Optimal High-Dimensional Shrinkage Covariance Estimation for Elliptical Distributions

Esa Ollila

Department of Signal Processing and Acoustics
Aalto University, Finland

Abstract—We derive an optimal shrinkage sample covariance matrix (SCM) estimator which is suitable for high dimensional problems and when sampling from an unspecified elliptically symmetric distribution. Specifically, we derive the optimal (oracle) shrinkage parameters that obtain the minimum mean-squared error (MMSE) between the shrinkage SCM and the true covariance matrix when sampling from an elliptical distribution. Subsequently, we show how the oracle shrinkage parameters can be consistently estimated under the random matrix theory regime. Simulations show the advantage of the proposed estimator over the conventional shrinkage SCM estimator due to Ledoit and Wolf (2004). The proposed shrinkage SCM estimator often provides significantly better performance than the Ledoit-Wolf estimator and has the advantage that consistency is guaranteed over the whole class of elliptical distributions with finite 4th order moments.

I. INTRODUCTION

We consider the problem of estimating the covariance matrix based on a sample $\mathbf{x}_1, \ldots, \mathbf{x}_n$ of independent and identically distributed (i.i.d.) random vectors from an unspecified p-variate distribution $\mathbf{x} \sim F$ with mean vector $\mathbb{E}[\mathbf{x}] = \mathbf{0}$ and $p \times p$ positive definite covariance matrix $\boldsymbol{\Sigma} = \mathbb{E}[\mathbf{x}\mathbf{x}^\top]$. The *sample covariance matrix* (SCM) $\mathbf{S} = \frac{1}{n}\sum_{i=1}^{n} \mathbf{x}_i \mathbf{x}_i^\top$ is the most commonly used estimator of the covariance matrix, and when random sampling from a multivariate Gaussian $\mathcal{N}_p(\mathbf{0}, \boldsymbol{\Sigma})$ distribution, it is also the optimal maximum likelihood estimator (MLE). Estimation of high-dimensional (HD) covariance matrix when the sample size n is smaller, or not much larger than the dimension p, has attracted a significant research interest in recent years. Indeed since such data problems are becoming increasingly common in finance [1], genomics or classification, for example. Insufficient number of samples causes significant estimation errors in the SCM. Moreover, if $p > n$, the SCM $\mathbf{S}$ is always singular, i.e., not invertible even if the true covariance matrix $\boldsymbol{\Sigma}$ is known to be positive definite and hence non-singular. The commonly used approach is then to use shrinkage regularization as in [1], [2], [3], [4], [5], [6], for example.

One of the most ommonly used estimator in "large p compared to sample size n problems" is the *regularized SCM (RSCM)*,

$$\mathbf{S}_{\alpha,\beta} = \beta \mathbf{S} + \alpha \mathbf{I}, \qquad (1)$$

where $\alpha, \beta > 0$ denotes the *shrinkage (regularization) parameters*. Optimal RSCM estimator is often defined as one that is based on *oracle shrinkage parameters* that minimize the mean squared error (MSE),

$$(\alpha_o, \beta_o) = \arg\min_{\alpha,\beta>0}\left\{ \mathrm{MSE}(\mathbf{S}_{\alpha,\beta}) = \mathbb{E}\left[\left\| \boldsymbol{\Sigma} - \mathbf{S}_{\alpha,\beta} \right\|_{\mathrm{F}}^2 \right] \right\}, \quad (2)$$

where $\|\cdot\|_{\mathrm{F}}$ denotes the Frobenius matrix norm ($\|\mathbf{A}\|_{\mathrm{F}}^2 = \mathrm{tr}(\mathbf{A}^\top \mathbf{A}) = \mathrm{tr}(\mathbf{A}\mathbf{A}^\top)$ for any matrix $\mathbf{A}$). The solution (α_o, β_o) are called "oracle" shrinkage parameters as they will obviously depend on the true unknown covariance matrix $\boldsymbol{\Sigma}$ and hence can not be used in practise. The widely popular Ledoit-Wolf (LW-)RSCM [1] is based on consistent estimators $(\hat{\alpha}_o^{\mathrm{LW}}, \hat{\beta}_o^{\mathrm{LW}})$ of (α_o, β_o) under the random matrix theory (RMT) regime. However, more accurate finite sample estimation performance can be obtained by assuming that the observations are from a specific p-variate distribution, e.g., the multivariate normal distribution, as has been shown in [4]. In this paper, we derive consistent estimators of the oracle shrinkage parameters (α_o, β_o) under the RMT regime when sampling from an *unspecified* elliptically symmetric distribution. Elliptical distributions (see [7], [8], [9]) constitute a large class of distributions that include e.g., the multivariate normal distribution, generalized Gaussian and all compound Gaussian distributions as special cases.

The RMT regime refers to the case that

(R1) $n, p \to \infty$ and $p/n \to c$, where $0 < c < \infty$.

