Linear pooling of sample covariance matrices

Elias Raninen, Student Member, IEEE, David E. Tyler, Esa Ollila, Senior Member, IEEE

Abstract—We consider the problem of estimating high-dimensional covariance matrices of \( K \)-populations or classes in the setting where the sample sizes are comparable to the data dimension. We propose estimating each class covariance matrix as a distinct linear combination of all class sample covariance matrices. This approach is shown to reduce the estimation error when the sample sizes are limited, and the true class covariance matrices share a somewhat similar structure.

We develop an effective method for estimating the coefficients in the linear combination that minimize the mean squared error under the general assumption that the samples are drawn from (unspecified) elliptically symmetric distributions possessing finite fourth-order moments. To this end, we utilize the spatial sign covariance matrix, which we show (under rather general conditions) to be an asymptotically unbiased estimator of the normalized covariance matrix as the dimension grows to infinity. We also show how the proposed method can be used in choosing the regularization parameters for multiple target matrices in a single class covariance matrix estimation problem. We assess the proposed method via numerical simulation studies including an application in global minimum variance portfolio optimization using real stock data.

Index Terms—covariance matrix, elliptical distribution, high-dimensional, multiclass, regularization, shrinkage, spatial sign covariance matrix

I. INTRODUCTION

HIGH-DIMENSIONAL covariance matrix estimation is a challenging problem as the dimension \( p \) of the observations can be much larger than the sample size \( n \). Such problems are increasingly common, for example, in finance [1], genomics [2], graphical models [3], chemometrics [4], wireless sensor networks [5], and adaptive filtering in array signal processing [6]. This paper considers a high-dimensional problem, where there are \( K \) distinct classes (populations). Since the population variables are in general the same, but are measured under different population conditions, it is reasonable to presume the \( K \) distinct covariance matrices share some common features or structure. In the small sample size setting, it is then advantageous to leverage on this presumption in the estimation of the population covariance matrices.

Consider \( K \) mutually independent classes, where each class \( k \in \{1, \ldots, K\} \) consists of independent and identically distributed (i.i.d.) \( p \)-dimensional observations \( X_k = \{x_{i,k} : i = 1, \ldots, n_k\} \) of size \( n_k \) with mean \( \mu_k = E[x_{i,k}] \) and positive definite symmetric covariance matrix

\[
\Sigma_k = E[(x_{i,k} - \mu_k)(x_{i,k} - \mu_k)^\top].
\]  

The ordinary estimators for the covariance matrix and the mean are the sample covariance matrix (SCM)

\[
S_k = \frac{1}{n_k - 1} \sum_{i=1}^{n_k} (x_{i,k} - \bar{x}_k)(x_{i,k} - \bar{x}_k)^\top,
\]  

and the sample mean \( \bar{x}_k = \frac{1}{n_k} \sum_{i=1}^{n_k} x_{i,k} \). When the sample size \( n_k \) and the data dimensionality \( p \) are comparable in size, the SCM can be highly variable, resulting in an unstable estimate of the population covariance matrix. Also, the SCM is positive definite only if \( n_k > p \) and \( \lambda_k \) spans \( \mathbb{R}^p \). Due to these problems, a commonly used approach in high-dimensional covariance matrix estimation is to use regularization (shrinkage).

In the one population case \( (K = 1) \), linear regularization usually refers to estimating the covariance matrix by a linear or convex combination of the SCM (or some other primary estimator) and a (usually positive definite) target matrix. Multiple examples can be found in finance [1], [7]–[9], genomics [2], and signal processing [10]–[15]. The target matrix is chosen based on prior knowledge or assumptions about the true population covariance matrix. Sometimes using more than one target matrix can further reduce the estimation error. In the double shrinkage approach of [16] and [17] there are two convex regularization steps: the SCM is first regularized toward a diagonal matrix consisting of the sample variances after which the resulting estimator is further regularized toward a scaled identity matrix. Recently, also multi-target shrinkage methods have been proposed that are able to incorporate a larger number of simultaneous target matrices [18]–[23].

In the multiple population setting, regularization via pooling the information in the different class samples is also possible. For example, [24] considered covariance matrix estimation from two independent data sets, whose covariance matrices are different but close to each other. This type of problems are encountered in radar processing as well as in hyperspectral imaging applications, where additional data sets may have been acquired with slightly different measurement configurations [24]. In the context of wireless sensor networks, [25] considered linear parameter estimation from independent and non-identically distributed scalar sample statistics. In discriminant analysis classification, the pooled SCM, \( S_{pool} = \frac{1}{n} \sum_{k=1}^{K} n_k S_k \), is often used as a shrinkage target and the class covariance matrices are estimated via a convex combination \( \Sigma_k = a S_k + (1 - a) S_{pool} \), where \( a \in [0, 1] \). This was studied in a Bayesian framework in [26] and [27], and in the Regularized Discriminant Analysis (RDA) framework in [28]. Furthermore, the optimal tuning parameter for this setting under elliptically distributed data was derived in our earlier work [29], [30]. For applications using RDA see, e.g., [31], [32].

E. Raninen and E. Ollila are with the Department of Signal Processing and Acoustics, Aalto University, P.O. Box 15400, FI-00076 Aalto, Finland. D. E. Tyler is with the Department of Statistics, Rutgers, The State University of New Jersey, Piscataway NJ 08854, U.S.A.

The research work by E. Raninen and E. Ollila was supported in part by the Academy of Finland under Grant 298118. Research work by D. E. Tyler was supported in part by the National Science Foundation under Grant DMS-1812198.
As noted previously, at least some of the $K$ population covariance matrices can be similar (close to each other in terms of suitable distance metric) and so it would be beneficial to use regularization to reduce the variance of the final estimates of the covariance matrices. Following this idea, we propose to estimate each class covariance matrix as a nonnegative linear combination of the SCMs of all classes. For $a \geq 0$, define
\[
S(a) = \sum_{i=1}^{K} a_i S_i. \tag{3}
\]
Restricting the coefficients to be nonnegative ensures that the estimator (3) is positive semidefinite. The aim is to find a $K \times K$ nonnegative coefficient matrix $A = (a_{ij}) = (a_1 \cdots a_K) \geq 0$, that minimizes the total mean squared error (MSE),
\[
A^* = \arg\min_{A \geq 0} \sum_{k=1}^{K} E \left[ ||S(a_k) - \Sigma_k||_F^2 \right] \tag{4}
\]
\[
\leftrightarrow a^*_k = \arg\min_{a \geq 0} E \left[ ||S(a) - \Sigma_k||_F^2 \right], \tag{5}
\]
with the estimate of $\Sigma_k$ then taken to be $\hat{\Sigma}_k = S(a^*_k)$. The equivalence of (4) and (5) is evident from the fact that the optimization problems for each class are separable. The solution to this problem is given in Section II. It is easy to see that the RDA based estimators form a subset of the more general form (3), which permits using individual weights for each SCM in the sum.

Below we summarize the main contributions of the paper.

- We propose covariance matrix estimators for multiclass problems, based on linearly pooling the class SCMs. Several aspects and properties of the estimator are discussed including possible modifications and an extension for complex-valued data.
- We show how the optimal linear coefficients can be estimated by assuming that the data is elliptically distributed. To this end, we use the spatial sign covariance matrix (SSCM), which we show under rather general assumptions to be asymptotically unbiased with respect to growing dimension.
- We show how the estimator can be used as a multi-target shrinkage estimator in a single class problem.
- Numerical simulations are conducted including a portfolio optimization problem using real stock data. The simulations show promising performance of the proposed estimator compared to competing estimators.
- Code is available at github.com/EliasRaninen, which works both for real and complex-valued data sets.

The rest of the paper is organized as follows. In Section II, we derive the optimal coefficients for the linear pooling estimator and study some of its properties. In Section III, we propose methods for estimating the statistical parameters needed for the estimation of the optimal coefficients. This section also presents the results regarding the SSCM. In Section IV, we discuss possible extensions and modifications to the estimator. In Section V, we discuss the similarities and differences between the proposed method and closely related multi-target shrinkage covariance matrix estimation methods. Furthermore, we show how our proposed method can be used as a multi-target shrinkage covariance matrix estimator in a single class problem with arbitrary positive semidefinite target matrices. Section VI provides numerical simulation studies and Section VII provides an application to investment portfolio selection using historical stock data. Lastly, Section VIII concludes.

Notation: Matrices are denoted by upper case boldface letters ($A$ or $\Delta$), vectors are denoted by lower case boldface letters ($a$ or $\delta$), and scalars are denoted by lower case letters ($a$ or $\delta$). For a matrix $A = (a_{ij})$, the notation $A \geq 0$ means that the matrix is nonnegative, that is, $a_{ij} \geq 0$, for all $i$ and $j$. Similarly, for a vector $a = (a_i)$ the notation $a \geq 0$ means that $a_i \geq 0$, for all $i$. The notation $A > 0$ ($A \geq 0$) means that $A$ is positive definite (positive semidefinite). The notation $\text{diag}(a)$ denotes a diagonal matrix with the entries of $a$ on the main diagonal. The identity matrix is denoted by $I$ and the vector of all ones is denoted by $1$. The notation $e_i$ denotes the $i$th Euclidean basis vector, i.e., a vector whose main diagonal. The identity matrix is denoted by $I$ and the vector of all ones is denoted by $1$. The notation $e_i$ denotes the $i$th Euclidean basis vector, i.e., a vector whose

II. LINEAR POOLING OF SCMs

In this section, we address solving for the coefficients of the linear combination of SCMs in (3). First, define the scaled $MSE$ of the SCM $S_k$ as
\[
\delta_k = p^{-1} \text{MSE}(S_k) = p^{-1} E[||S_k - \Sigma_k||_F^2]
\]
and the scaled inner products of the covariance matrices as
\[
c_{ij} = p^{-1} \text{tr}(\Sigma_i \Sigma_j),
\]
where $p$ is the dimension of the data. Then define
\[
\Delta = \text{diag}(\delta_1, \ldots, \delta_K) \quad \text{and} \quad C = (c_{11} \cdots c_{KK}) = (c_{ij}).
\]
We can then state the following result.

Theorem 1. (The $MSE$ of the linearly pooled estimator) For class $k$, the $MSE$ in (5) can be written as the strictly convex quadratic function
\[
p(a^T (\Delta + C) a - 2c_k^T a + c_k)
\]
where $\Delta + C$ is a positive definite symmetric matrix.

Proof. See Appendix A

As a consequence of Theorem 1, the optimal coefficients can be computed in the following way.

