\bar{\mathbf{x}}_i^T}{(\zeta-1)+\bar{\mathbf{x}}_i^T\tilde{\mathbf{\Sigma}}^{-1}\bar{\mathbf{x}}_i} + \frac{2\gamma}{N+2\alpha}\frac{\bar{\mathbf{t}}\bar{\mathbf{t}}^T}{(\zeta-1)+\bar{\mathbf{t}}^T\tilde{\mathbf{\Sigma}}^{-1}\bar{\mathbf{t}}} + \frac{2\alpha K}{N+2\alpha}\frac{\mathbf{S}\mathbf{T}\mathbf{S}^T}{\mathrm{Tr}\left(\mathbf{S}^T\tilde{\mathbf{\Sigma}}^{-1}\mathbf{S}\mathbf{T}\right)}. \tag{30}$$

Rewriting the equality above in terms of the original variables yields the system of equations as follows:

$$\boldsymbol{\mu} = \frac{(K+1)\sum w_i\mathbf{x}_i + 2\gamma w_{\mathbf{t}}\mathbf{t}}{(K+1)\sum w_i + 2\gamma w_{\mathbf{t}}} \tag{31}$$

$$\mathbf{R} = \frac{\zeta}{(K+1)\sum w_i + 2\gamma w_{\mathbf{t}}} \cdot \left\{(K+1)\sum_{i=1}^{N} w_i(\mathbf{x}_i-\boldsymbol{\mu})(\mathbf{x}_i-\boldsymbol{\mu})^T + 2\gamma w_{\mathbf{t}}(\mathbf{t}-\boldsymbol{\mu})(\mathbf{t}-\boldsymbol{\mu})^T + \frac{2\alpha\mathbf{T}}{\mathrm{Tr}(\mathbf{R}^{-1}\mathbf{T})}\right\} \tag{32}$$

$$\zeta = \frac{(K+1)\sum w_i + 2\gamma w_{\mathbf{t}}}{N+2\alpha}, \tag{33}$$

where

$$w_i = \frac{1}{1+(\mathbf{x}_i-\boldsymbol{\mu})^T\mathbf{R}^{-1}(\mathbf{x}_i-\boldsymbol{\mu})}$$

and

$$w_{\mathbf{t}} = \frac{1}{1+(\mathbf{t}-\boldsymbol{\mu})^T\mathbf{R}^{-1}(\mathbf{t}-\boldsymbol{\mu})}.$$

By substituting (33) into (32) we get exactly the condition (10) that a stationary point of problem (9) should satisfy. Namely, if $(\boldsymbol{\mu}, \mathbf{R})$ solves (10) then there exists a corresponding $\tilde{\mathbf{\Sigma}}$ that solves (30).

Multiplying both sides of (30) by $\tilde{\mathbf{\Sigma}}$ and taking the trace gives the identity

$$\frac{2\gamma-2\alpha}{N+2\alpha} = \frac{\zeta-1}{\zeta}\cdot\frac{(K+1)\sum w_i + 2\gamma w_{\mathbf{t}}}{N+2\alpha},$$

combined with (33), ζ can be solved as $\zeta = \frac{N+2\gamma}{N+2\alpha} > 0$.

Under the existence condition provided in Theorem 2, the global minimal of (9) is a solution of (10) with $\mathbf{R}^\star \succ \mathbf{0}$, which implies (30) has at least one symmetric positive definite solution $\tilde{\mathbf{\Sigma}}^\star$. It suffices to prove that the solution is unique on $\mathbb{S}_{++}^{K+1}$ when $\gamma \ge \alpha$.

First consider the case $\gamma > \alpha$, which implies $\zeta - 1 > 0$. Without loss of generality we can assume $\tilde{\mathbf{\Sigma}} = \mathbf{I}$ is a solution, since if $\mathbf{\Sigma}$ is a solution we can always define $\tilde{\bar{\mathbf{x}}}_i = \mathbf{\Sigma}^{-\frac{1}{2}}\bar{\mathbf{x}}_i$, $\tilde{\bar{\boldsymbol{\mu}}} = \mathbf{\Sigma}^{-\frac{1}{2}}\bar{\boldsymbol{\mu}}$ and $\tilde{\mathbf{S}} = \mathbf{\Sigma}^{-\frac{1}{2}}\mathbf{S}$. Now assume there is another solution $\tilde{\mathbf{\Sigma}} = \mathbf{\Sigma}_1$ and its largest eigenvalue $\lambda_1 > 1$, then

