Coupled Regularized Sample Covariance Matrix Estimator for Multiple Classes

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

Abstract—The estimation of covariance matrices of multiple classes with limited training data is a difficult problem. The sample covariance matrix (SCM) is known to perform poorly when the number of variables is large compared to the available number of samples. In order to reduce the mean squared error (MSE) of the SCM, regularized (shrinkage) SCM estimators are often used. In this work, we consider regularized SCM (RSCM) estimators for multiclass problems that couple together two different target matrices for regularization: the pooled (average) SCM of the classes and the scaled identity matrix. Regularization toward the pooled SCM is beneficial when the population covariances are similar, whereas regularization toward the identity matrix guarantees that the estimators are positive definite. We derive the MSE optimal tuning parameters for the estimators as well as propose a method for their estimation under the assumption that the class populations follow (unspecified) elliptical distributions with finite fourth-order moments. The MSE performance of the proposed coupled RSCMs are evaluated with simulations and in a regularized discriminant analysis (RDA) classification set-up on real data. The results based on three different real data sets indicate comparable performance to cross-validation but with a significant speed-up in computation time.

Index Terms—Covariance matrix estimation, regularization, shrinkage, elliptical distribution, regularized discriminant analysis.

I. INTRODUCTION

AN INCREASINGLY common scenario in modern supervised learning problems is that the dimension p of the data is large compared to the number of available training samples n or exceed it multifold ($p \gg n$). Such scenarios are commonly referred to as high-dimensional or insufficient sample support problems. In this paper, we address the problem of high-dimensional covariance matrix estimation in a multiclass setup, where there are K different classes or populations, each comprising n_k, $k = 1, \ldots, K$, independent and identically distributed (i.i.d.) p-dimensional samples. Estimates of the covariance matrices are needed in many multivariate analysis problems, such as in principal component analysis and canonical correlation

analysis [1], discriminant analysis [2], Gaussian mixture models (GMM) [3], as well as in many engineering applications, e.g., in array signal processing [4], genomics [5], portfolio optimization in finance [6], and graphical models [7]. The success of the analysis is often directly lined with the accuracy of the estimated covariance matrix.

The covariance matrix of class $k \in \{1, \ldots, K\}$ is defined as

$$\boldsymbol{\Sigma}_k = \mathbb{E}[(\mathbf{x}_{ik} - \boldsymbol{\mu}_k)(\mathbf{x}_{ik} - \boldsymbol{\mu}_k)^\top],$$

where $\mathbf{x}_{ik}$ denotes the ith sample from class k and $\boldsymbol{\mu}_k = \mathbb{E}[\mathbf{x}_{ik}]$ is the mean of class k. The conventional estimate for the covariance matrix is the unbiased sample covariance matrix (SCM) defined for class k by

$$\mathbf{S}_k = \frac{1}{n_k - 1} \sum_{i=1}^{n_k} (\mathbf{x}_{ik} - \overline{\mathbf{x}}_k)(\mathbf{x}_{ik} - \overline{\mathbf{x}}_k)^\top,$$

where $\overline{\mathbf{x}}_k = (1/n_k) \sum_i \mathbf{x}_{ik}$ is the sample mean of class k. In high-dimensional settings, the SCM is known to work poorly due to its high variability. Furthermore, if $n_k < p$, then the SCM is singular, and hence, its inverse cannot be computed.

Better estimators can be developed by using regularization, where the key idea is to shift or shrink the estimator toward a predetermined target or model. This can significantly decrease the variance of the estimator and improve the overall performance by reducing its mean squared error (MSE). This phenomenon can be understood via the following well-known bias-variance decomposition of the MSE. For an estimator $\hat{\boldsymbol{\Sigma}}_k$ of $\boldsymbol{\Sigma}_k$, the MSE can be written as

$$\mathrm{MSE}(\hat{\boldsymbol{\Sigma}}_k) = \mathbb{E}\left[\left\|\boldsymbol{\Sigma}_k - \hat{\boldsymbol{\Sigma}}_k\right\|_{\mathrm{F}}^2\right]$$

$$= \mathbb{E}\left[\left\|\hat{\boldsymbol{\Sigma}}_k - \mathbb{E}[\hat{\boldsymbol{\Sigma}}_k]\right\|_{\mathrm{F}}^2\right] + \left\|\boldsymbol{\Sigma}_k - \mathbb{E}[\hat{\boldsymbol{\Sigma}}_k]\right\|_{\mathrm{F}}^2,$$

where the first term on the right-hand side is the variance and the second term is the squared bias of the estimator. Since the SCM is unbiased, its MSE is equal to its variance. By using a regularized SCM (RSCM), however, it is possible to reduce the MSE significantly at the cost of introducing some bias.

In general, regularization combines an unstructured estimate with a predefined model or target. The target in regularization can be decided based on a prior knowledge, assumptions, or on a property, which we want to enforce. Regularization can be accomplished different ways. For instance, the SCM can be combined linearly with a target matrix as, e.g., in [5], [6], [8]–[13]. The approaches for parameter tuning differs between the

methods. A popular approach taken in [8] is based on estimating the asymptotically optimal (in terms of minimizing the MSE) tuning parameters. Under additional distributional assumptions, the method of [8] has later been improved for Gaussian samples in [10] and elliptically distributed samples in [13]. Other approaches for tuning parameter selection are for example the expected likelihood approach [14], [15]. Alternatively, the eigenvalues of the SCM can be transformed non-linearly toward a specific structure as in [16] and [17]. Multiple targets can also be used. For instance, in [18], a double shrinkage covariance matrix estimator was considered, which shrinks simultaneously toward a spherical matrix and a diagonal matrix. A somewhat related RSCM formulation had been proposed in [3] in the setting of Gaussian mixture models. In [19], the estimator of [8] was extended to multiple simultaneous target matrices, which satisfy a certain target structure. In [20], a linear multi-target shrinkage covariance matrix estimator was proposed, which optimizes the tuning parameters using low-complexity leave-one-out cross-validation. There are also methods tailored for multiclass problems. For instance, [21] considered covariance matrix estimation from two possibly mismatched data sets using the maximum likelihood principle. In our previous work [22], linear pooling of SCMs was considered and was applied to portfolio optimization. In the context of discriminant analysis classification, a popular RSCM was proposed in [2]. Inspired by [2], this paper focuses on two specific target matrices: the pooled SCM and the spherical matrix.