Proposition 1. (Optimal nonnegative coefficients) The solution to (5), $a^*_k$, is found by solving the strictly convex quadratic programming (QP) problem
\[
\text{minimize} \quad \frac{1}{2} a^T (\Delta + C) a - c_k^T a
\]
subject to $a \geq 0$. \hfill \qed
Proof. Follows from Theorem [1]

Many efficient algorithms exist for solving constrained convex QPs [33]. The optimization problem (6) requires knowledge of the matrices $C$ and $\Delta$, which depend on the unknown population parameters. We can nonetheless estimate the solution by using estimates $\hat{C}$ and $\hat{\Delta}$, which can be computed from the data as explained in Section III.

It is instructive to consider an unconstrained version of the optimization problem (6), where the weights are allowed to take negative values. For this case, we have the following closed form solution.

**Proposition 2. (Optimal unconstrained coefficients)** The unconstrained solution, which minimizes the MSE in (4) is

$$a^*_k = (\Delta + C)^{-1}c_k \iff A^* = (\Delta + C)^{-1}C. \quad (7)$$

**Proof.** Follows from Theorem [1] 

Note that if the closed form solution $a^*_k \geq 0$ in (7), then it is also the solution to (6).

Consider the single class case, for which the problem reduces to finding an optimal scaling parameter $a^*_1$ such that $\mathbb{E}[(a^*_1S_1 - \Sigma_1)\|z\|^2]$ is minimized. Proposition 2 above then states that the optimal parameter that minimizes the MSE is

$$a^*_1 = \frac{c_{11}}{\delta_1 + c_{11}} = \frac{1}{\text{NMSE}(\Sigma_1) + 1}, \quad \text{where } \text{NMSE}(\Sigma_1) = \text{MSE}(\Sigma_1)/\|\Sigma_1\|^2$$

is small. This implies that the contribution of $S_j$ in the linear combination $S(a^*_j)$ is small.

If one further assumes that all populations have the same distribution and the sample sizes are equal, then $\delta_1 = \cdots = \delta_K = \delta$ and

$$a^* = a^*_{jk} = \frac{1}{\text{NMSE}(S_j) + \frac{1}{K}} < \frac{1}{K} \quad \forall j, k.$$ 

That is, the pooled SCM is shrunk by the factor $Ka^* < 1$.

Due to the positivity of the coefficients, the conclusions of these special cases naturally also hold for the constrained case, where the coefficients are constrained to be nonnegative.

### III. Estimation

In this section, we address the estimation of $\Delta$ and $C$. We review elliptically symmetric distributions, introduce the relevant statistical parameters as well as show how to estimate them. Regarding an estimate for the sphericity parameter, we use the SSCM for which we then prove an asymptotic unbiasedness result in Theorem [2].

#### A. Elliptically symmetric distributions

We will assume that the samples are generated from unspecified elliptically symmetric distributions with finite fourth-order moments. That is, an absolutely continuous random vector $x \in \mathbb{R}^p$ from the $k$th population is assumed to have a density function up to a constant of the form

$$|\Sigma_k|^{-1/2}g_k((x - \mu_k)^\top \Sigma_k^{-1}(x - \mu_k)),$$

where $g_k : \mathbb{R}_+ \rightarrow \mathbb{R}_+$ is called the density generator [37]. Here, we let $g_k$ to be defined so that $\Sigma_k$ represents the covariance matrix of $x$, which means that $C_{k,0}^{-1} \int_0^\infty t^{p/2-1}g_k(t)dt = p$, where $C_k = \int_0^\infty \int_0^\infty g_k(t)dt$. We write $x \sim E_{p}(\mu_k, \Sigma_k, g_k)$ to denote this case. For example, the multivariate normal (MVN) distribution is a particular instance of the elliptical distribution for which $g_k(t) = \exp(-t/2)$. We write $x \sim N_p(\mu_k, \Sigma_k)$ to denote this case. An elliptically distributed random vector $x \sim E_p(\mu_k, \Sigma_k, g_k)$ can be expressed by the stochastic representation as

$$x = \mu_k + r_k\Sigma_k^{1/2}u,$$

where $r_k$ is a random variable called the modular variate, verifying $\mathbb{E}[r_k^2] = p$, and $u \in \mathbb{R}$ is a random variable distributed uniformly on the unit sphere, i.e., $u \in \{z \in \mathbb{R}^p : \|z\| = 1\}$. Furthermore, $u$ and $r_k$ are independent. More generally, we note that any random vector $x$ which satisfies (3) is said to have an elliptical distribution, even if it is not absolutely continuous, i.e., does not have a density. The relationship between the modular variate $r_k$ and $x$ is readily seen from (3) to be $r_k^2 = (x - \mu_k)^\top \Sigma_k^{-1}(x - \mu_k)$.

Sometimes we are only interested in the covariance matrix up to a scaling constant. Hence, we define the shape matrix:

$$\Lambda_k = p \Sigma_k / \text{tr}(\Sigma_k).$$
which verifies \( \text{tr}(\Lambda_k) = p \). Additionally, we define three statistical parameters that describe the elliptical distribution. First, we define the scale:

\[
\eta_k = \frac{\text{tr}(\Sigma_k)}{p},
\]

which is equal to the mean of the eigenvalues of \( \Sigma_k \). Note that, \( \Sigma_k = \eta_k A_k \). Second, we define the sphericity:

\[
\gamma_k = \frac{p \text{tr}(\Lambda_k^2)}{\text{tr}(\Sigma_k)^2} = \frac{\text{tr}(\Lambda_k)}{p},
\]

which equals the ratio of the mean of the squared eigenvalues relative to the mean of the squared eigenvalues squared. The sphericity parameter gets values in the range \([1, p]\) and attains its minimum for the scaled identity matrix and its maximum for a rank one matrix. Third, we define the elliptical kurtosis:

\[
\kappa_k = \frac{1}{3} \kurt(x_i) = \frac{1}{3} \left( \frac{\text{E}[|x_i - \mu|^4]}{\text{E}[(x_i - \mu)^2]^2} - 3 \right),
\]

where \( x_i \) denotes any marginal variable of \( x = (x_i) \sim \mathcal{E}_p(\mu_k, \Sigma_k, g_k) \) and \( \mu = \text{E}[x_i] \). For example, if the sample is from a MVN distribution, then \( \kappa_k = 0 \). The kurtosis parameter also satisfies \( \kappa_k = \mathcal{E}[r^4_k]/(p(p + 2) - 1) \), and hence, not only does \( \kappa_k \) represent the kurtosis of each of the variables \( x_i \), but it also represents the kurtosis of any univariate linear combination \( b^\top x \), where \( b \in \mathbb{R}^p \setminus \{0\} \). The lower bound for the kurtosis parameter is \( \kappa^{LB} = -2/(p + 2) \) [38].

For elliptical populations, the scaled MSE \( \delta_k \) of the SCM obtains an explicit form [9] Lemma 1:

\[
\delta_k = \eta_k^2 \left( \left( \frac{1}{n_k - 1} + \frac{\kappa_k}{n_k} \right) \eta_k + \frac{\kappa_k}{n_k} \gamma_k \right),
\]

which depends on the known sample size \( n_k \) as well as the unknown scale \( \eta_k \) [9], the unknown sphericity \( \gamma_k \) [10], and the unknown elliptical kurtosis \( \kappa_k \) [11].

### B. Estimates of the scale and the elliptical kurtosis

We estimate the scale using the SCM via

\[
\hat{\eta}_k = \frac{\text{tr}(S_k)}{p}.
\]

The kurtosis \( \hat{\kappa}_k \) is estimated via the (bias-corrected) average sample kurtosis of the marginal variables,

\[
\hat{\kappa}_k = \frac{N}{3} \left( (n_k + 1) g_2 + 6 \right),
\]

\[
g_2 = \frac{1}{p} \sum_{i,k} \left( \frac{1}{n_k} \sum_{j=1}^{n_k} (x_{i,j,k} - \bar{x}_{i,j,k})^4 \right)^{1/2} - 3,
\]

\[
N = \frac{n_k - 1}{(n_k - 2)(n_k - 3)}.
\]

where \( \bar{x}_{i,j,k} = \frac{1}{n_k} \sum_{j=1}^{n_k} x_{i,j,k} \) [39]. In case \( \kappa^{LB} \) is less than \( \kappa^{LB} \), we set \( \hat{\kappa}_k = 0.99 \kappa^{LB} \). Note that, although \( \kappa_k \) is invariant under affine transformations of \( x \), the estimator \( \hat{\kappa}_k \) in (14) is not. An alternative estimator of \( \kappa_k \) that is affine equivariant is

\[
\hat{\kappa}_k = \frac{1}{n_k} \sum_{i=1}^{n_k} \left[ (x_{i,k} - \bar{x}_k)^\top S_k^{-1}(x_{i,k} - \bar{x}_k) \right]^2 - 1.
\]

This estimator requires \( n_k > p \), and hence we use \( \hat{\kappa}_k \).

### C. Estimate of the sphericity using the SSCM

Regarding the sphericity, it would be natural to develop an estimator using the SCM as well. However, a simple and particularly well performing estimator of the sphericity is based on the robust spatial sign covariance matrix (SSCM) [16] and it has been used, e.g., in [40]–[42] and [9]. Particularly, in [9], both a SCM and a SSCM based estimator of the sphericity was compared and, except for the case where the samples were MVN, the simulations suggested the superiority of the SSCM based sphericity estimator. Before introducing the estimator for the sphericity, we will discuss the properties of the SSCM.

For \( x \sim \mathcal{E}_p(\mu, \Sigma, g) \), we define the population SSCM as

\[
\Sigma_{\text{sgn}} = E \left[ \frac{(x - \mu)(x - \mu)^\top}{\|x - \mu\|^2} \right],
\]

and since \( \text{tr}(\Sigma_{\text{sgn}}) = 1 \), its corresponding shape matrix is given by \( \Lambda_{\text{sgn}} = p \Sigma_{\text{sgn}} \).

It is known that \( \Lambda_{\text{sgn}} \) and \( \Lambda \) have the same eigenvectors as well as the multiplicities and the orders of the eigenvalues, but the eigenvalues themselves are different [43]. The sample SSCM of the \( k \)th population and its corresponding shape estimator are defined as

\[
S_{\text{sgn},k} = \frac{1}{n_k} \sum_{i=1}^{n_k} \left( x_{i,k} - \mu_k \right) (x_{i,k} - \mu_k)^\top \quad \text{and} \quad \hat{\Lambda}_k = p S_{\text{sgn},k},
\]

where the mean \( \mu_k \) is replaced with the sample spatial median [44], \( \hat{\mu}_k = \min_{m} \sum_{i=1}^{n_k} |x_{i,k} - m| \), when it is unknown. When the mean is known \( E[\Lambda_k] = \Lambda_{\text{sgn},k} \), and \( \hat{\Lambda}_k \) is distribution-free over the class of elliptical distributions. The latter statement can be proved by writing (15) in terms of the stochastic representation [6] and observing that the modular variate \( r_k \) cancels out in each summand.

