Optimal Shrinkage Covariance Matrix Estimation Under Random Sampling From Elliptical Distributions

Esa Ollila, *Member, IEEE*, and Elias Raninen, *Student Member, IEEE*

***Abstract*—This paper considers the problem of estimating a high-dimensional covariance matrix in a low sample support situation where the sample size is smaller, or not much larger, than the dimensionality of the data, which could potentially be very large. We develop a regularized sample covariance matrix (RSCM) estimator which can be applied in commonly occurring high-dimensional data problems. The proposed RSCM estimator is based on estimators of the unknown optimal (oracle) shrinkage parameters that yield the minimum mean squared error between the RSCM and the true covariance matrix when the data is sampled from an unspecified elliptically symmetric distribution. We propose two variants of the RSCM estimator which differ in the approach in which they estimate the underlying sphericity parameter involved in the theoretical optimal shrinkage parameter. The performance of the proposed RSCM estimators are evaluated with numerical simulation studies. In particular, when the sample sizes are low, the proposed RSCM estimators often show a significant improvement over the conventional RSCM estimator by Ledoit and Wolf (2004). We further evaluate the performance of the proposed estimators in a portfolio optimization problem with real data wherein the proposed methods are able to outperform the benchmark methods.**

***Index Terms*—Sample covariance matrix, shrinkage estimation, regularization, elliptical distribution.**

I. Introduction

ESTIMATING high-dimensional covariance matrices in a low sample support situation where the sample size n is smaller or not much larger than the dimension p of the samples is a problem that has attracted significant research interest in the recent years [1]–[9]. This is due to the fact that high-dimensional data analysis problems have become increasingly common in a wide spectrum of fields, such as in finance [2], bioinformatics, and classification [10].

We consider the problem of estimating the high-dimensional covariance matrix based on a sample $\mathbf{x}_1, \ldots, \mathbf{x}_n$ of independent and identically distributed (i.i.d.) random vectors. The observations are assumed to be generated from an unspecified p-variate distribution $\mathbf{x} \sim F$ with a mean vector $\boldsymbol{\mu} = \mathbb{E}[\mathbf{x}]$ and a $p \times p$ positive definite covariance matrix

$$\boldsymbol{\Sigma} = \mathrm{var}(\mathbf{x}) = \mathbb{E}\big[(\mathbf{x}-\boldsymbol{\mu})(\mathbf{x}-\boldsymbol{\mu})^\top\big] \in \mathbb{S}_{++}^{p\times p}.$$

The most commonly used estimators of the unknown parameters $(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \in \mathbb{R}^p \times \mathbb{S}_{++}^{p\times p}$ are the *sample mean vector* and the *sample covariance matrix* (SCM),

$$\bar{\mathbf{x}} = \frac{1}{n}\sum_{i=1}^n \mathbf{x}_i,$$

$$\mathbf{S} = \frac{1}{n-1}\sum_{i=1}^n (\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})^\top.$$

They have desirable properties, such as being unbiased estimators and in the case when the samples are generated from the multivariate normal distribution they are sufficient statistics. However, the SCM does not perform well in high-dimensional problems for several reasons. Foremost, significant estimation errors result from having an insufficient number of samples. Moreover, if $p > n$, the SCM is always singular, i.e., it is not invertible even though the true covariance matrix is known to be positive definite and hence non-singular. In these situations, a frequently used approach for improving the estimation accuracy is to use shrinkage regularization.

One of the most commonly used estimators in low sample support problems is the *regularized* SCM (RSCM) of the form

$$\mathbf{S}_{\alpha,\beta} = \beta\mathbf{S} + \alpha\mathbf{I}, \tag{1}$$

where $\alpha, \beta > 0$ denote the *shrinkage parameters* or *regularization parameters*. In signal processing, an estimator of the form (1) was proposed in [11], [12] and is often referred to as the diagonal loading estimator. Another line of research has been to consider robust regularized covariance matrix estimators, e.g., [2]–[8], and [9]. In this paper, the focus is on determining the optimal (in MSE sense) shrinkage parameters for the RSCM.

We define the optimal RSCM estimator as the one that is based on the *oracle shrinkage parameters* minimizing the mean squared error (MSE), that is,

$$(\alpha_o, \beta_o) = \underset{\alpha,\beta>0}{\arg\min}\, \mathbb{E}\Big[\big\|\mathbf{S}_{\alpha,\beta} - \boldsymbol{\Sigma}\big\|_{\mathrm{F}}^2\Big], \tag{2}$$

where $\|\cdot\|_{\mathrm{F}}$ denotes the Frobenius matrix norm, i.e., $\|\mathbf{A}\|_{\mathrm{F}}^2 = \mathrm{tr}(\mathbf{A}^\top\mathbf{A}) = \mathrm{tr}(\mathbf{A}\mathbf{A}^\top)$ for any matrix $\mathbf{A}$. We use the prefix *oracle* for the shrinkage parameters (α_o, β_o) as they depend on

Manuscript received August 29, 2018; revised December 14, 2018, February 12, 2019, and March 13, 2019; accepted March 14, 2019. Date of publication March 28, 2019; date of current version April 22, 2019. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Francois Desbouvries. This work was supported by the Academy of Finland under Grant 298118. *(Corresponding author: Esa Ollila.)*

The authors are with the Department of Signal Processing and Acoustics, Aalto University, Espoo 02150, Finland (e-mail: esa.ollila@aalto.fi; elias.raninen@aalto.fi).

Digital Object Identifier 10.1109/TSP.2019.2908144

the true unknown covariance matrix $\mathbf{\Sigma}$. Although, the oracle shrinkage parameters cannot be used in practice, they have the theoretical significance for being a benchmark for best possible performance w.r.t. the MSE metric.

The widely popular Ledoit-Wolf (LW-)RSCM [2] estimator is based on the mean square consistent estimators $(\hat{\alpha}_o^{\mathrm{LW}}, \hat{\beta}_o^{\mathrm{LW}})$ of the oracle parameters (α_o, β_o) under the random matrix theory (RMT) regime R1, i.e., when both p and n increase such that

(R1) $c = p/n \to c_0$, $0 < c_0 < \infty$, as $n, p \to \infty$.

It is, however, possible to improve upon the LW-estimator and obtain a more accurate finite sample estimation performance by assuming that the observations are generated from a specific p-variate distribution, e.g., the multivariate normal (MVN) distribution. For example, in [7, Theorem 1], the authors derived an optimal shrinkage parameter assuming that the samples have a Gaussian distribution with a known location ($\boldsymbol{\mu}$). Such a strict assumption on the distribution of the data implies performance loss if the assumption does not hold. Another somewhat related approach has been taken for example in [13, Proposition 3], where the authors considered robust M-estimators and looked for an asymptotically optimal shrinkage parameter in the RMT regime which minimizes the squared Frobenius distance between normalized regularized M-estimators of scatter matrix and a normalized covariance matrix.

In this paper, we instead assume that the observations are from an unspecified elliptically symmetric (ES) distribution and derive estimators of the optimal oracle shrinkage parameters (α_o, β_o) that are able to perform reliably under the RMT regime. ES distributions form a large class of distributions comprising, e.g., the MVN distribution, generalized Gaussian, and all compound Gaussian distributions as special cases, see e.g., [14], [15], and [16].

The rest of this 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 p-variate distribution and an elliptical distribution with finite fourth order moments, respectively. In Section IV, we develop estimators of (α_o, β_o) under the RMT regime and when sampling from an unspecified elliptically symmetric distribution. In Section V, we conduct several simulation studies and compare the proposed estimators with the popular LW-RSCM estimator. In Section VI, we evaluate the performance of the proposed estimators in a portfolio optimization application using the Global Minimum Variance Portfolio (GMVP) framework, where we use real data of historical (daily) stock returns from Hong Kong's Hang Seng Index (HSI) and Standard and Poor's 500 (S&P 500) index. In the application, the proposed methods show better performance than the benchmark methods. Finally, Section VII concludes.

Notation: We denote the open cone of $p \times p$ positive definite symmetric matrices by $\mathbb{S}_{++}^{p\times p}$. The *vectorization* of an $n \times p$ matrix $\mathbf{A} = (\mathbf{a}_1, \dots \mathbf{a}_p)$ is denoted by $\mathrm{vec}(\mathbf{A}) = (\mathbf{a}_1^\top, \dots, \mathbf{a}_p^\top)^\top$. The *matrix trace* of a square matrix $\mathbf{A}$ is denoted by $\mathrm{tr}(\mathbf{A})$. The *Kronecker product* $\mathbf{A} \otimes \mathbf{B}$ of any matrices $\mathbf{A}$ and $\mathbf{B}$ is a block matrix with its ijth block being equal to $a_{ij}\mathbf{B}$. The Kronecker product has the useful property: $(\mathbf{A} \otimes \mathbf{B})(\mathbf{C} \otimes \mathbf{D}) = (\mathbf{AC} \otimes \mathbf{BD})$ for the matrices $\mathbf{A}$, $\mathbf{B}$, $\mathbf{C}$, and $\mathbf{D}$ of appropriate dimensions. We denote the *identity matrix* of proper dimension by $\mathbf{I}$ and the *centering matrix* of proper dimension by $\mathbf{H} = \mathbf{I} - \mathbf{1}\mathbf{1}^\top/n$, where $\mathbf{1}$ is a vector of ones. The *standard basis vector*, which has its ith element equal to 1 and all other elements zero is denoted by $\mathbf{e}_i$. The *commutation matrix* $\mathbf{K}_p$ is a $p^2 \times p^2$ block matrix with its ijth block equal to a $p \times p$ matrix that has a 1 at element ji and zeros elsewhere, i.e., $\mathbf{K}_p = \sum_{i,j} \mathbf{e}_i\mathbf{e}_j^\top \otimes \mathbf{e}_j\mathbf{e}_i^\top$. It also has the following important properties [17]: $\mathbf{K}_p\mathrm{vec}(\mathbf{A}) = \mathrm{vec}(\mathbf{A}^\top)$ and $\mathbf{K}_p(\mathbf{A} \otimes \mathbf{B})\mathbf{K}_p = (\mathbf{B} \otimes \mathbf{A})$ for any $p \times p$ matrices $\mathbf{A}$ and $\mathbf{B}$. Throughout the paper, we will also use the following identities: $\mathrm{tr}(\mathbf{A} \otimes \mathbf{B}) = \mathrm{tr}(\mathbf{A})\mathrm{tr}(\mathbf{B})$, $\mathrm{tr}(\mathrm{vec}(\mathbf{A})\mathrm{vec}(\mathbf{B})^\top) = \mathrm{tr}(\mathbf{A}^\top\mathbf{B}) = \mathrm{vec}(\mathbf{B})^\top\mathrm{vec}(\mathbf{A})$ for any square matrices $\mathbf{A}$ and $\mathbf{B}$ of same order. Notation "$=_d$" reads "has the same distribution as".

II. Optimal Oracle Shrinkage Parameters

In this section, we derive the oracle shrinkage parameters (α_o, β_o) for any p-variate distribution. First, we define the *scale* and *sphericity* parameters of $\mathbf{\Sigma} \in \mathbb{S}_{++}^{p\times p}$ as

$$\eta = \frac{\mathrm{tr}(\mathbf{\Sigma})}{p} \quad \text{and} \quad \gamma = \frac{p\mathrm{tr}(\mathbf{\Sigma}^2)}{\mathrm{tr}(\mathbf{\Sigma})^2}. \tag{3}$$

Note that η equals the mean of the eigenvalues of $\mathbf{\Sigma}$ whereas γ is equal to the ratio of the mean of the squared eigenvalues relative to the mean of the eigenvalues squared. The sphericity γ [1], [18] measures how close the covariance matrix is to a scaled identity matrix. Furthermore, the values for the sphericity are in the range $1 \le \gamma \le p$. This can be seen by applying the Cauchy-Schwartz inequality:

$$\mathrm{tr}(\mathbf{\Sigma})^2 = \left(\sum_{i=1}^{p} \lambda_i \cdot 1\right)^2 \le p \cdot \sum_{i=1}^{p} \lambda_i^2 = p\mathrm{tr}(\mathbf{\Sigma}^2).$$

By dividing the right-hand side of the equation by the left-hand side, we have $\gamma \ge 1$ with equality if and only if $\mathbf{\Sigma} = \eta\mathbf{I}$ for some $\eta > 0$. Furthermore, the upper bound $\gamma = p$ is achieved for rank one matrices in which case $\mathbf{\Sigma}$ has only one non-zero eigenvalue.

The scale and sphericity, η and γ, are elemental in our developments. As is shown in Theorem 3, the optimal shrinkage parameter pair (α_o, β_o) for a given elliptical distribution depends on the true covariance matrix $\mathbf{\Sigma}$ only through η and γ. Simple plug-in estimates of (α_o, β_o) can then be obtained by replacing (η, γ) with their estimates. If the elliptical distribution is unknown, an additional elliptical kurtosis parameter needs to be estimated.

The next theorem provides the expressions for the oracle shrinkage parameters in the case of sampling from an unspecified p-variate distribution with finite fourth order moments. Write $\mathrm{MSE}(\mathbf{S}) = \mathbb{E}\left[\left\|\mathbf{S} - \mathbf{\Sigma}\right\|_{\mathrm{F}}^2\right]$ for the mean squared error (MSE) and

$$\mathrm{NMSE}(\mathbf{S}) = \frac{\mathbb{E}\left[\|\mathbf{S} - \mathbf{\Sigma}\|_{\mathrm{F}}^2\right]}{\|\mathbf{\Sigma}\|_{\mathrm{F}}^2}$$

for the *normalized MSE*.