Furthermore, we assume that the set of eigenvalues of $\boldsymbol{\Sigma}$ converge to a fixed spectrum, and that

(R2) As $p \to \infty$, $\eta_i = \mathrm{tr}(\boldsymbol{\Sigma}^i)/p \to \eta_i^o$, $0 < \eta_i^0 < \infty$ for $i = 1, \ldots, 4$

Our numerical examples illustrate that the RSCM estimator that is based on the proposed consistent estimators $(\hat{\alpha}_o^{Ell}, \hat{\beta}_o^{Ell})$ outperform its competitors, e.g., the LW-RSCM estimator, when sampling from an elliptical population.

The paper is organized as follows. In Section II and Section III we derive the optimal shrinkage parameters (α_o, β_o) under the general assumption of sampling from any general p-variate distribution and an elliptical distribution with finite 4th order moments, respectively. In Section IV, consistent estimators of (α_o, β_o) are proposed under assumptions (R1) and (R2) when sampling from an unspecified elliptical distribution. Simulation studies of Section V illustrate that the proposed shrinkage estimator always outperforms the LW estimator when the samples are drawn from an elliptical population.

Notation: Let $\mathcal{S}_p$ be the open cone of positive definite $p \times p$ symmetric matrices, and let $\mathbf{I}$ be the identity matrix of proper

dimension, $\mathrm{vec}(\cdot)$ denotes an operator that transforms a matrix into a vector by stacking the columns of the matrix, $\mathrm{tr}(\cdot)$ denotes the matrix trace operator, and $\otimes$ denote the *Kronecker product*: for any matrix $\mathbf{A}$ and $\mathbf{B}$, $\mathbf{A} \otimes \mathbf{B}$ is a block matrix with (i,j)-block being equal to $a_{ij}\mathbf{B}$. A *commutation matrix* $\mathbf{K}_p$ is a $p^2 \times p^2$ block matrix with (i,j)-block equal to a $p \times p$ matrix that has a 1 at entry (j,i) and 0's elsewhere. It has the following important property [10]: $\mathbf{K}_p\mathrm{vec}(\mathbf{A}) = \mathrm{vec}(\mathbf{A}^\top)$ for any $p \times p$ matrix $\mathbf{A}$.

II. OPTIMAL ORACLE SHRINKAGE PARAMETERS

Define *scale* measures of $\boldsymbol{\Sigma} \in \mathcal{S}_p$ as

$$\eta = \mathrm{tr}(\boldsymbol{\Sigma})/p \quad \text{and} \quad \eta_2 = \mathrm{tr}(\boldsymbol{\Sigma}^2)/p. \tag{3}$$

An important measure in our future developments is the following measure of sphericity [11],

$$\gamma = \frac{\eta_2}{\eta^2} = \frac{p\,\mathrm{tr}(\boldsymbol{\Sigma}^2)}{\mathrm{tr}(\boldsymbol{\Sigma})^2}. \tag{4}$$

Statistic γ measures how close the covariance matrix is to a scaled identity matrix. It verifies $\gamma \geq 1$ and $\gamma = 1$ if and only if $\boldsymbol{\Sigma} = c\mathbf{I}$ for some $c > 0$.

The parameters η and γ are elemental in our developments. As is shown in Theorem 2, the optimal shrinkage parameter pair (α_o, β_o) for elliptical distributions depends on the true covariance matrix $\boldsymbol{\Sigma}$ only through η and γ. Simple "plug-in" estimates of (α_o, β_o) can then be obtained by simply replacing (η, γ) with their estimates. Finding accurate and consistent estimators of the shrinkage parameters is then a considerably simpler task than in the general case of Theorem 1.

Next theorem provides the expression for the oracle shrinkage parameters in the case of sampling from an unspecified p-variate distribution with finite 4th order moments.

Theorem 1. *Let* $\{\mathbf{x}_i\}_{i=1}^n$ *denote a random sample from any p-variate distribution (not necessarily elliptical distribution) with finite 4th order moments. Then the oracle parameters in (2) are*

$$\beta_o = \frac{p(\gamma - 1)\eta^2}{\mathbb{E}\big[\,\mathrm{tr}\,(\mathbf{S}^2)\big] - p\eta^2} \quad \text{and} \quad \alpha_o = (1 - \beta_o)\eta \tag{5}$$

where η and γ are defined in (3) *and* (4)*, respectively. The value of MSE at the optimum is*

$$\mathrm{MSE}(\mathbf{S}_{\alpha_o,\beta_o}) = \|\boldsymbol{\Sigma} - \eta\mathbf{I}\|_{\mathrm{F}}^2 (1 - \beta_o). \tag{6}$$

The optimal β_o is always in the range $[0, 1)$.