$$\begin{aligned}\tilde{\mathbf{\Sigma}} &\le \frac{K+1}{N+2\alpha}\sum_{i=1}^{N}\frac{\bar{\mathbf{x}}_i\bar{\mathbf{x}}_i^T}{(\zeta-1)+\lambda_1^{-1}\bar{\mathbf{x}}_i^T\bar{\mathbf{x}}_i} + \frac{2\gamma}{N+2\alpha}\frac{\bar{\mathbf{t}}\bar{\mathbf{t}}^T}{(\zeta-1)+\lambda_1^{-1}\bar{\mathbf{t}}^T\bar{\mathbf{t}}} + \frac{2\alpha K}{N+2\alpha}\frac{\mathbf{S}\mathbf{T}\mathbf{S}^T}{\lambda_1^{-1}\mathrm{Tr}(\mathbf{S}\mathbf{T}\mathbf{S}^T)} \\ &< \frac{K+1}{N+2\alpha}\sum_{i=1}^{N}\frac{\lambda_1\bar{\mathbf{x}}_i\bar{\mathbf{x}}_i^T}{(\zeta-1)+\bar{\mathbf{x}}_i^T\bar{\mathbf{x}}_i} + \frac{2\alpha K}{N+2\alpha}\frac{\lambda_1\mathbf{S}\mathbf{T}\mathbf{S}^T}{\mathrm{Tr}\left(\mathbf{S}\mathbf{T}\mathbf{S}^T\right)} + \frac{2\gamma}{N+2\alpha}\frac{\lambda_1\bar{\mathbf{t}}\bar{\mathbf{t}}^T}{(\zeta-1)+\bar{\mathbf{t}}^T\bar{\mathbf{t}}} \\ &= \lambda_1\mathbf{I},\end{aligned}$$

where the first inequality follows from the fact that for any positive semidefinite matrix $\mathbf{A}$, $\mathrm{Tr}(\mathbf{\Sigma}_1^{-1}\mathbf{A}) \ge \mathrm{Tr}(\lambda_1^{-1}\mathbf{A})$, the second strict inequality follows from the assumption that $\lambda_1 > 1$

and the last equality follows from the assumption that $\mathbf{I}$ is a solution to the equation (30). We have a contradiction $\lambda_1 < \lambda_1$, hence $\lambda_1 \leq 1$. Similarly we can prove $\lambda_{K+1} \geq 1$, therefore $\mathbf{\Sigma}_1 = \mathbf{I}$.

Next consider the case $\gamma = \alpha$, which implies $\zeta - 1 = 0$. The equation simplifies to

$$\left(\frac{N}{2} + \alpha\right)\tilde{\mathbf{\Sigma}} = \frac{K+1}{2}\sum_{i=1}^{N}\frac{\bar{\mathbf{x}}_i\bar{\mathbf{x}}_i^T}{\bar{\mathbf{x}}_i^T\tilde{\mathbf{\Sigma}}^{-1}\bar{\mathbf{x}}_i} + \alpha K\frac{\mathbf{S}\mathbf{T}\mathbf{S}^T}{\mathrm{Tr}\left(\tilde{\mathbf{\Sigma}}^{-1}\mathbf{S}\mathbf{T}\mathbf{S}^T\right)} + \alpha\frac{\bar{\mathbf{t}}\bar{\mathbf{t}}^T}{\bar{\mathbf{t}}^T\tilde{\mathbf{\Sigma}}^{-1}\bar{\mathbf{t}}}.$$

By the same reasoning as Theorem 6 in [22] and the fact that $\tilde{\mathbf{\Sigma}}_{K+1,K+1} = 1$, we conclude that the solution to the above equation is unique.

REFERENCES

[1] Y. Sun, P. Babu, and D. Palomar, "Regularized robust estimation of mean and covariance matrix under heavy tails and outliers," in *Proc. IEEE 8th Sens. Array Multichannel Signal Process. Workshop (SAM)*, Jun. 2014, pp. 125–128.

[2] S. Visuri, H. Oja, and V. Koivunen, "Subspace-based direction-of-arrival estimation using nonparametric statistics," *IEEE Trans. Signal Process.*, vol. 49, no. 9, pp. 2060–2073, 2001.