Theorem 1: Let $\mathbf{x}_1, \dots, \mathbf{x}_n$ denote an i.i.d. random sample from any p-variate distribution with finite fourth order moments,

mean vector $\boldsymbol{\mu}$, and covariance matrix $\boldsymbol{\Sigma}$. Then the oracle shrinkage parameters in (2) are

$$\beta_o = \frac{p(\gamma-1)\eta^2}{\mathbb{E}\left[\mathrm{tr}\left(\mathbf{S}^2\right)\right] - p\eta^2} \tag{4}$$

$$= \frac{(\gamma-1)}{(\gamma-1)+\gamma\cdot \mathrm{NMSE}(\mathbf{S})} \tag{5}$$

and

$$\alpha_o = (1-\beta_o)\eta, \tag{6}$$

where η and γ are defined in (3). Furthermore, the optimal β_o is always in the range $[0,1)$ and the value of the MSE at the optimum is

$$\mathrm{MSE}(\mathbf{S}_{\alpha_o,\beta_o}) = (1-\beta_o)\|\boldsymbol{\Sigma}-\eta\mathbf{I}\|_{\mathrm{F}}^2. \tag{7}$$

Proof: See appendix. ■

Theorem 1 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 as

$$\mathbf{S}_{\alpha_o,\beta_o} = \beta_o\mathbf{S} + (1-\beta_o)\eta\mathbf{I}.$$

The scale η can be estimated with

$$\hat{\eta} = \frac{\mathrm{tr}(\mathbf{S})}{p}, \tag{8}$$

which is a mean square consistent estimator both in the conventional (fixed p) regime and the RMT asymptotic regime. Therefore, the estimator of α_o is simply $\hat{\alpha}_o = (1-\hat{\beta}_o)\hat{\eta}$, and we can focus on finding an estimator $\hat{\beta}_o$ of β_o.

This is the approach also taken by Ledoit and Wolf [2] who developed an estimator $\hat{\beta}_o^{\mathrm{LW}}$ that converges to β_o in (4) under the RMT regime (R1) and some mild technical assumptions when sampling from a distribution $\mathbf{x}\sim F$ with finite 8th order moments. The estimate of α_o is then $\hat{\alpha}_o^{\mathrm{LW}} = (1-\hat{\beta}_o^{\mathrm{LW}})\hat{\eta}$. The RSCM based on the shrinkage parameter pair $(\hat{\alpha}_o^{\mathrm{LW}}, \hat{\beta}_o^{\mathrm{LW}})$ of [2] is referred hereafter as the LW-RSCM estimator.

III. Optimal Oracle Shrinkage Parameters: The Elliptical Case

We now derive the optimal oracle shrinkage parameters for the case in which the data can be assumed elliptically distributed. For a review of elliptical distributions, see [14], [15], and [16].

The probability density function (p.d.f.) of an elliptically distributed random vector $\mathbf{x}\sim\mathcal{E}_p(\boldsymbol{\mu},\boldsymbol{\Sigma},g)$ is

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

where $\mathbb{E}[\mathbf{x}]=\boldsymbol{\mu}$ is the mean vector, $\boldsymbol{\Sigma}$ is the positive definite covariance matrix, $g:[0,\infty)\to[0,\infty)$ is the *density generator*, which is a fixed function that is independent of $\mathbf{x}$, $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$, and $C_{p,g}$ is a normalizing constant ensuring that $f(\mathbf{x})$ integrates to 1. Here, we let g to be defined so that $\boldsymbol{\Sigma}$ represents the covariance matrix of $\mathbf{x}$, which means that $\int_0^\infty t^{p/2}\,g(t)\mathrm{d}t = p$. The density generator g determines the elliptical distribution. For example, the multivariate normal (MVN) distribution, denoted $\mathbf{x}\sim\mathcal{N}_p(\boldsymbol{\mu},\boldsymbol{\Sigma})$, is obtained when $g(t)=\exp(-t/2)$. As in Theorem 1, we assume that the elliptical population possesses finite fourth order moments. Technically, this implies that

$$\int_0^\infty t^{p/2+1}\,g(t)\mathrm{d}t < \infty. \tag{9}$$

For example, the MVN and the multivariate t-distribution with degrees of freedom $\nu>4$ verify the above condition.

The *kurtosis* of a random variable x is defined as

$$\mathrm{kurt}(x) = \frac{\mathbb{E}[(x-\mu)^4]}{(\mathbb{E}[(x-\mu)^2])^2} - 3,$$

where $\mu=\mathbb{E}[x]$. The *elliptical kurtosis parameter* [14] κ of a random vector $\mathbf{x}=(x_1,\ldots,x_p)^\top\sim\mathcal{E}_p(\boldsymbol{\mu},\boldsymbol{\Sigma},g)$ is defined as

$$\kappa = \frac{\mathbb{E}[r^4]}{p(p+2)} - 1 = \frac{1}{3}\cdot\mathrm{kurt}(x_i), \tag{10}$$

where r is the *generating variate* or *second order modular variate* of the elliptical distribution, which is defined as the square-root of the quadratic form $r^2=(\mathbf{x}-\boldsymbol{\mu})^\top\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})$. Above $\mathrm{kurt}(x_i)$ denotes the kurtosis of (any) marginal variable x_i. Later, in Section IV, we use the relationship of κ with $\mathrm{kurt}(x_i)$ in devising an estimator $\hat{\kappa}$. Indeed this relationship allows us to estimate κ without the need to estimate the fourth order moment of r from the data.

The elliptical kurtosis shares properties similar to the kurtosis of a real random variable. Especially, if $\mathbf{x}\sim\mathcal{N}_p(\boldsymbol{\mu},\boldsymbol{\Sigma})$, then $\kappa=0$. This is obvious since the marginal distributions are Gaussian and hence $\kappa=(1/3)\,\mathrm{kurt}(x_i)=0$. Another way to derive this is by noting that the quadratic form r^2 has a chi-squared distribution with p degrees of freedom, i.e., $r^2\sim\chi_p^2$, and hence $\mathbb{E}[r^4]=p(p+2)$.

The importance of the elliptical kurtosis parameter κ is due to the fact that the $p^2\times p^2$ covariance matrix of $\mathrm{vec}(\mathbf{S})$ depends on the underlying elliptical distribution g only through κ. This result is established in Theorem 2.

Theorem 2: Let $\mathbf{x}_1,\ldots,\mathbf{x}_n$ denote an i.i.d. random sample from an elliptical distribution with finite fourth order moments, mean vector $\boldsymbol{\mu}$, and covariance matrix $\boldsymbol{\Sigma}$. Then

$$\mathrm{var}(\mathrm{vec}(\mathbf{S})) = \left(\frac{1}{n-1}+\frac{\kappa}{n}\right)(\mathbf{I}+\mathbf{K}_p)(\boldsymbol{\Sigma}\otimes\boldsymbol{\Sigma}) + \frac{\kappa}{n}\mathrm{vec}(\boldsymbol{\Sigma})\mathrm{vec}(\boldsymbol{\Sigma})^\top. \tag{11}$$

Proof: See appendix. ■

Theorem 2 reveals that the elliptical kurtosis parameter κ along with the true covariance matrix $\boldsymbol{\Sigma}$ provide a complete description of the covariances between the elements s_{ij} and s_{kl} of the SCM $\mathbf{S}=(s_{ij})$. The mathematics underlying Theorem 2 is so rich that we are able to relate it to at least three fundamental results in the field of statistics given below.

First, consider the one-dimensional case, $p=1$, where we have a univariate sample $x_1,\ldots,x_n$ from a distribution of a random variable $x\in\mathbb{R}$. Then the SCM reduces to the unbiased

sample variance $s^2 = (1/(n-1))\sum_{i=1}^n (x_i - \bar{x})^2$ and equation (11) reduces to $\mathrm{var}(s^2)$. We can now compute $\mathrm{var}(s^2)$ using (11), which states that

$$\begin{aligned}\mathrm{var}(s^2) &= \left(\frac{1}{n-1} + \frac{\kappa}{n}\right) 2\sigma^4 + \frac{\kappa}{n}\sigma^4 \\ &= \sigma^4 \left(\frac{2}{n-1} + \frac{\mathrm{kurt}(x)}{n}\right), \end{aligned} \tag{12}$$

where we used that $\boldsymbol{\Sigma} \equiv \sigma^2 = \mathbb{E}[(x - \mathbb{E}[x])^2]$, $\boldsymbol{\Sigma} \otimes \boldsymbol{\Sigma} = \sigma^4$ and $\kappa = \mathrm{kurt}(x)/3$ due to (10). Hence, we obtained the classic formula for $\mathrm{var}(s^2)$ often encountered in elementary statistics textbooks. Under the Gaussian distribution, $\mathrm{kurt}(x) = 0$, in which case Theorem 2 states that $\mathrm{var}(s^2) = 2\sigma^4/(n-1)$. This is an expected result since $(n-1)s^2/\sigma^2 = \sum_i (x_i - \bar{x})^2/\sigma^2 \sim \chi^2_{n-1}$.

Secondly, we can connect Theorem 2 with the well-known covariance matrix of the Wishart distribution. Let $W_p(m, \mathbf{M})$ denote the Wishart distribution of a random symmetric positive definite $p \times p$ matrix where $m > p-1$ denotes the degrees of freedom parameter and $\mathbf{M} \in \mathbb{S}_{++}^{p\times p}$ denotes the scale matrix parameter of the Wishart distribution. Under the MVN assumption, it is well-known that $(n-1)\mathbf{S} \sim W_p(n-1, \boldsymbol{\Sigma})$ and consequently $\mathrm{var}(\mathrm{vec}(\mathbf{S}))$ has the famous covariance matrix form

$$\mathrm{var}(\mathrm{vec}(\mathbf{S})) = \frac{1}{n-1}(\mathbf{I} + \mathbf{K}_p)(\boldsymbol{\Sigma} \otimes \boldsymbol{\Sigma}). \tag{13}$$

Suppose now that the elliptical distribution in Theorem 2 is the multivariate normal, thus, $\mathbf{x}_1, \ldots, \mathbf{x}_n \overset{iid}{\sim} \mathcal{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. Since in this case $\kappa = 0$, we have that (11) reduces to (13).

Lastly, notice that

$$\begin{aligned}\mathrm{var}(\sqrt{n}\mathrm{vec}(\mathbf{S})) \to (1+\kappa)(\mathbf{I} + \mathbf{K}_p)(\boldsymbol{\Sigma} \otimes \boldsymbol{\Sigma}) \\ + \kappa\, \mathrm{vec}(\boldsymbol{\Sigma})\mathrm{vec}(\boldsymbol{\Sigma})^\top\end{aligned}$$

as $n \to \infty$. The right-hand side of the previous equation equals the well-known asymptotic covariance matrix of the limiting normal distribution of $\sqrt{n}(\mathrm{vec}(\mathbf{S}) - \mathrm{vec}(\boldsymbol{\Sigma}))$ when sampling from an elliptical distribution $\mathcal{E}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}, g)$ with finite fourth order moments. This is a famous result in multivariate statistics [14].

In the next Lemma, we derive the MSE and normalized MSE (NMSE) of the SCM.

Lemma 1: Let $\mathbf{x}_1, \ldots, \mathbf{x}_n$ denote an i.i.d. random sample from a p-variate elliptical distribution with finite fourth order moments, mean $\boldsymbol{\mu}$, and covariance matrix $\boldsymbol{\Sigma}$. Then the MSE and the NMSE of $\mathbf{S}$ are

$$\mathrm{MSE}(\mathbf{S}) = \left(\frac{1}{n-1} + \frac{\kappa}{n}\right)\mathrm{tr}(\boldsymbol{\Sigma})^2 + \left(\frac{1}{n-1} + \frac{2\kappa}{n}\right)\mathrm{tr}(\boldsymbol{\Sigma}^2)$$

$$\mathrm{NMSE}(\mathbf{S}) = \left(1 + \frac{p}{\gamma}\right)\left(\frac{1}{n-1} + \frac{\kappa}{n}\right) + \frac{\kappa}{n}$$

where γ and κ are defined in (3) and (10), respectively.

Proof: see Appendix. ∎

The 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 3: Let $\mathbf{x}_1, \ldots, \mathbf{x}_n$ denote an i.i.d. random sample from an elliptical distribution with finite fourth order moments, mean $\boldsymbol{\mu}$, and covariance matrix $\boldsymbol{\Sigma}$. Then the oracle parameters (α_o, β_o) that minimize the MSE are

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

and $\alpha_o^{\text{Ell}} = (1 - \beta_o^{\text{Ell}})\eta$, where the parameters η, γ and κ are defined in (3) and (10), respectively.