The rest of the paper is organized as follows. In Section II, we give some background and motivation for the proposed estimator. In Section III, we analyze some properties of the proposed estimator as well as discuss the optimization of the tuning parameters. In Section IV, we show how the theoretical tuning parameters can be estimated in practice when the unknown class populations follow unspecified elliptically symmetric distributions. Section V discusses some practical considerations and how to use the method in choosing the tuning parameters in RDA. In Section VI, synthetic simulation studies are conducted in order to assess the MSE performance of the estimator. The method is then compared to cross-validation in choosing the tuning parameters for RDA classification using three different real data sets. Lastly, Section VII concludes.

Notation: Throughout the paper, unless stated otherwise, all norms are Frobenius norms defined by $\|\mathbf{A}\|_{\mathrm{F}}^2 = \langle \mathbf{A}, \mathbf{A} \rangle_{\mathrm{F}} = \mathrm{tr}(\mathbf{A}^\top \mathbf{A})$, where the inner product of two matrices (of appropriate dimensions) $\mathbf{A}$ and $\mathbf{B}$ is defined by $\langle \mathbf{A}, \mathbf{B} \rangle_{\mathrm{F}} = \mathrm{tr}(\mathbf{A}^\top \mathbf{B})$. For any square matrix $\mathbf{A}$, we frequently use the notation $\mathbf{I_A} = (1/p)\mathrm{tr}(\mathbf{A})\mathbf{I}$ and $\mathbf{A^I} = \mathbf{A} - \mathbf{I_A}$. For a vector $\mathbf{a} \in \mathbb{R}^p$, the Euclidean norm is defined as $\|\mathbf{a}\| = \sqrt{\mathbf{a}^\top \mathbf{a}}$. We define $\mathbb{R}_{\geq 0} = \{a \in \mathbb{R} : a \geq 0\}$. For a scalar variable a, we use the following shorthand notation $\partial_a = \partial/\partial a$ for the partial derivative with respect to a. For scalars $a < b$ and c, we define the *clamp* or *clip* function $[c]_a^b = \max\{a, \min\{b, c\}\}$, which projects c onto the interval $[a, b]$. Lastly, $\mathrm{Unif}\{a, b\}$ and $\mathrm{Unif}(a, b)$ denote the discrete and continuous univariate uniform distributions on the set $\{a, a+1, \ldots, b-1, b\}$ and on the open interval (a, b), respectively.

II. BACKGROUND AND MOTIVATION

In multiclass problems, when the classes can be assumed to have a similar covariance structure, it is beneficial to shrink the individual class covariance matrix estimates toward the pooled (average) SCM of the classes,

$$\mathbf{S} = \sum_{k=1}^K \pi_k \mathbf{S}_k, \tag{1}$$

where

$$\pi_k = \frac{n_k}{n_1 + n_2 + \cdots + n_K}.$$

For example, the methods proposed in [2], [23], and [24] used the convex combination

$$\hat{\mathbf{\Sigma}}_k(\beta) = \beta \mathbf{S}_k + (1 - \beta)\mathbf{S}, \tag{2}$$

where $\beta \in [0, 1]$, as an estimate for the class covariance matrix. The methods in [2], [23], and [24] were designed for discriminant analysis (DA) classification in which the problem is to classify a new sample $\mathbf{x}$ to one of the K classes using the *discriminant rule*

$$\hat{k} = \underset{k \in \{1, \ldots, K\}}{\arg \min} \ (\mathbf{x} - \hat{\boldsymbol{\mu}}_k)^\top \hat{\mathbf{\Sigma}}_k^{-1}(\mathbf{x} - \hat{\boldsymbol{\mu}}_k) + \log |\hat{\mathbf{\Sigma}}_k|, \tag{3}$$

where $\hat{\boldsymbol{\mu}}_k$ and $\hat{\mathbf{\Sigma}}_k$ denote estimates of the mean and the covariance matrix of class k, respectively. Different DA methods differ in the approach used to estimate the class means and class covariance matrices. For example, *quadratic discriminant analysis* (QDA) uses sample means and SCMs in (3). In *linear discriminant analysis* (LDA) the pooled SCM $\mathbf{S}$ in (1) is used for all classes. If the true covariance matrices are equal, $\mathbf{\Sigma}_1 = \mathbf{\Sigma}_2 = \cdots = \mathbf{\Sigma}_K \equiv \mathbf{\Sigma}$, then LDA is well justified since the pooled SCM is an unbiased estimator of $\mathbf{\Sigma}$, i.e., $\mathbb{E}[\mathbf{S}] = \sum_j \pi_j \mathbb{E}[\mathbf{S}_j] = (\sum_j \pi_j)(\mathbf{\Sigma}) = \mathbf{\Sigma}$. However, even in the case that the true population covariance matrices differ substantially, LDA often outperforms QDA when the sample size is small compared to the dimension [23], [25]. This is due to the high variance of the class SCMs compared to the pooled SCM. The partially pooled estimator (2) includes QDA and LDA as special cases when $\beta = 1$ and $\beta = 0$, respectively.

In a very high-dimensional case, when $p > \sum_j n_j$, the partially pooled estimator (2) will no longer be positive definite. In *regularized discriminant analysis* [2] (RDA), this problem is solved by coupling the pooled estimator with a spherical target matrix, that is, by regularization toward a scaled identity matrix via

$$\hat{\mathbf{\Sigma}}_k(\alpha, \beta) = \alpha \hat{\mathbf{\Sigma}}_k(\beta) + (1 - \alpha)\mathbf{I}_{\hat{\mathbf{\Sigma}}_k(\beta)}, \tag{4}$$

where $\hat{\mathbf{\Sigma}}_k(\beta)$ is given in (2) and $\mathbf{I}_{\hat{\mathbf{\Sigma}}_k(\beta)} = (1/p)\mathrm{tr}(\hat{\mathbf{\Sigma}}_k(\beta))\mathbf{I}$.

The success of regularization depends on proper selection of the tuning parameters. In classification problems, cross-validating the classification error is the standard method for choosing the tuning parameters. Cross-validation can, however, be computationally very costly for large data sets. In certain high-dimensional binary classification settings, it is possible to leverage on results from random matrix theory in order to

estimate the (asymptotically) optimal tuning parameters, which minimize the classification error [26], [27].