Since the eigenvalues of \( \Lambda_{\text{sgn}} \) and \( \Lambda \) are different, \( \hat{\Lambda}_k \) is biased [45]. Surprisingly, as shown in Theorem 2 below, this bias becomes more negligible in higher dimensions provided the sequence of covariance matrix structures being considered with increasing \( p \) satisfies

\[
\gamma = o(p) \quad \text{as} \quad p \to \infty.
\]

\textbf{Theorem 2.} Let \( x \sim \mathcal{E}_p(\mu, \Sigma, g) \). Then,

\[
\Lambda_{\text{sgn}} = \Lambda + O(\gamma).
\]

Furthermore, if \( \{\gamma\} \) holds, then \( \Lambda_{\text{sgn}} = \Lambda + o(\|\Lambda\|_F) \).

\textbf{Proof.} See Appendix B.

The central condition \( \{\gamma\} \) holds for many common sequences of covariance models, as shown in Proposition 3 below. We first present, in the following lemma, a simple general condition under which \( \{\gamma\} \) holds. This lemma is seen to hold in particular for the case when the eigenvalues of \( \Lambda \) are bounded as \( p \to \infty \).

\textbf{Lemma 1.} If \( \Lambda \) is a shape matrix for which \( \lambda_{\max}(\Lambda) = O(p^{\tau/2}) \), where \( \tau < 1 \), as \( p \to \infty \), then \( \{\gamma\} \) holds.

\textbf{Proof.} See Appendix C.
Proposition 3. The following sequences, in $p$, of covariance models satisfy (17).

- **First order autoregressive** (AR(1)) covariance matrices: $(\Sigma)_{ij} = \sigma^2 \rho^{i-j}$, where $|\rho| < 1$ and $\sigma^2 > 0$ are both fixed, i.e., they do not depend on the dimension $p$.
- **1-banded Toeplitz covariance matrices**: $(\Sigma)_{ij} = \sigma^2 \rho^{i-j}$ for $|i-j| \leq 1$ and $(\Sigma)_{ij} = 0$ otherwise, where $|\rho| < -(2\cos(p\pi/(p+1))^{-1}$ and $\sigma^2 > 0$ are both fixed.
- **Spiked covariance matrices**: $\Sigma = \Sigma_r + \alpha I$, where $\Sigma_r$ is positive semidefinite with rank $r \leq p$ and $(r/p)(\lambda_{\max}(\Sigma_r)/\alpha)^2 = o(p)$. Here, $\alpha, r$, and $\lambda_{\max}(\Sigma_r)$ may vary with $p$.

However, for the **compound symmetric (CS)** covariance matrix $(\Sigma)_{ij} = \sigma^2 \rho$, for $i \neq j$ and $(\Sigma)_{ii} = \sigma^2$, where $\rho \in (-(p-1)^{-1}, 1)$ and $\sigma^2 > 0$ are both fixed, one obtains $\gamma = O(p)$.

**Proof.** See Appendix C

The restrictions on $\rho$ for the covariance models in Proposition 3 are needed to ensure that $\Sigma$ is positive definite.

Let us illustrate this result in the case that $\Sigma$ has an AR(1) covariance structure. In this case $\gamma \rightarrow (1 + \rho^2)/(1 - \rho^2)$ (see Appendix C-A). For $\rho = 0.5$, $\gamma \rightarrow 5/3 = O(1)$. From Theorem 2 we then have that the relative error $||A||^{-1}_F ||\Lambda_{sym} \alpha - A||_F$ is of the order $O(\sqrt{\gamma/p}) = O(p^{-1/2})$.

We also have the following proposition about the normalized mean squared distance between $\Lambda$ and a scaled SCM.

**Proposition 4.** Let $X = (x_i) \sim \mathcal{N}_p(0, \Sigma)$ and assume that (17) holds. Then,

$$\mathbb{E}\left[\|\hat{A} - \Lambda_{SCM}\|_F^2\right] \xrightarrow{p \to \infty} \frac{2}{n},$$

where $\Lambda_{SCM} = \frac{1}{n} \sum_{i=1}^n x_i x_i^\top$ and $\eta = \text{tr}(\Sigma)/p$, so that $\mathbb{E}[\Lambda_{SCM}] = \Lambda$.

**Proof.** See Appendix D

We may now focus on the sphericity estimator in (20). It can be shown by straightforward calculation (see Appendix D) that

$$\mathbb{E}[\text{tr}(\hat{A}^2)] = \frac{p}{n} + \frac{n-1}{n} \frac{\text{tr}(\mathbb{E}[A]^2)}{p}.$$  

If (17) holds, then by Theorem 2 one has that

$$\text{tr}(\mathbb{E}[A]^2)/p \to \gamma \text{ as } p \to \infty.$$  

By (18) and (19), a natural estimator for the sphericity of the $k$th class is then

$$\hat{\gamma}_k = \frac{n_k}{n_k - 1}\left(\frac{\text{tr}(\hat{A}_k^2)}{p} - \frac{p}{n_k}\right).$$  

In a high-dimensional setting, using the spatial median $\hat{\mu}$ in (16) results in nonnegligible error in the sphericity estimate. This was shown in (40), which considered SSCM based hypothesis testing of sphericity of the covariance matrix (i.e., $H_0: \Sigma \propto I$), and where they used a similar sphericity statistic as in (20). They also proposed a method for estimating and correcting for this error. Hence, we use the corrected estimator of the sphericity (40):

$$\hat{\gamma}_k = \hat{\gamma}_k - pd_k,$$  

where

$$d_k = \frac{1}{n_k^2} \left(1 - \frac{2 q_{2,k}}{q_{1,k}^2}\right) + \frac{1}{n_k^2} \left(\frac{q_{2,k}}{q_{1,k}} - 6 \frac{(q_{2,k})^2}{q_{1,k}^2} + 2 \frac{q_{2,k} q_{3,k}}{q_{1,k}^3} - 2 \frac{q_{3,k}}{q_{1,k}^3}\right)$$

and $q_{m,k} = (1/n_k) \sum_{i=1}^{n_k} \|x_{i,k} - \hat{\mu}_k\|^{-m}$. When computational simplicity is desired, it is also possible to use $d_k \approx n_k^{-2} + 2n_k^{-3}$, which is often a good approximation (see (40)).

**D. Final estimates**

The inner products $c_{ij} = p^{-1} \text{tr}(\Sigma_i \Sigma_j)$, for $i \neq j$, of the matrix $C$ are also estimated using SSCMs. In this case, however, no error correction due to using the spatial median is needed (see (41)). The estimates for $C$ and $\Delta$ are thus $\hat{C} = (\hat{c}_{ij})$ and $\Delta = \text{diag}(\hat{\delta}_1, \ldots, \hat{\delta}_K)$, where

$$\hat{c}_{ij} = \left\{\begin{array}{ll}
\text{tr}(\hat{\eta}_i \hat{A}_i \hat{\eta}_j \hat{A}_j)/p, & \text{for } i \neq j, \\
\hat{\gamma}_i^* \hat{\gamma}_j^*, & \text{for } i = j.
\end{array}\right.$$  

and $\hat{\delta}_k$ is obtained from (12) via $\hat{\eta}_k$ (13), $\hat{\kappa}_k$ (14), and $\hat{\gamma}_k^*$ (21).

**IV. EXTENSIONS AND MODIFICATIONS**

In this section, we first show how to incorporate regularization towards the identity matrix in the estimator. Then, we show how the optimal coefficients can alternatively be solved via a semidefinite optimization problem, which enables relaxing the nonnegativity constraint. Lastly, we extend the estimator to complex elliptically symmetric distributions.

**A. Additional regularization towards the identity matrix**

It is often beneficial to incorporate regularization towards the identity matrix. For example, if $p > n = \sum_k n_k$, then all of the SCMs are singular. Regularization towards the identity can easily be added by using the estimator

$$\hat{S}(a) = \sum_{j=1}^K a_j S_j + a_I I,$$  

where $a = (a_1, \ldots, a_K, a_I)^T \in \mathbb{R}^{K+1}$. By constraining $a_i \geq 0$, $1 \leq j \leq K$, and $a_I \geq \epsilon$, where $\epsilon > 0$ is a chosen lower bound for the identity regularization, the estimator (23) will be positive definite. Then

$$a_k^* = \arg \min_{a_j \geq 0, a_I \geq \epsilon} \mathbb{E}[\|\hat{S}(a) - \Sigma_k\|_F^2]$$

can be solved via the strictly convex QP problem

$$\text{minimize } \frac{1}{2} a^T (\hat{\Delta} + \hat{C}) a - \hat{c}^T a$$

subject to $a_j \geq 0$, $1 \leq j \leq K$, $a_I \geq \epsilon_k$, (24)

where

$$\hat{\Delta} = \text{diag}(\delta_1, \ldots, \delta_K, 0) \text{ and } \hat{C} = \left(\begin{array}{c}
\Sigma \eta \\
\eta^\top
\end{array}\right),$$  

(25)
and \( \eta = (\eta_k) \in \mathbb{R}^K \) is a vector consisting of scales \([3]\). The unconstrained optimal solution for this case is \( A^* = (\Delta + \hat{C})^{-1} (C \eta) \). The positive definiteness of \( \Delta + \hat{C} \) is shown in Appendix \([A]\). Algorithm \([1]\) summarizes the procedure for computing the linearly pooled estimates of the class covariance matrices.

Algorithm 1: Linear pooling of SCMs with identity

\[
\begin{align*}
\text{input} &: \text{Data } X_1, \ldots, X_K \text{ of the classes and } \epsilon_k > 0, \\
& \quad k = 1, \ldots, K. \\
\text{output} &: \hat{\Sigma}_1, \ldots, \hat{\Sigma}_K.
\end{align*}
\]

1. Compute SCMs \( S_1, \ldots, S_K \) of the classes.
2. Compute \( C \) and \( \Delta \) of \([25]\) (estimate as in Section \([III]\)).
3. Compute \( \hat{A} = (\hat{a}_1 \cdots \hat{a}_K) = (\Delta + \hat{C})^{-1} (C \eta) \).
4. for \( k = 1 \) to \( K \) do
   5. if \( \exists i \in \{1, \ldots, K\} : (\hat{a}_k)_i < 0 \) or \( \hat{a}_{1k} < \epsilon_k \) then
   6. Set \( \hat{a}_k \) as the solution of \([24]\).
   else
   7. \( \hat{a}_k \leftarrow \hat{a}_k. \)
   8. \( \hat{\Sigma}_k \leftarrow \hat{S}(\hat{a}_k) \) of \([23]\).

Remark 1. The QP formulation of the problem makes it easy to incorporate additional constraints if needed. For example, in order to find a convex combination of the SCMs the equality constraint \( \bar{a}^T \bar{a} = 1 \) should be added to the QP \([24]\).