Proof: Follows from (5) and Lemma 1. ∎

It is not surprising that β_o, and hence also α_o, depend on the density generator g of the elliptical distribution only via the elliptical kurtosis parameter κ. Specifying the elliptical distribution also specifies the value of κ. For example, when sampling from the Gaussian distribution, the elliptical kurtosis parameter is $\kappa = 0$ and β_o^{Ell} in Theorem 3 reduces to

$$\beta_o^{\text{Gau}} = \frac{(\gamma-1)}{(\gamma-1) + (\gamma+p)/(n-1)}. \tag{14}$$

Consequently, an estimator of β_o^{Gau} is obtained by substituting an estimator $\hat{\gamma}$ in place of γ in (14). Recall that an estimator of α_o^{Gau} is then obtained as $\hat{\alpha}_o^{\text{Gau}} = (1 - \hat{\beta}_o^{\text{Gau}})\mathrm{tr}(\mathbf{S})/p$.

Since in this paper we do not assume any particular elliptical distribution, we need to find an estimator $\hat{\kappa}$ of the elliptical kurtosis parameter κ as well. Naturally, if the assumption on multivariate normality of the data is valid, then (14) should be used for estimating the optimal oracle value.

When the mean vector of the population is known, the unbiased SCM is $\mathbf{S} = \frac{1}{n}\sum_{i=1}^n \mathbf{x}_i\mathbf{x}_i^\top$ (as one can assume without loss of generality that $\boldsymbol{\mu} = \mathbf{0}$). In this case the optimal shrinkage parameter β_o of the RSCM stated in Theorem 3 remains unchanged apart from the last term in the denominator of β_o, that is, $(\gamma+p)/(n-1)$ is replaced by $(\gamma+p)/n$. This centered case was addressed in [19].

IV. Estimation of the Oracle Parameters

In this section, we develop estimators $\hat{\gamma}$ and $\hat{\kappa}$ of the unknown parameters γ and κ that determine the shrinkage parameter β_o ($cf.$ Theorem 3). These are used to obtain *plug-in estimators* of the shrinkage parameters as

$$\hat{\beta}_o^{\text{Ell}} = \frac{(\hat{\gamma}-1)}{(\hat{\gamma}-1) + \hat{\kappa}(2\hat{\gamma}+p)/n + (\hat{\gamma}+p)/(n-1)},$$

$$\hat{\alpha}_o^{\text{Ell}} = (1 - \hat{\beta}_o^{\text{Ell}})\hat{\eta}.$$

We will first address the estimation of κ. Regarding γ, we found two different well performing estimators, and hence, we will address its estimation last.

A natural estimate of κ is the conventional sample average

$$\hat{\kappa} = \max\left(-\frac{2}{p+2}, \frac{1}{3p}\sum_{j=1}^p \hat{K}_j\right), \tag{15}$$

where $\hat{K}_j$ is an estimate of the kurtosis of the jth variable and defined as

$$\hat{K}_j = \frac{n-1}{(n-2)(n-3)}\left((n+1)\hat{k}_j + 6\right).$$

Here $\hat{k}_j = m_j^{(4)}/\left(m_j^{(2)}\right)^2 - 3$ denotes the conventional sample estimate of the kurtosis of the jth variable, where $m_j^{(q)} = \frac{1}{n}\sum_{i=1}^n (x_{ij} - \bar{x}_j)^q$ denotes the qth order sample moment. Note that $\hat{k}_j$ is a biased estimator of the kurtosis even for Gaussian samples [20]. We use a popular estimator of the kurtosis $\hat{K}_j$ used by many statistical software packages such as SAS which corrects for the bias of the conventional sample kurtosis $\hat{k}_j$. Kurtosis $\mathrm{kurt}(x_j)$ can be expressed as a ratio of the fourth order cumulant to the squared variance. The estimator $\hat{K}_j$ is formed by taking the ratio of the unbiased sample estimators of these two statistics. It is an unbiased estimator of the kurtosis when the samples are Gaussian distributed [20]. Furthermore, in the simulation study of [20] it obtained a smaller MSE than $\hat{k}_j$ when the sample length n was small and the samples were generated from a non-Gaussian distribution. To ensure that the final estimate $\hat{\kappa}$ does not go below the theoretical lower bound of $-2/(p+2)$ [21], a maximum constraint is used in (15). The constructed estimate of κ is consistent both in the conventional and the RMT regime.

Note that, if the estimates $(\hat{\gamma}, \hat{\kappa})$ are restricted to be within their theoretical ranges, i.e., $1 \le \hat{\gamma} \le p$ and $\hat{\kappa} \ge -2/(p+2)$, then it is straightforward to verify that the plug-in estimator satisfies $\hat{\beta}_o^{\mathrm{Ell}} \in [0, 1)$.

In the following subsections, we consider two options for estimating the sphericity γ under the RMT regime (R1). We denote the estimators by $\hat{\gamma}^{\mathrm{Ell1}}$ and $\hat{\gamma}^{\mathrm{Ell2}}$. Both estimators have their own benefits and disadvantages. The first estimator, $\hat{\gamma}^{\mathrm{Ell1}}$, enjoys statistical robustness with respect to heavier-tailed distributions. The second estimator, $\hat{\gamma}^{\mathrm{Ell2}}$, is computationally more efficient and can easily be used and tuned for very high-dimensional set-ups such as microarray studies where p is often tens of thousands but n is of few tens [22]. It is also highly efficient under Gaussianity, or for mild departures from Gaussianity. Its obvious disadvantage is that it is not robust to heavier-tailed elliptical distributions.

A. Ell1-RSCM Estimator

The first estimator of the sphericity γ uses the *sample spatial sign covariance matrix* [23] defined as

$$\mathbf{S}_{\mathrm{sgn}} = \frac{1}{n}\sum_{i=1}^n \frac{(\mathbf{x}_i - \hat{\boldsymbol{\mu}})(\mathbf{x}_i - \hat{\boldsymbol{\mu}})^\top}{\|\mathbf{x}_i - \hat{\boldsymbol{\mu}}\|^2}, \tag{16}$$

where $\hat{\boldsymbol{\mu}} = \arg\min_{\boldsymbol{\mu}} \sum_{i=1}^n \|\mathbf{x}_i - \boldsymbol{\mu}\|$ is the *sample spatial median* [24]. The sample spatial sign covariance matrix is well-known to be highly robust although it is not a consistent estimator of the covariance matrix [25], [26]. Namely, it does provide consistent estimators of the eigenvectors of the covariance matrix but not of the eigenvalues.

Consider an estimator of the form,

$$\begin{aligned}\hat{\gamma}^{\mathrm{Ell1}*} &= \frac{n}{n-1}\left(p\mathrm{tr}\left(\mathbf{S}_{\mathrm{sgn}}^2\right) - \frac{p}{n}\right)\\ &= \frac{p}{n(n-1)}\sum_{i\neq j}(\mathbf{v}_i^\top \mathbf{v}_j)^2\\ &= p\,\mathrm{ave}_{i\neq j}\{\cos^2(\sphericalangle(\mathbf{x}_i, \mathbf{x}_j))\},\end{aligned}$$

where $\mathrm{ave}_{i\neq j}$ denotes the arithmetic average over indices, $i, j \in \{1, \ldots, n\}$, $i \neq j$, and $\mathbf{v}_i = (\mathbf{x}_i - \hat{\boldsymbol{\mu}})/\|\mathbf{x}_i - \hat{\boldsymbol{\mu}}\|$.

In [5, Lemma 4.1] it was shown that $\hat{\gamma}^{\mathrm{Ell1}*}$ is a consistent estimator of γ when sampling from a centered elliptical distribution $\mathcal{E}_p(\mathbf{0}, \boldsymbol{\Sigma}, g)$ and when the eigenvalues of $\boldsymbol{\Sigma}$ converge to a fixed spectrum, verifying the assumption:

(R2) $\mathrm{tr}(\boldsymbol{\Sigma}^i)/p$ has a finite and positive limit for $i = 1, 2, 3, 4$ under the RMT regime (R1), that is, when $n, p \to \infty$ and $p/n \to c_0 \in (0, \infty)$.

In their paper, it was assumed that the location (symmetry center) is known to be zero, which is why they do not have the centering of the samples by the sample spatial median in (16). We also remark that our estimator $\hat{\gamma}^{\mathrm{Ell1}*}$ differs from [5, Lemma 4.1] in that we scale their estimator by $\frac{n}{n-1}$. This scaling is used for correcting bias for small samples and is needed to ensure that $\mathbb{E}[\hat{\gamma}^{\mathrm{Ell1}*}] \in [1, p]$. In order to guarantee that the estimate remains inside the valid interval $[1, p]$, as a final estimator, we use

$$\hat{\gamma}^{\mathrm{Ell1}} = \min(p, \max(1, \hat{\gamma}^{\mathrm{Ell1}*})). \tag{17}$$

We can now define the *Ell1-RSCM estimator* as the RSCM based on the estimators of the optimal shrinkage parameters using the plug-in estimates $\hat{\eta}$ of (8), $\hat{\kappa}$ of (15) and $\hat{\gamma}^{\mathrm{Ell1}}$ of (17).

B. Ell2-RSCM Estimator

In order to develop the second estimator $\hat{\gamma}^{\mathrm{Ell2}}$ of γ, we need to find the expressions for $\mathbb{E}[\mathrm{tr}(\mathbf{S}^2)]$ and $\mathbb{E}[\mathrm{tr}(\mathbf{S})^2]$, which are given in the following Lemma 2.

Lemma 2: Let $\mathbf{x}_1, \ldots, \mathbf{x}_n$ denote an i.i.d. random sample from an elliptical distribution with finite fourth order moments, mean vector $\boldsymbol{\mu}$, and covariance matrix $\boldsymbol{\Sigma}$. Then

$$\mathbb{E}[\mathrm{tr}(\mathbf{S}^2)] = \left(\frac{1}{n-1} + \frac{\kappa}{n}\right)\mathrm{tr}(\boldsymbol{\Sigma})^2 + \left(1 + \frac{1}{n-1} + \frac{2\kappa}{n}\right)\mathrm{tr}(\boldsymbol{\Sigma}^2)$$

and

$$\mathbb{E}[\mathrm{tr}(\mathbf{S})^2] = \left(1 + \frac{\kappa}{n}\right)\mathrm{tr}(\boldsymbol{\Sigma})^2 + 2\left(\frac{1}{n-1} + \frac{\kappa}{n}\right)\mathrm{tr}(\boldsymbol{\Sigma}^2).$$

Proof: see Appendix. ■

Next, we construct an estimator for $\vartheta = \mathrm{tr}(\boldsymbol{\Sigma}^2)/p$. The natural plug-in estimate, $\mathrm{tr}(\mathbf{S}^2)/p$, is not a mean square consistent estimator in the RMT regime (R1) and assumption (R2). This follows at once from Lemma 2 as it shows that $\mathrm{tr}(\mathbf{S}^2)/p$ is not asymptotically unbiased since

$$\lim_{\substack{n,p\to\infty \\ p/n\to c_o}} \frac{\mathbb{E}[\mathrm{tr}(\mathbf{S}^2)]}{p} = c_0(1+\kappa)\eta_o^2 + \vartheta_0,$$

where $\eta_o > 0$ and $\vartheta_o > 0$ denote finite limit values of $\mathrm{tr}(\boldsymbol{\Sigma})/p$ and $\mathrm{tr}(\boldsymbol{\Sigma}^2)/p$, respectively, under (R1) and (R2).

In the next Theorem 4, a proper estimator $\hat{\vartheta}$ of ϑ under the RMT regime is developed. Theorem 4 extends [18, Lemma 2.1] to the elliptical case.

Theorem 4: Let $\mathbf{x}_1, \ldots, \mathbf{x}_n$ denote an i.i.d. random sample from a p-variate elliptical distribution with finite fourth order moments, mean vector $\boldsymbol{\mu}$, and covariance matrix $\boldsymbol{\Sigma}$. Then an

unbiased estimate of $\vartheta = \mathrm{tr}(\mathbf{\Sigma}^2)/p$ for any finite n and any p is

$$\hat{\vartheta} = b_n \left(\frac{\mathrm{tr}(\mathbf{S}^2)}{p} - a_n \frac{p}{n} \left[\frac{\mathrm{tr}(\mathbf{S})}{p} \right]^2 \right),$$

where

$$a_n = \left(\frac{n}{n+\kappa} \right) \left(\frac{n}{n-1} + \kappa \right)$$

$$b_n = \frac{(\kappa+n)(n-1)^2}{(n-2)(3\kappa(n-1)+n(n+1))}.$$

Furthermore, under the RMT regime (R1) and assumption (R2), the estimator is asymptotically unbiased, i.e., $\mathbb{E}[\hat{\vartheta}] \to \vartheta_o$, where $\vartheta_o > 0$ denotes the finite limit of ϑ.