The main contribution of this work is to develop a computationally efficient method for choosing the tuning parameters in a multiclass covariance matrix estimation setting, where there can be more than two classes, and where the RSCMs are defined in (4). Specifically, we find estimates of the tuning parameters that minimize the class-specific MSE:

$$(\alpha_k^\star, \beta_k^\star) = \arg \min_{\alpha, \beta \in [0,1]} \mathbb{E}\left[\left\|\hat{\boldsymbol{\Sigma}}_k(\alpha, \beta) - \boldsymbol{\Sigma}_k\right\|_{\mathrm{F}}^2\right] \quad (5)$$

for each population $k = 1, \ldots, K$. The expressions for the optimal tuning parameters will depend on unknown population parameters since the MSE expression involves the unknown covariance matrix. However, we show that the MSE can be estimated fairly easily by assuming that the class populations follow unspecified elliptically symmetric distributions.

It is worth noting that, in RDA [2] the tuning parameters are common across the classes, whereas in this paper we use class-specific tuning parameters. However, in cases when it is useful to use common tuning parameters, they can easily be acquired by averaging as explained in Subsection V-B. Hence, our proposed method can be used to obtain tuning parameters for the original RDA [2] framework, for which there already exist widely established toolboxes in programming languages such as R (see e.g., [28], [29]). Lastly, an important distinction to [2] is that our method is not only targeted for classification problems but is suitable for other applications as well. For example, [21] considered the problem of estimating a covariance matrix from two data sets, where the population covariance matrix of the first data set is different but close to the population covariance matrix of the second data set. This type of problems are encountered in radar processing as well as in hyperspectral imaging applications, where the additional data sets may have been acquired with slightly different measurement configurations, and hence, have slightly different population parameters.

The current paper extends our earlier preliminary work in [30], which considered the estimation of the MSE optimal tuning parameters for the partially pooled estimator in (2). Here, we consider the more general estimator in (4), which includes additional shrinkage toward the scaled identity matrix, and is therefore applicable also in the cases when $p > \sum_j n_j$.

III. ESTIMATOR

Let us first consider four special cases of the estimator (4):

C1) *The unpooled regularized SCM estimator* omits the pooled SCM and only shrinks toward the scaled identity matrix:

$$\hat{\boldsymbol{\Sigma}}_k(\alpha_k, \beta_k = 1) = \alpha_k \mathbf{S}_k + (1 - \alpha_k)\mathbf{I}_{\mathbf{S}_k}.$$

This type of shrinkage is typically considered in single class covariance matrix estimation (see e.g., [8] and [13]).

C2) *The partially pooled estimator* omits regularization toward the scaled identity and only shrinks toward the
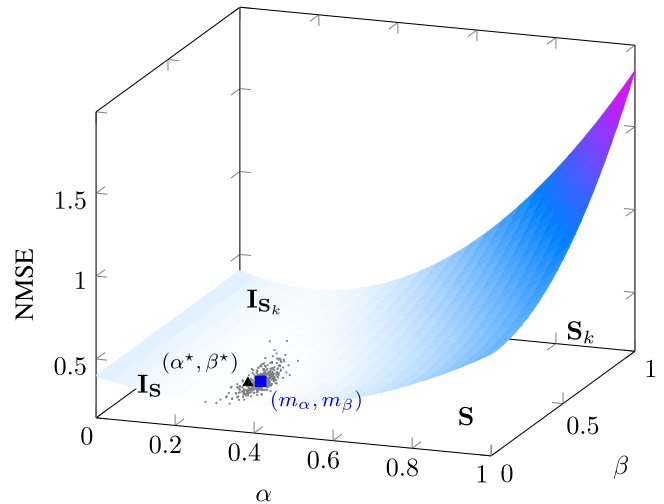

Fig. 1. The theoretical NMSE as a function of the tuning parameters (α, β) for class 4 of the setup A in Section VI.

pooled SCM:

$$\hat{\boldsymbol{\Sigma}}_k(\alpha_k = 1, \beta_k) = \hat{\boldsymbol{\Sigma}}_k(\beta_k) = \beta_k \mathbf{S}_k + (1 - \beta_k)\mathbf{S}.$$

C3) *The fully pooled estimator* uses the pooled SCM for every class k and shrinks it toward the scaled identity matrix:

$$\hat{\boldsymbol{\Sigma}}_k(\alpha_k, \beta_k = 0) = \alpha_k \mathbf{S} + (1 - \alpha_k)\mathbf{I}_{\mathbf{S}}.$$

Such shrinkage can be considered if all classes have an identical distribution.

C4) *The scaled identity estimator* uses the partially pooled estimator to scale the identity matrix:

$$\hat{\boldsymbol{\Sigma}}_k(\alpha_k = 0, \beta_k) = \mathbf{I}_{(\beta_k \mathbf{S}_k + (1 - \beta_k)\mathbf{S})}.$$

Since it is clear that the tuning parameters are class-specific, we drop the subscripts from α_k and β_k and denote them from now on simply by α and β.

Fig. 1 depicts the theoretical normalized MSE (NMSE), $\mathrm{NMSE}(\hat{\boldsymbol{\Sigma}}_k(\alpha, \beta)) = \mathrm{MSE}(\hat{\boldsymbol{\Sigma}}_k(\alpha, \beta))/\|\boldsymbol{\Sigma}_k\|_{\mathrm{F}}^2$, of the estimator (4) as a function of the tuning parameters (α, β). The figure corresponds to setup A of the numerical study of Section VI. In the figure, the small gray dots depict the estimated tuning parameters (showing 400 realizations of the 4000 Monte Carlo trials). The blue square (■) denotes the mean (m_α, m_β) of the estimated tuning parameters over the Monte Carlo runs. The optimal tuning parameter pair $(\alpha^\star, \beta^\star)$ is denoted by the black triangle (▲). The special cases (C1–C4) correspond to the edges of the (α, β)-plane.

As can be observed from Fig. 1, the optimal tuning parameter pair is in this case closest to $\mathbf{I}_{\mathbf{S}}$. Hence, in this particular case it would be a good strategy to use the pooled SCM and regularize it toward a scaled identity matrix, i.e., to use the fully pooled estimator (C3). As can be noted, the proposed method is able to automatically choose the tuning parameters in a near optimal way.

