Linear Pooling of Sample Covariance Matrices

Elias Raninen [ID], *Student Member, IEEE*, David E. Tyler, and Esa Ollila [ID], *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 $\mathcal{X}_k = \{\mathbf{x}_{1,k}, \ldots, \mathbf{x}_{n_k,k}\}$ of size n_k with mean $\boldsymbol{\mu}_k = \mathbb{E}[\mathbf{x}_{i,k}]$ and positive definite symmetric covariance matrix

$$\boldsymbol{\Sigma}_k = \mathbb{E}[(\mathbf{x}_{i,k} - \boldsymbol{\mu}_k)(\mathbf{x}_{i,k} - \boldsymbol{\mu}_k)^\top]. \tag{1}$$

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

$$\mathbf{S}_k = \frac{1}{n_k - 1} \sum_{i=1}^{n_k} (\mathbf{x}_{i,k} - \bar{\mathbf{x}}_k)(\mathbf{x}_{i,k} - \bar{\mathbf{x}}_k)^\top, \tag{2}$$

and the sample mean $\bar{\mathbf{x}}_k = \frac{1}{n_k} \sum_{i=1}^{n_k} \mathbf{x}_{i,k}$. 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 $\mathcal{X}_k$ 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, $\mathbf{S}_{\text{pool}} = \frac{1}{n} \sum_{k=1}^{K} n_k \mathbf{S}_k$, $n = \sum_{k=1}^{K} n_k$, is often used as a shrinkage target and the class

covariance matrices are estimated via a convex combination $\hat{\boldsymbol{\Sigma}}_k = a\mathbf{S}_k + (1-a)\mathbf{S}_{\mathrm{pool}}$, 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]

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 $\mathbf{a} \geq \mathbf{0}$, define

$$\mathbf{S}(\mathbf{a}) = \sum_{i=1}^{K} a_i \mathbf{S}_i. \tag{3}$$

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 $\mathbf{A} = (a_{ij}) = (\mathbf{a}_1 \cdots \mathbf{a}_K) \geq \mathbf{0}$, that minimizes the total mean squared error (MSE),

$$\mathbf{A}^\star = \arg\min_{\mathbf{A} \geq \mathbf{0}} \sum_{k=1}^{K} \mathbb{E}\left[\|\mathbf{S}(\mathbf{a}_k) - \boldsymbol{\Sigma}_k\|_{\mathrm{F}}^2\right] \tag{4}$$

$$\Leftrightarrow \mathbf{a}_k^\star = \arg\min_{\mathbf{a} \geq \mathbf{0}} \mathbb{E}\left[\|\mathbf{S}(\mathbf{a}) - \boldsymbol{\Sigma}_k\|_{\mathrm{F}}^2\right], \ k = 1, \ldots, K, \tag{5}$$

with the estimate of $\boldsymbol{\Sigma}_k$ then taken to be $\hat{\boldsymbol{\Sigma}}_k = \mathbf{S}(\mathbf{a}_k^\star)$. 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 https://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 ($\mathbf{A}$ or $\boldsymbol{\Delta}$), vectors are denoted by lower case boldface letters ($\mathbf{a}$ or $\boldsymbol{\delta}$), and scalars are denoted by lower case letters (a or δ). For a matrix $\mathbf{A} = (a_{ij})$, the notation $\mathbf{A} \geq \mathbf{0}$ means that the matrix is nonnegative, that is, $a_{ij} \geq 0$, for all i and j. Similarly, for a vector $\mathbf{a} = (a_i)$ the notation $\mathbf{a} \geq \mathbf{0}$ means that $a_i \geq 0$, for all i. The notation $\mathbf{A} \succ \mathbf{0}$ ($\mathbf{A} \succeq \mathbf{0}$) means that $\mathbf{A}$ is positive definite (positive semidefinite). The notation $\mathrm{diag}(\mathbf{a})$ denotes a diagonal matrix with the entries of $\mathbf{a}$ on the main diagonal. The identity matrix is denoted by $\mathbf{I}$ and the vector of all ones is denoted by $\mathbf{1}$. The notation $\mathbf{e}_i$ denotes the ith Euclidean basis vector, i.e., a vector whose ith coordinate is 1 and all other coordinates are 0. For real sequences a_p and b_p, as $p \to \infty$, the notation $a_p = o(b_p)$ means that the sequence $a_p/b_p \to 0$, and the notation $a_p = O(b_p)$ means that the sequence a_p/b_p is bounded. For a matrix-valued sequence $\mathbf{A}_p$, we write $\mathbf{A}_p = o(b_p)$ and $\mathbf{A}_p = O(b_p)$ if and only if $\|\mathbf{A}_p\|_{\mathrm{F}} = o(b_p)$ and $\|\mathbf{A}_p\|_{\mathrm{F}} = O(b_p)$, respectively. The Frobenius norm is defined as $\|\mathbf{A}\|_{\mathrm{F}} = \sqrt{\mathrm{tr}(\mathbf{A}^\top \mathbf{A})}$ while $\|\cdot\|$ denotes the Euclidean norm for vectors. The largest eigenvalue of $\mathbf{A}$ is $\lambda_{\max}(\mathbf{A})$. The determinant of $\mathbf{A}$ is denoted by $|\mathbf{A}|$.

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 $\mathbf{S}_k$ as

$$\delta_k = p^{-1}\mathrm{MSE}(\mathbf{S}_k) = p^{-1}\mathbb{E}[\|\mathbf{S}_k - \boldsymbol{\Sigma}_k\|_{\mathrm{F}}^2]$$

and the scaled inner products of the covariance matrices as

$$c_{ij} = p^{-1}\mathrm{tr}(\boldsymbol{\Sigma}_i\boldsymbol{\Sigma}_j),$$

where p is the dimension of the data. Then define

$$\boldsymbol{\Delta} = \mathrm{diag}(\delta_1, \ldots, \delta_K) \text{ and } \mathbf{C} = \left(\mathbf{c}_1 \cdots \mathbf{c}_K\right) = (c_{ij}).$$

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

$$p(\mathbf{a}^\top(\boldsymbol{\Delta} + \mathbf{C})\mathbf{a} - 2\mathbf{c}_k^\top\mathbf{a} + c_{kk}),$$

where $\boldsymbol{\Delta} + \mathbf{C}$ 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), $\mathbf{a}_k^\star$, is found by solving the strictly convex quadratic programming (QP) problem

$$\begin{aligned} \text{minimize } & \tfrac{1}{2}\mathbf{a}^\top(\mathbf{\Delta} + \mathbf{C})\mathbf{a} - \mathbf{c}_k^\top\mathbf{a} \\ \text{subject to } & \mathbf{a} \geq \mathbf{0}. \end{aligned} \quad (6)$$

Proof: Follows from Theorem 1. □

Many efficient algorithms exist for solving constrained convex QPs [33].[1] The optimization problem (6) requires knowledge of the matrices $\mathbf{C}$ and $\mathbf{\Delta}$, which depend on the unknown population parameters. We can nonetheless estimate the solution by using estimates $\hat{\mathbf{C}}$ and $\hat{\mathbf{\Delta}}$, 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

$$\mathbf{a}_k^\star = (\mathbf{\Delta} + \mathbf{C})^{-1}\mathbf{c}_k \Leftrightarrow \mathbf{A}^\star = (\mathbf{\Delta} + \mathbf{C})^{-1}\mathbf{C}. \quad (7)$$

Proof: Follows from Theorem 1. □

Note that if the closed form solution $\mathbf{a}_k^\star \geq \mathbf{0}$ 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 $\mathbb{E}[\|a_1\mathbf{S}_1 - \mathbf{\Sigma}_1\|_{\mathrm{F}}^2]$ is minimized. Proposition 2 above then states that the optimal parameter that minimizes the MSE is

$$a_1^\star = \frac{c_{11}}{\delta_1 + c_{11}} = \frac{1}{\mathrm{NMSE}(\mathbf{S}_1) + 1},$$

where $\mathrm{NMSE}(\mathbf{S}_1) = \mathrm{MSE}(\mathbf{S}_1)/\|\mathbf{\Sigma}_1\|_{\mathrm{F}}^2$ is referred to as the normalized MSE (NMSE). It can easily be shown that $\mathrm{MSE}(a_1^\star\mathbf{S}_1) = a_1^\star\mathrm{MSE}(\mathbf{S}_1) < \mathrm{MSE}(\mathbf{S}_1)$ since $a_1^\star < 1$. Therefore, the (oracle) estimator $\hat{\mathbf{\Sigma}}_1 = a_1^\star\mathbf{S}_1$ is always more efficient than $\mathbf{S}_1$. For the univariate normal case, one obtains $a_1^\star = (n_1 - 1)/(n_1 + 1)$. This result was first obtained in [35]. A corresponding result for the general (non-normal) univariate case was obtained in [36], and it can be written as $a_1^\star = ((n_1 + 1)/(n_1 - 1) + 3\kappa_1/n_1)^{-1}$, where κ_1 is the symmetric kurtosis of the population (see (11)).

Consider next the special case when all population covariance matrices are equal, i.e., $\mathbf{\Sigma}_1 = \cdots = \mathbf{\Sigma}_K \equiv \mathbf{\Sigma}$. In this case, $\mathbf{C} = c\mathbf{1}\mathbf{1}^\top$ with $c = \mathrm{tr}(\mathbf{\Sigma}^2)/p$. Using the Sherman-Morrison formula, $(\mathbf{\Delta} + c\mathbf{1}\mathbf{1}^\top)^{-1} = \mathbf{\Delta}^{-1} - \mu\mathbf{\Delta}^{-1}\mathbf{1}\mathbf{1}^\top\mathbf{\Delta}^{-1}$, where $\mu = c/(1 + c\mathbf{1}^\top\mathbf{\Delta}^{-1}\mathbf{1})$, we obtain the solution (7) as

$$\mathbf{A}^\star = \mu\mathbf{\Delta}^{-1}\mathbf{1}\mathbf{1}^\top.$$

Hence, all columns of $\mathbf{A}^\star$ coincide, and the coefficients in each column are $a_{jk}^\star = \mu/\delta_j$. These coefficients can be written in an equivalent form

$$a_{jk}^\star = \frac{\mathrm{NMSE}(\mathbf{S}_j)^{-1}}{1 + \sum_{i=1}^{K}\mathrm{NMSE}(\mathbf{S}_i)^{-1}}.$$

Thus, the weights are positive and proportional to the inverses of the NMSE of the SCMs. If $\mathbf{S}_j$ has a large NMSE relative to others, which occurs for example when the sample size n_j is small relative to others, then the weight $a_{jk}^\star$ assigned for $\mathbf{S}_j$ is small. This implies that the contribution of $\mathbf{S}_j$ in the linear combination $\mathbf{S}(\mathbf{a}_k^\star)$ is small.