Proof: See appendix. ■

Note that a_n and b_n depend on the elliptical distribution via the elliptical kurtosis parameter κ. Using the estimate of the kurtosis by defining $\hat{a}_n = a_n(\hat{\kappa})$ and $\hat{b}_n = b_n(\hat{\kappa})$ one obtains an estimate of ϑ which does not require knowing the underlying elliptical distribution. Thus based on Theorem 4, we propose an estimator of the form

$$\hat{\gamma}^{\mathrm{Ell2}*} = \hat{b}_n \left(\frac{p\mathrm{tr}(\mathbf{S}^2)}{\mathrm{tr}(\mathbf{S})^2} - \hat{a}_n c \right). \tag{18}$$

Note that, if n is reasonably large (e.g., $n > 100$), then $\hat{a}_n \approx 1 + \hat{\kappa}$ and $b_n \approx 1$ and then one may use

$$\hat{\gamma}^{\mathrm{Ell2}*} \approx \left(\frac{p\mathrm{tr}(\mathbf{S}^2)}{\mathrm{tr}(\mathbf{S})^2} - (1+\hat{\kappa})c \right).$$

In order to guarantee that the estimator remains in the valid interval, $1 \leq \gamma \leq p$, we use

$$\hat{\gamma}^{\mathrm{Ell2}} = \min(p, \max(1, \hat{\gamma}^{\mathrm{Ell2}*})) \tag{19}$$

as our final estimator, where $\hat{\gamma}^{\mathrm{Ell2}*}$ is defined in (18). We can now define the *Ell2-RSCM estimator* as the RSCM based on the shrinkage parameter estimators $\hat{\alpha}_o^{\mathrm{Ell}}$, and $\hat{\beta}_o^{\mathrm{Ell}}$ where the estimates $\hat{\eta}$, $\hat{\kappa}$, and $\hat{\gamma}$ are obtained using (8), (15), and (19), respectively.

Finally, we wish to note that albeit $\hat{\gamma}^{\mathrm{Ell2}}$ does not require knowledge of the underlying elliptically symmetric distribution of the data, it is not a robust estimator. This is due to the fact that $\mathrm{tr}(\mathbf{S}^2)$ contains fourth order moments of the data, and the 8th order moments of the elliptically symmetric distribution need to exist in order for $\mathrm{tr}(\mathbf{S}^2)/p$ to be asymptotically normal. Consequently, the Ell2-RSCM estimator is not well suited for heavier-tailed distributions.

C. Ell3-RSCM Estimator

Ell3-RSCM is a hybrid of the Ell1-RSCM and Ell2-RSCM estimators. Ell3-RSCM uses the estimator which has a smaller estimated sphericity $\hat{\gamma}$. Thus, it will always favor more shrinkage over less shrinkage. This rule can be summarized as: if $\hat{\gamma}^{\mathrm{Ell1}} < \hat{\gamma}^{\mathrm{Ell2}}$, then choose Ell1-RSCM, otherwise choose Ell2-RSCM.

V. Simulation Study

We conduct a simulation study to investigate the performance of the RSCM estimators in terms of their finite sample NMSE.

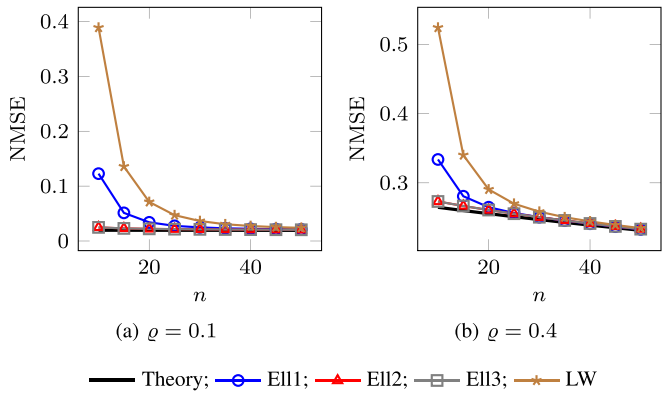


Fig. 1. AR(1) process: Comparison of covariance estimators when $p = 100$, $\varrho \in \{0.1, 0.4\}$, and the samples are drawn from a Gaussian distribution.

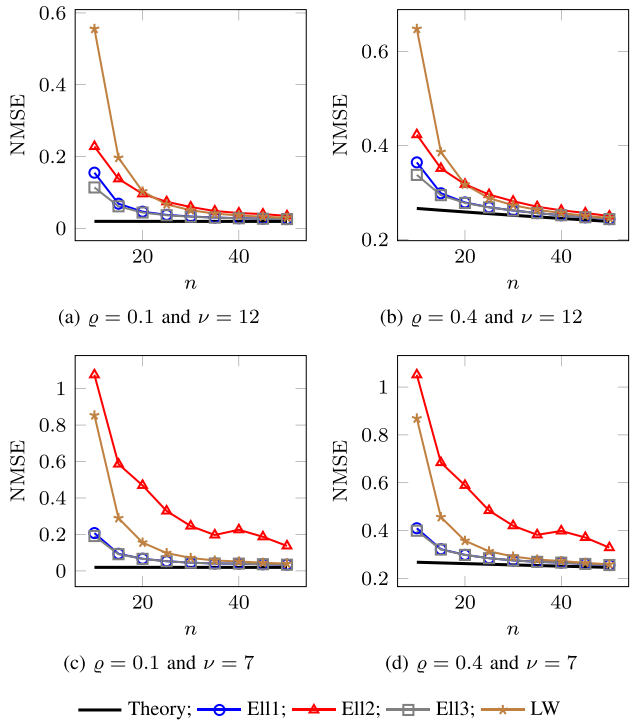


Fig. 2. AR(1) process: Comparison of covariance estimators when $p = 100$, $\varrho \in \{0.1, 0.4\}$, and the samples are drawn from a t_ν-distribution.

Each simulation is repeated 10 000 times and the NMSE is averaged over the Monte Carlo runs for each RSCM estimator. The theoretical NMSE value can be computed by normalizing (7) by $\|\mathbf{\Sigma}\|_{\mathrm{F}}^2$. The theoretical NMSE curves are depicted in Figures 1 and 2 and in Figures 4–6 using a solid black line and associated with a legend label Theory. These are then compared with the empirical NMSE values obtained with different RSCM estimators. The mean vector $\boldsymbol{\mu}$ is held fixed over the Monte Carlo trials and generated randomly as $\{\mu_i\}_{i=1}^p \overset{iid}{\sim} \mathcal{N}(0, 4)$.

A. AR(1) Covariance Matrix

In the first experiment, we consider the covariance structure of a first order autoregressive (AR) Gaussian stochastic process.

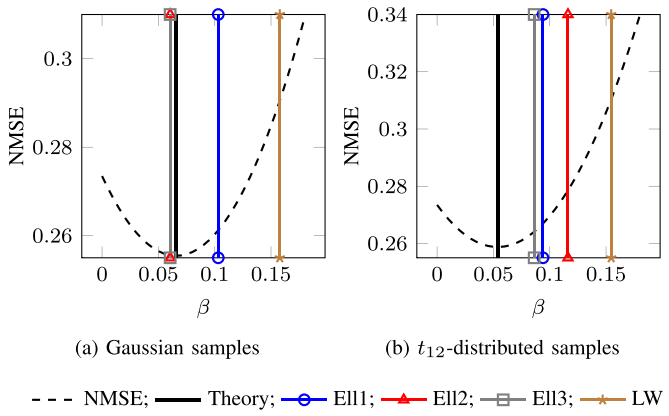


Fig. 3. The theoretical NMSE of the shrinkage estimator $\mathbf{S}_{(1-\beta)\eta,\beta}$ as a function of β when the covariance matrix has an AR(1) structure with $\varrho = 0.4$, $p = 100$ and $n = 20$. The minimum NMSE is obtained at β_o^{Ell} which is indicated by a solid vertical line. The average estimated value of the shrinkage parameter obtained by LW-, Ell1-, Ell2-, and Ell3-estimators are also indicated by vertical lines.

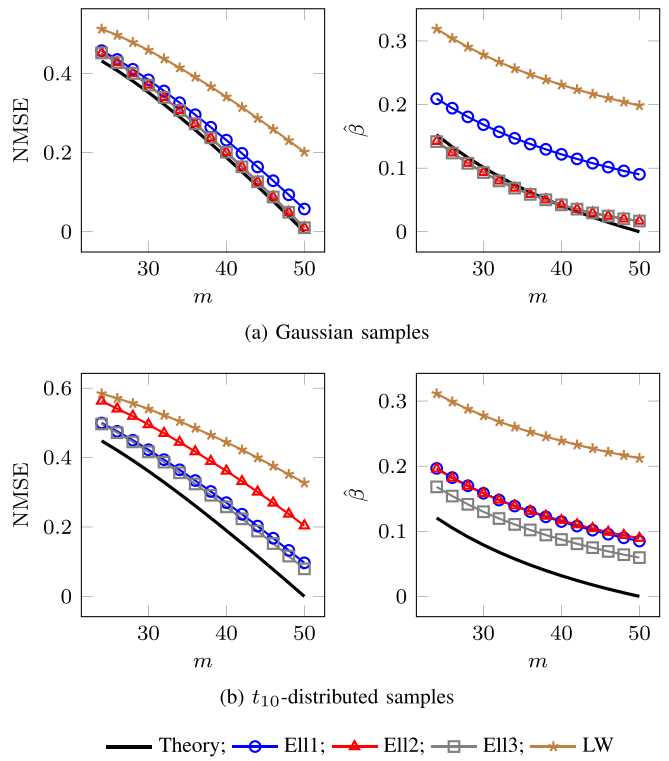


Fig. 4. The covariance matrix $\mathbf{\Sigma}$ has m eigenvalues equal to 1 and $50 - m$ eigenvalues equal to 0.01. Here $p = 50$ and $n = 10$.

Let $\mathbf{\Sigma}$ be the covariance matrix of such an AR(1) process, so that

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

Note that, $\mathbf{\Sigma}$ verifies $\eta = \text{tr}(\mathbf{\Sigma})/p = 1$. Also, when $\varrho \downarrow 0$, then $\mathbf{\Sigma}$ is close to an identity matrix, and when $\varrho \uparrow 1$, $\mathbf{\Sigma}$ tends to a singular matrix of rank 1. The dimension is fixed at $p = 100$ and n varies from 10 to 50 in steps of 5 samples. Figure 1 depicts the

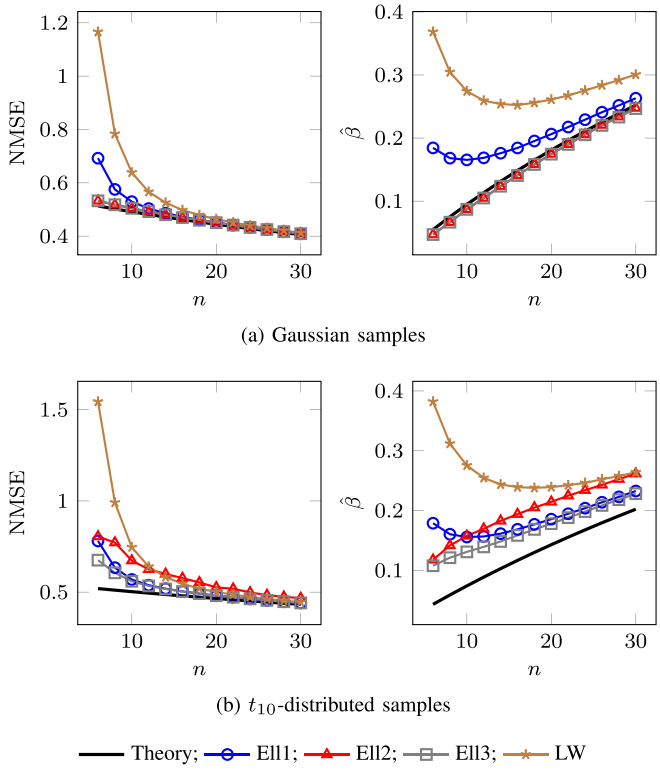


Fig. 5. The covariance matrix $\mathbf{\Sigma}$ has 30 eigenvalues equal to 100, 40 eigenvalues equal to 1, and 30 eigenvalues equal to 0.01 ($p = 100$).

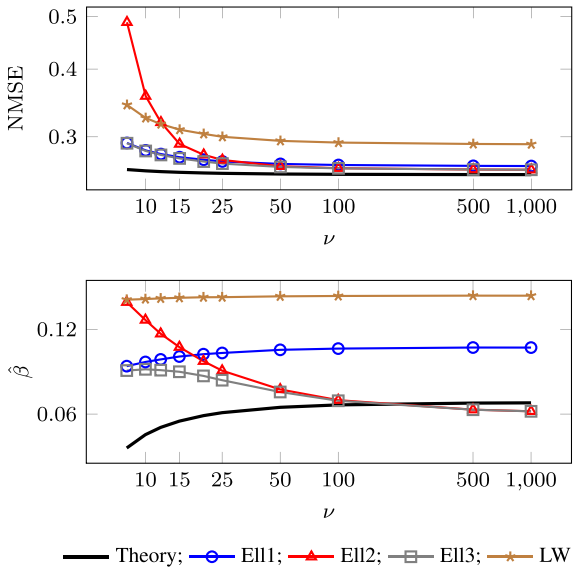


Fig. 6. AR(1) process: The NMSE and $\hat{\beta}$ as the t_ν-distribution changes with $\nu = 8$, 10, 12, 15, 25, 50, 100, 500, and 1000. Here $\varrho = 0.4$, $p = 100$ and $n = 20$.

NMSE performance as a function of the sample length n when the samples are drawn from a Gaussian distribution.

It can be noted that when the sample sizes were small, both the Ell1-RSCM estimator and the Ell2-RSCM estimator outperformed the LW-RSCM estimator with a significant margin. We also notice that the performance of the Ell2-RSCM and Ell3-RSCM estimators were almost overlapping with the

theoretical optimal value for all values of n and for both values of ϱ.