B. Semidefinite programming formulation

Constraining the coefficients in \([6]\) to be nonnegative is sufficient to ensure positive semidefiniteness of the estimator \([3]\). In some cases this approach can be suboptimal since it may be possible to obtain a positive semidefinite estimator with a lower MSE by allowing some of the coefficients to be negative. Using semidefinite programming (SDP), the nonnegativity constraint can be replaced with a positive semidefiniteness constraint as in \([46]\) as follows. The objective function in \([6]\) can be rewritten as \( (a - B^{-1} c_k)^T B(a - B^{-1} c_k) + \text{const.} \), where \( B = \Delta + C \). By introducing an auxiliary variable \( t \) and constraint \( (a - B^{-1} c_k)^T B(a - B^{-1} c_k) \leq t \), the problem is converted into a minimization of \( t \). Using the Schur complement, the problem is reformulated as a semidefinite program (SDP) in the variables \( a \) and \( t \):

\[
\begin{align*}
\text{minimize} & \quad t \\
\text{subject to} & \quad \begin{pmatrix} t & (a - B^{-1} c_k)^T \\ a - B^{-1} c_k & B^{-1} \end{pmatrix} \succeq 0 \quad (26)
\end{align*}
\]

This is a convex optimization problem, which can be solved in polynomial time with software such as CVX, which is a package for specifying and solving convex programs \([47], [48]\). There are cases where the SDP \([26]\), nonnegativity constrained QP \([6]\), and unconstrained problem \([7]\) all give different solutions. In these cases only the SDP and QP solutions yield positive semidefinite estimators. In theory (when using the true \( C \) and \( \Delta \)), the SDP solution will have a lower MSE. However, with estimation error in \( C \) and \( \Delta \), this is necessarily not the case. The computational complexity of the SDP problem is significantly greater than that of the QP problem. The QP can be computed with 2–3 orders of magnitude faster (depending on the problem dimension). Due to the above reasons, we recommend the QP approach.

C. Extension to complex-valued data

Extending the results to complex-valued data requires using the complex-valued definitions of the elliptical kurtosis and the scaled MSE of the SCM. For a review on complex elliptical symmetric (CES) distributions, see e.g., \([49]\). A CES distribution (absolutely continuous) random vector \( x \in \mathbb{C}^p \) from the \( k \)th population has a density function up to a constant of the form

\[
|\Sigma_k|^{-1} h_k((x - \mu_k)^H \Sigma_k^{-1}(x - \mu_k)),
\]

where \( h_k : \mathbb{R}^p \to \mathbb{R}^> \) is the density generator, \( \mu = E[x] \) is the mean vector and \( \Sigma_k = E[(x - \mu)(x - \mu)^H] \) denotes the Hermitian positive definite covariance matrix of \( x \). Above \((\cdot)^H\) denotes the Hermitian (complex-conjugate) transpose. We denote this case by \( x \sim \mathcal{C}\mathcal{E}_p(\mu_k, \Sigma_k, h_k) \). The definitions of the SCM \([2]\) and the SSCM \([16]\) stay unchanged except that in their definitions \((\cdot)^T\) is replaced with \((\cdot)^H\). Furthermore, the inner products between Hermitian matrices \( (A = A^H) \) remain unchanged since \( \text{tr}(AA^H) = \text{tr}(A^2) \). Hence, the definitions of the spherialy parameter \( \gamma_k \) as well as the scale \( \eta_k \) remain unchanged. The kurtosis of a complex marginal variable \( x_i \) of \( x = (x_i) \in \mathbb{C}^p \) is defined as

\[
\kurt(x_i) = \frac{E[|x_i - \mu_i|^4]}{E[|x_i - \mu_i|^2]^2} - 2,
\]

where \( \mu_i = E[x_i] \). The elliptical kurtosis is then

\[
\kappa = (1/2)\kurt(x_i).
\]

The elliptical kurtosis \( \kappa_k \) of class \( k \) is estimated using the average sample kurtosis of the marginal variables

\[
\hat{\kappa}_k = \frac{1}{2p} \sum_{j=1}^{p} \frac{1}{\frac{1}{n_k} \sum_{i=1}^{n_k} |x_{ij,k} - \bar{x}_{j,k}|^4}{\frac{1}{n_k} \sum_{i=1}^{n_k} |x_{ij,k} - \bar{x}_{j,k}|^2} - 1, \quad (27)
\]

where \( \bar{x}_{j,k} = \frac{1}{n_k} \sum_{i=1}^{n_k} x_{ij,k} \). The theoretical lower bound of the kurtosis in the complex-valued case is \( \kappa_{LB} = -1/(p + 1) \) \([49]\). In case \((27)\) is less than \( \kappa_{LB} \), we set \( \hat{\kappa}_k = 0.99\kappa_{LB} \).

The matrix \( \Delta \) is estimated via \([25]\), given in the next lemma, which is an extension of \([9]\) Lemma 1) to the complex case.

Lemma 2. \([50]\) Theorem 3) Let \( x_k = \{x_{1}, \ldots, x_n\} \in \mathbb{C}^p \) be an i.i.d. random sample from \( \mathbb{C}\mathcal{E}_p(\mu_k, \Sigma_k, h_k) \) with finite fourth-order moments. Then, the scaled MSE of the SCM is

\[
\delta_k = \eta_k^2 \left( \frac{1}{n_k - 1} + \frac{\kappa_k}{n_k} \right) p + \kappa_k n_k \gamma_k. \quad (28)
\]

V. MULTI-TARGET SHRINKAGE ESTIMATORS

Multi-target shrinkage covariance matrix estimators are capable of simultaneously regularizing towards several target matrices. They are mainly designed for the single population
setting. However, some of the multi-target shrinkage estimators can also be used in a multiclass setting for pooling SCMs. Therefore, we give a brief review of the existing methods. Lastly, we show how our proposed method can be applied to a single class multi-target shrinkage covariance matrix estimation problem.

### A. Overview of multi-target shrinkage estimators

There exist multi-target shrinkage covariance matrix estimators, which can be used with user-defined target matrices and therefore also for pooling SCMs. In this section, we discuss these estimators and in Section [VI] we compare their performance to our proposed method by simulations.

Since the multi-target shrinkage estimators are developed for the single class covariance matrix estimation setting, we define the data set as \( \mathcal{X} = \{x_1, \ldots, x_n\} \), where \( (x_i)_j = x_{ij} \). Let \( \Sigma \) denote the covariance matrix and let \( S \) denote the SCM computed from \( \mathcal{X} \). Multi-target shrinkage covariance matrix estimators are often defined by

\[
\hat{\Sigma}(\mathbf{a}) = a_0 S + \sum_{k=1}^{K} a_k T_k,
\]

where \( T_k, k = 1, \ldots, K \), are linearly independent target matrices and \( a_j, j = 0, \ldots, K \), are the regularization coefficients. In [18] and [19], convex multi-target shrinkage covariance matrix estimators were proposed, where \( a_k \geq 0 \) and \( \sum_{k=1}^{K} a_k \leq 1 \) for \( k = 1, \ldots, K \) and \( a_0 = 1 - \sum_{k=1}^{K} a_k \). The coefficients \( \mathbf{a} = (a_k)_{k=1}^{K} \) were chosen as the minimizers of the MSE loss function

\[
L(\mathbf{a}; \mathbf{Q}, \mathbf{b}) = \mathbb{E}[\|\hat{\Sigma}(\mathbf{a}) - \Sigma\|^2] = \mathbf{a}^T \mathbf{Q} \mathbf{a} - 2 \mathbf{a}^T \mathbf{b} + \mathbb{E}[\|\Sigma - \Sigma\|^2]
\]

resulting in the constrained QP problem

\[
\mathbf{a}^* = \arg \min_{\mathbf{a} \geq 0, \mathbf{a}^T = 1} L(\mathbf{a}; \mathbf{Q}, \mathbf{b}),
\]

where \( \mathbf{Q}_{ij} = q_{ij} = \mathbb{E}[\mathbf{tr}((T_i - \Sigma)(T_j - \Sigma))] \) and \( \mathbf{b}_i = b_i = \mathbb{E}[\mathbf{tr}((T_i - \Sigma)(\Sigma - S))] = \mathbb{E}[\|\Sigma - \Sigma\|^2] - \mathbb{E}[\mathbf{tr}((\Sigma - \Sigma)(T_i - \Sigma T_i))] \). As the elements in \( \mathbf{Q} \) and \( \mathbf{b} \) depend on the unknown covariance matrix \( \Sigma \), they have to be estimated. In [19], the proposed estimates were \( q_{ij} = \mathbb{E}[\mathbf{tr}((T_i - \Sigma)(T_j - \Sigma))] \) and

\[
b_i = b_{ij} = \frac{1}{n(n-1)} \sum_{s=1}^{n} \sum_{t=1}^{n} \left(x_{si}x_{sj} - \frac{1}{n} \sum_{t=1}^{n} x_{ti}x_{tj}\right)^2,
\]

where the latter was obtained by approximating \( \mathbb{E}[\mathbf{tr}((\Sigma - \Sigma)(T_i - \Sigma T_i))] \approx \mathbb{E}[\mathbf{tr}((\Sigma - \Sigma)(T_i - \Sigma T_i))] \). We compare our method with this method in the simulations of Section [VI] where it is denoted by BARTZ. Regarding [18], a specific structural condition [18] (10) and (21) was imposed on the target matrices, which prevents using it for pooling SCMs.

A leave-one-out-cross-validation (LOOCV) approach was considered in [20] for different scenarios. Their proposition for SCM based linear shrinkage estimation used the LOOCV loss function

\[
L_{\text{CV}}(\mathbf{a}; \mathbf{Q}_{\text{CV}}, \mathbf{b}_{\text{CV}}) = \frac{1}{n} \sum_{i=1}^{n} \|a_i S_{-i} + \sum_{k=1}^{K} a_k T_k - x_i x_i^T\|^2
\]

\[
= \mathbf{a}^T \mathbf{Q}_{\text{CV}} \mathbf{a} - 2 a^T \mathbf{b}_{\text{CV}} + \text{const.}
\]

where \( x_i \) is assumed to have zero mean, \( S_{-i} = \frac{1}{n} \sum_{j \neq i} x_j x_j^T \) is the SCM computed without the \( i \)th sample, and \( \mathbf{a} = (a_0, a_1, \ldots, a_K)^T \geq 0 \). They showed how the elements in \( \mathbf{Q}_{\text{CV}} \) and \( \mathbf{b}_{\text{CV}} \) can be computed analytically. They also proposed a corresponding convex SCM based shrinkage estimator, which requires that the targets have the same trace as the estimated SCM. In the simulations of Section [VI] we denote the linear estimator by LOOCV.