Proof. It was shown in [1, Theorem 2.1] that

$$\beta_o = \frac{\|\boldsymbol{\Sigma} - \eta\mathbf{I}\|_{\mathrm{F}}^2}{\|\boldsymbol{\Sigma} - \eta\mathbf{I}\|_{\mathrm{F}}^2 + \mathbb{E}\big[\|\mathbf{S} - \boldsymbol{\Sigma}\|_{\mathrm{F}}^2\big]} \tag{7}$$

and $\alpha_o = (1 - \beta_o)\eta$. The form of β_o in (7) implies that $\beta_o \in [0, 1)$. We now show that (7) can be expressed in the form (5). First, we observe that

$$a_1 = \mathbb{E}\big[\|\mathbf{S} - \boldsymbol{\Sigma}\|_{\mathrm{F}}^2\big] = \mathbb{E}\big[\,\mathrm{tr}(\mathbf{S}^2)\big] - 2\mathbb{E}\big[\,\mathrm{tr}\,(\mathbf{S}\boldsymbol{\Sigma})\big] + \mathrm{tr}\,(\boldsymbol{\Sigma}^2)$$
$$= \mathbb{E}\big[\,\mathrm{tr}\,(\mathbf{S}^2)\big] - \mathrm{tr}(\boldsymbol{\Sigma}^2) \tag{8}$$

where we used that $\mathbb{E}[\mathrm{tr}(\mathbf{S}\boldsymbol{\Sigma})] = \mathrm{tr}(\mathbb{E}[\mathbf{S}]\boldsymbol{\Sigma}) = \mathrm{tr}(\boldsymbol{\Sigma}^2)$. The numerator of β_o in (7) is

$$a_2 = \|\boldsymbol{\Sigma} - \eta\mathbf{I}\|_{\mathrm{F}}^2 = \mathrm{tr}(\boldsymbol{\Sigma}^2) - (1/p)\big\{\,\mathrm{tr}(\boldsymbol{\Sigma})\big\}^2$$
$$= p(\eta_2 - \eta^2) = p(\gamma - 1)\eta^2 \tag{9}$$

which shows that denominator of β_o is $a_1 + a_2 = \mathbb{E}\big[\,\mathrm{tr}(\mathbf{S}^2)\big] - (1/p)\big\{\,\mathrm{tr}(\boldsymbol{\Sigma})\big\}^2 = \mathbb{E}\big[\,\mathrm{tr}(\mathbf{S}^2)\big] - p\eta^2$. These expressions for numerator and denominator of β_o yield the assertion (5) for β_o. Write $L(\alpha, \beta) = \mathbb{E}\big[\|\mathbf{S}_{\alpha,\beta} - \boldsymbol{\Sigma}\|_{\mathrm{F}}^2\big]$ for the MSE. Note that

$$L(\alpha, \beta) = \mathbb{E}\Big[\|\alpha\mathbf{I} + \beta(\mathbf{S} - \boldsymbol{\Sigma}) - (1 - \beta)\boldsymbol{\Sigma}\|_{\mathrm{F}}^2\Big]$$
$$= \alpha^2 p + \beta^2 a_1 + (1 - \beta)^2 \eta_2 p - 2\alpha(1 - \beta)\eta p. \tag{10}$$

The MSE at the optimum is

$$\begin{aligned}
\mathrm{MSE}(\mathbf{S}_{\alpha_o,\beta_o}) &= L((1 - \beta_o)\eta, \beta_o) \\
&= \beta_o^2 a_1 + (1 - \beta_o)^2 \eta_2 p - (1 - \beta_o)^2 \eta^2 p \\
&= (1 - \beta_o)^2 \underbrace{\{p(\eta_2 - \eta^2)\}}_{= \, a_2 \text{ by (9)}} + \beta_o^2 a_1 \\
&= (1 - \beta_o)^2 a_2 + (1 - \beta_o)\beta_o a_2 \\
&= (1 - \beta_o)a_2,
\end{aligned}$$

where the 3rd identity follows as $\beta_o a_1 = (1 - \beta_o)a_2$. This completes the proof. $\qquad\square$

Theorem has important implications. First, since $\alpha_o = (1 - \beta_o)\eta$ is determined by the value of $\beta_o \in [0, 1)$, the optimal RSCM can be expressed simply as

$$\mathbf{S}_{\alpha_o,\beta_o} = \beta_o \frac{1}{n} \sum_{i=1}^n \mathbf{x}_i\mathbf{x}_i^\top + (1 - \beta_o)\eta\mathbf{I}.$$

Since $\hat{\eta} = \mathrm{tr}(\mathbf{S})/p$ is a consistent estimator of $\eta = \mathrm{tr}(\boldsymbol{\Sigma})/p$ both in the conventional (fixed p) and RMT asymptotic regime, we need to simply focus on finding a consistent estimator $\hat{\beta}_o$ of β_o. Consistent estimator of α_o is determined simply as $\hat{\alpha}_o = (1 - \hat{\beta}_o)\,\mathrm{tr}(\mathbf{S})/p$.

Ledoit and Wolf [1] showed that the following estimate

$$\hat{\beta}_{\mathrm{LW}}^* = 1 - \frac{\sum_{i=1}^n \|\mathbf{x}_i\mathbf{x}_i^\top - \mathbf{S}\|_{\mathrm{F}}^2}{pn^2(\hat{\gamma} - 1)} = 1 - \frac{\frac{1}{np}\sum_{i=1}^n \|\mathbf{x}_i\|_2^4 - \hat{\eta}_2}{n(\hat{\gamma} - 1)}$$

where $\hat{\gamma} = \hat{\eta}_2/\hat{\eta} = p\,\mathrm{tr}(\mathbf{S}^2)/\mathrm{tr}(\mathbf{S})$ and $\hat{\eta}_2 = \mathrm{tr}(\mathbf{S}^2)/p$, converges to β_o in (5) in probability under RMT regime (R1) and (R2) when sampling from a distribution $\mathbf{x} \sim F$ with finite 4th-order moments. The authors of [1] then proposed to estimate the shrinkage parameters using

$$\hat{\beta}_0^{\mathrm{LW}} = \max(0, \hat{\beta}_{\mathrm{LW}}^*) \quad \text{and} \quad \hat{\alpha}_0^{\mathrm{LW}} = (1 - \hat{\beta}_0^{\mathrm{LW}})\,\mathrm{tr}(\mathbf{S})/p,$$

where the $\mathtt{max}$ constraint ensures that the final estimate remains on the interval $[0, 1]$. The RSCM based on the above penalty parameters is referred hereafter as *LW-RSCM estimator*.