Let us now give an expression for the MSE of the estimator.

Theorem 1: The MSE of the estimator (4) is a bivariate polynomial of the form

$$\text{MSE}(\hat{\boldsymbol{\Sigma}}_k(\alpha, \beta)) = \alpha^2\beta^2 C_{22} + \alpha^2\beta C_{21} + \alpha^2 C_{20} + \beta^2 C_{02}$$
$$+ \alpha\beta C_{11} + \alpha C_{10} + \beta C_{01} + C_{00}, \quad (6)$$

where the coefficients, C_{ij}, depend on the scalars $\text{tr}(\boldsymbol{\Sigma}_j)$, $\mathbb{E}[\|\mathbf{S}_j\|_{\text{F}}^2]$, $\mathbb{E}[\|\mathbf{I}_{\mathbf{S}_j}\|_{\text{F}}^2]$, and $\langle\boldsymbol{\Sigma}_i, \boldsymbol{\Sigma}_j\rangle_{\text{F}}$.

Proof: See Appendix A for the proof and the expressions for the coefficients C_{ij}. ∎

The estimation of the coefficients C_{ij} is deferred to Section IV. Given estimates of the coefficients, the critical points of (6) can be solved numerically (see Appendix A).

A useful property of the MSE polynomial in (6) is that given a fixed value of either α or β, the optimization of the remaining tuning parameter is a convex problem. This is known as biconvexity.

Theorem 2: The MSE (6) is a biconvex function. The optimal tuning parameter $\alpha^\star$ given a fixed value of $\beta \in [0, 1]$ is

$$\alpha^\star = \left[-\frac{1}{2}\frac{\beta C_{11} + C_{10}}{\beta^2 C_{22} + \beta C_{21} + C_{20}}\right]_0^1. \quad (7)$$

Likewise, the optimal tuning parameter $\beta^\star$ given a fixed value of $\alpha \in [0, 1]$ is

$$\beta^\star = \left[-\frac{1}{2}\frac{\alpha^2 C_{21} + \alpha C_{11} + C_{01}}{\alpha^2 C_{22} + C_{02}}\right]_0^1. \quad (8)$$

Proof: See Appendix A–C. ∎

The optimal tuning parameter for the special cases (C1–C4) can be solved using (7) and (8) of Theorem 2.

A. Partially Pooled Estimator Without Identity Shrinkage

Let us next highlight certain properties of the partially pooled estimator (2) of (C2) corresponding to the case $\hat{\boldsymbol{\Sigma}}_k(\alpha = 1, \beta)$. The propositions below provide additional insights on the optimal parameter $\beta^\star$ in this case.

Proposition 1: The optimal tuning parameter $\beta^\star$ given $\alpha = 1$ satisfies $\beta^\star < 1$.

Proof: The proof follows from showing that the numerator of (8) is always less than the denominator, which is always positive. By setting $\alpha = 1$, after some algebra, we have that $-1/2$ times the numerator, i.e., $(-1/2)(C_{21} + C_{11} + C_{01})$, is equal to $\mathbb{E}[\langle\mathbf{S} - \mathbf{S}_k, \mathbf{S} - \boldsymbol{\Sigma}_k\rangle_{\text{F}}]$ and the denominator $(C_{22} + C_{02})$ is equal to $\mathbb{E}[\|\mathbf{S} - \mathbf{S}_k\|_{\text{F}}^2]$ (See Appendix A). Subtracting the numerator from the denominator, we get

$$\mathbb{E}[\|\mathbf{S} - \mathbf{S}_k\|_{\text{F}}^2] - \mathbb{E}[\langle\mathbf{S} - \mathbf{S}_k, \mathbf{S} - \boldsymbol{\Sigma}_k\rangle_{\text{F}}]$$
$$= \mathbb{E}[\langle\mathbf{S} - \mathbf{S}_k, \boldsymbol{\Sigma}_k - \mathbf{S}_k\rangle_{\text{F}}]$$
$$= (1 - \pi_k)(\mathbb{E}[\|\mathbf{S}_k\|_{\text{F}}^2] - \|\boldsymbol{\Sigma}_k\|_{\text{F}}^2)$$
$$= (1 - \pi_k)\mathbb{E}[\|\mathbf{S}_k - \boldsymbol{\Sigma}_k\|_{\text{F}}^2] > 0,$$

which holds almost surely (a.s.) for any continuous distribution. ∎

This has the important implication that in a multiclass problem it is always possible to reduce the MSE of the SCM by using regularization toward the pooled SCM.

In the special case that all of the population covariance matrices are equal, we would like to only use the pooled SCM. This is exactly what happens as is shown in the next proposition.

Proposition 2: Consider multiple populations generated by the same distribution such that they have equal covariance matrices $\boldsymbol{\Sigma}_1 = \ldots = \boldsymbol{\Sigma}_K$ and equal sample sizes $\pi_1 = \ldots = \pi_K$. Given that $\alpha = 1$ then $\beta^\star = 0$.

Proof: The proof follows from showing that if the true population covariance matrices and the sample sizes are equal, then the numerator of (8) is zero and the denominator is nonzero. Observe that if $\boldsymbol{\Sigma}_k = \boldsymbol{\Sigma}_j$ and $\pi_k = \pi_j$, $\forall j$, then we have $\mathbb{E}[\|\mathbf{S}_k\|_{\text{F}}^2] = \mathbb{E}[\|\mathbf{S}_j\|_{\text{F}}^2]$ and $\langle\boldsymbol{\Sigma}_j, \boldsymbol{\Sigma}_k\rangle_{\text{F}} = \|\boldsymbol{\Sigma}_k\|_{\text{F}}^2$, $\forall j$. Hence, opening up the numerator $\mathbb{E}[\langle\mathbf{S} - \mathbf{S}_k, \mathbf{S} - \boldsymbol{\Sigma}_k\rangle_{\text{F}}]$ yields