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

$$a^\star \equiv a_{jk}^\star = \frac{1}{\mathrm{NMSE}(\mathbf{S}_j) + K} < \frac{1}{K} \quad \forall j, k.$$

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

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 $\mathbf{\Delta}$ and $\mathbf{C}$. 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 $\mathbf{x} \in \mathbb{R}^p$ from the kth population is assumed to have a density function up to a constant of the form

$$|\mathbf{\Sigma}_k|^{-1/2}g_k((\mathbf{x} - \boldsymbol{\mu}_k)^\top\mathbf{\Sigma}_k^{-1}(\mathbf{x} - \boldsymbol{\mu}_k)),$$

where $g_k : \mathbb{R}_{\geq 0} \to \mathbb{R}_{>0}$ is called the *density generator* [37]. Here, we let g_k to be defined so that $\mathbf{\Sigma}_k$ represents the covariance matrix of $\mathbf{x}$, which means that $C_k^{-1}\int_0^\infty t^{p/2}g_k(t)\mathrm{d}t = p$, where $C_k = \int_0^\infty t^{p/2-1}g_k(t)\mathrm{d}t$. We write $\mathbf{x} \sim \mathcal{E}_p(\boldsymbol{\mu}_k, \mathbf{\Sigma}_k, g_k)$ 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 $\mathbf{x} \sim \mathcal{N}_p(\boldsymbol{\mu}_k, \mathbf{\Sigma}_k)$ to denote this case. An elliptically distributed random vector $\mathbf{x} \sim \mathcal{E}_p(\boldsymbol{\mu}_k, \mathbf{\Sigma}_k, g_k)$ can be expressed by the *stochastic representation* as

$$\mathbf{x} = \boldsymbol{\mu}_k + r_k\mathbf{\Sigma}_k^{1/2}\mathbf{u}, \quad (8)$$

where r_k is a random variable called the *modular variate*, verifying $\mathbb{E}[r_k^2] = p$, and $\mathbf{u}$ is a random vector distributed uniformly on the unit sphere, i.e., $\mathbf{u} \in \{\mathbf{z} \in \mathbb{R}^p : \|\mathbf{z}\| = 1\}$. Furthermore, $\mathbf{u}$ and r_k are independent. More generally, we note that any random vector $\mathbf{x}$ which satisfies (8) is said to have an elliptical distribution, even if it is not absolutely continuous, i.e., does not have a

---

[1]In [34] it was shown that using the ellipsoidal method, the strictly convex QP can be solved in polynomial time. In the simulations, we use the MATLAB function `quadprog`, which uses an interior-point method.

density. The relationship between the modular variate r_k and $\mathbf{x}$ is readily seen from (8) to be $r_k^2 = (\mathbf{x} - \boldsymbol{\mu}_k)^\top \boldsymbol{\Sigma}_k^{-1} (\mathbf{x} - \boldsymbol{\mu}_k)$.

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

$$\boldsymbol{\Lambda}_k = p \frac{\boldsymbol{\Sigma}_k}{\mathrm{tr}(\boldsymbol{\Sigma}_k)},$$

which verifies $\mathrm{tr}(\boldsymbol{\Lambda}_k) = p$. Additionally, we define three statistical parameters that describe the elliptical distribution. First, we define the *scale:*

$$\eta_k = \mathrm{tr}(\boldsymbol{\Sigma}_k)/p, \tag{9}$$

which is equal to the mean of the eigenvalues of $\boldsymbol{\Sigma}_k$. Note that, $\boldsymbol{\Sigma}_k = \eta_k \boldsymbol{\Lambda}_k$. Second, we define the *sphericity:*

$$\gamma_k = \frac{p \, \mathrm{tr}(\boldsymbol{\Sigma}_k^2)}{\mathrm{tr}(\boldsymbol{\Sigma}_k)^2} = \frac{\mathrm{tr}(\boldsymbol{\Lambda}_k^2)}{p}, \tag{10}$$

which equals the ratio of the mean of the squared eigenvalues relative to the mean of the 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:*

$$\kappa_k = \frac{1}{3} \mathrm{kurt}(x_i) = \frac{1}{3} \left(\frac{\mathbb{E}[(x_i - \mu_i)^4]}{\mathbb{E}[(x_i - \mu_i)^2]^2} - 3 \right), \tag{11}$$

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

For elliptical populations, the scaled MSE δ_k of the SCM obtains an explicit form [9, Lemma 1]:

$$\delta_k = \eta_k^2 \left(\left(\frac{1}{n_k - 1} + \frac{\kappa_k}{n_k} \right) (p + \gamma_k) + \frac{\kappa_k}{n_k} \gamma_k \right), \tag{12}$$

which depends on the known sample size n_k as well as the unknown scale η_k (9), the unknown sphericity γ_k (10), and the unknown elliptical kurtosis κ_k (11).

B. Estimates of the Scale and the Elliptical Kurtosis

We estimate the scale using the SCM via

$$\hat{\eta}_k = \mathrm{tr}(\mathbf{S}_k)/p. \tag{13}$$

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

$$\hat{\kappa}_k = \frac{N}{3} ((n_k + 1)g_2 + 6), \tag{14}$$

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

$$N = \frac{n_k - 1}{(n_k - 2)(n_k - 3)},$$

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

$$\tilde{\kappa}_k = \frac{1}{n_k} \sum_{i=1}^{n_k} \frac{[(\mathbf{x}_{i,k} - \bar{\mathbf{x}}_k)^\top \mathbf{S}_k^{-1} (\mathbf{x}_{i,k} - \bar{\mathbf{x}}_k)]^2}{p(p+2)} - 1.$$

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

$$\boldsymbol{\Sigma}_{\mathrm{sgn}} = \mathbb{E}\left[\frac{(\mathbf{x} - \boldsymbol{\mu})(\mathbf{x} - \boldsymbol{\mu})^\top}{\|\mathbf{x} - \boldsymbol{\mu}\|^2} \right], \tag{15}$$

and since $\mathrm{tr}(\boldsymbol{\Sigma}_{\mathrm{sgn}}) = 1$, its corresponding shape matrix is given by $\boldsymbol{\Lambda}_{\mathrm{sgn}} = p \boldsymbol{\Sigma}_{\mathrm{sgn}}$.

It is known that $\boldsymbol{\Lambda}_{\mathrm{sgn}}$ and $\boldsymbol{\Lambda}$ 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 kth population and its corresponding shape estimator are defined as

$$\mathbf{S}_{\mathrm{sgn},k} = \frac{1}{n_k} \sum_{i=1}^{n_k} \frac{(\mathbf{x}_{i,k} - \boldsymbol{\mu}_k)(\mathbf{x}_{i,k} - \boldsymbol{\mu}_k)^\top}{\|\mathbf{x}_{i,k} - \boldsymbol{\mu}_k\|^2} \text{ and } \hat{\boldsymbol{\Lambda}}_k = p\mathbf{S}_{\mathrm{sgn},k}, \tag{16}$$

where the mean $\boldsymbol{\mu}_k$ is replaced with the *sample spatial median* [44], $\hat{\boldsymbol{\mu}}_k = \arg\min_{\mathbf{m}} \sum_{i=1}^{n_k} \|\mathbf{x}_{i,k} - \mathbf{m}\|$, when it is unknown. When the mean is known $\mathbb{E}[\hat{\boldsymbol{\Lambda}}_k] = \boldsymbol{\Lambda}_{\mathrm{sgn},k}$, and $\hat{\boldsymbol{\Lambda}}_k$ is distribution-free over the class of elliptical distributions. The latter statement can be proved by writing (16) in terms of the stochastic representation (8) and observing that the modular variate r_k cancels out in each summand.

Since the eigenvalues of $\boldsymbol{\Lambda}_{\mathrm{sgn}}$ and $\boldsymbol{\Lambda}$ are different, $\hat{\boldsymbol{\Lambda}}_k$ 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

$$\gamma = o(p) \text{ as } p \to \infty. \tag{17}$$

Theorem 2: Let $\mathbf{x} \sim \mathcal{E}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}, g)$. Then,

$$\boldsymbol{\Lambda}_{\mathrm{sgn}} = \boldsymbol{\Lambda} + O(\gamma).$$

Furthermore, if (17) holds, then $\boldsymbol{\Lambda}_{\mathrm{sgn}} = \boldsymbol{\Lambda} + o(\|\boldsymbol{\Lambda}\|_{\mathrm{F}})$.

Proof: See Appendix B.   □

The central condition (17) 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 (17) holds. This lemma is seen to hold in particular for the case when the eigenvalues of $\mathbf{\Lambda}$ are bounded as $p \to \infty$.

Lemma 1: If $\mathbf{\Lambda}$ is a shape matrix for which $\lambda_{\max}(\mathbf{\Lambda}) = O(p^{\tau/2})$, where $\tau < 1$, as $p \to \infty$, then (17) holds.

Proof: See Appendix C. □

Proposition 3: The following sequences, in p, of covariance models satisfy (17).

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

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

Proof: See Appendix C. □

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

Let us illustrate this result in the case that $\mathbf{\Sigma}$ has an AR(1) covariance structure. In this case $\gamma \to (1 + \varrho^2)/(1 - \varrho^2)$ (see Appendix C-A). For $\varrho = 0.5$, $\gamma \to 5/3 = O(1)$. From Theorem 2, we then have that the relative error $\|\mathbf{\Lambda}\|_{\mathrm{F}}^{-1}\|\mathbf{\Lambda}_{\mathrm{sgn}} - \mathbf{\Lambda}\|_{\mathrm{F}}$ is of the order $O(\gamma/\|\mathbf{\Lambda}\|_{\mathrm{F}}) = O(\sqrt{\gamma/p}) = O(p^{-1/2})$.

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

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

$$\frac{\mathbb{E}\left[\|\hat{\mathbf{\Lambda}} - \mathbf{\Lambda}_{\mathrm{SCM}}\|_{\mathrm{F}}^2\right]}{\|\mathbf{\Lambda}\|_{\mathrm{F}}^2} \xrightarrow{p \to \infty} \frac{2}{n},$$

where $\mathbf{\Lambda}_{\mathrm{SCM}} = \eta^{-1} \frac{1}{n} \sum_{i=1}^{n} \mathbf{x}_i \mathbf{x}_i^\top$ and $\eta = \mathrm{tr}(\mathbf{\Sigma})/p$, so that $\mathbb{E}[\mathbf{\Lambda}_{\mathrm{SCM}}] = \mathbf{\Lambda}$.