Next, we consider heavier-tailed distributions than the Gaussian. Namely, the Student's t_ν-distribution with $\nu = 12$ and $\nu = 7$ degrees of freedom; the kurtosis of the marginal variable being $\mathrm{kurt}(x_i) = 0.75$ and $\mathrm{kurt}(x_i) = 2$, respectively. The results are given in Figure 2.

First, we notice that Ell1-RSCM and Ell3-RSCM outperformed Ell2-RSCM and LW-RSCM for all values of n, ν and ϱ. In the case of $\nu = 7$, the performance of the Ell2-RSCM estimator declined due to its non-robustness, and it is performing the worst among the shrinkage estimators. In the case of $\nu = 12$, the LW-RSCM estimator and the Ell2-RSCM estimator had similar performances for larger values of n, but Ell2-RSCM performed better at small values of n. Since Ell1-RSCM and Ell2-RSCM differ only in the way they estimate the sphericity γ, the performance loss of Ell2-RSCM over Ell1-RSCM can be attributed to a larger variability and the non-robustness of the estimator $\hat{\gamma}^{\mathrm{Ell2}}$ as compared to $\hat{\gamma}^{\mathrm{Ell1}}$. Also note that when the samples were drawn from the t_7-distribution, the performance loss of LW-RSCM to Ell1-RSCM and Ell3-RSCM increased. Indeed, this difference in performance can be attributed to better robustness properties of the Ell1-RSCM estimator over the LW-RSCM estimator when sampling from a heavier-tailed elliptical distribution.

Next, for the case $\varrho = 0.4$ and $n = 20$, the dotted line in Figure 3 depicts the theoretical NMSE of $\mathbf{S}_{(1-\beta)\eta,\beta}$ as a function of β. Notice that the minimum NMSE is obtained at β_o^{Ell} which is shown by a black vertical line. The average estimates of the optimal regularization parameter β_o given by the different estimators are also indicated by vertical lines.

As can be seen, for Gaussian data, the Ell2-RSCM estimator of β_o was very close to the theoretical minimum and significantly better than the Ell1-RSCM estimator. The LW-RSCM estimator was far apart from the minimum compared to Ell1-RSCM and Ell2-RSCM. In the case of the t_ν-distribution with $\nu = 12$ degrees of freedom, the Ell1-RSCM estimator was performing better than Ell2-RSCM due to its robustness, and both were significantly closer to the minimum than the LW-RSCM estimator.

In the last AR(1) simulation study, $\varrho = 0.4$, $p = 100$, and the sample size is held constant at $n = 20$. The degrees of freedom of the t_ν-distribution of the samples is varied from $\nu = 8$ up to $\nu = 1000$. The results are shown in Figure 6. One can observe that the Ell3-RSCM estimator is able to attain the lowest empirical NMSE among all of the estimators.

B. Largely Varying Spectrum

The next study follows the set-up in [5], where $\mathbf{\Sigma}$ 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 equal to 0.01. For the case $n = 10$, Figure 4 depicts the NMSE as a function of m averaged over 10 000 Monte Carlo runs when sampling from a Gaussian distribution and a t_ν-distribution with $\nu = 10$ degrees of freedom.

In the Gaussian case, the Ell2-RSCM estimator had excellent performance as its NMSE curve is essentially overlapping with the theoretical NMSE curve. This is attested also in the right-hand side plot which depicts the graph of the average estimate $\hat{\beta}_o$ and the theoretical optimal value β_o as a function of m. As can be seen, the Ell2-RSCM estimator was essentially performing at the oracle level, whereas the shrinkage parameter corresponding to the LW-RSCM estimator was somewhat far from the theoretical optimum. The NMSE curves show that the Ell2-RSCM estimator performed better than the Ell1-RSCM estimator for Gaussian samples, however, with a rather small margin. In the case of t_{10}-distribution, as expected, Ell1-RSCM performed better than Ell2-RSCM due to its robustness in estimating the sphericity. The hybrid estimator Ell3-RSCM was able to perform slightly better than the other estimators in both cases.

The next simulation set-up considers a very challenging scenario in which the spectrum of $\mathbf{\Sigma}$ consists of several different 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 equal to 0.01. The samples were drawn from a Gaussian distribution and a t_ν-distribution with $\nu = 10$ degrees of freedom. The NMSE curves are plotted as a function of the sample length n in Figure 5.

It can be seen that under Gaussian sampling, the Ell2-RSCM and the Ell3-RSCM estimators achieved near optimal performance for all n considered. Indeed, this behavior was already seen in the other simulation studies. The more robust Ell1-RSCM estimator performed slightly worse than the Ell2-RSCM estimator in the Gaussian case for small n. It can be noticed that the performance of the LW-RSCM estimator degrades for small n. In the case when the samples are from a t_{10}-distribution, we observe that the more robust Ell1-RSCM estimator starts dominating the non-robust Ell2-RSCM estimator. Again, we note that the Ell3-RSCM estimator performed the best.

From these simulations, we can conclude that the Ell1-RSCM estimator is better suited for heavier-tailed distributions than the Ell2-RSCM estimator, which then again works well for Gaussian or close to Gaussian distributions. The Ell3-RSCM estimator is, however, able to perform the best in all of the cases. This is due to the fact that it has the freedom of choosing among two different estimates of the sphericity; and in the conducted simulations, the rule of choosing the smaller estimate of the sphericity turns out to work well. In the synthetic simulations, all three proposed estimators outperformed the LW-RSCM estimator apart from the Ell2-RSCM estimator in the case of t_7-distributed samples.

VI. Portfolio Optimization

Portfolio selection and optimization is one of the most important topics in investment theory. It is a mathematical framework wherein one seeks portfolio allocations which balance the return-risk tradeoff such that it satisfies the investor's needs. Some historical key references are [28]–[31], and [32].

Consider a portfolio consisting of p assets. The objective is to find optimal portfolio weights which determine the proportion of wealth that is to be invested in each particular stock. That is,

a fraction $w_i \in \mathbb{R}$ of the total wealth is invested in the ith asset, $i = 1, \ldots, p$, and the portfolio with p assets is described by the portfolio *weight* or *allocation vector* $\mathbf{w} \in \mathbb{R}^p$ which satisfies the constraint $\mathbf{1}^\top \mathbf{w} = 1$. The *global mean variance portfolio* (GMVP) aims at finding the weight vector that minimizes the portfolio variance (risk or volatility), and hence does not require specifying the mean vector. Let $\mathbf{r}_t \in \mathbb{R}^p$ denote the net returns of the p assets at time t. The GMVP optimization problem is

$$\underset{\mathbf{w}\in\mathbb{R}^p}{\text{minimize}}\ \mathbf{w}^\top \boldsymbol{\Sigma} \mathbf{w} \quad \text{subject to} \quad \mathbf{1}^\top \mathbf{w} = 1,$$

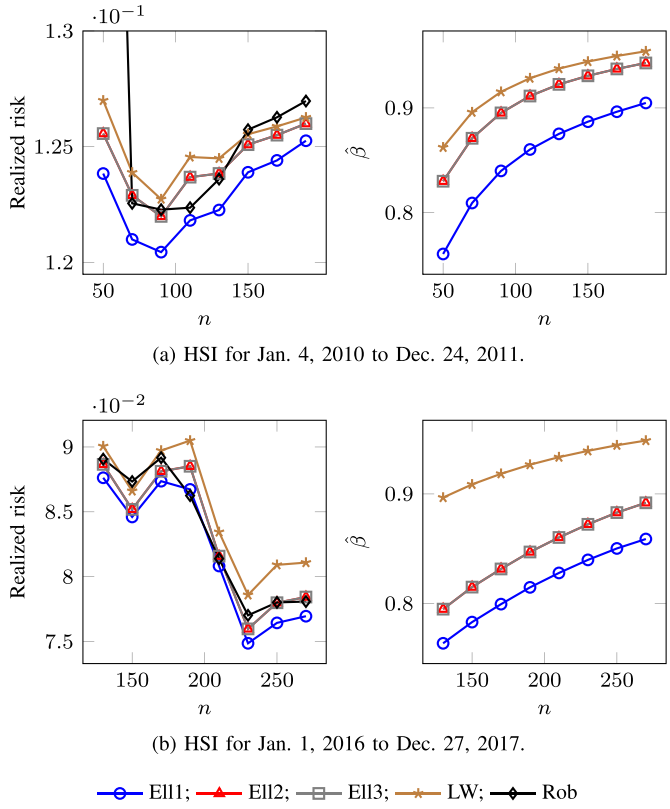
where $\mathbf{1}$ denotes is a p-vector of ones and $\boldsymbol{\Sigma}$ denotes the covariance matrix of the vector $\mathbf{r}_t$ of returns. The problem is straightforward to solve and the well-known solution is

$$\mathbf{w}_o = \frac{\boldsymbol{\Sigma}^{-1}\mathbf{1}}{\mathbf{1}^\top \boldsymbol{\Sigma}^{-1}\mathbf{1}}. \tag{20}$$

Naturally, the covariances of the net returns cannot be foreseen, and hence, the covariance matrix needs to be estimated from the historical data. Next, we apply the proposed RSCM estimators in the GMVP optimization application.

We investigate the out-of-sample portfolio performance of different RSCM estimators. In particular, we use the divident adjusted daily closing prices downloaded from the Yahoo! Finance (http://finance.yahoo.com) database to obtain the net returns for 50 stocks that are currently included in the Hang Seng Index (HSI) for two different time periods, from Jan. 4, 2010 to Dec. 24, 2011, and from Jan. 1, 2016 to Dec. 27, 2017 (excluding the weekends and public holidays). In both cases, the time series contains $T = 491$ trading days. For the first period (2010-2011), we had full length time series for only $p = 45$ stocks, whereas in the latter case we had full length time series for all stocks, so $p = 50$. Our third time series contains the net returns of $p = 396$ stocks from Standard and Poor's 500 (S&P 500) index for the time period from Jan. 4, 2016 to Apr. 27, 2018 (excluding the weekends and public holidays). In this case, the time series contains $T = 583$ trading days.

At a particular day t, we used the previous n days (i.e., from $t - n$ to $t - 1$) as the training window to estimate the covariance matrix, and the portfolio weight vector. The obtained weight vector $\hat{\mathbf{w}}_0$ was then used to compute the portfolio returns for the following 20 days. Next, the window was shifted 20 trading days forward, a new weight vector was computed, and the portfolio returns for another 20 days were computed. Hence, this scenario corresponds to the case that the portfolio manager holds the assets for approximately a month (20 trading days), after which they are liquidated and new weights are computed. In this manner, we obtained $T - n$ daily returns from which the realized risk was computed as the sample standard deviation of the obtained portfolio returns. To obtain the annualized realized risk, the sample standard deviations of the daily returns were multiplied by $\sqrt{250}$. In our tests, different training window lengths n were considered. Figure 7 depicts the annualized realized risks for the different RSCM estimators for both time periods of the HSI data. We also included in our study the robust GMVP weight estimator proposed in [27] that uses a robust regularized Tyler's M-estimator with a tuning parameter selection that is optimized



(a) HSI for Jan. 4, 2010 to Dec. 24, 2011.

(b) HSI for Jan. 1, 2016 to Dec. 27, 2017.

Fig. 7. Annualized realized portfolio risk and average $\hat{\beta}$ achieved out-of-sample for a portfolio consisting of $p = 45$ stocks in HSI for Jan. 4, 2010 to Dec. 24, 2011 (upper panel); and $p = 50$ stocks for Jan. 1, 2016 to Dec. 27, 2017 (lower panel). Both time-series contain 491 trading days. The portfolio allocations are estimated by GMVP using different RSCM estimators and different training window lengths n. The method of [27] that uses a robust regularized covariance estimator is also included and referred to as Rob.

for the GMVP problem. In [27], it was illustrated that their estimator outperforms a large array of regularized covariance matrix estimators both for simulated and real financial data.

As can be seen from Figure 7, for period 2010-2011, the Ell1-RSCM estimator achieved the smallest realized risk, outperforming all the other estimators for all window lengths. The robust method of [27] performed slightly better than the Ell2-RSCM and Ell3-RSCM estimators only for certain window lengths ($n = 70$ and $n = 110$), but it was also the worst method for a very small window length ($n = 50$). For period 2016-2017, the differences between the estimators were not as large as in the period 2010-2011. Here we observed that for some window lengths, the Ell1- and the Ell2-RSCM estimators and the robust method of [27] had rather identical behaviour (e.g., when $n = 210$). Overall, however, the Ell1-RSCM method was the best performing method.