III. OPTIMAL ORACLE SHRINKAGE PARAMETERS: THE ELLIPTICAL CASE

Assume now that $\mathbf{x}_1, \ldots, \mathbf{x}_n$ are independent and identically distributed (i.i.d.) random vectors from a centered elliptical distribution with mean vector $\mathbb{E}[\mathbf{x}] = \mathbf{0}$ and positive definite covariance matrix $\boldsymbol{\Sigma} = \mathbb{E}[\mathbf{x}\mathbf{x}^\top]$, denoted $\mathcal{E}_p(\mathbf{0}, \boldsymbol{\Sigma}, g)$. For a review of elliptical distributions, see [7], [8], [9]. The probability density function (p.d.f.) of $\mathbf{x} \sim \mathcal{E}_p(\mathbf{0}, \boldsymbol{\Sigma}, g)$ is

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

where $g : [0, \infty) \to [0, \infty)$ is a fixed function, called the *density generator*, that is independent of $\mathbf{x}$ and $\boldsymbol{\Sigma}$, and $C_{p,g}$ is a normalizing constant ensuring that $f(\mathbf{x})$ integrates to 1. Let g be defined so that $\boldsymbol{\Sigma}$ represents the covariance matrix of $\mathbf{x}$. For example, the p-variate Gaussian distribution, denoted $\mathbf{x} \sim \mathcal{N}_p(\mathbf{0}, \boldsymbol{\Sigma})$, is a member in this class with density generator $g(t) = \exp(-t/2)$. As earlier in Theorem 1, we assume that the elliptical population possesses finite 4th-order moments.

Recall that the *kurtosis* of a zero mean random variable x is defined as

$$\text{kurt}(x) = \frac{\mathbb{E}[x^4]}{(\mathbb{E}[x^2])^2} - 3.$$

The *elliptical kurtosis parameter* [7] κ of a random vector $\mathbf{x} = (x_1, \ldots, x_p)^\top \sim \mathcal{E}_p(\mathbf{0}, \boldsymbol{\Sigma}, g)$ is defined as

$$\kappa = \frac{\mathbb{E}[r^4]}{p(p+2)} - 1 = \frac{1}{3} \cdot \text{kurt}(x_i), \quad (11)$$

where r denotes the (2nd order) modular variate of the elliptical distribution, defined as $r = \sqrt{\mathbf{x}^\top\boldsymbol{\Sigma}^{-1}\mathbf{x}}$. The elliptical kurtosis shares properties similar to kurtosis of a real random variable. Especially, if $\mathbf{x} \sim \mathcal{N}_p(\mathbf{0}, \boldsymbol{\Sigma})$, then $\kappa = 0$. This is obvious since the marginal distributions are Gaussian and hence $\kappa = (1/3)\,\text{kurt}(x_i) = 0$. Another way to derive this is by noting that $r^2 = \mathbf{x}^\top\boldsymbol{\Sigma}^{-1}\mathbf{x} \sim \chi_p^2$ and hence $\mathbb{E}[r^4] = p(p+2)$. The importance of elliptical kurtosis parameter κ is due to the fact that the $p^2 \times p^2$ covariance matrix of $\text{vec}(\mathbf{S})$ can be expressed as [7]:

$$\text{cov}(\text{vec}(\mathbf{S})) =$$
$$\frac{(1+\kappa)}{n}(\mathbf{I} + \mathbf{K}_p)(\boldsymbol{\Sigma} \otimes \boldsymbol{\Sigma}) + \frac{\kappa}{n}\text{vec}(\boldsymbol{\Sigma})\text{vec}(\boldsymbol{\Sigma})^\top, \quad (12)$$

where $\mathbf{K}_p$ denotes the commutation matrix defined in the Introduction. Thus the elliptical kurtosis parameter κ along with the true covariance matrix $\boldsymbol{\Sigma}$ provide a complete description of the covariances between elements S_{ij} and S_{kl} of the SCM $\mathbf{S}$.

In the next Lemma we derive the MSE of the SCM.