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

$$\frac{\mathbb{E}[\mathrm{tr}(\hat{\mathbf{\Lambda}}^2)]}{p} = \frac{p}{n} + \frac{n-1}{n} \frac{\mathrm{tr}(\mathbb{E}[\hat{\mathbf{\Lambda}}]^2)}{p}. \tag{18}$$

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

$$\mathrm{tr}(\mathbb{E}[\hat{\mathbf{\Lambda}}]^2)/p \to \gamma \quad \text{as } p \to \infty. \tag{19}$$

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

$$\hat{\gamma}_k = \frac{n_k}{n_k - 1}\left(\frac{\mathrm{tr}(\hat{\mathbf{\Lambda}}_k^2)}{p} - \frac{p}{n_k}\right). \tag{20}$$

In a high-dimensional setting, using the spatial median $\hat{\boldsymbol{\mu}}$ 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.,

$H_0 : \mathbf{\Sigma} \propto \mathbf{I}$), 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]:

$$\hat{\gamma}_k^* = \hat{\gamma}_k - p d_k, \tag{21}$$

where

$$d_k = \frac{1}{n_k^2} \cdot \left(2 - 2\frac{q_{2,k}}{q_{1,k}^2} + \left(\frac{q_{2,k}}{q_{1,k}^2}\right)^2\right)$$
$$+ \frac{1}{n_k^3} \cdot \left(8\frac{q_{2,k}}{q_{1,k}^2} - 6\left(\frac{q_{2,k}}{q_{1,k}^2}\right)^2 + 2\frac{q_{2,k}q_{3,k}}{q_{1,k}^5} - 2\frac{q_{3,k}}{q_{1,k}^3}\right)$$

and $q_{m,k} = (1/n_k)\sum_{i=1}^{n_k} \|\mathbf{x}_{i,k} - \hat{\boldsymbol{\mu}}_k\|^{-m}$. 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 $c_{ij} = p^{-1}\mathrm{tr}(\mathbf{\Sigma}_i\mathbf{\Sigma}_j)$, for $i \neq j$, of the matrix $\mathbf{C}$ 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 $\mathbf{C}$ and $\mathbf{\Delta}$ are thus $\hat{\mathbf{C}} = (\hat{c}_{ij})$ and $\hat{\mathbf{\Delta}} = \mathrm{diag}(\hat{\delta}_1, \ldots, \hat{\delta}_K)$, where

$$\hat{c}_{ij} = \begin{cases} \mathrm{tr}((\hat{\eta}_i\hat{\mathbf{\Lambda}}_i)(\hat{\eta}_j\hat{\mathbf{\Lambda}}_j))/p, & \text{for } i \neq j, \\ \hat{\gamma}_i^*\hat{\eta}_i^2, & \text{for } i = j, \end{cases} \tag{22}$$

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

$$\tilde{\mathbf{S}}(\mathbf{a}) = \sum_{j=1}^{K} a_j \mathbf{S}_j + a_I \mathbf{I}, \tag{23}$$

where $\mathbf{a} = (a_1, \ldots, a_K, a_I)^\top \in \mathbb{R}^{K+1}$. 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

$$\mathbf{a}_k^\star = \underset{a_j \geq 0,\, a_I \geq \epsilon_k}{\arg\min} \; \mathbb{E}\left[\|\tilde{\mathbf{S}}(\mathbf{a}) - \mathbf{\Sigma}_k\|_{\mathrm{F}}^2\right]$$

Algorithm 1: Linear Pooling of SCMs With Identity.

> **input** : Data $\mathcal{X}_1, \ldots, \mathcal{X}_K$ of the classes and $\epsilon_k > 0$, $k = 1, \ldots, K$.
>
> 1 Compute SCMs $\mathbf{S}_1, \ldots, \mathbf{S}_K$ of the classes.
> 2 Compute $\tilde{\mathbf{C}}$ and $\tilde{\boldsymbol{\Delta}}$ of (25) (estimate as in Section III).
> 3 Compute $\hat{\mathbf{A}} = (\hat{\mathbf{a}}_1 \cdots \hat{\mathbf{a}}_K) = (\tilde{\boldsymbol{\Delta}} + \tilde{\mathbf{C}})^{-1} \begin{pmatrix} \mathbf{C} & \boldsymbol{\eta} \end{pmatrix}^\top$.
> 4 **for** $k = 1$ **to** K **do**
> 5   **if** $\exists i \in \{1, \ldots, K\} : (\hat{\mathbf{a}}_k)_i < 0$ *or* $\hat{a}_{Ik} < \epsilon_k$ **then**
> 6   Set $\hat{\mathbf{a}}_k^\star$ as the solution of (24).
> 7   **else**
> 8 $\hat{\mathbf{a}}_k^\star \leftarrow \hat{\mathbf{a}}_k$.
> 9 $\hat{\boldsymbol{\Sigma}}_k \leftarrow \tilde{\mathbf{S}}(\hat{\mathbf{a}}_k^\star)$ of (23).
>
> **output** : $\hat{\boldsymbol{\Sigma}}_1, \ldots, \hat{\boldsymbol{\Sigma}}_K$.

---

can be solved via the strictly convex QP problem

$$\begin{aligned} \text{minimize } & \tfrac{1}{2}\mathbf{a}^\top(\tilde{\boldsymbol{\Delta}} + \tilde{\mathbf{C}})\mathbf{a} - \tilde{\mathbf{c}}_k^\top\mathbf{a} \\ \text{subject to } & a_j \geq 0, \ 1 \leq j \leq K, a_I \geq \epsilon_k, \end{aligned} \quad (24)$$

where

$$\tilde{\boldsymbol{\Delta}} = \text{diag}(\delta_1, \ldots, \delta_K, 0) \quad \text{and} \quad \tilde{\mathbf{C}} = \begin{pmatrix} \mathbf{C} & \boldsymbol{\eta} \\ \boldsymbol{\eta}^\top & 1 \end{pmatrix}, \quad (25)$$

and $\boldsymbol{\eta} = (\eta_k) \in \mathbb{R}^K$ is a vector consisting of *scales* (9). The unconstrained optimal solution for this case is $\mathbf{A}^\star = (\tilde{\boldsymbol{\Delta}} + \tilde{\mathbf{C}})^{-1} \begin{pmatrix} \mathbf{C} & \boldsymbol{\eta} \end{pmatrix}^\top$. The positive definiteness of $\tilde{\boldsymbol{\Delta}} + \tilde{\mathbf{C}}$ is shown in Appendix A. Algorithm 1 summarizes the procedure for computing the linearly pooled estimates of the class covariance matrices.

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 $\mathbf{1}^\top\mathbf{a} = 1$ 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 $(\mathbf{a} - \mathbf{B}^{-1}\mathbf{c}_k)^\top\mathbf{B}(\mathbf{a} - \mathbf{B}^{-1}\mathbf{c}_k) + \text{const.}$, where $\mathbf{B} = \boldsymbol{\Delta} + \mathbf{C}$. By introducing an auxiliary variable t and constraint $(\mathbf{a} - \mathbf{B}^{-1}\mathbf{c}_k)^\top\mathbf{B}(\mathbf{a} - \mathbf{B}^{-1}\mathbf{c}_k) \leq t$, 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 $\mathbf{a}$ and t:

$$\begin{aligned} \text{minimize } & t \\ \text{subject to } & \begin{pmatrix} t & (\mathbf{a} - \mathbf{B}^{-1}\mathbf{c}_k)^\top \\ \mathbf{a} - \mathbf{B}^{-1}\mathbf{c}_k & \mathbf{B}^{-1} \end{pmatrix} \succeq \mathbf{0} \\ & \mathbf{S}(\mathbf{a}) \succeq \mathbf{0}. \end{aligned} \quad (26)$$

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 $\mathbf{C}$ and $\boldsymbol{\Delta}$), the SDP solution will have a lower MSE. However, with estimation error in $\hat{\mathbf{C}}$ and $\hat{\mathbf{A}}$, 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 distributed (absolutely continuous) random vector $\mathbf{x} \in \mathbb{C}^p$ from the kth population has a density function up to a constant of the form

$$|\boldsymbol{\Sigma}_k|^{-1} h_k((\mathbf{x} - \boldsymbol{\mu}_k)^\mathsf{H}\boldsymbol{\Sigma}_k^{-1}(\mathbf{x} - \boldsymbol{\mu}_k)),$$

where $h_k : \mathbb{R}_{\geq 0} \to \mathbb{R}_{>0}$ is the density generator, $\boldsymbol{\mu} = \mathbb{E}[\mathbf{x}]$ is the mean vector and $\boldsymbol{\Sigma}_k = \mathbb{E}[(\mathbf{x} - \boldsymbol{\mu})(\mathbf{x} - \boldsymbol{\mu})^\mathsf{H}]$ denotes the Hermitian positive definite covariance matrix of $\mathbf{x}$. Above $(\cdot)^\mathsf{H}$ denotes the Hermitian (complex-conjugate) transpose. We denote this case by $\mathbf{x} \sim \mathbb{C}\mathcal{E}_p(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k, h_k)$. The definitions of the SCM (2) and the SSCM (16) stay unchanged except that in their definitions $(\cdot)^\top$ is replaced with $(\cdot)^\mathsf{H}$. Furthermore, the inner products between Hermitian matrices $(\mathbf{A} = \mathbf{A}^\mathsf{H})$ remain unchanged since $\text{tr}(\mathbf{A}\mathbf{A}^\mathsf{H}) = \text{tr}(\mathbf{A}^2)$. Hence, the definitions of the sphericity parameter γ_k as well as the scale η_k remain unchanged. The kurtosis of a complex marginal variable x_i of $\mathbf{x} = (x_i) \in \mathbb{C}^p$ is defined as

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

where $\mu_i = \mathbb{E}[x_i]$. The elliptical kurtosis is then

$$\kappa = (1/2)\text{kurt}(x_i).$$

The elliptical kurtosis κ_k of class k is estimated using the average sample kurtosis of the marginal variables

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

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 $\kappa^{\text{LB}} = -1/(p+1)$ [49]. In case (27) is less than κ^{LB}, we set $\hat{\kappa}_k = 0.99\kappa^{\text{LB}}$.

The matrix $\boldsymbol{\Delta}$ is estimated via (28) given in the next lemma, which is an extension of [9, Lemma 1] to the complex case.