Next, we wish to point out that while Ell1-RSCM was observed to have the best performance in general, also the Ell2-RSCM estimator outperformed the LW-RSCM over the entire span of the estimation windows considered for both periods. Also, note that the optimal training window length which yielded the smallest realized risk was $n = 90$ for the period 2010-2011, but much larger ($n = 230$) for the period 2016-2017. This could be explained by the fact that the stock markets were more

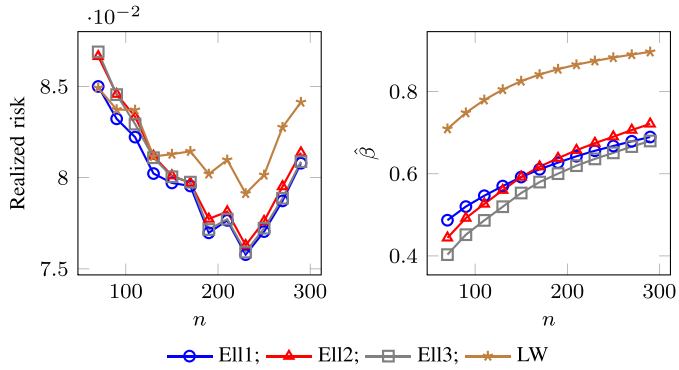


Fig. 8. Annualized realized portfolio risk achieved out-of-sample over 583 trading days for a portfolio consisting of $p = 396$ stocks in S&P 500 index for Jan. 4, 2016 to Apr. 27, 2018. The portfolio allocations are estimated by GMVP using different regularized SCM estimators and different training window lengths n. The right panel shows the average $\hat{\beta}$ of the RSCM estimators for different training window lengths.

turbulent in the first period, and hence, the realized risks were much higher.

Figure 8 depicts the annualized realized risks for the different RSCM estimators for the time period from Jan. 4, 2016 to Apr. 27, 2018 of the S&P 500 data. We have excluded the method of [27] from this study as it is not well suited for very high-dimensional problems because of its large computational cost due to the grid search method it uses in finding the optimal tuning parameter.

With the S&P 500 data, Ell1-RSCM achieved the smallest realized risk and outperformed the other estimators for all training window lengths n. The Ell2-RSCM estimator outperformed LW-RSCM when $n \geq 130$. The Ell3-RSCM estimator had similar performance as Ell2-RSCM when $n \leq 170$ and it performed similar to Ell1-RSCM for $n > 170$. The optimal training window length which produced the smallest realized risk was $n = 230$ for all methods. Note that, the same result was achieved with the HSI data for the period 2016-2017.

VII. Conclusion

This paper proposed a regularized sample covariance matrix (RSCM) estimator Ell-RSCM, which is suitable for high-dimensional problems, where the data can be considered as generated from an unknown elliptically symmetric distribution. The proposed estimator is based on the estimation of the optimal shrinkage parameters which minimize the mean squared error. The estimation of the optimal shrinkage parameters was shown to reduce to a simpler problem of estimating three statistical population parameters: the scale η, the sphericity γ, and the elliptical kurtosis κ. The paper showed alternative ways of how to estimate these parameters under the random matrix theory regime. In the construction of the proposed estimator Ell-RSCM, elliptical distribution theory was used in the derivation of the analytical form of the mean squared error of the SCM. The conducted synthetic simulation studies showed an advantage of using the proposed Ell-RSCM estimator over the widely popular Ledoit-Wolf (LW-RSCM) estimator. Furthermore, we tested the performance of the proposed Ell-RSCM estimator using real data in a portfolio optimization problem, wherein the proposed methods were able to outperform the benchmark methods. MATLAB codes of the proposed Ell-RSCM methods and the datasets used in Section VI are available at http://users.spa.aalto.fi/esollila/regscm/ along with an additional example of supervised classification of real data, where the proposed RSCM estimators are used in the regularized discriminant analysis (RDA) classification framework [10].

Appendix

A. Proof of Theorem 1

Write $L(\alpha, \beta) = \mathbb{E}\left[\|\beta\mathbf{S} + \alpha\mathbf{I} - \boldsymbol{\Sigma}\|_{\mathrm{F}}^2\right]$ and note that

$$L(\alpha, \beta) \tag{21}$$

$$= \mathbb{E}\left[\left\|\alpha\mathbf{I} + \beta(\mathbf{S} - \boldsymbol{\Sigma}) - (1-\beta)\boldsymbol{\Sigma}\right\|_{\mathrm{F}}^2\right]$$

$$= \alpha^2 p + \beta^2 a_1 + (1-\beta)^2 \mathrm{tr}(\boldsymbol{\Sigma}^2) - 2\alpha(1-\beta)\mathrm{tr}(\boldsymbol{\Sigma}), \tag{22}$$

where $a_1 = \mathrm{MSE}(\mathbf{S}) = \mathbb{E}[\|\mathbf{S} - \boldsymbol{\Sigma}\|_{\mathrm{F}}^2]$. Here, we used the fact that $\mathbb{E}[\mathrm{tr}(\mathbf{S} - \boldsymbol{\Sigma})]$ as well as $\mathbb{E}[\mathrm{tr}\{(\mathbf{S} - \boldsymbol{\Sigma})\boldsymbol{\Sigma}\}]$ vanish as $\mathbf{S}$ is unbiased, i.e., $\mathbb{E}[\mathbf{S}] = \boldsymbol{\Sigma}$. The Hessian matrix of the quadratic objective function $L(\alpha, \beta)$ is

$$\nabla^2 L = 2\begin{pmatrix} p & \mathrm{tr}(\boldsymbol{\Sigma}) \\ \mathrm{tr}(\boldsymbol{\Sigma}) & a_1 + \mathrm{tr}(\boldsymbol{\Sigma}^2) \end{pmatrix} = 2\begin{pmatrix} p & \mathrm{tr}(\boldsymbol{\Sigma}) \\ \mathrm{tr}(\boldsymbol{\Sigma}) & \mathbb{E}[\mathrm{tr}(\mathbf{S}^2)] \end{pmatrix}, \tag{23}$$

where we used that

$$\begin{aligned} a_1 = \mathrm{MSE}(\mathbf{S}) &= \mathbb{E}\left[\|\mathbf{S} - \boldsymbol{\Sigma}\|_{\mathrm{F}}^2\right] \\ &= \mathbb{E}\left[\mathrm{tr}\left(\mathbf{S}^2\right)\right] - 2\mathbb{E}\left[\mathrm{tr}\left(\mathbf{S}\boldsymbol{\Sigma}\right)\right] + \mathrm{tr}\left(\boldsymbol{\Sigma}^2\right) \\ &= \mathbb{E}\left[\mathrm{tr}\left(\mathbf{S}^2\right)\right] - \mathrm{tr}\left(\boldsymbol{\Sigma}^2\right). \end{aligned} \tag{24}$$

The last identity follows from $\mathbb{E}[\mathrm{tr}(\mathbf{S}\boldsymbol{\Sigma})] = \mathrm{tr}(\mathbb{E}[\mathbf{S}]\boldsymbol{\Sigma}) = \mathrm{tr}(\boldsymbol{\Sigma}^2)$. The positive definiteness of the Hessian matrix can be observed by noting that $\nabla^2 L \succ 0$ if and only if $pa_1 + p\mathrm{tr}(\boldsymbol{\Sigma}^2) - \mathrm{tr}(\boldsymbol{\Sigma})^2 > 0$. Indeed, this is true, given that $a_1 = \mathrm{MSE}(\mathbf{S}) \neq 0$, since $p\mathrm{tr}(\boldsymbol{\Sigma}^2) \geq \mathrm{tr}(\boldsymbol{\Sigma})^2$ (i.e., $\gamma \geq 1$). It follows that the objective function L is strictly convex and a unique minimum can be found by solving the stationary point.

The optimal solution for α_0 is found by solving $\frac{\partial}{\partial\alpha}L(\alpha, \beta) = 0$ which yields $\alpha_o = (1 - \beta_o)\mathrm{tr}(\boldsymbol{\Sigma})/p$. Substituting the optimum in place of α in (22) and solving for β_o yields

$$\beta_o = \frac{\|\boldsymbol{\Sigma} - \eta\mathbf{I}\|_{\mathrm{F}}^2}{\|\boldsymbol{\Sigma} - \eta\mathbf{I}\|_{\mathrm{F}}^2 + \mathrm{MSE}(\mathbf{S})}. \tag{25}$$

Note that, $\beta_o = 1$ only if $\mathbf{S} = \boldsymbol{\Sigma}$, and thus $\mathrm{MSE}(\mathbf{S}) = 0$, which has zero probability when sampling from a continuous distribution. The form of β_o in (25) therefore implies that $\beta_o \in [0, 1)$. We now show that (25) can be expressed in the form (4).

The numerator of β_o in (25) is

$$\begin{aligned} a_2 = \|\boldsymbol{\Sigma} - \eta\mathbf{I}\|_{\mathrm{F}}^2 &= \mathrm{tr}(\boldsymbol{\Sigma}^2) - (1/p)\mathrm{tr}(\boldsymbol{\Sigma})^2 \\ &= p(\vartheta - \eta^2) = p(\gamma - 1)\eta^2, \end{aligned} \tag{26}$$

where we denote $\vartheta = \mathrm{tr}(\boldsymbol{\Sigma}^2)/p$. This shows that the denominator of β_o is $a_1 + a_2 = \mathbb{E}\left[\mathrm{tr}\left(\mathbf{S}^2\right)\right] - (1/p)\mathrm{tr}(\boldsymbol{\Sigma})^2 = \mathbb{E}\left[\mathrm{tr}\left(\mathbf{S}^2\right)\right] - p\eta^2$. These expressions for the numerator and the

denominator of β_o yield the assertion (4) for β_o. Substituting (26) into (25) and multiplying both the numerator and denominator by $1/(p\eta^2)$ gives (5).

Next, we derive the expression for the MSE of the RSCM $\mathbf{S}_{\alpha,\beta}$. By using the variance and bias decomposition of the MSE, we have

$$\begin{aligned}\mathrm{MSE}(\mathbf{S}_{\alpha,\beta}) &= \mathrm{tr}(\mathrm{var}(\mathrm{vec}\,(\mathbf{S}_{\alpha,\beta}))) + \|\mathbb{E}[\mathbf{S}_{\alpha,\beta}] - \boldsymbol{\Sigma}\|_{\mathrm{F}}^2 \\ &= \beta^2 \mathrm{tr}(\mathrm{var}(\mathrm{vec}\,(\mathbf{S}))) + \|\alpha\mathbf{I} - (1-\beta)\boldsymbol{\Sigma}\|_{\mathrm{F}}^2 \\ &= \beta^2 \mathrm{MSE}(\mathbf{S}) + \|\alpha\mathbf{I} - (1-\beta)\boldsymbol{\Sigma}\|_{\mathrm{F}}^2.\end{aligned}$$

We used the fact that from the unbiasedness of $\mathbf{S}$ it follows that $\mathrm{MSE}(\mathbf{S}) = \mathrm{tr}(\mathrm{var}(\mathrm{vec}\,(\mathbf{S}))) = a_1$. At the optimum, we have $\beta_o a_1 = (1-\beta_o)a_2$, which can be seen from (25), and $\alpha_o = (1-\beta_o)\eta$. The MSE at the optimum is therefore

$$\begin{aligned}\mathrm{MSE}(\mathbf{S}_{(1-\beta_o)\eta,\beta_o}) &= \beta_o^2 \mathrm{MSE}(\mathbf{S}) + (1-\beta_o)^2 \|\boldsymbol{\Sigma} - \eta\mathbf{I}\|_{\mathrm{F}}^2 \\ &= \beta_o(1-\beta_o)a_2 + (1-\beta_o)^2 a_2 \\ &= (1-\beta_o)a_2,\end{aligned}$$

which concludes the proof.

B. Proof of Theorem 2

We will use the following matrix decomposition in our proof. Let $\mathbf{X} = (\mathbf{x}_1 \cdots \mathbf{x}_n)^\top$ denote the $n \times p$ data matrix with the ith transposed observation as its row vector. Then the SCM can be written as

$$\mathbf{S} = \frac{1}{n-1}\mathbf{X}^\top \mathbf{H}\mathbf{X},$$

where $\mathbf{H}$ is the centering matrix.

For elliptically distributed observations $\{\mathbf{x}_i\}_{i=1}^n \overset{iid}{\sim} \mathcal{E}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}, g)$, we have the following stochastic decomposition $\mathbf{x}_i =_d \boldsymbol{\Sigma}^{1/2}\mathbf{z}_i + \boldsymbol{\mu}$, where $\mathbf{z}_i \sim \mathcal{E}_p(\mathbf{0}, \mathbf{I}, g)$. Let $\mathbf{Z} = (\mathbf{z}_1 \cdots \mathbf{z}_n)^\top$ denote the $n \times p$ data matrix collecting the random vectors $\mathbf{z}_i$ as its row vectors. Then the stochastic decomposition implies that

$$\mathbf{X}^\top \mathbf{H}\mathbf{X} =_d \boldsymbol{\Sigma}^{1/2}\mathbf{Z}^\top \mathbf{H}\mathbf{Z}\boldsymbol{\Sigma}^{1/2}.$$

Hence,

$$\begin{aligned}&\mathrm{var}\,(\mathrm{vec}\,(\mathbf{S})) = \mathrm{var}\left(\frac{1}{n-1}\mathrm{vec}\left(\boldsymbol{\Sigma}^{1/2}\mathbf{Z}^\top\mathbf{H}\mathbf{Z}\boldsymbol{\Sigma}^{1/2}\right)\right) \\ &= (\boldsymbol{\Sigma}^{1/2} \otimes \boldsymbol{\Sigma}^{1/2})\,\mathrm{var}\left(\frac{1}{n-1}\mathrm{vec}\left(\mathbf{Z}^\top\mathbf{H}\mathbf{Z}\right)\right)(\boldsymbol{\Sigma}^{1/2} \otimes \boldsymbol{\Sigma}^{1/2}).\end{aligned} \tag{27}$$

Since the matrix $\mathbf{Z}^\top\mathbf{H}\mathbf{Z}$ is radially distributed, we can apply [33, Theorem 1], which states

$$\mathrm{var}\left(\frac{1}{n-1}\mathrm{vec}\left(\mathbf{Z}^\top\mathbf{H}\mathbf{Z}\right)\right) = \tau_1(\mathbf{I} + \mathbf{K}_p) + \tau_2 \mathrm{vec}\,(\mathbf{I})\,\mathrm{vec}\,(\mathbf{I})^\top, \tag{28}$$

where the parameters τ_1 and τ_2 correspond to the variance of any off-diagonal element and the covariance of any two diagonal elements of the matrix $\frac{1}{n-1}\mathbf{Z}^\top\mathbf{H}\mathbf{Z}$, respectively.