$$\mathbb{E}[\|\mathbf{S}\|_{\text{F}}^2] - \mathbb{E}[\langle\mathbf{S}, \boldsymbol{\Sigma}_k\rangle_{\text{F}}] - \mathbb{E}[\langle\mathbf{S}_k, \mathbf{S}\rangle_{\text{F}}] + \mathbb{E}[\langle\mathbf{S}_k, \boldsymbol{\Sigma}_k\rangle_{\text{F}}]$$
$$= \mathbb{E}[\|\mathbf{S}\|_{\text{F}}^2] - \mathbb{E}[\langle\mathbf{S}_k, \mathbf{S}\rangle_{\text{F}}]$$
$$= \left(\sum_j \pi_j^2 - \pi_k\right)\mathbb{E}[\|\mathbf{S}_k\|_{\text{F}}^2] + \left(\sum_{i \neq j}\pi_i\pi_j - \sum_{j, j \neq k}\pi_j\right)\|\boldsymbol{\Sigma}_k\|_{\text{F}}^2$$
$$= 0,$$

which follows from $\sum_j \pi_j^2 = K\pi_k^2 = (K\pi_k)\pi_k = \pi_k$, $\sum_{i \neq j}\pi_i\pi_j = K(K-1)\pi_k^2 = (K\pi_k)(K-1)\pi_k = 1 - \pi_k$, and $\sum_{j, j \neq k}\pi_j = (K-1)\pi_k = 1 - \pi_k$. Furthermore, using this result we see that the denominator, $\mathbb{E}[\|\mathbf{S} - \mathbf{S}_k\|_{\text{F}}^2]$, is a.s. positive. Opening up the denominator, we have

$$\mathbb{E}[\|\mathbf{S}\|_{\text{F}}^2] + \mathbb{E}[\|\mathbf{S}_k\|_{\text{F}}^2] - 2\mathbb{E}[\langle\mathbf{S}, \mathbf{S}_k\rangle_{\text{F}}]$$
$$= \mathbb{E}[\|\mathbf{S}_k\|_{\text{F}}^2] - \mathbb{E}[\langle\mathbf{S}, \mathbf{S}_k\rangle_{\text{F}}]$$
$$= (1 - \pi_k)\mathbb{E}[\|\mathbf{S}_k\|_{\text{F}}^2] - \sum_{j, j \neq k}\pi_j\langle\boldsymbol{\Sigma}_k, \boldsymbol{\Sigma}_j\rangle_{\text{F}}$$
$$= (1 - \pi_k)\left(\mathbb{E}[\|\mathbf{S}_k\|_{\text{F}}^2] - \|\boldsymbol{\Sigma}_k\|_{\text{F}}^2\right) > 0 \text{ a.s.}$$

∎

B. Streamlined Analytical Estimator

An alternative estimator to (4) can be obtained if we change the α-regularization target so that the estimator becomes

$$\tilde{\boldsymbol{\Sigma}}_k(\alpha, \beta) = \alpha\hat{\boldsymbol{\Sigma}}_k(\beta) + (1 - \alpha)\mathbf{I}_{\mathbf{T}}, \quad (9)$$

where $\mathbf{T} \in \{\mathbf{S}_k, \mathbf{S}\}$ and $\hat{\boldsymbol{\Sigma}}_k(\beta)$ is defined in (2). This simplifies the expression for the MSE and allows for an analytical solution for the tuning parameters. In [3] and [18], somewhat similar formulations were used with different targets in a GMM and a single class covariance matrix estimation setting, respectively. Note that the difference between (9) and (4) is in the scale of the identity target. The trace of (9) (sum of the eigenvalues) is dependent on α, whereas in (4) this is not the case. However, when $\text{tr}(\mathbf{I}_{\mathbf{T}}) \approx \text{tr}(\mathbf{I}_{\hat{\boldsymbol{\Sigma}}_k(\beta)})$, the performance of the two estimators is expected to be similar. The MSE of (9) is given in the next theorem.

Theorem 3: The theoretical MSE of the (streamlined analytical) estimator (9) is a bivariate polynomial of the form

$$\mathrm{MSE}(\tilde{\boldsymbol{\Sigma}}_k(\alpha, \beta)) = \alpha^2 \beta^2 B_{22} + \alpha^2 \beta B_{21} + \alpha^2 B_{20}$$
$$+ \alpha \beta B_{11} + \alpha B_{10} + B_{00}.$$

The coefficients B_{ij} depend on the scalars $\mathrm{tr}(\boldsymbol{\Sigma}_j)$, $\mathbb{E}[\|\mathbf{S}_j\|_{\mathrm{F}}^2]$, $\mathbb{E}[\|\mathbf{I}_{\mathbf{S}_j}\|_{\mathrm{F}}^2]$, and $\langle \boldsymbol{\Sigma}_i, \boldsymbol{\Sigma}_j \rangle_{\mathrm{F}}$. If $(\alpha^\star, \beta^\star) \in (0, 1) \times (0, 1)$, the optimal tuning parameters $(\alpha^\star, \beta^\star)$ minimizing the MSE are

$$\alpha^\star = \frac{2 B_{10} B_{22} - B_{11} B_{21}}{B_{21}^2 - 4 B_{20} B_{22}} \text{ and } \beta^\star = \frac{2 B_{11} B_{20} - B_{10} B_{21}}{2 B_{10} B_{22} - B_{11} B_{21}}.$$

Otherwise, the optimal parameters are on the boundary of the feasible set $[0, 1] \times [0, 1]$, and are given by one of the following options

i) $\alpha^\star = \left[-\dfrac{1}{2} \dfrac{B_{10}}{B_{20}} \right]_0^1$ and $\beta^\star = 0$,

ii) $\alpha^\star = \left[-\dfrac{1}{2} \dfrac{B_{10} + B_{11}}{B_{22} + B_{21} + B_{20}} \right]_0^1$ and $\beta^\star = 1$,

iii) $\alpha^\star = 1$ and $\beta^\star = \left[-\dfrac{1}{2} \dfrac{B_{21} + B_{11}}{B_{22}} \right]_0^1$,

iv) $\alpha^\star = 0$, which implies $\tilde{\boldsymbol{\Sigma}} = \mathbf{I_T}$ and that the MSE does not depend on β.