### B. Multi-target shrinkage estimation via pooling SCMs

As discussed above, some of the multi-target shrinkage estimators can be used in order to linearly pool SCMs in a multi-class setting. Conversely, the proposed method of linearly pooling SCMs can be used in single class covariance matrix estimation as a multi-target shrinkage estimator as explained below.

Consider that there is only a single data set \( \mathcal{X} \), its corresponding SCM \( S \) and multiple positive definite symmetric target matrices \( \{T_m\}_{m=1}^{M} \). Our approach for multi-target shrinkage estimation is detailed in Algorithm 2. The idea is to generate artificial data sets \( \{Y_m\}_{m=1}^{M} \) with covariance matrices \( \{T_m\}_{m=1}^{M} \) so that \( \mathcal{X} \) and the generated data sets \( \{Y_m\}_{m=1}^{M} \) are approximately mutually independent. Specifically, \( Y_m \) is conditionally independent of \( \mathcal{X} \) given \( T_m \). Then, for the SCMs \( \{S_{T_m}\}_{m=1}^{M} \) computed from \( \{Y_m\}_{m=1}^{M} \), we make the approximations \( \mathbb{E}[\mathbf{tr}(S_{T_i}S_{T_j})] \approx \mathbb{E}[\mathbf{tr}(S_{T_i}S_{T_j})] = \mathbb{E}[\mathbf{tr}(S_{T_j}S_{T_j})] \approx \mathbb{E}[\mathbf{tr}(\Sigma T_j)] \).

We use the zero mean MVN distribution to generate the data sets, i.e., \( Y_m \sim \mathcal{N}_p(0, T_m) \), \( m = 1, \ldots, M \), each consisting of \( L \) samples. We can then apply the proposed method for pooling \( S \) and \( \{S_{T_m}\}_{m=1}^{M} \). We illustrate the usefulness of this method in a portfolio optimization problem in Section [VI] where it compares well with the other multi-target methods.

### VI. Simulation study

This section provides several simulation studies in order to assess the MSE performance of the proposed estimator as well as the accuracy of the plugin estimates of \( \Delta \) and \( C \) and the estimated coefficients \( \mathbf{A} \). We denote the proposed linear pooling estimator [23], which includes shrinkage towards identity, by LINPOOL. The LINPOOL estimator with an additional
had an CS structure. In the last mixed setup, classes
an AR(1) structure. In the second setup, all covariance matrices
different setups. In the first setup, all covariance matrices had
and held constant over the repetitions. We simulated three
the classes were generated from the standard MVN distribution
(see Section V for details).
uses a convex combination of the SCMs and the identity matrix,
and BARTZ from [19], which
are reported for the multi-target shrinkage method LOOCV
by LINPOOL-C. For LINPOOL and LINPOOL-C, we set the
1

TABLE I: The normalized mean squared error over 1000

replications and standard deviation in the parenthesis.

|                | class 1 | class 2 | class 3 | class 4 | total     |
|----------------|---------|---------|---------|---------|-----------|
| AR(1)         |         |         |         |         |           |
| LOOCV         | 0.14 (0.31) | 0.14 (0.10) | 0.24 (0.01) | 0.26 (0.02) | 0.78 (0.09) |
| BARTZ         | 0.15 (0.05) | 0.18 (0.01) | 0.29 (0.05) | 0.30 (0.02) | 0.91 (0.08) |
| LINPOOL       | 0.11 (0.05) | 0.14 (0.01) | 0.22 (0.05) | 0.26 (0.02) | 0.73 (0.08) |
| LINPOOL-C     | 0.12 (0.03) | 0.18 (0.01) | 0.27 (0.04) | 0.30 (0.02) | 0.87 (0.07) |
| CS            |         |         |         |         |           |
| LOOCV         | 0.18 (0.33) | 0.05 (0.05) | 0.18 (0.39) | 0.06 (0.07) | 0.47 (0.52) |
| BARTZ         | 0.20 (0.33) | 0.05 (0.05) | 0.13 (0.14) | 0.06 (0.04) | 0.43 (0.37) |
| LINPOOL       | 0.15 (0.29) | 0.04 (0.03) | 0.15 (0.24) | 0.05 (0.04) | 0.38 (0.38) |
| LINPOOL-C     | 0.16 (0.29) | 0.04 (0.03) | 0.12 (0.14) | 0.06 (0.04) | 0.38 (0.32) |
| Mixed: Σ₁ and Σ₂ are AR(1); Σ₃ and Σ₄ are CS |         |         |         |         |           |
| LOOCV         | 0.18 (0.06) | 0.21 (0.01) | 0.18 (0.38) | 0.06 (0.06) | 0.64 (0.39) |
| BARTZ         | 0.17 (0.04) | 0.31 (0.02) | 0.13 (0.14) | 0.05 (0.04) | 0.66 (0.15) |
| LINPOOL       | 0.16 (0.05) | 0.21 (0.01) | 0.15 (0.24) | 0.06 (0.05) | 0.58 (0.25) |
| LINPOOL-C     | 0.14 (0.02) | 0.31 (0.02) | 0.13 (0.14) | 0.05 (0.04) | 0.64 (0.15) |

Fig. 1 displays boxplots of the estimates \( \hat{\delta}_k \), \( \hat{c}_{ij} \), and \( \hat{a}_{i4} \) (coefficients for \( \Sigma_4 \) of LINPOOL). Left: AR(1) setup. Right: CS setup. The red triangles denote the theoretical values.

Table II tabulates the normalized mean squared error,

\[
\text{NMSE}(\hat{\Sigma}_k) = \text{Ave}|\hat{\Sigma}_k - \Sigma_k||^2 / ||\Sigma_k||^2,
\]

and total NMSE (sum of NMSEs of the classes) of the covariance matrix estimates for the three different setups over 1000 repetitions. Table II shows that the proposed method LINPOOL performed best in the AR(1) setup and the mixed setup, whereas LINPOOL-C performed best (slightly better than LINPOOL) in the CS setup.

A. Three different setups

In the first simulation, we considered two different covariance matrix structures: the AR(1), where \( (\Sigma_1)_{ij} = \sigma_k^2 \delta_{k}\delta_{k-i-j} \); and the CS structure, where \( (\Sigma_2)_{ij} = \sigma_k^2 \delta_{k} \) for \( i \neq j \) and \( (\Sigma_2)_{ij} = \sigma_k^2 \delta_{k} \) for \( i = j \). We generated four \( p = 100 \) dimensional random samples of sizes \( (n_1 = 20, n_2 = 100, n_3 = 20, n_4 = 100) \) from four independent multivariate \( t \)-distributions with \( \nu = 8 \) degrees of freedom. The means of the classes were generated from the standard MVN distribution and held constant over the repetitions. We simulated three different setups. In the first setup, all covariance matrices had an AR(1) structure. In the second setup, all covariance matrices had an CS structure. In the last mixed setup, classes 1 and 2 had an AR(1) structure and classes 3 and 4 had a CS structure. For all setups, we used \( \sigma_k^2 = \bar{k} \) and \( (\varphi_1 = 0.3, \varphi_2 = 0.4, \varphi_3 = 0.5, \varphi_4 = 0.6) \).
trials for each value of $K$ the NMSE decreases. When the added class has a different (i.e., doesn’t share the same structure as class 1). One can observe that every time an AR(1) structured class is added, when the added class covariance matrix has an CS structure (depending on the conditioning of the estimate) it can be useful for this may be that the total number of observations is still relatively low and including the fourth class helps in reducing the variance of the estimate.

Next, we examine the NMSE of the LINPOOL estimator of class 1 as the number of classes $K$ increase from 2 to 16 classes. The setup is as follows. The first class has an AR(1) covariance matrix structure with a fixed parameter $\varrho_1 = 0.5$. The other class covariance matrices also have an AR(1) structure except for the classes $k = 4, 8, 12, 16$, which have a CS structure. The parameter $\varrho_k$, for $k \geq 2$, is chosen uniformly at random from the interval $[0.1, 0.6]$ for each Monte Carlo trial. The means of the classes were generated from the standard MVN distribution for each Monte Carlo trial. The dimension is $p = 100$ and the data is multivariate t-distributed with $\nu = 8$ degrees of freedom.

Figure 3 depicts the results averaged over 1000 Monte Carlo trials for each value of $K$. The red vertical lines mark the spots when the added class covariance matrix has an CS structure (i.e., doesn’t share the same structure as class 1). One can observe that every time an AR(1) structured class is added, the NMSE decreases. When the added class has a different covariance structure (the CS structure), the NMSE does not decrease. An exception to this is when $K = 4$. A reason for this may be that the total number of observations is still relatively low and including the fourth class helps in reducing the variance of the estimate.

**Complex-valued case**

In the next simulation, the classes have an AR(1) covariance matrix structure, $\Sigma_{k} = \sigma_{k}^2 \rho_{k}^{\lfloor i-j \rfloor}$ for $i \leq j$, and ($\Sigma_{k})_{ji} = (\Sigma_{k}^*$)$_{ij}$, for $i > j$, where $(\cdot)^*$ denotes complex conjugation. The used parameters were $\sigma_k^2 = k$, $\varrho_1 = 0.3e^{j2\pi 0.3}$, $\varrho_2 = 0.4e^{j2\pi 0.4}$, $\varrho_3 = 0.5e^{j2\pi 0.5}$, and $\varrho_4 = 0.6e^{j2\pi 0.6}$. We simulated the NMSE as a function of the sample size from $n_k \in \{10, 15, \ldots, 40\}$ for all $k$. The data was generated from the complex multivariate $t$-distribution with $\nu = 8$ degrees of freedom and dimension $p = 100$. The results were averaged over 1000 Monte Carlo trials for each sample size and are shown in Figure 2. It can be observed that especially for small sample sizes ($n_k < 30$) LINPOOL performed better than LOOCV.

**VII. PORTFOLIO OPTIMIZATION**

We studied the performance of the proposed method in a portfolio optimization problem using dividend adjusted daily closing prices. Portfolio optimization is a central topic in investment theory, see, e.g., [51]–[54], and [55]. A focus in portfolio optimization has been on the estimation of the covariance matrix of the stock returns, commonly using shrinkage regularization techniques or random matrix theory, see, e.g., [1], [7], [56]–[58], and [59]. In a portfolio optimization problem, a particular investment portfolio is determined by a weight or allocation vector $w \in \mathbb{R}^p$ (verifying the constraint $1^T w = 1$) whose elements describe the fraction of the total wealth invested in each of the $p$ stocks. We considered different portfolios. First, we considered the global minimum variance portfolio (GMVP) in which one seeks a portfolio $w$ that minimizes the risk (variance). The optimization problem is thus

$$\text{minimize } w^T \Sigma w \text{ subject to } 1^T w = 1. \quad (30)$$

The well-known solution is $w^* = \Sigma_{-1} 1 / 1^T \Sigma 1$, where $\Sigma$ is the covariance matrix of the stock returns. We also considered a constrained portfolio, where the coefficients are constrained to be within the range $0 \leq w_i \leq 1$, for all $i$, i.e., shorting (negative weights) is not allowed and the portfolio manager is not allowed to put more than 10% of the wealth in one stock. The optimization problem for this case is the same as in (30) but having the additional constraint $0 \leq w \leq 0.1 \cdot 1$, which results in a QP problem.