We will first derive the expression for τ_1. For $q \neq r$, it holds that

$$\begin{aligned}(n-1)^2\tau_1 &= \mathrm{var}\left((\mathbf{Z}\mathbf{e}_q)^\top\mathbf{H}(\mathbf{Z}\mathbf{e}_r)\right) \\ &= \mathrm{var}\left(\mathrm{tr}\left(\mathbf{H}(\mathbf{Z}\mathbf{e}_r)(\mathbf{Z}\mathbf{e}_q)^\top\right)\right) \\ &= \mathrm{var}\left(\mathrm{vec}\,(\mathbf{H})^\top \mathrm{vec}\left((\mathbf{Z}\mathbf{e}_r)(\mathbf{Z}\mathbf{e}_q)^\top\right)\right) \\ &= \mathrm{vec}\,(\mathbf{H})^\top \mathrm{var}\left(\mathrm{vec}\left((\mathbf{Z}\mathbf{e}_r)(\mathbf{Z}\mathbf{e}_q)^\top\right)\right)\mathrm{vec}\,(\mathbf{H}).\end{aligned}$$

Next, we recall that $\mathbf{z}_i \sim \mathcal{E}_p(\mathbf{0}, \mathbf{I}, g)$ has a stochastic representation (*cf.* [15, Theorem 2.9]) $\mathbf{z}_i =_d r_i\mathbf{u}_i$, where r_i is the generating variate with a density $f(r) = C \cdot r^{p-1}g(r^2)$ (where C is a normalizing constant) and $\mathbf{u}_i = (u_{i1}, u_{i2}, \ldots, u_{ip})^\top$ is uniformly distributed on the unit hypersphere $\mathcal{S}^{p-1} = \{\mathbf{x} \in \mathbb{R}^p : \mathbf{x}^\top\mathbf{x} = 1\}$ and r_i is independent of $\mathbf{u}_i$. Using this stochastic representation for $\mathbf{z}_i$, we can write $\mathbf{Z}\mathbf{e}_q = (r_1u_{1q}, r_2u_{2q}, \ldots, r_nu_{nq})^\top$. The klth element of the ijth block (i.e., the $ijkl$th element) of the $n^2 \times n^2$ matrix $\mathrm{var}\left(\mathrm{vec}\left((\mathbf{Z}\mathbf{e}_r)(\mathbf{Z}\mathbf{e}_q)^\top\right)\right)$ can then be written as

$$\begin{aligned}&\mathrm{cov}\left((\mathbf{Z}\mathbf{e}_r)_k(\mathbf{Z}\mathbf{e}_q)_i, (\mathbf{Z}\mathbf{e}_r)_l(\mathbf{Z}\mathbf{e}_q)_j\right) \\ &= \mathbb{E}\left[r_ku_{kr} \cdot r_iu_{iq} \cdot r_lu_{lr} \cdot r_ju_{jq}\right] \\ &\quad - \mathbb{E}\left[r_ku_{kr} \cdot r_iu_{iq}\right]\mathbb{E}\left[r_lu_{lr} \cdot r_ju_{jq}\right].\end{aligned}$$

Using the following identities for $\forall i, j$ and $q \neq r$ (*cf.* [15, Section 3.1]) :

$$\begin{aligned}\mathbb{E}\left[u_{iq}u_{jr}\right] &= 0, & \mathbb{E}\left[u_{iq}^2u_{ir}^2\right] &= \frac{1}{p(p+2)}, \\ \mathbb{E}\left[u_{iq}^2\right] &= \frac{1}{p}, & \mathbb{E}\left[u_{iq}^4\right] &= \frac{3}{p(p+2)}, \\ \mathbb{E}\left[r_i^2\right] &= p, & \mathbb{E}\left[r_i^4\right] &= (1+\kappa)p(p+2),\end{aligned}$$

where $3\kappa = \mathrm{kurt}(z_{iq}) = \mathrm{kurt}(x_{iq})$, we find that the only nonzero elements of $\mathrm{var}\left(\mathrm{vec}\left((\mathbf{Z}\mathbf{e}_r)(\mathbf{Z}\mathbf{e}_q)^\top\right)\right)$ correspond to

$$\begin{aligned}\mathbb{E}[r_i^4]\mathbb{E}[u_{ir}^2u_{iq}^2] &= 1+\kappa \quad \text{for } i=j=k=l, \text{ and} \\ \mathbb{E}[r_i^2]\mathbb{E}[r_k^2]\mathbb{E}[u_{ir}^2]\mathbb{E}[u_{kq}^2] &= 1 \qquad \text{for } i=j, k=l, i \neq k.\end{aligned}$$

This implies that

$$\mathrm{var}\left(\mathrm{vec}\left((\mathbf{Z}\mathbf{e}_r)(\mathbf{Z}\mathbf{e}_q)^\top\right)\right) = \mathbf{I} + \kappa\sum_{i=1}^n \mathbf{e}_i\mathbf{e}_i^\top \otimes \mathbf{e}_i\mathbf{e}_i^\top. \tag{29}$$

Hence, we can write τ_1 as

$$\begin{aligned}\tau_1 &= \frac{1}{(n-1)^2}\mathrm{vec}\,(\mathbf{H})^\top\left(\mathbf{I} + \kappa\sum_{i=1}^n \mathbf{e}_i\mathbf{e}_i^\top \otimes \mathbf{e}_i\mathbf{e}_i^\top\right)\mathrm{vec}\,(\mathbf{H}) \\ &= \frac{1}{n-1} + \frac{\kappa}{n},\end{aligned} \tag{30}$$

where we used $\mathrm{vec}\,(\mathbf{H})^\top\,\mathrm{vec}\,(\mathbf{H}) = n-1$ and

$$\sum_{i=1}^n \mathrm{vec}\,(\mathbf{H})^\top(\mathbf{e}_i\mathbf{e}_i^\top \otimes \mathbf{e}_i\mathbf{e}_i^\top)\mathrm{vec}\,(\mathbf{H}) = \sum_{i=1}^n h_{ii}^2 = \frac{(n-1)^2}{n}.$$

Next, we find the expression for τ_2. For $q \neq r$, we have

$$\begin{aligned}(n-1)^2\tau_2 &= \operatorname{cov}\left((\mathbf{Z}\mathbf{e}_q)^\top \mathbf{H}(\mathbf{Z}\mathbf{e}_q), (\mathbf{Z}\mathbf{e}_r)^\top \mathbf{H}(\mathbf{Z}\mathbf{e}_r)\right)\\ &= \mathbb{E}\left[(\mathbf{Z}\mathbf{e}_q)^\top \mathbf{H}(\mathbf{Z}\mathbf{e}_q)(\mathbf{Z}\mathbf{e}_r)^\top \mathbf{H}(\mathbf{Z}\mathbf{e}_r)\right]\\ &\quad - \mathbb{E}[(\mathbf{Z}\mathbf{e}_q)^\top \mathbf{H}(\mathbf{Z}\mathbf{e}_q)]\mathbb{E}[(\mathbf{Z}\mathbf{e}_r)^\top \mathbf{H}(\mathbf{Z}\mathbf{e}_r)].\end{aligned}$$

By using basic algebraic properties of the trace and the vectorization transform and noting that $\mathbb{E}[(\mathbf{Z}\mathbf{e}_r)(\mathbf{Z}\mathbf{e}_r)^\top] = \mathbf{I}$, we arrive at the form

$$\begin{aligned}&(n-1)^2\tau_2 =\\ &\operatorname{tr}\left((\mathbf{H}\otimes\mathbf{H})\mathbb{E}\left[\operatorname{vec}\left((\mathbf{Z}\mathbf{e}_q)(\mathbf{Z}\mathbf{e}_r)^\top\right)\operatorname{vec}\left((\mathbf{Z}\mathbf{e}_q)(\mathbf{Z}\mathbf{e}_r)^\top\right)^\top\right]\right)\\ &\quad - \operatorname{tr}(\mathbf{H})^2.\end{aligned}$$

The expression involving the expectation is equal to $\operatorname{var}\left(\operatorname{vec}\left((\mathbf{Z}\mathbf{e}_q)(\mathbf{Z}\mathbf{e}_r)^\top\right)\right)$, which implies

$$\begin{aligned}&(n-1)^2\tau_2\\ &\quad = \operatorname{tr}\left\{(\mathbf{H}\otimes\mathbf{H})\left(\mathbf{I} + \kappa\sum_{i=1}^n \mathbf{e}_i\mathbf{e}_i^\top \otimes \mathbf{e}_i\mathbf{e}_i^\top\right)\right\} - \operatorname{tr}(\mathbf{H})^2.\end{aligned}$$

By noting that $\operatorname{tr}(\mathbf{H}\otimes\mathbf{H}) = \operatorname{tr}(\mathbf{H})^2$ and

$$\sum_{i=1}^n \operatorname{tr}\left((\mathbf{H}\otimes\mathbf{H})(\mathbf{e}_i\mathbf{e}_i^\top \otimes \mathbf{e}_i\mathbf{e}_i^\top)\right) = \sum_{i=1}^n h_{ii}^2,$$

we find that

$$\tau_2 = \frac{1}{(n-1)^2}\kappa\frac{(n-1)^2}{n} = \frac{\kappa}{n}. \tag{31}$$

By substituting (28) into (27), and noticing that

$$(\mathbf{\Sigma}^{1/2}\otimes\mathbf{\Sigma}^{1/2})\operatorname{vec}(\mathbf{I}) = \operatorname{vec}(\mathbf{\Sigma}) \text{ and}$$

$$(\mathbf{\Sigma}^{1/2}\otimes\mathbf{\Sigma}^{1/2})(\mathbf{I}+\mathbf{K}_p)(\mathbf{\Sigma}^{1/2}\otimes\mathbf{\Sigma}^{1/2}) = (\mathbf{I}+\mathbf{K}_p)(\mathbf{\Sigma}\otimes\mathbf{\Sigma}),$$

completes the proof.

C. Proof of Lemma 1

To obtain the stated expression of the $\mathrm{MSE}(\mathbf{S})$, we note that since $\mathbf{S}$ is unbiased, i.e., $\mathbb{E}[\mathbf{S}] = \mathbf{\Sigma}$, it holds that

$$\begin{aligned}\mathrm{MSE}(\mathbf{S}) &= \mathbb{E}\left[\|\mathbf{S}-\mathbf{\Sigma}\|_{\mathrm{F}}^2\right]\\ &= \mathbb{E}\left[\operatorname{vec}(\mathbf{S}-\mathbf{\Sigma})^\top\operatorname{vec}(\mathbf{S}-\mathbf{\Sigma})\right]\\ &= \operatorname{tr}\left(\mathbb{E}\left[\operatorname{vec}(\mathbf{S}-\mathbb{E}[\mathbf{S}])\operatorname{vec}(\mathbf{S}-\mathbb{E}[\mathbf{S}])^\top\right]\right)\\ &= \operatorname{tr}\left(\operatorname{var}(\operatorname{vec}(\mathbf{S}))\right).\end{aligned} \tag{32}$$

Then we substitute the expression stated in (11) for $\operatorname{var}(\operatorname{vec}(\mathbf{S}))$ into equation (32) and use the following identities

$$\begin{aligned}\operatorname{tr}(\mathbf{\Sigma}\otimes\mathbf{\Sigma}) &= \operatorname{tr}(\mathbf{\Sigma})^2,\\ \operatorname{tr}\left(\operatorname{vec}(\mathbf{\Sigma})\operatorname{vec}(\mathbf{\Sigma})^\top\right) &= \operatorname{tr}(\mathbf{\Sigma}^2), \text{ and}\\ \operatorname{tr}\big(\mathbf{K}_p(\mathbf{\Sigma}\otimes\mathbf{\Sigma})\big) &= \operatorname{tr}(\mathbf{\Sigma}^2),\end{aligned}$$

where the last identity follows from

$$\begin{aligned}\operatorname{tr}\big(\mathbf{K}_p(\mathbf{\Sigma}\otimes\mathbf{\Sigma})\big) &= \sum_{i,j}\operatorname{tr}\big(\left(\mathbf{e}_i\mathbf{e}_j^\top \otimes \mathbf{e}_j\mathbf{e}_i^\top\right)(\mathbf{\Sigma}\otimes\mathbf{\Sigma})\big)\\ &= \sum_{i,j}\operatorname{tr}\left(\mathbf{e}_j^\top\mathbf{\Sigma}\mathbf{e}_i\cdot\mathbf{e}_i^\top\mathbf{\Sigma}\mathbf{e}_j\right)\\ &= \operatorname{tr}(\mathbf{\Sigma}^2).\end{aligned}$$

The expression for the $\mathrm{NMSE}(\mathbf{S})$ is obtained by dividing the $\mathrm{MSE}(\mathbf{S})$ by $\|\mathbf{\Sigma}\|_{\mathrm{F}}^2 = \operatorname{tr}(\mathbf{\Sigma}^2)$.