Proof: See Appendix B for the proof and the expressions for the coefficients B_{ij}. ∎

IV. ESTIMATING THE TUNING PARAMETERS

The MSE and the optimal tuning parameters in Theorem 1 and Theorem 3 depend on the coefficients C_{ij} and B_{ij}, which in turn are functions of the unknown scalars

$$\mathrm{tr}(\boldsymbol{\Sigma}_j), \ \mathbb{E}[\|\mathbf{S}_j\|_{\mathrm{F}}^2], \ \mathbb{E}[\|\mathbf{I}_{\mathbf{S}_j}\|_{\mathrm{F}}^2] \text{ and } \langle \boldsymbol{\Sigma}_i, \boldsymbol{\Sigma}_j \rangle_{\mathrm{F}}. \quad (10)$$

Hence, estimates of the optimal tuning parameters can be formed by a plug-in method, i.e., replacing the above unknown parameters by their estimates. Such estimates are constructed in this section under the very general assumption that the samples are generated from unspecified elliptical distributions with finite fourth-order moments.

Elliptical distributions are a location-scale family of spherical distributions generalizing many common distributions, such as the multivariate normal distribution (MVN) and Student's t-distribution to name a few. The probability density function (p.d.f.) of an elliptically distributed sample $\mathbf{x}$ from class k is up to a normalizing constant of the form

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

where $\boldsymbol{\mu}_k$ and $\boldsymbol{\Sigma}_k$ are the mean and the positive definite covariance matrix of the distribution, respectively. The function $g_k : \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0}$ is called the *density generator*, which determines the distribution. For example, $g_k(t) = \exp(-t/2)$ defines the MVN distribution. For references about elliptical distributions see, e.g., [31] and [32].

Estimates of the unknown scalars, $\mathrm{tr}(\boldsymbol{\Sigma}_k)$ and $\|\boldsymbol{\Sigma}_k\|_{\mathrm{F}}^2$, can be obtained by finding estimates of the *scale* parameter η_k and the

sphericity parameter γ_k defined as

$$\eta_k = \frac{\mathrm{tr}(\boldsymbol{\Sigma}_k)}{p} \text{ and } \gamma_k = \frac{\|\boldsymbol{\Sigma}_k\|_{\mathrm{F}}^2}{\|\mathbf{I}_{\boldsymbol{\Sigma}_k}\|_{\mathrm{F}}^2} \in [1, p].$$

Observe that $\|\mathbf{I}_{\boldsymbol{\Sigma}_k}\|_{\mathrm{F}}^2 = p \eta_k^2$ and $\|\boldsymbol{\Sigma}_k\|_{\mathrm{F}}^2 = p \gamma_k \eta_k^2$. Under the ellipticity assumption, the unknowns $\mathbb{E}[\|\mathbf{S}_k\|_{\mathrm{F}}^2]$ and $\mathbb{E}[\|\mathbf{I}_{\mathbf{S}_k}\|_{\mathrm{F}}^2]$ are functions of the scale, sphericity, and the elliptical kurtosis as shown in the next lemma. The elliptical kurtosis is defined as $\kappa_k = (1/3)\mathrm{kurt}(x_i)$, where $\mathrm{kurt}(x_i)$ is the excess kurtosis of any marginal variable x_i of a random vector from the kth class.

Lemma 1: [13, Lemma 2.] Consider an elliptical distribution with covariance matrix $\boldsymbol{\Sigma}_k$ and finite fourth-order moments. Then,

$$\mathbb{E}[\|\mathbf{S}_k\|_{\mathrm{F}}^2] = p \eta_k^2 \left(\tau_{1k} p + (1 + \tau_{1k} + \tau_{2k}) \gamma_k \right) \text{ and}$$

$$\mathbb{E}[\|\mathbf{I}_{\mathbf{S}_k}\|_{\mathrm{F}}^2] = \eta_k^2 \left((1 + \tau_{2k}) p + 2 \tau_{1k} \gamma_k \right),$$

where $\tau_{1k} = 1/(n_k - 1) + \kappa_k / n_k$, and $\tau_{2k} = \kappa_k / n_k$.

In the following subsections, we show how to estimate the unknown parameters, η_k, γ_k, κ_k, and $\langle \boldsymbol{\Sigma}_i, \boldsymbol{\Sigma}_j \rangle_{\mathrm{F}}$. We use the same approach as in [22].

A. Estimation of the Elliptical Kurtosis

The elliptical kurtosis of class k is estimated by the sample average

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

where $\hat{k}_j = m_j^{(4)} / (m_j^{(2)})^2 - 3$ is the sample estimate of the kurtosis of the jth variable (of class k), and $m_j^{(q)} = \frac{1}{n_k} \sum_{i=1}^{n_k} ((\mathbf{x}_{ik})_j - (\bar{\mathbf{x}}_k)_j)^q$ is the qth order sample moment. Here, the notation $(\cdot)_j$ picks the jth variable from the vector. The maximum constraint ensures that $\hat{\kappa}_k$ respects the theoretical lower bound of $-2/(p+2)$ [33]. Note that, the finite fourth-order moments are required in order to have a finite elliptical kurtosis.

B. Estimation of the Sphericity

It would be natural to use the SCM in order to develop an estimator for γ_k. However, following [13], [22], [30], and [34], we use the following simple and well performing estimator

$$\hat{\gamma}_k = \left[\frac{p n_k}{n_k - 1} \left(\|\tilde{\mathbf{S}}_k\|_{\mathrm{F}}^2 - \frac{1}{n_k} \right) \right]_1^p,$$

which uses the *sample spatial sign covariance matrix* (SSCM) defined as

$$\tilde{\mathbf{S}}_k = \frac{1}{n_k} \sum_{i=1}^{n_k} \frac{(\mathbf{x}_{ik} - \boldsymbol{\mu}_k)(\mathbf{x}_{ik} - \boldsymbol{\mu}_k)^\top}{\|\mathbf{x}_{ik} - \boldsymbol{\mu}_k\|^2},$$

where the mean $\boldsymbol{\mu}_k$ is estimated by the spatial median [35] $\tilde{\boldsymbol{\mu}}_k = \arg\min_{\boldsymbol{\mu}} \sum_{i=1}^{n_k} \|\mathbf{x}_{ik} - \boldsymbol{\mu}\|$. Particularly, in [13], the SSCM based sphericity estimator was compared to a SCM based sphericity estimator, when sampling from elliptical distributions. The SSCM based estimator performed better in all cases

except when sampling from a MVN distribution. Furthermore, when $\boldsymbol{\mu}$ is known, in [22] it was shown that the expectation of the estimator is asymptotically unbiased as the dimension grows if $\gamma_k/p \to 0$ as $p \to \infty$. This assumption was shown to hold, for example, for the first order autoregressive (AR(1)) covariance matrix but not for the compound symmetry (CS) covariance matrix (see Section VI for the definitions of these covariance matrix structures) [22].