Lemma 1. *Let* $\{\mathbf{x}_i\}_{i=1}^n \overset{iid}{\sim} \mathcal{E}_p(\mathbf{0}, \boldsymbol{\Sigma}, g)$, *where* $\boldsymbol{\Sigma} = \text{cov}(\mathbf{x}_i)$ *and 4th-order moments exists. Then the MSE of* $\mathbf{S}$ *is*

$$\text{MSE}(\mathbf{S}) = \frac{p}{n} \cdot \eta^2 \left\{ \kappa(2\gamma + p) + \gamma + p \right\},$$

and the normalized mean squared error (NMSE) is

$$\text{NMSE}(\mathbf{S}) = \frac{\mathbb{E}[\|\mathbf{S} - \boldsymbol{\Sigma}\|_F^2]}{\|\boldsymbol{\Sigma}\|_F^2} = \frac{1}{\gamma} \cdot \frac{1}{n} \left\{ \kappa(2\gamma + p) + \gamma + p \right\}.$$

Furthermore,

$$\mathbb{E}[\text{tr}(\mathbf{S}^2)] = \text{MSE}(\mathbf{S}) + p\eta_2.$$

Above η, γ *and* κ *are defined in* (3), (4) *and* (11), *respectively.*

Proof. Since $\mathbf{S}$ is unbiased, so $\mathbb{E}[\mathbf{S}] = \boldsymbol{\Sigma}$, it holds that

$$\text{MSE}(\mathbf{S}) = \text{tr}\{\text{cov}(\text{vec}(\mathbf{S}))\}, \quad (13)$$

where $\text{cov}(\text{vec}(\mathbf{S}))$ has the expression stated in (12). Then recall the following results: $\text{tr}(\mathbf{A} \otimes \mathbf{B}) = \text{tr}(\mathbf{A})\,\text{tr}(\mathbf{B})$, $\text{tr}\{\text{vec}(\mathbf{A})\text{vec}(\mathbf{B})^\top\} = \text{tr}(\mathbf{A}\mathbf{B})$ for any square matrices $\mathbf{A}$ and $\mathbf{B}$ of same order; see e.g., [10]. These imply that

$$\text{tr}(\boldsymbol{\Sigma} \otimes \boldsymbol{\Sigma}) = \text{tr}(\boldsymbol{\Sigma})^2, \;\; \text{tr}\{\text{vec}(\boldsymbol{\Sigma})\text{vec}(\boldsymbol{\Sigma})^\top\} = \text{tr}(\boldsymbol{\Sigma}^2). \quad (14)$$

It is also easy to show that

$$\text{tr}\left\{\mathbf{K}_p(\boldsymbol{\Sigma} \otimes \boldsymbol{\Sigma})\right\} = \text{tr}(\boldsymbol{\Sigma}^2) \quad (15)$$

by recalling the definition of the commutation matrix and the property $\text{tr}(\mathbf{A} \otimes \mathbf{B}) = \text{tr}(\mathbf{A})\,\text{tr}(\mathbf{B})$. Using (13) - (15), then yield the stated expression for $\text{MSE}(\mathbf{S})$. The expression for NMSE is obtained by dividing $\text{MSE}(\mathbf{S})$ by $\text{tr}(\boldsymbol{\Sigma}^2) = p\eta_2$. The last results follows as

$$\text{MSE}(\mathbf{S}) = \mathbb{E}[\|\mathbf{S} - \boldsymbol{\Sigma}\|_F^2] = \mathbb{E}[\text{tr}\{(\mathbf{S} - \boldsymbol{\Sigma})(\mathbf{S} - \boldsymbol{\Sigma})\}]$$
$$= \mathbb{E}[\text{tr}(\mathbf{S}^2) - 2\,\text{tr}(\mathbf{S}\boldsymbol{\Sigma}) + \text{tr}(\boldsymbol{\Sigma}^2)]$$
$$= \mathbb{E}[\text{tr}(\mathbf{S}^2)] - p\eta_2$$

by using that $\mathbb{E}[\text{tr}(\mathbf{S}\boldsymbol{\Sigma})] = \text{tr}(\mathbb{E}[\mathbf{S}]\boldsymbol{\Sigma}) = \text{tr}(\boldsymbol{\Sigma}^2) = p\eta_2$. $\qquad\square$

Next theorem states that the oracle parameters derived in Theorem 1 can be written in a much simpler form when sampling from an elliptically symmetric distribution.

Theorem 2. *Let* $\{\mathbf{x}_i\}_{i=1}^n \overset{iid}{\sim} \mathcal{E}_p(\mathbf{0}, \boldsymbol{\Sigma}, g)$ *and assume that elliptical population possesses finite 4th-order moments. Then the oracle parameters* (α_o, β_o) *that minimize the MSE are*

$$\beta_o^{Ell} = \frac{\gamma - 1}{(\gamma - 1) + \gamma \cdot \text{NMSE}(\mathbf{S})}$$
$$= \frac{\gamma - 1}{\gamma - 1 + (1/n)\{\kappa(2\gamma + p) + \gamma + p\}}$$

and $\alpha_o^{Ell} = (1 - \beta_o^{Ell})\eta$.