D. Proof of Lemma 2

The first statement for $\mathbb{E}[\operatorname{tr}(\mathbf{S}^2)]$ follows from Lemma 1 by noting that $\mathbb{E}[\operatorname{tr}(\mathbf{S}^2)] = \mathrm{MSE}(\mathbf{S}) + \operatorname{tr}(\mathbf{\Sigma}^2)$, which was shown in (24).

Regarding the second statement, we write

$$\begin{aligned}\mathbb{E}\left[\operatorname{tr}(\mathbf{S})^2\right] &= \mathbb{E}\Big[\sum_i s_{ii}\sum_j s_{jj}\Big] = \sum_{i,j}\mathbb{E}\left[s_{ii}s_{jj}\right]\\ &= \sum_{i,j}\left(\operatorname{cov}\left(s_{ii},s_{jj}\right) + \mathbb{E}\left[s_{ii}\right]\mathbb{E}\left[s_{jj}\right]\right).\end{aligned}$$

Here, the covariance of s_{ii} and s_{jj} is the ijth element of the ijth block of $\operatorname{var}\left(\operatorname{vec}\left(\mathbf{S}\right)\right)$ in (11) since

$$\begin{aligned}\operatorname{cov}(s_{ii},s_{jj}) &= \operatorname{cov}(\mathbf{e}_i^\top\mathbf{S}\mathbf{e}_i, \mathbf{e}_j^\top\mathbf{S}\mathbf{e}_j)\\ &= \operatorname{cov}\left((\mathbf{e}_i\otimes\mathbf{e}_i)^\top\operatorname{vec}(\mathbf{S}), (\mathbf{e}_j\otimes\mathbf{e}_j)^\top\operatorname{vec}(\mathbf{S})\right)\\ &= (\mathbf{e}_i^\top\otimes\mathbf{e}_i^\top)\operatorname{var}(\operatorname{vec}\left(\mathbf{S}\right))(\mathbf{e}_j\otimes\mathbf{e}_j).\end{aligned}$$

Using the following identities:

$$\begin{aligned}(\mathbf{e}_i^\top\otimes\mathbf{e}_i^\top)(\mathbf{\Sigma}\otimes\mathbf{\Sigma})(\mathbf{e}_j\otimes\mathbf{e}_j) &= \mathbf{e}_i^\top\mathbf{\Sigma}\mathbf{e}_j\cdot\mathbf{e}_i^\top\mathbf{\Sigma}\mathbf{e}_j,\\ (\mathbf{e}_i^\top\otimes\mathbf{e}_i^\top)\mathbf{K}_p(\mathbf{\Sigma}\otimes\mathbf{\Sigma})(\mathbf{e}_j\otimes\mathbf{e}_j) &= \mathbf{e}_i^\top\mathbf{\Sigma}\mathbf{e}_j\cdot\mathbf{e}_i^\top\mathbf{\Sigma}\mathbf{e}_j,\\ (\mathbf{e}_i^\top\otimes\mathbf{e}_i^\top)\operatorname{vec}\left(\mathbf{\Sigma}\right)\operatorname{vec}\left(\mathbf{\Sigma}\right)^\top(\mathbf{e}_j\otimes\mathbf{e}_j) &= \mathbf{e}_i^\top\mathbf{\Sigma}\mathbf{e}_i\cdot\mathbf{e}_j^\top\mathbf{\Sigma}\mathbf{e}_j,\end{aligned}$$

and the fact that $\mathbf{S}$ is unbiased, i.e., $\mathbb{E}[s_{ii}] = \mathbf{e}_i^\top\mathbf{\Sigma}\mathbf{e}_i$, we can write

$$\begin{aligned}\mathbb{E}[s_{ii}s_{jj}] &= \operatorname{cov}(s_{ii},s_{jj}) + \mathbb{E}[s_{ii}]\mathbb{E}[s_{jj}]\\ &= 2\tau_1(\mathbf{e}_i^\top\mathbf{\Sigma}\mathbf{e}_j\cdot\mathbf{e}_i^\top\mathbf{\Sigma}\mathbf{e}_j) + (1+\tau_2)(\mathbf{e}_i^\top\mathbf{\Sigma}\mathbf{e}_i\cdot\mathbf{e}_j^\top\mathbf{\Sigma}\mathbf{e}_j),\end{aligned}$$

where τ_1 and τ_2 are given in (30) and (31), respectively. By summing all i and j, we have

$$\mathbb{E}[\operatorname{tr}(\mathbf{S})^2] = \sum_{i,j}\mathbb{E}[s_{ii}s_{jj}] = 2\tau_1\operatorname{tr}(\mathbf{\Sigma}^2) + (1+\tau_2)\operatorname{tr}(\mathbf{\Sigma})^2,$$

which completes the proof.

E. Proof of Theorem 4

Using Lemma 2, write

$$\begin{aligned}&b_n^{-1}p\mathbb{E}[\hat{\vartheta}] =\\ &\left(\tau_1 - \frac{a_n}{n}(1+\tau_2)\right)\operatorname{tr}(\mathbf{\Sigma})^2 + \left(1+\tau_1+\tau_2 - 2\tau_1\frac{a_n}{n}\right)\operatorname{tr}(\mathbf{\Sigma}^2),\end{aligned}$$

where τ_1 and τ_2 are defined in (30) and (31). By choosing $a_n = n\tau_1/(1+\tau_2)$, we see that $b_n = \left(1+\tau_1+\tau_2-2\tau_1 a_n/n\right)^{-1}$.

The terms a_n and b_n can equivalently be expressed in the form given in the theorem by using the equations for τ_1 and τ_2. The last statement is a consequence of the fact that ϑ converges to a finite limit value as $p \to \infty$ and that $\mathbb{E}[\hat{\vartheta}] = \vartheta$.

REFERENCES

[1] O. Ledoit *et al.*, "Some hypothesis tests for the covariance matrix when the dimension is large compared to the sample size," *Ann. Statist.*, vol. 30, no. 4, pp. 1081–1102, 2002.

[2] O. Ledoit and M. Wolf, "A well-conditioned estimator for large-dimensional covariance matrices," *J. Multivariate Anal.*, vol. 88, pp. 365–411, 2004.

[3] 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.*, 2007, pp. III-1105–III-1108.

[4] E. Ollila and D. E. Tyler, "Regularized M-estimators of scatter matrix," *IEEE Trans. Signal Process.*, vol. 62, no. 22, pp. 6059–6070, Nov. 2014.

[5] T. Zhang and A. Wiesel, "Automatic diagonal loading for Tyler's robust covariance estimator," in *Proc. IEEE Workshop Statistical Signal Process.*, 2016, pp. 1–5.

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

[7] 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, Oct. 2010.

[8] R. Couillet and M. McKay, "Large dimensional analysis and optimization of robust shrinkage covariance matrix estimators," *J. Multivariate Anal.*, vol. 131, pp. 99–120, 2014.

[9] 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, Nov. 2014.

[10] J. H. Friedman, "Regularized discriminant analysis," *J. Amer. Statistical Assoc.*, vol. 84, no. 405, pp. 165–175, 1989.

[11] B. D. Carlson, "Covariance matrix estimation errors and diagonal loading in adaptive arrays," *IEEE Trans. Aerosp. Electron. Syst.*, vol. 24, no. 4, pp. 397–401, Jul. 1988.

[12] Y. Abramovich, "Controlled method for adaptive optimization of filters using the criterion of maximum signal-to-noise ratio," *Radio Eng. Elect. Phys.*, vol. 26, no. 3, pp. 87–95, 1981.

[13] N. Auguin, D. Morales-Jimenez, M. McKay, and R. Couillet, "Robust shrinkage M-estimators of large covariance matrices," in *Proc. IEEE Workshop Statistical Signal Process.*, 2016, pp. 1–4.

[14] R. J. Muirhead, *Aspects of Multivariate Statistical Theory*. New York, NY, USA: Wiley, 1982.

[15] K.-T. Fang, S. Kotz, and K.-W. Ng, *Symmetric Multivariate and Related Distributions*. New York, NY, USA: Chapman and Hall/CRC, 1990.

[16] 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, Nov. 2012.

[17] J. R. Magnus and H. Neudecker, *Matrix Differential Calculus With Applications in Statistics and Econometrics*, 2nd ed. Chichester, U.K.: Wiley, 1999.

[18] M. S. Srivastava, "Some tests concerning the covariance matrix in high dimensional data," *J. Jpn. Statistical Soc.*, vol. 35, no. 2, pp. 251–272, 2005.

[19] E. Ollila, "Optimal high-dimensional shrinkage covariance estimation for elliptical distributions," in *Proc. 25th Eur. Signal Process. Conf.*, 2017, pp. 1639–1643.

[20] D. N. Joanes and C. A. Gill, "Comparing measures of sample skewness and kurtosis," *J. Roy. Statistical Soc.: Series D (The Statistician)*, vol. 47, no. 1, pp. 183–189, 1998.

[21] P. M. Bentler and M. Berkane, "Greatest lower bound to the elliptical theory kurtosis parameter," *Biometrika*, vol. 73, no. 1, pp. 240–241, 1986.

[22] M. N. Tabassum and E. Ollila, "Compressive regularized discriminant analysis of high-dimensional data with applications to microarray studies," *IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2018, pp. 4204–4208.

[23] S. Visuri, V. Koivunen, and H. Oja, "Sign and rank covariance matrices," *J. Statistical Planning Inference*, vol. 91, pp. 557–575, 2000.

[24] B. Brown, "Statistical uses of the spatial median," *J. Roy. Statistical Soc. Series B (Methodological)*, vol. 45, no. 1, pp. 25–30, 1983.

[25] 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 L1-Norm Related Methods*. Basel, Switzerland: Birkhäuser, 2002, pp. 257–269.

[26] A. F. Magyar and D. E. Tyler, "The asymptotic inadmissibility of the spatial sign covariance matrix for elliptically symmetric distributions," *Biometrika*, vol. 101, no. 3, pp. 673–688, 2014.

[27] L. Yang, R. Couillet, and M. R. McKay, "A robust statistics approach to minimum variance portfolio optimization," *IEEE Trans. Signal Process.*, vol. 63, no. 24, pp. 6684–6697, 2015.

[28] H. Markowitz, "Portfolio selection," *J. Finance*, vol. 7, no. 1, pp. 77–91, 1952.

[29] H. Markowitz, *Portfolio Selection, Efficient Diversification of Investments*. New York, NY, USA: Wiley, 1959.

[30] J. Tobin, "Liquidity preference as behavior towards risk," *Rev. Econ. Stud.*, vol. 25, no. 2, pp. 65–86, 1958.

[31] W. Sharpe, "Capital asset prices: A theory of market equilibrium under conditions of risk," *J. Finance*, vol. 19, no. 3, pp. 425–442, 1964.

[32] J. Lintner, "The valuation of risk assets and the selection of risky investments in stock portfolios and capital budgets," *Rev. Economics Statist.*, vol. 47, no. 1, pp. 13–37, 1965.

[33] D. E. Tyler, "Radial estimates and the test for sphericity," *Biometrika*, vol. 69, no. 2, pp. 429–436, 1982.

Esa Ollila (M'03) received the M.Sc. degree in mathematics from the University of Oulu, Oulu, Finland, in 1998, Ph.D. degree in statistics with honors from the University of Jyväskylä, Jyväskylä, Finland, in 2002, and the D.Sc.(Tech) degree with honors in signal processing from Aalto University, Aalto, Finland, in 2010. From 2004 to 2007, he was a Postdoctoral Fellow and from August 2010 to May 2015 an Academy Research Fellow of the Academy of Finland. He has also been a Senior Lecturer with the University of Oulu. Currently, since June 2015, he has been an Associate Professor of Signal Processing, Aalto University, Finland. He is also an Adjunct Professor (Statistics) of Oulu University. During the Fall-term 2001, he was a Visiting Researcher with the Department of Statistics, Pennsylvania State University, State College, PA while the academic year 2010–2011 he spent as a Visiting Postdoctoral Research Associate with the Department of Electrical Engineering, Princeton University, Princeton, NJ, USA. He is a member of EURASIP SAT in Theoretical and Methodological Trends in Signal Processing and co-author of a recent textbook, *Robust Statistics for Signal Processing*, published by Cambridge University Press in 2018. His research interests lie in the intersection of the fields of statistical signal processing, high-dimensional statistics, and machine learning with emphasis on robust statistical methods and modeling.

Elias Raninen (S'17) received the B.Sc. (Tech.) degree and M.Sc. (Tech.) degree from Aalto University School of Electrical Engineering, Espoo, Finland, in 2015 and 2017, respectively. He received the bachelor's degree in pop and jazz music education from Metropolia University of Applied Sciences, Helsinki, Finland, in 2009. He is currently working toward the D.Sc. (Tech.) degree at the Department of Signal Processing and Acoustics, Aalto University, Espoo, Finland. He received the Aalto ELEC doctoral school scholarship in 2017. His research interests are in signal processing, statistics, machine learning, and optimization.