C. Estimation of the Scale and the Inner Products

The scale η_k as well as the inner products $\langle \boldsymbol{\Sigma}_i, \boldsymbol{\Sigma}_j \rangle_{\mathrm{F}}$, for $i \neq j$, can be estimated by using the SCMs, i.e., by $\hat{\eta}_k = \mathrm{tr}(\mathbf{S}_k)/p$ and $\langle \mathbf{S}_i, \mathbf{S}_j \rangle_{\mathrm{F}}$, respectively. However, regarding the inner products $\langle \boldsymbol{\Sigma}_i, \boldsymbol{\Sigma}_j \rangle_{\mathrm{F}}$, we use the same SSCM based estimator as in [22], which is $\hat{\eta}_i \hat{\eta}_j p^2 \langle \tilde{\mathbf{S}}_i, \tilde{\mathbf{S}}_j \rangle_{\mathrm{F}}$. When $i = j$, we estimate $\langle \boldsymbol{\Sigma}_k, \boldsymbol{\Sigma}_k \rangle_{\mathrm{F}} = \|\boldsymbol{\Sigma}_k\|_{\mathrm{F}}^2$, by $p \hat{\gamma}_k \hat{\eta}_k^2$.

V. PRACTICAL CONSIDERATIONS

In this section, we first discuss how to compute the estimates of the optimal tuning parameters in practise. Then, we discuss regularizing the tuning parameters themselves, which will in effect make the method compatible with RDA [2].

A. Computation of the Optimal Tuning Parameters

A straightforward way of solving for the optimal tuning parameters (5) for the proposed estimator (4) is to form a two-dimensional grid of $(\alpha, \beta) \in [0, 1] \times [0, 1]$ and choose the point, which yields the minimum (estimated) MSE in (6). The solution can further be fine-tuned via an alternating convex minimization by iterating between (7) and (8). Convergence of the iterations is addressed in Appendix A–C. This method of finding the optimal tuning parameters was used in the simulations of Section VI, where the method is denoted by **POLY** (due to the polynomial structure of the MSE in Theorem 1). In the simulations, the particular (α, β)-grid was constructed from all pairs in the set $\{0, 0.05, \dots, 1\} \ni (\alpha, \beta)$.

An alternative way of solving for the optimal tuning parameters is to find the critical points of the MSE polynomial (6), which is addressed in Appendix A.

Given estimates for the coefficients C_{ij}, evaluating the equations (6), (7), and (8) is computationally very light. Most of the computation is due to estimating the coefficients and not in optimizing the tuning parameters.

The streamlined analytical estimator (9) with $\mathbf{T} = \mathbf{S}$, which is discussed in Section III-B, is denoted by **POLYs**. In this case, the tuning parameters are found using the plug-in estimates of the optimal values stated in Theorem 3.

B. Averaging of the Tuning Parameters

The estimation accuracy of the optimal tuning parameters $(\alpha_k^\star, \beta_k^\star)_{k=1}^K$ depend on the estimates of the statistical population parameters discussed in Section IV. If the estimation of these parameters is difficult for some data sets, which can happen if the elliptical distribution assumption does not fit the data well, regularization of the tuning parameters themselves can

be useful. A possible way to accomplish this is to average the tuning parameters of the classes by using $\bar{\hat{\alpha}} = \frac{1}{K} \sum \hat{\alpha}_k$ and $\bar{\hat{\beta}} = \frac{1}{K} \sum \hat{\beta}_k$ in place of the class-specific tuning parameters. Averaging the estimated class tuning parameters is reasonable as long as the optimal tuning parameters are not too different from each other, which is the case when the class covariance matrices are similar. If the population covariance matrices are very different from each other, averaging the estimated tuning parameters might degrade the performance. However, by using $(\bar{\hat{\alpha}}, \bar{\hat{\beta}})$ as common tuning parameters for all classes in (4), the resulting RSCM estimator will be compatible with RDA proposed in [2]. Thus, this method can be used instead of cross-validation as an alternative way for choosing the tuning parameters for RDA. In Section VI, we illustrate that this approach enables a significant speed up in the computation of RDA with no effective loss in performance.

The averaged version of the estimator (4), which uses the mean of the estimated tuning parameters for all classes is denoted by **POLY-Ave**. Regarding the streamlined analytical estimator of (9), the averaged version is denoted by **POLYs-Ave**.

VI. NUMERICAL STUDIES

In this section, the performance of the proposed estimators is reported and compared against other competing methods. First, in Subsection VI-A, the MSE performance of the method is examined with synthetic simulations. Then, in Subsection VI-B the method is applied to RDA classification and compared to cross-validation in choosing the tuning parameters.

A. Synthetic Simulations

We evaluated the empirical NMSE performance of the proposed methods as well as the accuracy of the plug-in estimates $\hat{\eta}_k$, $\hat{\kappa}_k$, $\hat{\gamma}_k$, and the estimates of the inner products $\langle \boldsymbol{\Sigma}_i, \boldsymbol{\Sigma}_j \rangle_{\mathrm{F}}$. The results were averaged over 4000 Monte Carlo trials. We simulated four different setups: A, B, C, and D. In each setup, we generated $K = 4$ classes, each of dimension $p = 200$. The data was generated from a multivariate Student's t_ν-distribution with various degrees of freedom ν and the means of the classes were generated from the standard normal distribution $\mathcal{N}_p(\mathbf{0}, \mathbf{I})$ and held fixed over the Monte Carlo trials for setups A, B, and C. Regarding setup D, the mean as well as the other parameters were randomly generated again for each Monte Carlo trial. The four different setups were as follows.