In the simulation, the covariance matrix $\Sigma$ was estimated via a sliding window method so that at day $t$ it was estimated using the daily net returns of the previous $n$ days from $t - n$ to $t - 1$. The portfolio weights were then computed via (30) with and without the additional weight constraints. These yielded the portfolio returns for the next 20 (trading) days. Then, the
sliding window shifted 20 days forward and the procedure was repeated. By denoting the total number of days in the data by \( T \), we obtained \( T - n \) daily returns from which we computed the realized risk as the empirical standard deviation of the daily portfolio returns.

We applied the proposed method (explained in Section V-B and Algorithm 2) for single class covariance matrix estimation using the same target matrices for regularization as in [1] and [8]. The target matrices were the single factor market index model \( \mathbf{T}_F \) of [1] and the constant correlation model \( \mathbf{T}_C \) of [8]. Their computation is explained in [1] and [8], respectively. At day \( t \), we used the \( n \) previous days to compute the SCM. However, due to the nature of our method, we were able to freely choose the amount of data used for computing the regularization target matrices. Hence, we chose to use the previous 40 days (\( t - 40 \)) to \( t - 1 \) corresponding to the previous two months) for computing \( \mathbf{T}_F \) and \( \mathbf{T}_C \), regardless of the window size \( n \) (\( n > 40 \)) used for estimating the SCM. This can be justified by the fact that the trend of the market is better captured by the most recent net returns. After computing \( \mathbf{T}_F \) and \( \mathbf{T}_C \), we generated 1000 i.i.d. samples both from \( \mathcal{Y}_C \sim \mathcal{N}(0, \mathbf{T}_C) \) and \( \mathcal{Y}_F \sim \mathcal{N}(0, \mathbf{T}_F) \) and estimated the coefficients for the proposed LINPOOL estimator \( \tilde{\mathbf{S}}(\mathbf{a}) = a_1 \mathbf{S}_F + a_2 \mathbf{S}_F + a_3 \mathbf{S}_C + a_4 \mathbf{I} \), where \( \mathbf{S}_F \) and \( \mathbf{S}_C \) denote the SCMs computed from \( \mathcal{Y}_F \) and \( \mathcal{Y}_C \), respectively. We also report the performance of the LINPOOL estimator (23) using a convex combination as explained in Remark 1. For both methods, we used the lowerbound constraint \( a_I \geq c = 10^{-8} \).

We also report the performance of the multi-target shrinkage estimation methods LOOCV of [20] and BARTZ of [19] using the same target matrices (\( \mathbf{T}_F, \mathbf{T}_C, \) and \( \mathbf{I} \)) as for our proposed method. Additionally, we report the performance of the following methods specifically tailored for portfolio optimization: LW-well of [7], LW-improved of [1], LW-honey of [8], and the random matrix theory based estimator LW-analytical of [56].

We used the same portfolio data sets as in [9] (obtained from https://finance.yahoo.com). That is, Standard & Poor’s 500 stock market index (S&P 500) tracking 396 stocks from from Jan. 4, 2016 to Apr. 27, 2018 consisting of \( T = 583 \) days. Hang Seng Index (HSI) from Jan. 4, 2010 to Dec. 24, 2011 (45 stocks during \( T = 491 \) days) and from Jan. 1, 2016 to Dec. 27, 2017 (50 stocks during \( T = 489 \) days after removing two zero return days).

Figure 4 shows the annualized risk obtained by the different estimators as a function of the training window length \( n \). Regarding S&P 500 2016-2018 data, there were large differences in the performances of the methods. LW-improved performed best and LW-well the worst. The differences between the methods were smaller in the constrained case, where BARTZ performed best. Regarding the HSI 2010-2011 data set, the proposed method LINPOOL achieved the lowest risk for the unconstrained portfolio with all window sizes \( n \). For the constrained portfolio, all of the methods performed nearly equally well with LW-improved having the lowest risk with the window length \( n = 100 \). Regarding the HSI 2016-2017 data, the proposed methods (LINPOOL and LINPOOL-C) achieved the lowest risk for both the constrained and unconstrained portfolios for all window sizes \( n > 120 \).

\section{Conclusions}

The paper proposed a regularized sample covariance matrix estimation method for high-dimensional multiclass problems. The proposed estimator was formed from a linear combination of the class SCMs. We derived the theoretically optimal coefficients that minimize the mean squared error. The optimal coefficients depend on unknown parameters, and their estimation was addressed under the assumption that the samples are generated from unknown elliptically symmetric populations with finite fourth-order moments. In constructing estimators for the unknown parameters, we utilized the sample spatial sign covariance matrix, which we showed in Theorem 2 to be an asymptotically unbiased estimator of the normalized covariance matrix in the case that the sphericity parameter of the distribution grows slower than the dimension. The effectiveness and usefulness of the proposed method was demonstrated via...
simulations and a portfolio optimization problem using real stock data. Codes are available at github.com/EliasRaninen

APPENDIX A

PROOF OF THEOREM \[1\]

Write \( f(a) = \mathbb{E}[\|S(a) - \Sigma_k\|^2] \) for the objective function. By expanding the expression for the squared error, we get

\[
\|S(a) - \Sigma_k\|^2 = \text{tr} \left( \left( \sum_{i=1}^{K} a_i S_i - \Sigma_k \right) \left( \sum_{j=1}^{K} a_j S_j - \Sigma_k \right)^\top \right)
\]

\[
= \sum_{i=1}^{K} \sum_{j=1}^{K} a_i a_j \text{tr}(S_i S_j) - 2 \sum_{j=1}^{K} a_j \text{tr}(S_j \Sigma_k) + \text{tr}(\Sigma_k^2).
\]

Taking the expectation and scaling by \( 1/p \) gives

\[
\left(1/p\right) f(a) = a^\top B a - 2 c_k^\top a + c_k c_k,
\]

where

\[
B = (b_{ij}) \in \mathbb{R}^{K \times K}, \quad b_{ij} = \mathbb{E}[\text{tr}(S_i S_j)]/p,
\]

\[
C = (c_{ij}) \in \mathbb{R}^{K \times K}, \quad c_{ij} = \mathbb{E}[\text{tr}(\Sigma_i \Sigma_j)]/p,
\]

and \( c_k \) corresponds to the \( k \)th column of \( C \). For \( i \neq j \),

\[
b_{ij} = p^{-1} \mathbb{E}[\text{tr}(S_i S_j)] = p^{-1} \mathbb{E}[\mathbb{E}[S_i] \mathbb{E}[S_j]] = p^{-1} \text{tr}(\Sigma_i \Sigma_j) = c_{ij}.
\]

Using that \( \mathbb{E}[\text{tr}(S_i^2)] = \text{MSE}(S_i) + \text{tr}(\Sigma_i^2) \), we get for \( i = j \),

\[
b_{ii} = p^{-1} \mathbb{E}[\text{tr}(S_i^2)] = \delta_i + c_{ii},
\]

and so \( B = \Delta + C \). By definition \( \Delta \) is symmetric and positive definite. Also, \( C \) is symmetric and positive semidefinite. This follows since for any \( m \in \mathbb{R}^K \), \( m^\top C m = \sum_{i=1}^{K} \sum_{j=1}^{K} m_i m_j \mathbb{E}[\text{tr}(\Sigma_i \Sigma_j)] = \text{tr}(\Sigma^2) \geq 0 \), where \( \Sigma = p^{-1/2} \sum_{j=1}^{K} m_j \Sigma_j \). Hence \( \Delta + C \geq 0 \) is invertible.

Regarding the extension discussed in Section \[14\], we now show the positive definiteness of \( \Delta + C \). Since we know that \( \Delta + C \succeq 0 \), due to the properties of the Schur complement \[60\] A.5.5, it holds that

\[
\bar{\Delta} + \bar{C} = \begin{pmatrix} \Delta + C & \eta \\ \eta^\top & 1 \end{pmatrix} \succ 0 \iff \Delta + C - \eta \eta^\top \succ 0.
\]

We can then show the positive definiteness of \( \Delta + C \) by showing that \( \Delta + C - \eta \eta^\top \) is positive definite. For \( m \in \mathbb{R}^K \setminus \{0\} \),

\[
m^\top (\Delta + C - \eta \eta^\top) m = \sum_k m_k^2 \delta_k + \sum_{i,j} m_i m_j (c_{ij} - \eta_i \eta_j)
\]

\[
= \sum_k m_k^2 \delta_k + \text{tr}(\Sigma^2) > 0,
\]

where \( \Sigma = p^{-1/2} \sum_{j=1}^{K} m_j (\Sigma_j - \eta_j I) \). So, \( \bar{\Delta} + \bar{C} \succ 0 \). \( \square \)

APPENDIX B

PROOF OF THEOREM \[2\]

Let \( x \sim \mathcal{N}(\mu, \Sigma; g) \) and assume \( \hat{\Lambda} \) is computed with a known mean \( \mu = 0 \) in \[16\]. Then \( \mathbb{E}[\hat{\Lambda}] = p \mathbb{E}[xx^\top/\|x\|^2] = \Lambda_{\text{sym}} \). Let \( u = (u_i) = \Sigma^{-1/2} / \| \Sigma^{-1/2} x \|. \) So, \( u \) is uniformly distributed on the unit sphere \( \{ u \in \mathbb{R}^p : \| u \| = 1 \} \). Denote \( \Lambda = (\Lambda_{ij}) \). We need the following lemma.