Lemma 2: [50, Theorem 3] Let $\mathcal{X}_k = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\} \subset \mathbb{C}^p$ be an i.i.d. random sample from $\mathbb{C}\mathcal{E}_p(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k, h_k)$ with finite

fourth-order moments. Then, the scaled MSE of the SCM is

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

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 $\mathcal{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$, where $(\mathbf{x}_i)_j = x_{ij}$. Let $\mathbf{\Sigma}$ denote the covariance matrix and let $\mathbf{S}$ denote the SCM computed from $\mathcal{X}$. Multi-target shrinkage covariance matrix estimators are often defined by

$$\hat{\mathbf{\Sigma}}(\mathbf{a}) = a_0 \mathbf{S} + \sum_{k=1}^{K} a_k \mathbf{T}_k, \tag{29}$$

where $\mathbf{T}_k$, $k = 1, \ldots, K$, 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

$$L(\mathbf{a}; \mathbf{Q}, \mathbf{b}) = \mathbb{E}[\|\hat{\mathbf{\Sigma}}(\mathbf{a}) - \mathbf{\Sigma}\|_{\mathrm{F}}^2] = \mathbf{a}^\top \mathbf{Q} \mathbf{a} - 2\mathbf{a}^\top \mathbf{b} + \mathbb{E}[\|\mathbf{S} - \mathbf{\Sigma}\|_{\mathrm{F}}^2]$$

resulting in the constrained QP problem

$$\mathbf{a}^\star = \underset{\mathbf{a} \geq \mathbf{0}, \mathbf{1}^\top \mathbf{a} \leq 1}{\arg \min} \; L(\mathbf{a}; \mathbf{Q}, \mathbf{b}), \text{ where}$$

$$(\mathbf{Q})_{ij} = q_{ij} = \mathbb{E}[\mathrm{tr}\left((\mathbf{T}_i - \mathbf{S})(\mathbf{T}_j - \mathbf{S})\right)]$$

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

$$\hat{b}_i \equiv \hat{b} = \sum_{i,j} \frac{1}{n(n-1)} \sum_{s=1}^{n} \left(x_{si} x_{sj} - \frac{1}{n} \sum_{t=1}^{n} x_{ti} x_{tj} \right)^2,$$

where the latter was obtained by approximating $\mathbb{E}[\mathrm{tr}((\mathbf{S} - \mathbf{\Sigma})(\mathbf{T}_i - \mathbb{E}[\mathbf{T}_i]))] \approx 0$. 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

$$L_{\mathrm{CV}}(\mathbf{a}; \mathbf{Q}_{\mathrm{CV}}, \mathbf{b}_{\mathrm{CV}}) = \frac{1}{n} \sum_{i=1}^{n} \left\| a_0 \mathbf{S}_{-i} + \sum_{k=1}^{K} a_k \mathbf{T}_k - \mathbf{x}_i \mathbf{x}_i^\top \right\|_{\mathrm{F}}^2$$

$$= \mathbf{a}^\top \mathbf{Q}_{\mathrm{CV}} \mathbf{a} - 2\mathbf{a}^\top \mathbf{b}_{\mathrm{CV}} + \mathrm{const.},$$

where $\mathbf{x}_i$ is assumed to have zero mean, $\mathbf{S}_{-i} = \frac{1}{n} \sum_{j, j \neq i} \mathbf{x}_j \mathbf{x}_j^\top$ is the SCM computed without the ith sample, and $\mathbf{a} = (a_0, a_1, \ldots, a_K)^\top \geq \mathbf{0}$. They showed how the elements in $\mathbf{Q}_{\mathrm{CV}}$ and $\mathbf{b}_{\mathrm{CV}}$ 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.

Multi-target shrinkage estimators have also been proposed, for example in [22] from a Bayesian perspective and in [23] by regularizing the Gaussian likelihood function. However, in both methods the target matrices are assumed to be positive definite, and hence, SCMs cannot be used as targets when $n_k < p$. In [21] a multi-target shrinkage estimator of the form (29) was proposed for space-time adaptive processing (STAP) problems, where the reliable estimation of the loss function required additional prior information.

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 $\mathbf{S}$ and multiple positive definite symmetric target matrices $\{\mathbf{T}_m\}_{m=1}^{M}$. Our approach for multi-target shrinkage estimation is detailed in Algorithm 2. The idea is to generate artificial data sets $\{\mathcal{Y}_m\}_{m=1}^{M}$ with covariance matrices $\{\mathbf{T}_m\}_{m=1}^{M}$ so that $\mathcal{X}$ and the generated data sets $\{\mathcal{Y}_m\}_{m=1}^{M}$ are approximately mutually independent. Specifically, $\mathcal{Y}_m$ is conditionally independent of $\mathcal{X}$ given $\mathbf{T}_m$. Then, for the SCMs $\{\mathbf{S}_{T_m}\}_{m=1}^{M}$ computed from $\{\mathcal{Y}_m\}_{m=1}^{M}$, we make the approximations $\mathbb{E}[\mathrm{tr}(\mathbf{S}_{T_i} \mathbf{S}_{T_j})] \approx \mathrm{tr}(\mathbb{E}[\mathbf{S}_{T_i}] \mathbb{E}[\mathbf{S}_{T_j}]) = \mathrm{tr}(\mathbf{T}_i \mathbf{T}_j)$, for $i \neq j$, as well as $\mathbb{E}[\mathrm{tr}(\mathbf{S} \mathbf{S}_{T_j})] \approx \mathrm{tr}(\mathbb{E}[\mathbf{S}] \mathbb{E}[\mathbf{S}_{T_j}]) = \mathrm{tr}(\mathbf{\Sigma} \mathbf{T}_j)$. We use the zero mean MVN distribution to generate the data sets, i.e., $\mathcal{Y}_m \sim \mathcal{N}_p(\mathbf{0}, \mathbf{T}_m)$, $m = 1, \ldots, M$, each consisting of L samples. We can then apply the proposed method for pooling $\mathbf{S}$ and $\{\mathbf{S}_{T_m}\}_{m=1}^{M}$. We illustrate the usefulness of this method in a portfolio optimization problem in Section VII, where it compares well with the other multi-target methods.

Algorithm 2: Multi-Target Shrinkage Estimator.

input:　Data $\mathcal{X}$, targets $\{\mathbf{T}_m\}_{m=1}^M$, sample size L,
　　　　and lowerbound $\epsilon > 0$ for identity target.

1:　Generate i.i.d. samples $\mathcal{Y}_m \sim \mathcal{N}_p(\mathbf{0}, \mathbf{T}_m)$ for
　　$m = 1, \ldots, M$ each of size L.

2:　Compute SCM $\mathbf{S}$ from $\mathcal{X}$ and SCMs $\mathbf{S}_{T_1}, \ldots, \mathbf{S}_{T_M}$
　　from $\mathcal{Y}_1, \ldots, \mathcal{Y}_M$.

3:　Compute $\tilde{\mathbf{C}}$ and $\tilde{\boldsymbol{\Delta}}$ of (25) (estimate as in Section III).

4:　$\hat{\mathbf{a}}^\star \leftarrow \underset{\mathbf{a} \geq \mathbf{0}, a_I \geq \epsilon}{\arg\min} \frac{1}{2}\mathbf{a}^\top(\tilde{\boldsymbol{\Delta}} + \tilde{\mathbf{C}})\mathbf{a} - \tilde{\mathbf{c}}_k^\top\mathbf{a}$.

5:　$\hat{\boldsymbol{\Sigma}} \leftarrow \tilde{\mathbf{S}}(\hat{\mathbf{a}}^\star)$ of (23).

output: $\hat{\boldsymbol{\Sigma}}$.

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 $\boldsymbol{\Delta}$ and $\mathbf{C}$ and the estimated coefficients $\hat{\mathbf{A}}$. We denote the proposed linear pooling estimator (23), which includes shrinkage towards identity, by LINPOOL. The LINPOOL estimator with an additional convexity constraint on the coefficients ($\mathbf{1}^\top\mathbf{a} = 1$) is denoted by LINPOOL-C. For LINPOOL and LINPOOL-C, we set the lowerbound for the identity target in all of the simulations to $\epsilon_k = 10^{-8}$. In addition to the proposed methods, the results are reported for the multi-target shrinkage method LOOCV from [20], which uses a nonnegative linear combination of the SCMs and the identity matrix, and BARTZ from [19], which uses a convex combination of the SCMs and the identity matrix (see Section V for details).

A. Three Different Setups

In the first simulation, we considered two different covariance matrix structures: the AR(1), where $(\boldsymbol{\Sigma}_k)_{ij} = \sigma_k^2 \varrho_k^{|i-j|}$; and the CS structure, where $(\boldsymbol{\Sigma}_k)_{ij} = \sigma_k^2 \varrho_k$, for $i \neq j$ and $(\boldsymbol{\Sigma}_k)_{ij} = \sigma_k^2$ 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 $\sigma_k^2 = k$ and $(\varrho_1 = 0.3, \varrho_2 = 0.4, \varrho_3 = 0.5, \varrho_4 = 0.6)$.

Table I tabulates the normalized mean squared error,

$$\mathrm{NMSE}(\hat{\boldsymbol{\Sigma}}_k) = \mathrm{Ave}\|\hat{\boldsymbol{\Sigma}}_k - \boldsymbol{\Sigma}_k\|_{\mathrm{F}}^2 / \|\boldsymbol{\Sigma}_k\|_{\mathrm{F}}^2,$$

and total NMSE (sum of NMSEs of the classes) of the covariance matrix estimates for the three different setups over 1000 repetitions. Table I 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.

TABLE I
THE NORMALIZED MEAN SQUARED ERROR OVER 1000 REPETITIONS AND
STANDARD DEVIATION IN THE PARENTHESIS

|	class 1	class 2	class 3	class 4	total
AR(1)					
LOOCV	0.14 (0.06)	0.14 (0.01)	0.24 (0.06)	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: $\boldsymbol{\Sigma}_1$ and $\boldsymbol{\Sigma}_2$ are AR(1); $\boldsymbol{\Sigma}_3$ and $\boldsymbol{\Sigma}_4$ 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 (0.01)	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)
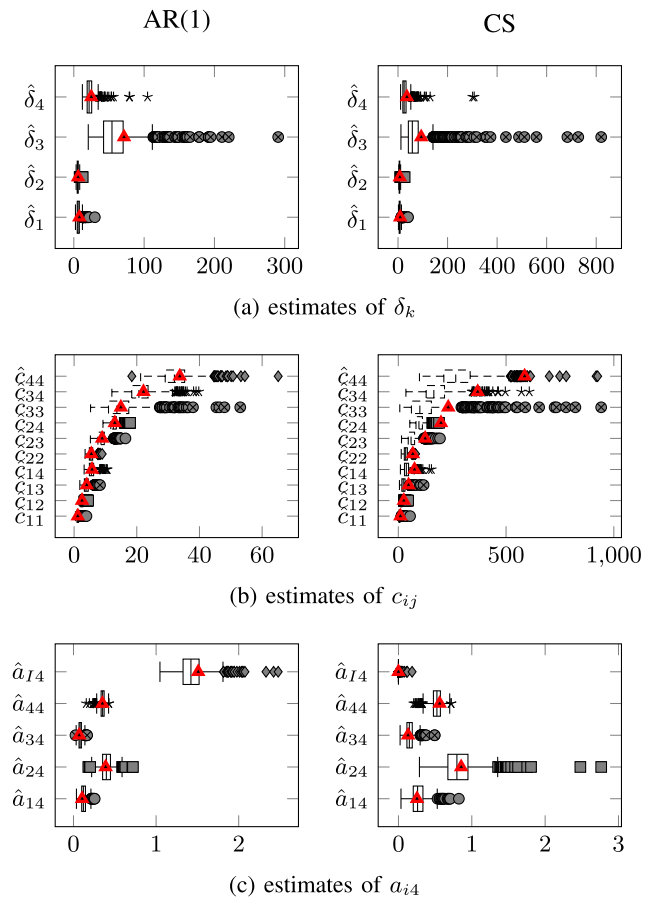
Fig. 1. Estimates of δ_k, c_{ij}, and a_{i4} (coefficients for $\hat{\boldsymbol{\Sigma}}_4$ of LINPOOL). *Left:* AR(1) setup. *Right:* CS setup. The red triangles denote the theoretical values.