Proof. Using Lemma 1, the denominator of β_o is

$$\mathbb{E}[\text{tr}(\mathbf{S}^2)] - p\eta^2$$
$$= \text{MSE}(\mathbf{S}) + p\eta_2 - p\eta^2$$
$$= p\eta^2\{\text{MSE}(\mathbf{S})/(p\eta^2) + \gamma - 1\}$$
$$= p\eta^2\{\gamma \cdot \text{NMSE}(\mathbf{S}) + \gamma - 1\},$$

where the last idenity follows as $\text{NMSE}(\mathbf{S}) = \text{MSE}(\mathbf{S})/\|\boldsymbol{\Sigma}\|_F^2 = \text{MSE}(\mathbf{S})/(p\eta_2)$ and recalling that $\gamma = \eta_2/\eta^2$. Substituting this expression into (5) yields the first assertion for β_o. The second assertion follows by recalling the expression for $\text{NMSE}(\mathbf{S})$ from Lemma 1. $\qquad\square$

It is not surprising that β_o and hence also α_o depend on the functional form of the elliptical distribution (i.e., on

density generator g) only via elliptical kurtosis parameter κ. Specifying the elliptical distribution (e.g., Gaussian, t-distribution, etc), also specifies the value of κ. For example, when sampling from the Gaussian distribution, the elliptical kurtosis parameter is $\kappa = 0$, but since we do not assume any particular elliptical distribution, we need to find a consistent estimator of the elliptical kurtosis parameter $\hat{\kappa}$ as well.

IV. CONSISTENT ESTIMATION OF THE ORACLE PARAMETERS

Let $\{\mathbf{x}_i\}_{i=1}^n \overset{iid}{\sim} \mathcal{E}_p(\mathbf{0}, \mathbf{\Sigma}, g)$, where $\mathrm{cov}(\mathbf{x}) = \mathbf{\Sigma}$ and assume that the 4th-order moments exists. In this section, we address the important topic of how to obtain consistent estimators of the unknown parameters η, γ and κ.

First we recall that the sample sign covariance matrix, defined as

$$\mathbf{S}_{sgn} = \frac{1}{n} \sum_{i=1}^n \frac{\mathbf{x}_i \mathbf{x}_i^\top}{\|\mathbf{x}_i\|^2},$$

is well-known to be highly robust although it is not a consistent estimator of the covariance matrix [12]. However, the following result from [3, Lemma 4.1] shows that it can be used to estimate the parameter γ.

Lemma 2. *Let $\{\mathbf{x}_i\}_{i=1}^n \overset{iid}{\sim} \mathcal{E}_p(\mathbf{0}, \mathbf{\Sigma}, g)$. Then*

$$\hat{\gamma} = p \, \mathrm{tr}\left(\mathbf{S}_{sgn}^2\right) - (p/n) \qquad (16)$$

is a consistent estimator of $\gamma = p \, \mathrm{tr}(\mathbf{\Sigma}^2)/\mathrm{tr}(\mathbf{\Sigma})^2$ under assumption (R1) and (R2).

Note that $\hat{\gamma}$ is a robust and distribution-free estimator of γ. The optimum parameter $\beta_o^{Ell} = \beta_o^{Ell}(\gamma, \kappa)$ depends on γ and κ. Hence a *plug-in estimator*,

$$\hat{\beta}_o^{Ell} = \beta_o^{Ell}(\hat{\gamma}, \hat{\kappa}),$$

where $\hat{\gamma}$ and $\hat{\kappa}$ are consistent estimators of γ and κ, is a consistent estimator of β_o as well. A natural estimate of κ is the conventional sample average,

$$\hat{\kappa} = \max\left(-\frac{2}{p+2}, \, \frac{1}{3p} \sum_{j=1}^p \hat{k}_j\right), \qquad (17)$$

where $\hat{k}_j = m_j^{(4)}/\left(m_j^{(2)}\right)^2 - 3$ is the sample kurtosis of the jth variable and $m_j^{(q)} = \frac{1}{n} \sum_{i=1}^n (x_{ij})^q$ denotes the qth order sample moment, $j = 1, \ldots, p$. Above the $\max$ constraint ensures that the final estimate $\hat{\kappa}$ does not exceed the theoretical lower bound [13], $-2/(p+2)$ of elliptical kurtosis parameter κ. The estimate $\hat{\kappa}$ is a consistent estimator of the elliptical kurtosis κ both in the conventional and RMT regime.

We can now define the *Ell-RSCM estimator* as the regularized SCM based on the following estimated optimal shrinkage parameters

$$\hat{\beta}_o^{Ell} = \max\left(0, \frac{T}{T + (1/n)\{\hat{\kappa}(2\hat{\gamma} + p) + \hat{\gamma} + p\}}\right) \qquad (18)$$

$$\hat{\alpha}_o^{Ell} = (1 - \hat{\beta}_0^{Ell}) \, \mathrm{tr}(\mathbf{S})/p$$

where $T = \hat{\gamma} - 1$ and $\hat{\gamma}$ and $\hat{\kappa}$ are defined in (16) and (17), respectively,

V. SIMULATION STUDY

We conduct a small simulation study to investigate the performance of RSCM estimators in terms of their finite sample NMSE. Each simulation is repeated 10000 times and the NMSE is computed (averaged of Monte-Carlo runs) for each RSCM estimator. Theoretical oracle MSE value derived in (6) and normalized by $\|\mathbf{\Sigma}\|_F^2$ is used as a benchmark lower bound for empirical NMSE values. This is shown in the figures as solid black line.