**Lemma 3.**

\[
\mathbb{E} \left[ \Lambda^{1/2} uu^\top \Lambda^{1/2} uu^\top \right] = p^{-1} (p + 2)^{-1} (\text{tr}(\Lambda) + 2 \Lambda^2).
\]

**Proof.**

\[
\mathbb{E} \left[ \Lambda^{1/2} uu^\top \Lambda^{1/2} uu^\top \right] = \sum_{k,l} \Lambda_{kl} \Lambda^{1/2} \mathbb{E}[u_k u_l uu^\top] \Lambda_{kl}^{1/2}
\]

\[
= \Lambda^{1/2} \left( \sum_l \Lambda_{kl} \mathbb{E}[u_k^2 uu^\top] + \sum_{k \neq l} \Lambda_{kl} \mathbb{E}[u_k u_l uu^\top] \right) \Lambda^{1/2}.
\]

Here (see \[61\] p. 3.12.2),

\[
(\mathbb{E}[u_i^2 uu^\top])_{jj} = \begin{cases} \mathbb{E}[u_i^2] = 3p^{-1}(p + 2)^{-1}, & l = j, \\ \mathbb{E}[u_i^2 u_j^2] = p^{-1}(p + 2)^{-1}, & l \neq j, \end{cases}
\]

and \( (\mathbb{E}[u_i^2 uu^\top])_{ij} = 0 \), for \( i \neq j \). Hence, \( \mathbb{E}[u_i^2 uu^\top] = p^{-1}(p + 2)^{-1}(I + 2e_i e_i^\top) \). Thus,

\[
\sum_l \Lambda_{kl} \mathbb{E}[u_k^2 uu^\top] = p^{-1}(p + 2)^{-1}((\sum_l \Lambda_{kl})I + 2 \sum_l \Lambda_{kl} e_l e_l^\top)
\]

\[
= p^{-1}(p + 2)^{-1}(\text{tr}(\Lambda)I + 2 \text{diag}(\Lambda)).
\]

Regarding the other term in \[31\], where \( k \neq l \),

\[
(\mathbb{E}[u_k u_l uu^\top])_{mn} = \begin{cases} \mathbb{E}[u_k^2 u_l^2] = p^{-1}(p + 2)^{-1}, & k = m, l = n, \\ \mathbb{E}[u_k^2 u_l^2] = p^{-1}(p + 2)^{-1}, & k = n, l = m, \\ 0, & \text{otherwise}. \end{cases}
\]

Hence, we have

\[
\sum_{k \neq l} \Lambda_{kl} \mathbb{E}[u_k u_l uu^\top] = p^{-1}(p + 2)^{-1}(\Lambda - \text{diag}(\Lambda))
\]

and the result follows by substituting \[32\] and \[33\] into \[31\]. \( \square \)

Note that, \[32\], \[33\], and \[31\] are valid more generally for any positive semidefinite symmetric matrix, not only for the shape matrix.

We are ready to prove Theorem \[2\] Since for any \( x > 0 \), \( x^2 \geq 2 - x \), we have

\[
\Lambda_{\text{sym}} = p \mathbb{E} \left[ \frac{x x^\top}{\|x\|^2} \right] = p \mathbb{E} \left[ \frac{\Lambda^{1/2} uu^\top \Lambda^{1/2}}{u^\top A u} \right]
\]

\[
\geq p \mathbb{E} \left[ \Lambda^{1/2} uu^\top \Lambda^{1/2} (2 - u^\top A u) \right]
\]

\[
= 2 p \mathbb{E}[\Lambda^{1/2} uu^\top \Lambda^{1/2}] - p \mathbb{E}[\Lambda^{1/2} uu^\top \Lambda^{1/2} uu^\top A u u^\top]
\]

\[
= 2 \Lambda - \frac{1}{p+2} (\text{tr}(\Lambda) I + 2 \Lambda^2) = \Lambda - \frac{2}{p+2} \Lambda.
\]

(34)
By scaling both sides of (33) by \( \epsilon = \| A \|_F^{-1} = 1/\sqrt{p} \), the first term will have unity norm \( (\| \epsilon A \|_F = 1) \). Let us then consider the second term on the right-hand side. Its trace is
\[
\frac{2 \text{tr}(A^2 - A)}{p + 2} = \frac{2p}{p + 2} (\gamma - 1) \\
\leq 2 \left( \sqrt{\frac{\gamma}{p}} - \frac{1}{p} \right) = O \left( \sqrt{\frac{\gamma}{p}} \right),
\]
and its norm is
\[
\epsilon \left\| \frac{2(A^2 - A)}{p + 2} \right\|_F \leq (2\epsilon/p) (\| A^2 \|_F + \| A \|_F) \leq (2/p) (\| A \|_F + 1) = O(\sqrt{\gamma/p}).
\]
Here, we used the triangle inequality and submultiplicativity properties of the Frobenius norm. By moving all the terms of (34) to the left-hand side and scaling them by \( \epsilon \), we can use (35) and (36) as well as the property that for any \( A \succeq 0 \), \( \text{tr}(A) \geq \| A \|_F \) to get
\[
O(\sqrt{\gamma/p}) \geq \left\| (A_{\text{sign}} - A) / \| v \| + O(\sqrt{\gamma/p}) \right\|_F \geq 0,
\]
which implies that \( A_{\text{sign}} = A + O(\gamma) \). Furthermore, if \( \gamma = o(p) \), then \( O(\sqrt{\gamma/p}) = o(1) \) implying \( A_{\text{sign}} = A + o(\| A \|_F) \).

**Appendix C**

**Proofs of Lemma 1 and Proposition 3**

Let \( \lambda_1 = \lambda_{\max}(A) \). Assume that as \( p \to \infty \), \( \lambda_1 = O(p^{\tau/2}) \), where \( \tau < 1 \). Then
\[
\gamma = \frac{\text{tr}(A^2)}{p} \leq \frac{p\lambda_1^2}{p} = \lambda_1^2 = O(p^{\tau}) = o(p).
\]
Particularly, \( \gamma = o(p) \) if \( \lambda_1 = O(1) \).

In the following, we assume that \( \phi \) is a fixed parameter that does not depend on the dimension \( p \). We also use \( \gamma = \text{tr}(A^2)/p = (1/p) \sum_{i,j} \Lambda_{ij}^2 \), where \( A = (\Lambda_{ij}) \).

**A. Sphericity of the AR(1) covariance matrix**

The shape matrix of the AR(1) covariance matrix with parameter \( \phi (|\phi| < 1) \) has \( p \) number of ones on the main diagonal, \( 2(p - 1) \) number of \( \phi \) on the first diagonals above and below the main diagonal, and \( 2(p - 2) \) number of \( \phi^2 \) on the second diagonals above and below the main diagonal, and so on. That gives,
\[
\gamma_{AR(1)} = 1 + \sum_{i=1}^{p-1} (p - i) \phi^2 = 2 \sum_{i=0}^{p-1} (\phi^2)^i - 2 \sum_{i=1}^{p-1} i (\phi^2)^i - 1.
\]
The first sum is the geometric series and the second sum is also well-known and its solution can be obtained by differentiating the geometric series. Hence, we get
\[
\gamma_{AR(1)} = 2 \left( \frac{1 - (\phi^2)^p}{1 - \phi^2} - \frac{2(p - 1)(\phi^2)^{p+1} - p(\phi^2)^p + \phi^2}{p(1 - \phi^2)^2} \right) = \frac{p - p\phi^2 - 2\phi^2 + 2(\phi^2)^{p+1}}{p\phi^2(1 - \phi^2)^2}.
\]
As \( p \to \infty \), \( \gamma_{AR(1)} \to (1 + \phi^2)/(1 - \phi^2) = O(1) = o(p) \).

**B. Sphericity of the 1-Banded Toeplitz covariance matrix**

The shape matrix corresponding to a 1-banded Toeplitz covariance matrix has \( p \) number of ones on the main diagonal and \( 2(p - 1) \) number of \( \phi \) on the off-diagonals while rest of the off-diagonals are zero. This implies that \( \gamma_{1B} = 1 + 2(1 - 1/p)\phi^2 \). As \( p \to \infty \), \( \gamma_{1B} \to 1 + 2\phi^2 = O(1) = o(p) \).

**C. Sphericity of the spiked covariance matrix**

Let \( \Sigma = \Sigma_r + \alpha I \), where \( \Sigma_r \) has rank \( r \leq p \). Let \( \alpha = \beta \eta_r \), where \( \eta_r = \text{tr}(\Sigma_r)/p \). Let \( \gamma_r = p \text{tr}(\Sigma_r^2)/\text{tr}(\Sigma_r^2) = \text{tr}(\Sigma_r^2)/(p\eta_r^2) \). Then, \( \text{tr}(\Sigma_r^2) = p^2 \eta_r^2 (1 + \beta + \beta^2) \) and
\[
\text{tr}(\Sigma_r^2) = \text{tr}((\Sigma_r + \beta \eta_r I)^2) = \text{tr}(\Sigma_r^2) + \beta^2 p\eta_r^2 + 2\beta p\eta_r^2 = p\eta_r^2 (\beta^2 + 2\beta + 2) = p\eta_r^2 (\gamma_r - 1 + (\beta + 1)^2).
\]
Then by computing \( \gamma_r = p \text{tr}(\Sigma_r^2)/\text{tr}(\Sigma_r^2) \), we get
\[
\gamma_r = \frac{\gamma_r - 1}{(\beta + 1)^2} + 1 \leq \frac{\eta_r^2 \gamma_r}{\alpha^2} + 1 = \frac{r\lambda_1^2}{p\alpha^2} + 1,
\]
where we used that the rank of \( \Sigma_r \) is \( r \) and \( \eta_r^2 \gamma_r = \text{tr}(\Sigma_r^2)/p \leq r\lambda_1^2/p \), where \( \lambda_1 = \lambda_{\max}(\Sigma_r) \). Therefore, \( \gamma = o(p) \), if \( \frac{r\lambda_1^2}{p\alpha^2} = o(p) \), which includes the cases \( \frac{\lambda_1}{\alpha} = O(p^{\tau/2}) \) and \( \frac{\lambda_1}{\alpha} = O(p^{\tau+1/2}) \) and \( r = O(1) \), where \( \tau < 1 \).

**D. Sphericity of the CS covariance matrix**

The shape matrix of the CS covariance matrix has \( p \) number of ones on the main diagonal and \( p - 1 \) number of \( \phi \) on the off-diagonals. Thus, \( \gamma_{CS} = 1 + (p - 1)\phi^2 = O(p) \).