Figure 1 displays boxplots of the estimates $\hat{\delta}_k$ and $\hat{c}_{ij}$ both for the AR(1) case as well as the CS case. The estimates of the optimal coefficients for LINPOOL for the fourth class $\hat{a}_{i4}$ are also shown. As can be seen from the boxplots, for the AR(1) case, the medians of the estimates were mostly correctly placed over the true values, which are denoted by the red triangles (▲). For the CS case, there was some significant bias in the estimation
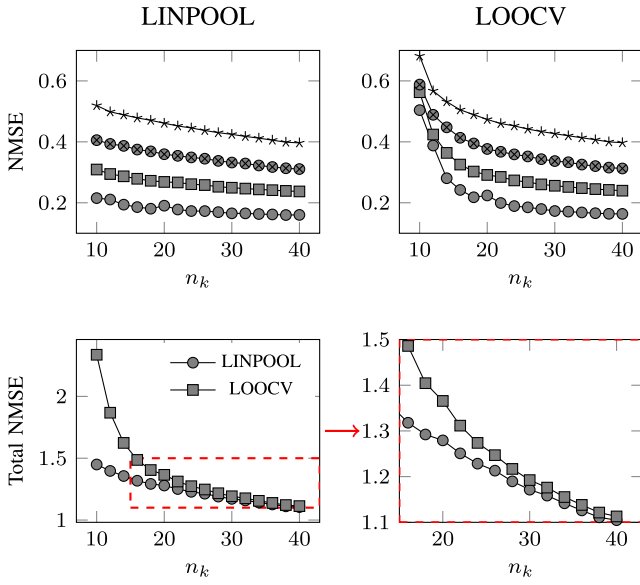
Fig. 2. NMSE as a function of sample size in the complex-valued AR(1) case with 4 classes. *Top:* NMSE of individual classes. *Bottom:* total combined NMSE.

Fig. 3. NMSE of $\hat{\boldsymbol{\Sigma}}_1$ as the number of classes increase.

of c_{ij}, which is due to the fact that the assumption (17) does not hold in this case as explained in Section III-C. Despite of this, the final coefficients $\hat{a}_{ij}$ were reasonable well estimated.

Here, it is good to note that, for the CS case, the coefficient corresponding to identity shrinkage $\hat{a}_{I4}$ is close to zero. When this happens, there is a possibility that (despite having a low MSE) the estimate is not well-conditioned resulting in high error when inverting the estimate. Therefore, in these cases (depending on the conditioning of the estimate) it can be useful to increase the lower bound ϵ_4. Generally, for class k, one could use $\epsilon_k = \alpha \eta_k$, where $\alpha \in [0, 1]$.

B. Increasing the Number of Classes

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 ϱ_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 sample sizes are equal with $n_k = 40$ for all classes k. 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
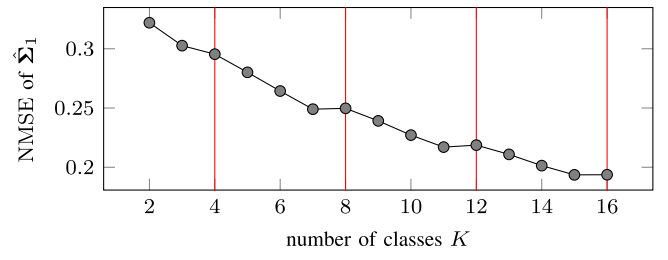
total number of observations is still relatively low and including the fourth class helps in reducing the variance of the estimate.

C. Complex-Valued Case

In the next simulation, the classes have an AR(1) covariance matrix structure, $(\boldsymbol{\Sigma}_k)_{ij} = \sigma_k^2 \varrho_k^{|i-j|}$ for $i \leq j$, and $(\boldsymbol{\Sigma}_k)_{ji} = (\boldsymbol{\Sigma}_k^*)_{ij}$, 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 for $n_k \in \{10, 15, \ldots 40\}$ 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 divident 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 $\mathbf{w} \in \mathbb{R}^p$ (verifying the constraint $\mathbf{1}^\top \mathbf{w} = 1$) whose elements describe the fraction of the total wealth invested in each of the p stocks. We considered two different portfolios. First, we considered the *global minimum variance portfolio* (GMVP) in which one seeks a portfolio $\mathbf{w}$ that minimizes the risk (variance). The optimization problem is thus

$$\underset{\mathbf{w} \in \mathbb{R}^p}{\text{minimize}} \; \mathbf{w}^\top \boldsymbol{\Sigma} \mathbf{w} \quad \text{subject to } \mathbf{1}^\top \mathbf{w} = 1. \quad (30)$$

The well-known solution is $\mathbf{w}^\star = \frac{\boldsymbol{\Sigma}^{-1} \mathbf{1}}{\mathbf{1}^\top \boldsymbol{\Sigma}^{-1} \mathbf{1}}$, where $\boldsymbol{\Sigma}$ is the covariance matrix of the stock returns. We also considered a *constrained* portfolio, where the coefficients are constrained to be within the range $0 \leq w_i \leq 0.1$, 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 $\mathbf{0} \leq \mathbf{w} \leq 0.1 \cdot \mathbf{1}$, which results in a QP problem.

In the simulation, the covariance matrix $\mathbf{\Sigma}$ 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 $\mathcal{Y}_C \sim \mathcal{N}(\mathbf{0}, \mathbf{T}_C)$ and $\mathcal{Y}_F \sim \mathcal{N}(\mathbf{0}, \mathbf{T}_F)$ and estimated the coefficients for the proposed LINPOOL estimator $\tilde{\mathbf{S}}(\mathbf{a}) = a_1\mathbf{S} + a_2\mathbf{S}_F + a_3\mathbf{S}_C + a_I\mathbf{I}$, 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 $a_I \geq \epsilon = 10^{-8}$. 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 the 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
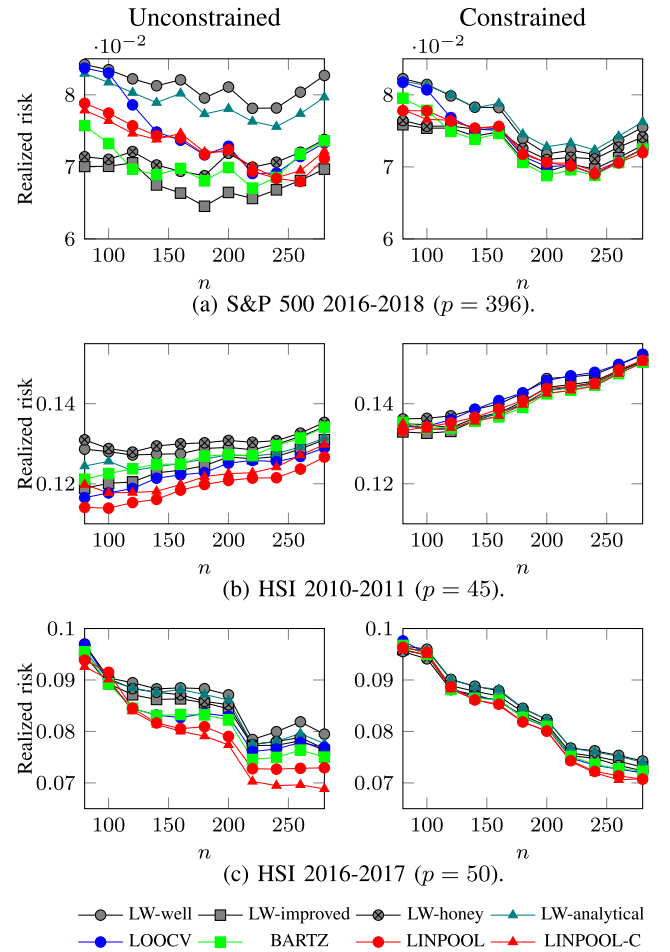

Fig. 4. Annualized realized GMVP risk achieved out-of-sample for different covariance matrix estimators and different training window lengths n. *Left:* unconstrained portfolio. *Right:* constrained portfolio (nonnegative weights and maximum single asset weight 0.1).

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 $n > 120$.

VIII. CONCLUSION

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 https://github.com/EliasRaninen.

APPENDIX A
PROOF OF THEOREM 1

Write $f(\mathbf{a}) = \mathbb{E}[\|\mathbf{S}(\mathbf{a}) - \boldsymbol{\Sigma}_k\|_{\mathrm{F}}^2]$ for the objective function. By expanding the expression for the squared error, we get

$$\|\mathbf{S}(\mathbf{a}) - \boldsymbol{\Sigma}_k\|_{\mathrm{F}}^2 = \mathrm{tr}\left(\left(\sum_{i=1}^K a_i \mathbf{S}_i - \boldsymbol{\Sigma}_k\right)\left(\sum_{j=1}^K a_j \mathbf{S}_j - \boldsymbol{\Sigma}_k\right)^\top\right)$$

$$= \sum_{i=1}^K \sum_{j=1}^K a_i a_j \, \mathrm{tr}(\mathbf{S}_i \mathbf{S}_j) - 2\sum_{j=1}^K a_j \, \mathrm{tr}(\mathbf{S}_j \boldsymbol{\Sigma}_k) + \mathrm{tr}(\boldsymbol{\Sigma}_k^2).$$

Taking the expectation and scaling by $1/p$ gives

$$(1/p)f(\mathbf{a}) = \mathbf{a}^\top \mathbf{B}\mathbf{a} - 2\mathbf{c}_k^\top \mathbf{a} + c_{kk},$$

where

$$\mathbf{B} = (b_{ij}) \in \mathbb{R}^{K\times K}, \; b_{ij} = \mathbb{E}[\mathrm{tr}(\mathbf{S}_i \mathbf{S}_j)]/p,$$

$$\mathbf{C} = (c_{ij}) \in \mathbb{R}^{K\times K}, \; c_{ij} = \mathrm{tr}(\boldsymbol{\Sigma}_i \boldsymbol{\Sigma}_j)/p,$$

and $\mathbf{c}_k$ corresponds to the kth column of $\mathbf{C}$. For $i \neq j$,

$$b_{ij} = p^{-1}\,\mathbb{E}[\mathrm{tr}(\mathbf{S}_i \mathbf{S}_j)] = p^{-1}\,\mathrm{tr}(\mathbb{E}[\mathbf{S}_i]\,\mathbb{E}[\mathbf{S}_j])$$

$$= p^{-1}\,\mathrm{tr}(\boldsymbol{\Sigma}_i \boldsymbol{\Sigma}_j) = c_{ij}.$$

Using that $\mathbb{E}[\mathrm{tr}(\mathbf{S}_i^2)] = \mathrm{MSE}(\mathbf{S}_i) + \mathrm{tr}(\boldsymbol{\Sigma}_i^2)$, we get for $i = j$,

$$b_{ii} = p^{-1}\,\mathbb{E}[\mathrm{tr}(\mathbf{S}_i^2)] = \delta_i + c_{ii},$$

and so $\mathbf{B} = \boldsymbol{\Delta} + \mathbf{C}$. By definition $\boldsymbol{\Delta}$ is symmetric and positive definite. Also, $\mathbf{C}$ is symmetric and positive semidefinite. This follows since for any $\mathbf{m} \in \mathbb{R}^K$, $\mathbf{m}^\top \mathbf{C}\mathbf{m} = \sum_{i=1}^K \sum_{j=1}^K m_i m_j p^{-1}\,\mathrm{tr}(\boldsymbol{\Sigma}_i \boldsymbol{\Sigma}_j) = \mathrm{tr}(\boldsymbol{\Sigma}_*^2) \geq 0$, where $\boldsymbol{\Sigma}_* = p^{-1/2}\sum_{j=1}^K m_j \boldsymbol{\Sigma}_j$. Hence $\boldsymbol{\Delta} + \mathbf{C} \succ \mathbf{0}$ is invertible.