A. AR(1) covariance matrix

In the first experiment, an autoregressive covariance structured is used. We let $\mathbf{\Sigma}$ be the covariance matrix of a Gaussian AR(1) process,

$$[\mathbf{\Sigma}]_{ij} = \varrho^{|i-j|}, \quad r \in (0, 1).$$

Note that $\mathbf{\Sigma}$ verifies $\eta = \mathrm{tr}(\mathbf{\Sigma})/p = 1$. When ϱ is close to 0, then $\mathbf{\Sigma}$ is close to an identity matrix and when ϱ tends to 1, $\mathbf{\Sigma}$ tends to a singular matrix of rank 1. Thus the theoretical value β_o is close to 0 for small values of ϱ, i.e., when the true covariance matrix is close to the target $\mathbf{I}$, and $\beta_0 \approx 1$ for ϱ close to 1. Dimension is fixed at $p = 100$ and n is allowed to vary from $0.2 \cdot p$ to $1.2 \cdot p$.

Figure 1 depicts the NMSE performance when the samples are drawn from a Gaussian distribution (upper panel) and a multivariate t_ν-distribution with $\nu = 8$ degrees of freedom (lower panel). Several conclusions can be drawn from these figures. First, when $\varrho = 0.1$ and thus $\mathbf{\Sigma}$ is close to the shrinkage target matrix $\mathbf{I}$, Ell-RSCM estimators outperform the LW-RSCM estimator. Especially, when the ratio n/p is small (i.e., p larger than n), we observe the largest performance differences in favor of Ell-RSCM. Second, when the true $\mathbf{\Sigma}$ starts to deviate significantly from the identity target matrix $\mathbf{I}$ (i.e., $\varrho = 0.4$), LW-RSCM and Ell-RSCM estimator have similar performance especially for large values of n/p. Third, when the samples are drawn from t_8-distribution, the performance of LW-RSCM estimator is seen to deterioritate in comparison to the proposed Ell-RSCM estimator. Indeed very large differences are witnessed in NMSE between the estimators especially when $n/p < 0.5$.

B. Largely varying spectrum

Our next study follows the set-up in [3] in which $\mathbf{\Sigma}$ has one (or a few) large eigenvalues. In the first set-up, $\mathbf{\Sigma}$ is a diagonal matrix of size 50×50, where m eigenvalues are equal to 1 and the remaining $50 - m$ eigenvalues are 0.01. For the case $n = p = 50$, Figure 2 depicts the NMSE as a function of m when sampling from a t_ν distribution with $\nu = 8$ degrees of freedom. Ell-RSCM has excellent performance as its NMSE curve is essentially overlapping with the theoretical NMSE curve. LW-RSCM estimator is performing poorly for all values of m except at the extremes, i.e, when m is either small or large, in which case the covariance matrix $\mathbf{\Sigma}$ is close to an (scaled) identity matrix.

Next simulation set-up considers a very challenging scenario in which the spectrum of $\mathbf{\Sigma}$ consists of several different
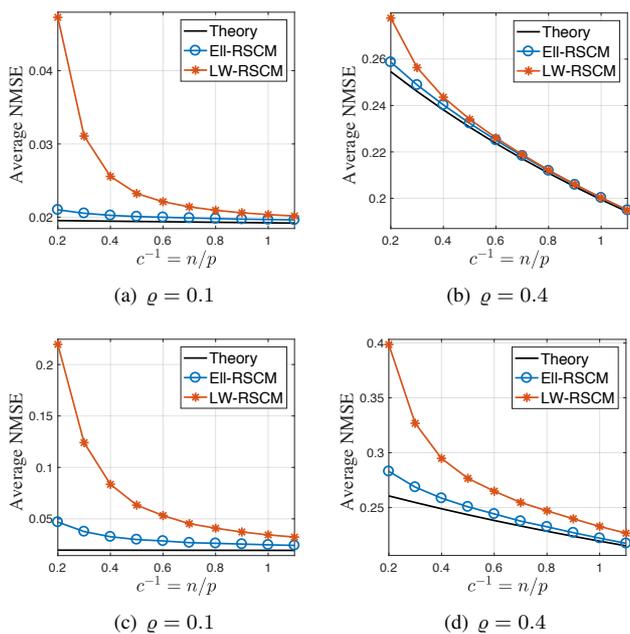
Fig. 1.   AR(1) process: Comparison of covariance estimators when $p = 100$ and $\varrho \in \{0.1, 0.4\}$ and the samples are from Gaussian distribution (upper panel) and t_ν-distribution with $\nu = 8$ degrees of freedom (lower panel).
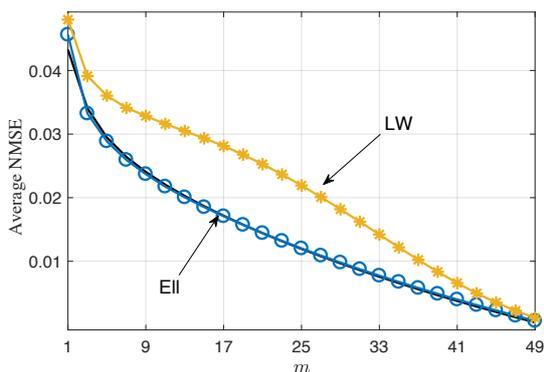