[3] Y. Chen, A. Wiesel, and A. O. Hero, "Robust shrinkage estimation of high-dimensional covariance matrices," *IEEE Trans. Signal Process.*, vol. 59, no. 9, pp. 4097–4107, 2011.

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

[5] V. Koivunen, "Nonlinear filtering of multivariate images under robust error criterion," *IEEE Trans. Image Process.*, vol. 5, no. 6, pp. 1054–1060, 1995.

[6] S. Visuri, H. Oja, and V. Koivunen, "Nonparametric statistics for subspace based frequency estimation," presented at the 10th Eur. Signal Process. Conf. (EUSIPCO 2000), Tampere, Finland, 2000.

[7] F. Rubio, X. Mestre, and D. P. Palomar, "Performance analysis and optimal selection of large minimum variance portfolios under estimation risk," *IEEE J. Sel. Topics Signal Process.*, vol. 6, no. 4, pp. 337–350, 2012.

[8] A. M. Zoubir, V. Koivunen, Y. Chakhchoukh, and M. Muma, "Robust estimation in signal processing: A tutorial-style treatment of fundamental concepts," *IEEE Signal Process. Mag.*, vol. 29, no. 4, pp. 61–80, 2012.

[9] U. A. Müller, M. M. Dacorogna, and O. V. Pictet, "Heavy tails in high-frequency financial data," in *A Practical Guide to Heavy Tails: Statistical Techniques and Applications*. Boston, MA, USA: Birkhaüser, 1998, pp. 55–77.

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

[11] 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.

[12] 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.

[13] 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.

[14] 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.

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

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

[17] 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.

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

[19] P. J. Bickel and E. Levina, "Regularized estimation of large covariance matrices," *Ann. Statist.*, pp. 199–227, 2008.

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

[21] 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, 2013.

[22] Y. Sun, P. Babu, and D. Palomar, "Regularized Tyler's scatter estimator: Existence, uniqueness, algorithms," *IEEE Trans. Signal Process.*, vol. 62, no. 19, pp. 5143–5156, Oct. 2014.

[23] E. Ollila and D. E. Tyler, Regularized M-Estimators of Scatter Matrix 2014 [Online]. Available: arXiv preprint arXiv:1405.2528

[24] C. Bishop, *Pattern Recognition and Machine Learning*. New York, USA: Springer, 2006.

[25] F. J. Fabozzi, P. N. Kolm, D. Pachamanova, and S. M. Focardi, *Robust Portfolio Optimization and Management*. New York, NY, USA: Wiley, 2007.

[26] 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.

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

[28] G. Frahm, "Generalized elliptical distributions: Theory and applications," Ph.D. dissertation, Universität zu Köln, Köln, 2004.

[29] K. L. Lange, R. J. A. Little, and J. M. G. Taylor, "Robust statistical modeling using the t distribution," *J. Amer. Statist. Assoc.*, vol. 84, no. 408, pp. 881–896, 1989.

[30] A. Lucas, "Robustness of the student t based M-estimator," *Commun. Statist.—Theory Methods*, vol. 26, no. 5, pp. 1165–1182, 1997.

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

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

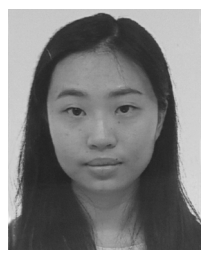
[33] R. Maronna, D. Martin, and V. Yohai, *Robust Statistics: Theory and Methods*, ser. Wiley Ser. Probabil. Statist.. Hoboken, NJ, USA: Wiley, 2006.

[34] K. S. Tatsuoka and D. E. Tyler, "On the uniqueness of S-functionals and M-functionals under nonelliptical distributions," *Ann. Statist.*, pp. 1219–1243, 2000.

[35] J. T. Kent, D. E. Tyler, and Y. Vard, "A curious likelihood identity for the multivariate t-distribution," *Commun. Statist.—Simul. Computat.*, vol. 23, no. 2, pp. 441–453, 1994.

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

[37] 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.



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 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 joined the Department of Electronic and Computer Engineering, Hong Kong University of Science and Technology (HKUST), Hong Kong, in 2006, where he is currently a Professor. Since 2013, he has been 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, USA. 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 a recipient of a 2004/2006 Fulbright Research Fellowship, the 2004 Young Author Best Paper Award by the IEEE Signal Processing Society, the 2002/2003 best Ph.D. prize in Information Technologies and Communications by the Technical University of Catalonia (UPC), the 2002/2003 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."