Regarding the extension discussed in Section IV, we now show the positive definiteness of $\tilde{\boldsymbol{\Delta}} + \tilde{\mathbf{C}}$. Since we know that $\boldsymbol{\Delta} + \mathbf{C} \succ \mathbf{0}$, due to the properties of the Schur complement [60, A.5.5], it holds that

$$\tilde{\boldsymbol{\Delta}} + \tilde{\mathbf{C}} = \begin{pmatrix} \boldsymbol{\Delta} + \mathbf{C} & \boldsymbol{\eta} \\ \boldsymbol{\eta}^\top & 1 \end{pmatrix} \succ \mathbf{0} \Leftrightarrow \boldsymbol{\Delta} + \mathbf{C} - \boldsymbol{\eta}\boldsymbol{\eta}^\top \succ \mathbf{0}.$$

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

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

$$= \sum_k m_k^2 \delta_k + \mathrm{tr}(\tilde{\boldsymbol{\Sigma}}_*^2) > 0,$$

where $\tilde{\boldsymbol{\Sigma}}_* = p^{-1/2}\sum_{j=1}^K m_j(\boldsymbol{\Sigma}_j - \eta_j \mathbf{I})$. So, $\tilde{\boldsymbol{\Delta}} + \tilde{\mathbf{C}} \succ \mathbf{0}$. $\square$

APPENDIX B
PROOF OF THEOREM 2

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

Lemma 3:

$$\mathbb{E}\left[\boldsymbol{\Lambda}^{1/2}\mathbf{u}\mathbf{u}^\top\boldsymbol{\Lambda}^{1/2}\mathbf{u}^\top\boldsymbol{\Lambda}\mathbf{u}\right] = p^{-1}(p+2)^{-1}\left(\mathrm{tr}(\boldsymbol{\Lambda})\boldsymbol{\Lambda} + 2\boldsymbol{\Lambda}^2\right).$$

Proof:

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

$$= \boldsymbol{\Lambda}^{1/2}\left(\sum_l \Lambda_{ll}\,\mathbb{E}[u_l^2\mathbf{u}\mathbf{u}^\top] + \sum_{k\neq l}\Lambda_{kl}\,\mathbb{E}[u_k u_l \mathbf{u}\mathbf{u}^\top]\right)\boldsymbol{\Lambda}^{1/2}. \tag{31}$$

Here (see [61, 3.1.2]),

$$(\mathbb{E}[u_l^2\mathbf{u}\mathbf{u}^\top])_{jj} = \begin{cases} \mathbb{E}[u_l^4] = 3p^{-1}(p+2)^{-1}, & l = j, \\ \mathbb{E}[u_l^2 u_j^2] = p^{-1}(p+2)^{-1}, & l \neq j, \end{cases}$$

and $(\mathbb{E}[u_l^2\mathbf{u}\mathbf{u}^\top])_{ij} = 0$, for $i \neq j$. Hence, $\mathbb{E}[u_l^2\mathbf{u}\mathbf{u}^\top] = p^{-1}(p+2)^{-1}(\mathbf{I} + 2\mathbf{e}_l\mathbf{e}_l^\top)$. Thus,

$$\sum_l \Lambda_{ll}\,\mathbb{E}[u_l^2\mathbf{u}\mathbf{u}^\top]$$

$$= p^{-1}(p+2)^{-1}\left(\left(\sum_l \Lambda_{ll}\right)\mathbf{I} + 2\sum_l \Lambda_{ll}\mathbf{e}_l\mathbf{e}_l^\top\right)$$

$$= p^{-1}(p+2)^{-1}(\mathrm{tr}(\boldsymbol{\Lambda})\mathbf{I} + 2\,\mathrm{diag}(\boldsymbol{\Lambda})). \tag{32}$$

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

$$(\mathbb{E}[u_k u_l \mathbf{u}\mathbf{u}^\top])_{mn}$$

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

Hence, we have

$$\sum_{k\neq l}\Lambda_{kl}\,\mathbb{E}[u_k u_l \mathbf{u}\mathbf{u}^\top] = 2p^{-1}(p+2)^{-1}(\boldsymbol{\Lambda} - \mathrm{diag}(\boldsymbol{\Lambda})) \tag{33}$$

and the result follows by substituting (32) and (33) into (31).

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 $x > 0$, $x^{-1} \geq 2 - x$, we have

$$\boldsymbol{\Lambda}_{\mathrm{sgn}} = p\,\mathbb{E}\left[\frac{\mathbf{x}\mathbf{x}^\top}{\|\mathbf{x}\|^2}\right] = p\,\mathbb{E}\left[\frac{\boldsymbol{\Lambda}^{1/2}\mathbf{u}\mathbf{u}^\top\boldsymbol{\Lambda}^{1/2}}{\mathbf{u}^\top\boldsymbol{\Lambda}\mathbf{u}}\right]$$

$$\succeq p\,\mathbb{E}\left[\boldsymbol{\Lambda}^{1/2}\mathbf{u}\mathbf{u}^\top\boldsymbol{\Lambda}^{1/2}(2 - \mathbf{u}^\top\boldsymbol{\Lambda}\mathbf{u})\right]$$

$$= 2p\,\mathbb{E}[\boldsymbol{\Lambda}^{1/2}\mathbf{u}\mathbf{u}^\top\boldsymbol{\Lambda}^{1/2}] - p\,\mathbb{E}[\boldsymbol{\Lambda}^{1/2}\mathbf{u}\mathbf{u}^\top\boldsymbol{\Lambda}^{1/2}\mathbf{u}^\top\boldsymbol{\Lambda}\mathbf{u}]$$

$$= 2\mathbf{\Lambda} - \frac{1}{p+2}\left(\mathrm{tr}(\mathbf{\Lambda})\mathbf{\Lambda} + 2\mathbf{\Lambda}^2\right) = \mathbf{\Lambda} - \frac{2(\mathbf{\Lambda}^2 - \mathbf{\Lambda})}{p+2}.$$
$$(34)$$

By scaling both sides of (34) by $\epsilon = \|\mathbf{\Lambda}\|_{\mathrm{F}}^{-1} = 1/\sqrt{p\gamma}$, the first term will have unity norm ($\|\epsilon\mathbf{\Lambda}\|_{\mathrm{F}} = 1$). Let us then consider the second term on the right-hand side. Its trace is

$$\epsilon\frac{2\,\mathrm{tr}(\mathbf{\Lambda}^2 - \mathbf{\Lambda})}{p+2} = \epsilon\frac{2p}{p+2}(\gamma - 1)$$
$$\leq 2\left(\sqrt{\frac{\gamma}{p}} - \sqrt{\frac{1}{p\gamma}}\right) = O\left(\sqrt{\frac{\gamma}{p}}\right), \quad (35)$$

and its norm is

$$\epsilon\left\|\frac{2(\mathbf{\Lambda}^2 - \mathbf{\Lambda})}{p+2}\right\|_{\mathrm{F}} \leq (2\epsilon/p)(\|\mathbf{\Lambda}^2\|_{\mathrm{F}} + \|\mathbf{\Lambda}\|_{\mathrm{F}})$$
$$\leq (2/p)(\|\mathbf{\Lambda}\|_{\mathrm{F}} + 1) = O(\sqrt{\gamma/p}). \quad (36)$$

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 ϵ, we can use (35) and (36) as well as the property that for any $\mathbf{A} \succeq \mathbf{0}$, $\mathrm{tr}(\mathbf{A}) \geq \|\mathbf{A}\|_{\mathrm{F}}$ to get

$$O(\sqrt{\gamma/p}) \geq \left\|(\mathbf{\Lambda}_{\mathrm{sgn}} - \mathbf{\Lambda})/\|\mathbf{\Lambda}\|_{\mathrm{F}} + O(\sqrt{\gamma/p})\right\|_{\mathrm{F}} \geq 0,$$

which implies that $\mathbf{\Lambda}_{\mathrm{sgn}} = \mathbf{\Lambda} + O(\gamma)$. Furthermore, if $\gamma = o(p)$, then $O(\sqrt{\gamma/p}) = o(1)$ implying $\mathbf{\Lambda}_{\mathrm{sgn}} = \mathbf{\Lambda} + o(\|\mathbf{\Lambda}\|_{\mathrm{F}})$. $\square$

APPENDIX C
PROOFS OF LEMMA 1 AND PROPOSITION 3

Let $\lambda_1 = \lambda_{\max}(\mathbf{\Lambda})$. Assume that as $p \to \infty$, $\lambda_1 = O(p^{\tau/2})$, where $\tau < 1$. Then

$$\gamma = \frac{\mathrm{tr}(\mathbf{\Lambda}^2)}{p} \leq \frac{p\lambda_1^2}{p} = \lambda_1^2 = O(p^{\tau}) = o(p).$$

Particularly, $\gamma = o(p)$ if $\lambda_1 = O(1)$.

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

A. Sphericity of the AR(1) Covariance Matrix

The shape matrix of the AR(1) covariance matrix with parameter ϱ ($|\varrho| < 1$) has p number of ones on the main diagonal, $2(p-1)$ number of ϱ on the first diagonals above and below the main diagonal, and $2(p-2)$ number of ϱ^2 on the second diagonals above and below the main diagonal, and so on. That gives,

$$\gamma_{\mathrm{AR}(1)} = 1 + \frac{2}{p}\sum_{i=1}^{p-1}(p-i)\varrho^{2i} = 2\sum_{i=0}^{p-1}(\varrho^2)^i - \frac{2}{p}\sum_{i=1}^{p-1}i(\varrho^2)^i - 1.$$

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

$$\gamma_{\mathrm{AR}(1)} = 2\frac{1 - (\varrho^2)^p}{1 - \varrho^2} - \frac{2}{p}\frac{(p-1)(\varrho^2)^{p+1} - p(\varrho^2)^p + \varrho^2}{(1 - \varrho^2)^2} - 1$$

$$= \frac{p - p\varrho^4 - 2\varrho^2 + 2(\varrho^2)^{p+1}}{p(\varrho^2 - 1)^2}.$$

As $p \to \infty$, $\gamma_{\mathrm{AR}(1)} \to (1 + \varrho^2)/(1 - \varrho^2) = O(1) = o(p)$.