Fig. 2.   The covariance matrix $\mathbf{\Sigma}$ has m eigenvalues equal to 1 and $50 - m$ eigenvalues equal to 0.01. The samples are from t_ν-distribution with $\nu = 8$ degrees of freedom and $n = p = 50$.

eigenvalues. We consider the case that $p = 100$ and the covariance matrix $\mathbf{\Sigma}$ has 30 eigenvalues equal to 100, 40 eigenvalues equal to 1 and 30 eigenvalues of 0.01. Samples are drawn from t_ν distribution with $\nu = 8$ degrees of freedom. The NMSE curves shown in Figure 2 illustrate the huge advantage of the proposed Ell-RSCM over the LW-RSCM estimator. In fact, in this scenario the LW estimator fails and it assigns $\hat{\beta}_o = 0$ for all values of n. Again the Ell-RSCM estimator reaches near oracle performance and thus there is not much space for improvements.

VI. CONCLUSION

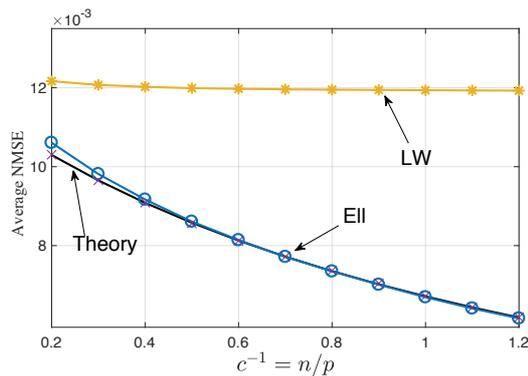We proposed an optimal regularized sample covariance matrix estimator, called Ell-RSCM estimator, which is suitable

Fig. 3.   The covariance matrix $\mathbf{\Sigma}$ has 30 eigenvalues equal to 100, 40 eigenvalues equal to 1 and 30 eigenvalues equal to 0.01. The samples are from t_ν-distribution with $\nu = 8$ degrees of freedom and $p = 100$.

for high-dimensional problems and when sampling from an unspecified elliptically symmetric distribution. The estimator is based on consistent estimators (under RMT regime) of the optimal shrinkage parameters that minimize the MSE. It smartly exploits elliptical theory such as the knowledge of the form of MSE of the SCM when sampling from an elliptical population. Our simulation studies illustrated the advantage of the proposed Ell-RSCM over the Ledoit-Wolf (LW-)RSCM estimator. The performance differences were often significant.

REFERENCES

[1] O. Ledoit and M. Wolf, "A well-conditioned estimator for large-dimensional covariance matrices," *J. Mult. Anal.*, vol. 88, pp. 365–411, 2004.
[2] E. Ollila and D. E. Tyler, "Regularized M-estimators of scatter matrix," *IEEE Trans. Signal Process.*, vol. 62, no. 22, pp. 6059–6070, 2014.
[3] T. Zhang and A. Wiesel, "Automatic diagonal loading for tyler's robust covariance estimator," in *IEEE Statistical Signal Processing Workshop (SSP'16)*, 2016, pp. 1–5.
[4] Y. Chen, A. Wiesel, Y. C. Eldar, and A. O. Hero, "Shrinkage algorithms for mmse covariance estimation," *IEEE Trans. Signal Process.*, vol. 58, no. 10, pp. 5016–5029, 2010.
[5] R. Couillet and M. McKay, "Large dimensional analysis and optimization of robust shrinkage covariance matrix estimators," *J. Mult. Anal.*, vol. 131, pp. 99–120, 2014.
[6] F. Pascal, Y. Chitour, and Y. Quek, "Generalized robust shrinkage estimator and its application to stap detection problem," *IEEE Trans. Signal Process.*, vol. 62, no. 21, pp. 5640–5651, 2014.
[7] R. J. Muirhead, *Aspects of Multivariate Statistical Theory*.   New York: Wiley, 1982, 704 pages.
[8] K.-T. Fang, S. Kotz, and K. W. Ng, *Symmetric Multivariate and Related Distributions*.   London: Chapman and hall, 1990.
[9] E. Ollila, D. E. Tyler, V. Koivunen, and H. V. Poor, "Complex elliptically symmetric distributions: survey, new results and applications," *IEEE Trans. Signal Process.*, vol. 60, no. 11, pp. 5597–5625, 2012.
[10] J. R. Magnus and H. Neudecker, *Matrix Differential Calculus with Applications in Statistics and Econometrics*, 2nd ed.   Chichester: Wiley, 1999, 422 pages.
[11] M. S. Srivastava, "Some tests concerning the covariance matrix in high dimensional data," *Journal of the Japan Statistical Society*, vol. 35, no. 2, pp. 251–272, 2005.
[12] C. Croux, E. Ollila, and H. Oja, "Sign and rank covariance matrices: statistical properties and application to principal components analysis," in *Statistical data analysis based on the L1-norm and related methods*. Birkhäuser Basel, 2002, pp. 257–269.
[13] P. M. Bentler and M. Berkane, "Greatest lower bound to the elliptical theory kurtosis parameter," *Biometrika*, vol. 73, no. 1, pp. 240–241, 1986.