**Appendix D**

**Proof of Proposition 4**

Expanding the mean squared distance gives
\[
E[\| \hat{A} - A_{SCM} \|_F^2] = E[\text{tr}(\hat{A}^2)] + E[\text{tr}(A_{SCM}^2)] - 2E[\text{tr}(\hat{A}A_{SCM})].
\]
The first term is
\[
E[\text{tr}(\hat{A}^2)] = \text{tr} \left( \left( \frac{p}{n} \sum_i x_i x_i^T \right)^2 \right) = \frac{p^2}{n^2} \text{tr} \left( \sum_i x_i x_i^T x_i x_i^T + \sum_{i \neq j} x_i x_i^T x_j x_j^T \right) = \frac{p^2}{n} + \frac{n - 1}{n} \text{tr}(E[A]^2).
\]
Now, let \( S = \frac{1}{n} \sum_{i=1}^n x_i x_i^T \). Since \( S \) is unbiased, using Wishart theory, we have \( \text{MSE}(S) = \text{tr} \left( \text{var}(\text{vec}(S)) \right) = (1/n) \text{tr}((I + K)(S \otimes S)) = (1/n) \text{tr}(S^2)(\text{tr}(S^2) + 2\text{tr}(S^2)) \), where \( K = \sum_{i=1}^n e_i e_i^T \otimes e_i e_i^T \) is the commutation matrix and \( \otimes \) denotes the Kronecker product. Using that \( \text{E}[\text{tr}(S^2)] = \text{MSE}(S) + \text{tr}(S^2) \),
\[
E[\text{tr}(\hat{A}_{SCM}^2)] = \frac{p^2}{\text{tr}(S)^2} E[\text{tr}(S^2)] = \frac{p^2}{n} + \left( \frac{n + 1}{n} \right) \text{tr}(A^2).
\]
The last term of (37) is
\[
E[\text{tr}(\Lambda \Lambda_{\text{SCM}})] = E\left[\frac{p^2}{n} \sum_{i} \sum_{j} x_i x_j^\top \frac{p}{n} \text{tr}(\Sigma) \sum_{j} x_i x_j^\top\right] = \frac{p^2}{n^2 \text{tr}(\Sigma)} \left(\sum_{i} \sum_{j} E[x_i x_i^\top] + \sum_{i \neq j} E[x_i x_j^\top] E[x_j x_i^\top]\right)
\]
\[
= \frac{p^2}{n} + \frac{n - 1}{n} \text{tr}(E[\Lambda]) \frac{1}{n} \cdot \text{tr}(E[\Lambda^2]) \tag{40}
\]
Now, substituting (38), (39), and (40) into (37) gives
\[
\frac{n - 1}{n} \text{tr}(E[\Lambda^2]) = 2 \text{tr}(E[\Lambda]) + \frac{n + 1}{n} \text{tr}(E[\Lambda^2]).
\]
Dividing this by $\|\Lambda\|^2$ and applying assumption $\gamma = o(p)$ and Theorem 2 we get the result $E[\|\Lambda - A_{\text{SCM}}\|^2]|E[\|\Lambda\|^2] = o_p(n^{-1/2})$.

References

[1] O. Ledoit and M. Wolf, “Improved estimation of the covariance matrix of stock returns with an application to portfolio selection,” Journal of Empirical Finance, vol. 10, no. 5, pp. 603–621, 2003.

[2] J. O. Engel and K. Strimmer, “A Shrinkage Approach to Large-Scale Covariance Matrix Estimation and Implications for Functional Genomics,” Statistical Applications in Genetics and Molecular Biology, vol. 4, no. 1, Nov. 2005.

[3] T. Zhang, A. Wiesel, and M. S. Greco, “Multivariate generalized gaussian distribution: Convexity and graphical models,” IEEE Transactions on Signal Processing, vol. 61, no. 16, pp. 4141–4148, Jun. 2013.

[4] J. Engel, L. Buydens, and L. Blanchet, “An overview of large-dimensional covariance and precision matrix estimators with applications in chemometrics,” Journal of Chemometrics, vol. 31, no. 4, e2880, Apr. 2017.

[5] Y. Chen, A. Wiesel, and A. O. Hero, “Robust Shrinkage Estimation of High-Dimensional Covariance Matrices,” IEEE Transactions on Signal Processing, vol. 59, no. 9, pp. 4097–4107, Sep. 2011.

[6] D. E. Tyler, O. Besson, and Y. I. Abramovich, “Regularized Covariance Matrix Estimation in Complex Elliptically Symmetric Distributions Using the Expected Likelihood Approach—Part 2: The Under-Sampled Case,” IEEE Transactions on Signal Processing, vol. 61, no. 23, pp. 5819–5829, Dec. 2013.

[7] R. Koenker and G. W. Battese, “Fast convexification of adaptive processor or a structured covariance matrix,” IEEE Transactions on Aerospace and Electronic Systems, vol. 36, no. 4, pp. 1115–1126, Oct. 2000.

[8] Y. I. Abramovich and O. Besson, “Regularized Covariance Matrix Estimation in Complex Elliptically Symmetric Distributions Using the Expected Likelihood Approach—Part 1: The Over-Sampled Case,” IEEE Transactions on Signal Processing, vol. 61, no. 23, pp. 5807–5818, Dec. 2013.

[9] A. A. Ayadi, A. D. Maio, and L. Pallotta, “A Geometric Approach to Covariance Matrix Estimation and its Applications to Radar Problems,” IEEE Transactions on Signal Processing, vol. 66, no. 4, pp. 907–922, Feb. 2018.

[10] A. A. Ayadi, A. D. Maio, and L. Pallotta, “A Geometric Approach to Covariance Matrix Estimation and its Applications to Radar Problems,” IEEE Transactions on Signal Processing, vol. 66, no. 4, pp. 907–922, Feb. 2018.

[11] A. D. Maio, A. Ayadi, A. C. Eldar, and A. O. Hero, “Shrinkage algorithms for MMSE covariance estimation,” IEEE Transactions on Signal Processing, vol. 58, no. 10, pp. 5016–5029, Oct. 2010.

[12] J. Li, J. Zhou, and B. Zhang, “Estimation of Large Covariance Matrices by Shrinkage to Structured Target in Normal and Non-Normal Distributions,” IEEE Access, vol. 6, pp. 2158–2169, 2018.

[13] M. Steiner and K. Geilbach, “Fast converging adaptive processor or a structured covariance matrix,” IEEE Transactions on Aerospace and Electronic Systems, vol. 36, no. 4, pp. 1115–1126, Oct. 2000.

[14] O. Ledoit and M. Wolf, “A well-conditioned estimator for large-dimensional covariance matrices,” Journal of Multivariate Analysis, vol. 88, no. 2, pp. 365–411, 2004.

[15] O. Ledoit and M. Wolf, “Honey, I Shrank the Sample Covariance Matrix,” The Journal of Portfolio Management, vol. 32, no. 4, pp. 116–129, Jul. 2006.

[16] E. Ollila and O. Ledoit, “Optimal shrinkage Covariance Matrix Estimation Under Random Sampling From Elliptical Distributions,” IEEE Transactions on Signal Processing, vol. 67, no. 10, pp. 2707–2719, May 2019.

[17] Y. Chen, A. Wiesel, and A. O. Hero, “Shrinkage algorithms for MMSE covariance estimation,” IEEE Transactions on Signal Processing, vol. 58, no. 10, pp. 5016–5029, Oct. 2010.

[18] J. Li, J. Zhou, and B. Zhang, “Estimation of Large Covariance Matrices by Shrinkage to Structured Target in Normal and Non-Normal Distributions,” IEEE Access, vol. 6, pp. 2158–2169, 2018.

[19] M. Steiner and K. Geilbach, “Fast converging adaptive processor or a structured covariance matrix,” IEEE Transactions on Aerospace and Electronic Systems, vol. 36, no. 4, pp. 1115–1126, Oct. 2000.

[20] E. Ollila and O. Ledoit, “Optimal shrinkage Covariance Matrix Estimation Under Random Sampling From Elliptical Distributions,” IEEE Transactions on Signal Processing, vol. 67, no. 10, pp. 2707–2719, May 2019.

[21] Y. Chen, A. Wiesel, and A. O. Hero, “Shrinkage algorithms for MMSE covariance estimation,” IEEE Transactions on Signal Processing, vol. 58, no. 10, pp. 5016–5029, Oct. 2010.

[22] J. Li, J. Zhou, and B. Zhang, “Estimation of Large Covariance Matrices by Shrinkage to Structured Target in Normal and Non-Normal Distributions,” IEEE Access, vol. 6, pp. 2158–2169, 2018.

[23] M. Steiner and K. Geilbach, “Fast converging adaptive processor or a structured covariance matrix,” IEEE Transactions on Aerospace and Electronic Systems, vol. 36, no. 4, pp. 1115–1126, Oct. 2000.

[24] Y. I. Abramovich and O. Besson, “Regularized Covariance Matrix Estimation in Complex Elliptically Symmetric Distributions Using the Expected Likelihood Approach—Part 2: The Under-Sampled Case,” IEEE Transactions on Signal Processing, vol. 61, no. 23, pp. 5819–5829, Dec. 2013.

[25] E. Ollila and O. Ledoit, “Optimal shrinkage Covariance Matrix Estimation Under Random Sampling From Elliptical Distributions,” IEEE Transactions on Signal Processing, vol. 67, no. 10, pp. 2707–2719, May 2019.
[50] E. Raninen, E. Ollila, and D. E. Tyler, “On the Variability of the Sample Covariance Matrix Under Complex Elliptical Distributions,” IEEE Signal Processing Letters, vol. 28, pp. 2092–2096, Oct. 4, 2021.
[51] H. Markowitz, “Portfolio selection,” The Journal of Finance, vol. 7, no. 1, pp. 77–91, 1952.
[52] ———, Portfolio Selection, Efficient Diversification of Investments. Wiley, 1959.
[53] J. Tobin, “ Liquidity preference as behavior towards risk,” The Review of Economic Studies, vol. 25, no. 2, pp. 65–86, 1958.
[54] W. Sharpe, “Capital asset pricing: A theory of market equilibrium under conditions of risk,” The Journal of Finance, vol. 19, no. 3, pp. 425–442, 1964.
[55] J.Lintner, “The valuation of risk assets and the selection of risky investments in stock portfolios and capital budgets,” The Review of Economics and Statistics, vol. 47, no. 1, pp. 13–37, 1965.
[56] O. Ledoit and M. Wolf, “Analytical nonlinear shrinkage of large-dimensional covariance matrices,” Annals of Statistics, vol. 48, no. 5, pp. 3043–3065, 2020.
[57] ———, “ Spectrum estimation: A unified framework for covariance matrix estimation and PCA in large dimensions,” Journal of Multivariate Analysis, vol. 139, pp. 360–384, 2015.
[58] L. Yang, R. Couillet, and M. R. McKay, “A robust statistics approach to minimum variance portfolio optimization,” IEEE Transactions on Signal Processing, vol. 63, no. 24, pp. 6684–6697, Aug. 2015.
[59] Y. Feng and D. P. Palomar, “A signal processing perspective on financial engineering,” Foundations and Trends® in Signal Processing, vol. 9, no. 1–2, pp. 1–231, 2016.
[60] S. Boyd and L. Vandenberghe, Convex Optimization. Cambridge University Press, 2004.
[61] K.-T. Fang, S. Kotz, and K. W. Ng, Symmetric Multivariate and Related Distributions. Boca Raton: Chapman and Hall/CRC, 1990, 230 pp.
[62] J. R. Magnus and H. Neudecker, “The commutation matrix: Some properties and applications,” Annals of Statistics, vol. 7, no. 2, pp. 381–394, 1979.