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 ϱ on the off-diagonals while rest of the off-diagonals are zero. This implies that $\gamma_{\mathrm{1B}} = 1 + 2(1 - 1/p)\varrho^2$. As $p \to \infty$, $\gamma_{\mathrm{1B}} \to 1 + 2\varrho^2 = O(1) = o(p)$.

C. Sphericity of the Spiked Covariance Matrix

Let $\mathbf{\Sigma} = \mathbf{\Sigma}_r + \alpha\mathbf{I}$, where $\mathbf{\Sigma}_r$ has rank $r \leq p$. Let $\alpha = \beta\eta_r$, where $\eta_r = \mathrm{tr}(\mathbf{\Sigma}_r)/p$. Let $\gamma_r = p\,\mathrm{tr}(\mathbf{\Sigma}_r^2)/\mathrm{tr}(\mathbf{\Sigma}_r)^2 = \mathrm{tr}(\mathbf{\Sigma}_r^2)/(p\eta_r^2)$. Then, $\mathrm{tr}(\mathbf{\Sigma})^2 = p^2\eta_r^2(1 + \beta)^2$ and

$$\mathrm{tr}(\mathbf{\Sigma}^2) = \mathrm{tr}\left((\mathbf{\Sigma}_r + \beta\eta_r\mathbf{I})^2\right) = \mathrm{tr}(\mathbf{\Sigma}_r^2) + \beta^2 p\eta_r^2 + 2\beta p\eta_r^2$$

$$= p\eta_r^2(\gamma_r + \beta^2 + 2\beta) = p\eta_r^2(\gamma_r - 1 + (\beta + 1)^2).$$

Then by computing $\gamma = p\,\mathrm{tr}(\mathbf{\Sigma}^2)/\mathrm{tr}(\mathbf{\Sigma})^2$, we get

$$\gamma = \frac{\gamma_r - 1}{(\beta + 1)^2} + 1 = \frac{\gamma_r - 1}{(\alpha/\eta_r + 1)^2} + 1 \leq \frac{\eta_r^2\gamma_r}{\alpha^2} + 1 \leq \frac{r\lambda_1^2}{p\alpha^2} + 1,$$

where we used that the rank of $\mathbf{\Sigma}_r$ is r and $\eta_r^2\gamma_r = \mathrm{tr}(\mathbf{\Sigma}_r^2)/p \leq r\lambda_1^2/p$, where $\lambda_1 = \lambda_{\max}(\mathbf{\Sigma}_r)$. Therefore, $\gamma = o(p)$, if $\frac{r\lambda_1^2}{p\alpha^2} = o(p)$, which includes the cases

$$\begin{cases} \frac{\lambda_1}{\alpha} = O(p^{\tau/2}), \\ \frac{\lambda_1}{\alpha} = O(p^{(\tau+1)/2}) \text{ and } r = O(1), \end{cases} \quad \text{where } \tau < 1.$$

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 $p(p-1)$ number of ϱ on the off-diagonals. Thus, $\gamma_{\mathrm{CS}} = 1 + (p-1)\varrho^2 = O(p)$. $\square$

APPENDIX D
PROOF OF PROPOSITION 4

Expanding the mean squared distance gives

$$\mathbb{E}[\|\hat{\mathbf{\Lambda}} - \mathbf{\Lambda}_{\mathrm{SCM}}\|_{\mathrm{F}}^2] = \mathbb{E}[\mathrm{tr}(\hat{\mathbf{\Lambda}}^2)] + \mathbb{E}[\mathrm{tr}(\mathbf{\Lambda}_{\mathrm{SCM}}^2)] - 2\,\mathbb{E}[\mathrm{tr}(\hat{\mathbf{\Lambda}}\mathbf{\Lambda}_{\mathrm{SCM}})].$$
$$(37)$$

The first term is

$$\mathbb{E}[\mathrm{tr}(\hat{\mathbf{\Lambda}}^2)] = \mathbb{E}\left[\mathrm{tr}\left(\left(\frac{p}{n}\sum_i\frac{\mathbf{x}_i\mathbf{x}_i^{\top}}{\|\mathbf{x}_i\|^2}\right)^2\right)\right]$$

$$= \frac{p^2}{n^2}\,\mathbb{E}\left[\mathrm{tr}\left(\sum_i\frac{\mathbf{x}_i\mathbf{x}_i^{\top}\mathbf{x}_i\mathbf{x}_i^{\top}}{\|\mathbf{x}_i\|^4} + \sum_{i \neq j}\frac{\mathbf{x}_i\mathbf{x}_i^{\top}\mathbf{x}_j\mathbf{x}_j^{\top}}{\|\mathbf{x}_i\|^2\|\mathbf{x}_j\|^2}\right)\right]$$

$$= \frac{p^2}{n} + \frac{n-1}{n}\,\mathrm{tr}(\mathbb{E}[\hat{\mathbf{\Lambda}}]^2). \quad (38)$$

Now, let $\mathbf{S} = \frac{1}{n}\sum_{i=1}^n\mathbf{x}_i\mathbf{x}_i^{\top}$. Since $\mathbf{S}$ is unbiased, using Wishart theory, we have $\mathrm{MSE}(\mathbf{S}) = \mathrm{tr}(\mathrm{var}(\mathrm{vec}(\mathbf{S}))) =$

$(1/n) \operatorname{tr}((\mathbf{I} + \mathbf{K})(\boldsymbol{\Sigma} \otimes \boldsymbol{\Sigma})) = (1/n)(\operatorname{tr}(\boldsymbol{\Sigma})^2 + \operatorname{tr}(\boldsymbol{\Sigma}^2))$, where $\mathbf{K} = \sum_{i=1}^{p} \sum_{j=1}^{p} \mathbf{e}_i \mathbf{e}_j^\top \otimes \mathbf{e}_j \mathbf{e}_i^\top$ is the commutation matrix [62] and $\otimes$ denotes the Kronecker product. Using that $\mathbb{E}[\operatorname{tr}(\mathbf{S}^2)] = \operatorname{MSE}(\mathbf{S}) + \operatorname{tr}(\boldsymbol{\Sigma}^2)$,

$$\mathbb{E}[\operatorname{tr}(\boldsymbol{\Lambda}_{\mathrm{SCM}}^2)] = \frac{p^2}{\operatorname{tr}(\boldsymbol{\Sigma})^2} \mathbb{E}[\operatorname{tr}(\mathbf{S}^2)] = \frac{p^2}{n} + \left(\frac{n+1}{n}\right) \operatorname{tr}(\boldsymbol{\Lambda}^2). \tag{39}$$

The last term of (37) is

$$\mathbb{E}[\operatorname{tr}(\hat{\boldsymbol{\Lambda}} \boldsymbol{\Lambda}_{\mathrm{SCM}})] = \mathbb{E}\left[\operatorname{tr}\left(\frac{p}{n}\sum_i \frac{\mathbf{x}_i \mathbf{x}_i^\top}{\|\mathbf{x}_i\|^2} \frac{p}{n\operatorname{tr}(\boldsymbol{\Sigma})}\sum_j \mathbf{x}_j \mathbf{x}_j^\top\right)\right]$$

$$= \frac{p^2}{n^2 \operatorname{tr}(\boldsymbol{\Sigma})} \operatorname{tr}\left(\sum_i \mathbb{E}[\mathbf{x}_i \mathbf{x}_i^\top] + \sum_{i \neq j} \mathbb{E}\left[\frac{\mathbf{x}_i \mathbf{x}_i^\top}{\|\mathbf{x}_i\|^2}\right]\mathbb{E}[\mathbf{x}_j \mathbf{x}_j^\top]\right)$$

$$= \frac{p^2}{n} + \frac{n-1}{n} \operatorname{tr}(\mathbb{E}[\hat{\boldsymbol{\Lambda}}]\boldsymbol{\Lambda}). \tag{40}$$

Now, substituting (38), (39), and (40) into (37) gives

$$\frac{n-1}{n}\left(\operatorname{tr}(\mathbb{E}[\hat{\boldsymbol{\Lambda}}]^2) - 2\operatorname{tr}(\mathbb{E}[\hat{\boldsymbol{\Lambda}}]\boldsymbol{\Lambda})\right) + \frac{n+1}{n}\operatorname{tr}(\boldsymbol{\Lambda}^2).$$

Dividing this by $\|\boldsymbol{\Lambda}\|_{\mathrm{F}}^2$ and applying assumption $\gamma = o(p)$ and Theorem 2, we get the result $\mathbb{E}[\|\hat{\boldsymbol{\Lambda}} - \boldsymbol{\Lambda}_{\mathrm{SCM}}\|_{\mathrm{F}}^2]/\|\boldsymbol{\Lambda}\|_{\mathrm{F}}^2 \xrightarrow{p \to \infty} \frac{2}{n}$. $\square$

REFERENCES

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

[2] J. Schäfer and K. Strimmer, "A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics," *Stat. Appl. Genet. Mol. Biol.*, vol. 4, no. 1, pp. 1–30, 2005.

[3] T. Zhang, A. Wiesel, and M. S. Greco, "Multivariate generalized Gaussian distribution: Convexity and graphical models," *IEEE Trans. Signal Process.*, vol. 61, no. 16, pp. 4141–4148, Aug. 2013.

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

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

[6] O. Besson and Y. I. Abramovich, "Regularized covariance matrix estimation in complex elliptically symmetric distributions using the expected likelihood approach 2: The under-sampled case," *IEEE Trans. Signal Process.*, vol. 61, no. 23, pp. 5819–5829, Dec. 2013.

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

[8] O. Ledoit and M. Wolf, "Honey, I shrunk the sample covariance matrix," *J. Portfolio Manage.*, vol. 30, no. 4, pp. 110–119, Jul. 2004.

[9] E. Ollila and E. Raninen, "Optimal shrinkage covariance matrix estimation under random sampling from elliptical distributions," *IEEE Trans. Signal Process.*, vol. 67, no. 10, pp. 2707–2719, May 2019.

[10] 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.

[11] J. Li, J. Zhou, and B. Zhang, "Estimation of large covariance matrices by shrinking to structured target in normal and non-normal distributions," *IEEE Access*, vol. 6, pp. 2158–2169, 2018.

[12] M. Steiner and K. Gerlach, "Fast converging adaptive processor or a structured covariance matrix," *IEEE Trans. Aerosp. Electron. Syst.*, vol. 36, no. 4, pp. 1115–1126, Oct. 2000.

[13] 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 Trans. Signal Process.*, vol. 61, no. 23, pp. 5807–5818, Dec. 2013.

[14] A. Aubry, A. D. Maio, and L. Pallotta, "A geometric approach to covariance matrix estimation and its applications to radar problems," *IEEE Trans. Signal Process.*, vol. 66, no. 4, pp. 907–922, Feb. 2018.

[15] A. De Maio, L. Pallotta, J. Li, and P. Stoica, "Loading factor estimation under affine constraints on the covariance eigenvalues with application to radar target detection," *IEEE Trans. Aerosp. Electron. Syst.*, vol. 55, no. 3, pp. 1269–1283, Jun. 2019.

[16] Y. Ikeda, T. Kubokawa, and M.S. Srivastava, "Comparison of linear shrinkage estimators of a large covariance matrix in normal and non-normal distributions," *Comput. Statist. Data Anal.*, vol. 95, pp. 95–108, 2016.

[17] Z. Halbe, M. Bortman, and M. Aladjem, "Regularized mixture density estimation with an analytical setting of shrinkage intensities," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 24, no. 3, pp. 460–470, Mar. 2013.

[18] T. Lancewicki and M. Aladjem, "Multi-target shrinkage estimation for covariance matrices," *IEEE Trans. Signal Process.*, vol. 62, no. 24, pp. 6380–6390, Dec. 2014.

[19] D. Bartz, J. Höhne, and K.-R. Müller, "Multi-target shrinkage," Dec. 2014, *arXiv: 1412.2041 [stat.ME]*.

[20] J. Tong, R. Hu, J. Xi, Z. Xiao, Q. Guo, and Y. Yu, "Linear shrinkage estimation of covariance matrices using low-complexity cross-validation," *Signal Process.*, vol. 148, pp. 223–233, 2018.

[21] M. Riedl and L. C. Potter, "Multimodel shrinkage for knowledge-aided space-time adaptive processing," *IEEE Trans. Aerosp. Electron. Syst.*, vol. 54, no. 5, pp. 2601–2610, Oct. 2018.

[22] H. Gray, G. G. R. Leday, C. A. Vallejos, and S. Richardson, "Shrinkage estimation of large covariance matrices using multiple shrinkage targets," Sep. 2018, *arXiv: 1809.08024*.

[23] B. Zhang, J. Zhou, and J. Li, "Improved covariance matrix estimators by multi-penalty regularization," in *Proc. 22nd Int. Conf. Inf.*, Ottawa, ON, Canada, 2019, pp. 1–7.

[24] O. Besson, "Maximum likelihood covariance matrix estimation from two possibly mismatched data sets," *Signal Process.*, vol. 167, 2020, Art. no 107285.

[25] A. Coluccia, "Robust opportunistic inference from non-homogeneous distribution-free measurements," *IEEE Trans. Signal Process.*, vol. 64, no. 15, pp. 3945–3954, Aug. 2016.

[26] T. Greene and W. S. Rayens, "Partially pooled covariance matrix estimation in discriminant analysis," *Commun. Statist. - Theory Methods*, vol. 18, no. 10, pp. 3679–3702, 1989.

[27] W. Rayens and T. Greene, "Covariance pooling and stabilization for classification," *Comput. Statist. Data Anal.*, vol. 11, no. 1, pp. 17–42, 1991.

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

[29] E. Raninen and E. Ollila, "Optimal pooling of covariance matrix estimates across multiple classes," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2018, pp. 4224–4228.

[30] E. Raninen and E. Ollila, "Coupled regularized sample covariance matrix estimator for multiple classes," *IEEE Trans. Signal Process.*, vol. 69, pp. 5681–5692, Oct. 2021.

[31] A. Krasoulis, K. Nazarpour, and S. Vijayakumar, "Use of regularized discriminant analysis improves myoelectric hand movement classification," in *Proc. 8th Int. IEEE/EMBS Conf. Neural Eng.*, 2017, pp. 395–398.

[32] J. Lu, K. Plataniotis, and A. Venetsanopoulos, "Regularized discriminant analysis for the small sample size problem in face recognition," *Pattern Recognit. Lett.*, vol. 24, no. 16, pp. 3079–3087, Dec. 2003.

[33] J. Nocedal and S. J. Wright, *Numerical Optimization*. New York, NY, USA: Springer, 2006.

[34] M. K. Kozlov, S. P. Tarasov, and L. G. Khachiyan, "The polynomial solvability of convex quadratic programming," *USSR Comput. Math. Math. Phys.*, vol. 20, no. 5, pp. 223–228, Jan. 1980.

[35] L. A. Goodman, "A simple method for improving some estimators," *Ann. Math. Statist.*, vol. 24, no. 1, pp. 114–117, Mar. 1953.

[36] D. T. Searls and P. Intarapanich, "A note on an estimator for the variance that utilizes the kurtosis," *Amer. Statistician*, vol. 44, no. 4, pp. 295–296, Nov. 1990.

[37] M. Bilodeau and D. Brenner, *Theory of Multivariate Statistics*. New York, NY, USA: Springer-Verlag, 1999.

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

[39] D. N. Joanes and C. A. Gill, "Comparing measures of sample skewness and kurtosis," *J. Roy. Stat. Society: Ser. D. (Statistician)*, vol. 47, no. 1, pp. 183–189, Mar. 1998.

[40] C. Zou, L. Peng, L. Feng, and Z. Wang, "Multivariate sign-based high-dimensional tests for sphericity," *Biometrika*, vol. 101, no. 1, pp. 229–236, 2014.

[41] G. Cheng, B. Liu, L. Peng, B. Zhang, and S. Zheng, "Testing the equality of two high-dimensional spatial sign covariance matrices," *Scand. J. Statist.*, vol. 46, no. 1, pp. 257–271, 2019.

[42] T. Zhang and A. Wiesel, "Automatic diagonal loading for Tyler's robust covariance estimator," in *Proc. IEEE Workshop Stat. Signal Process.*, Palma de Mallorca, Spain, 2016.

[43] A. Dürre, D. E. Tyler, and D. Vogel, "On the eigenvalues of the spatial sign covariance matrix in more than two dimensions," *Statist. Probability Lett.*, vol. 111, pp. 80–85, Apr. 2016.

[44] B. M. Brown, "Statistical uses of the spatial median," *J. Roy. Stat. Soc.: Ser. B. (Methodological)*, vol. 45, no. 1, pp. 25–30, Sep. 1983.

[45] 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, Sep. 2014.

[46] L. Du, J. Li, and P. Stoica, "Fully automatic computation of diagonal loading levels for robust adaptive beamforming," *IEEE Trans. Aerosp. Electron. Syst.*, vol. 46, no. 1, pp. 449–458, Jan. 2010.

[47] M. Grant and S. Boyd, "CVX: Matlab software for disciplined convex programming, version 2.1," Mar. pp. 95–110, 2014. [Online]. Available: http://cvxr.com/cvx

[48] M. Grant and S. Boyd, "Graph implementations for nonsmooth convex programs," in *Recent Advances in Learning and Control* (*Ser. Lecture Notes in Control Inf. Sci.*), V. Blondel, S. Boyd, and H. Kimura, Eds. London, U.K.: Springer-Verlag, 2008, vol. 371, pp. 95–110.

[49] 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.

[50] E. Raninen, E. Ollila, and D. E. Tyler, "On the variability of the sample covariance matrix under complex elliptical distributions," *IEEE Signal Process. Lett.*, vol. 28, pp. 2092–2096, Oct. 2021.

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

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

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

[54] W. Sharpe, "Capital asset prices: A theory of market equilibrium under conditions of risk," *J. 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," *Rev. Econ. Statist.*, vol. 47, no. 1, pp. 13–37, 1965.

[56] O. Ledoit and M. Wolf, "Analytical nonlinear shrinkage of large-dimensional covariance matrices," *Ann. Statist.*, vol. 48, no. 5, pp. 3043–3065, 2020.

[57] O. Ledoit and M. Wolf, "Spectrum estimation: A unified framework for covariance matrix estimation and PCA in large dimensions," *J. Multivariate Anal.*, 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 Trans. Signal Process.*, vol. 63, no. 24, pp. 6684–6697, Dec. 2015.

[59] Y. Feng and D. P. Palomar, "A signal processing perspective on financial engineering," *Found. Trends Signal Process.*, vol. 9, no. 1, pp. 1–231, 2016.

[60] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge: Cambridge Univ. Press, 2004.

[61] K.-T. Fang, S. Kotz, and K. W. Ng, *Symmetric Multivariate and Related Distributions*. London, U.K.: Chapman and Hall/CRC, 2018.

[62] J. R. Magnus and H. Neudecker, "The commutation matrix: Some properties and applications," *Ann. Statist.*, vol. 7, no. 2, pp. 381–394, 1979.

Elias Raninen (Student Member, IEEE) received the bachelor's degree in pop and jazz music education from the Metropolia University of Applied Sciences, Helsinki, Finland, in 2009, and the M.Sc. (Tech.) degree in computer, communication and information sciences (with Hons.) from the Aalto University School of Electrical Engineering, Espoo, Finland, in 2017. He is currently working toward the D.Sc. (Tech.) degree with the Department of Signal Processing and Acoustics, Aalto University, Espoo, Finland. His research interests include statistical signal processing, machine learning, and high-dimensional statistics.
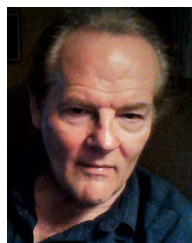
David E. Tyler received the Ph.D. degree from the Department of Statistics, Princeton University, Princeton, NJ, USA, 1979. He is currently a Distinguished Professor with the Department of Statistics and Biostatistics, Rutgers – The State University of New Jersey, New Jersey, NJ, USA, where he has been on the faculty since 1983, and was the Chair of the Department from 1993–1996. His current research interests in statistics include multivariate analysis, directional data analysis and robust statistics, while in the past he has worked in such diverse areas as psychometrics, spectral analysis of time series, functional data analysis, and computer vision.

Esa Ollila (Senior Member, IEEE) received the M.Sc. degree in mathematics from the University of Oulu, Oulu, Finland, in 1998, the Ph.D. degree in statistics (with Hons.) from the University of Jyväskylä, Jyväskylä, Finland, in 2002, and the D.Sc.(Tech) degree (with Hons.) 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. The academic year 2010–2011, he spent as a Visiting Postdoctoral Research Associate with the Department of Electrical Engineering, Princeton University, Princeton, NJ, USA. Since June 2015, he has been an Associate Professor of signal processing, Aalto University. He is also an Adjunct Professor of statistics with Oulu University. He is the coauthor of a recent textbook, *Robust Statistics for Signal Processing*, published by Cambridge University Press in 2018. His research interests include statistical signal processing, high-dimensional statistics, and robust statistics. He is currently an Associate Editor for the *Scandinavian Journal of Statistics*.