A) *AR(1) process*
- $n_1 = 25$, $n_2 = 50$, $n_3 = 75$, and $n_4 = 100$.
- $\nu = 8$ for all classes.
- $(\boldsymbol{\Sigma}_k)_{ij} = \varrho_k^{|i-j|}$, where $\varrho_1 = 0.2$, $\varrho_2 = 0.3$, $\varrho_3 = 0.4$, and $\varrho_4 = 0.5$.

B) *Compound symmetry*
- $n_1 = 25$, $n_2 = 50$, $n_3 = 75$, and $n_4 = 100$.
- $\nu = 8$ for all classes.
- $(\boldsymbol{\Sigma}_k)_{ii} = 1$ and $(\boldsymbol{\Sigma}_k)_{ij} = \varrho_k$ for $i \neq j$, where $\varrho_1 = 0.2$, $\varrho_2 = 0.3$, $\varrho_3 = 0.4$, and $\varrho_4 = 0.5$.

C) *Mixed structures*
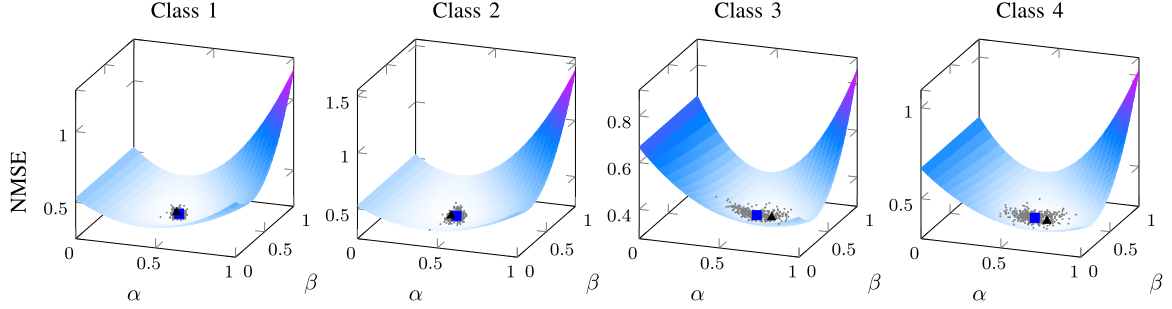- $n_k = 100$, for all $k = 1, 2, 3, 4$.

Fig. 2. The theoretical NMSE as a function of the tuning parameters (α, β) for the setup C (mixed setup). The black triangle (▲) denotes the optimal tuning parameter pair. The first 400 realizations of the estimated tuning parameters over the 4000 Monte Carlo trials are shown as the gray dots. The blue square (■) denotes the mean of the estimated tuning parameters.

- $\nu_1 = 12$, $\nu_2 = 8$, $\nu_3 = 12$, and $\nu_4 = 8$.
- Σ_1 and Σ_2 are as in setup A with $\varrho_1 = \varrho_2 = 0.6$ and Σ_3 and Σ_4 are as in setup B with $\varrho_3 = \varrho_4 = 0.1$.

D) *Randomized*

For each $k = 1, 2, 3, 4$ and each Monte Carlo trial:
- $n_k = \text{Unif}\{10, 200\}$.
- $\nu_k = \text{Unif}\{5, 12\}$.
- $\mu_k = \mathcal{N}_p(\mathbf{0}, \mathbf{I})$.
- Σ_k either AR(1) or CS with equal probability and $\varrho_k = \text{Unif}(0, 0.9)$.

In addition to the proposed estimators, in the simulations we include the sample covariance matrix (SCM), the pooled SCM (1) (POOL), the partially pooled estimator (2) using the method from [30] (PPOOL) for choosing β_k, the method from [13] (ELL1), which is designed for single class covariance matrix estimation and uses a convex combination $\beta_k \mathbf{S}_k + (1 - \beta_k)\hat{\eta}_k \mathbf{I}$. We also include the estimators proposed in [22] (LIN1 and LIN2), which use a nonnegative linear combination of the class SCMs. LIN2 differs from LIN1 in that it incorporates additional shrinkage toward the identity matrix. In addition, we include the shrinkage covariance matrix estimator proposed in [20] (LOOCV1 and LOOCV2). The method is based on estimating the covariance matrix via a nonnegative linear combination of the SCM and target matrices using (low-complexity) leave-one-out cross-validation. In the simulations, for each class k, LOOCV1 uses the pooled SCM and the identity matrix as target matrices for regularizing the SCM of class k. Respectively, for each class k, LOOCV2 uses the individual SCMs of the other classes as well as the identity matrix as target matrices for regularizing the SCM of class k.

The empirical NMSE, $\text{Ave}\|\hat{\Sigma}_k - \Sigma_k\|_{\text{F}}^2/\|\Sigma_k\|_{\text{F}}^2$, is reported for each class in Table I. As can be noted, in the setups A, B, and C, the proposed methods: POLY, POLYs, POLY-Ave, and POLYs-Ave yielded significantly lower NMSE than the methods: SCM, POOL, PPOOL, ELL1, LIN1, and LOOCV1. The only methods, which performed comparably well to POLY and POLYs were LIN2 and LOOCV2. In setup A (AR(1) case), the best performing methods were POLY, POLYs, and LIN2, which all had a similar NMSE. In setup B (CS case), the methods POLY, POLYs performed best (LIN2 and LOOCV2 had a slightly higher NMSE or higher standard deviation). In the setup C (mixed case), LIN2 and LOOCV2 had similar performance

TABLE I
EMPIRICAL NMSE ($\times 10$) OF SETUPS A, B, C, AND D. THE SAMPLE STANDARD DEVIATION ($\times 10$) IS GIVEN IN THE PARENTHESIS

	Class 1	Class 2	Class 3	Class 4	Sum
Setup A. AR(1)					
SCM	115.4 (73.8)	51.5 (32.8)	29.7 (12.8)	18.4 (8.2)	214.9 (81.5)
POOL	12.5 (3.1)	10.6 (2.8)	8.9 (2.3)	7.6 (1.9)	39.6 (10.1)
PPOOL	12.4 (2.6)	10.4 (2.3)	8.8 (1.8)	7.5 (1.6)	39.1 (8.2)
Ell-RSCM	1.1 (0.4)	1.7 (0.2)	2.6 (0.1)	3.3 (0.1)	8.6 (0.5)
LOOCV1	10.6 (20.9)	9.6 (19.5)	8.8 (6.9)	8.3 (5.6)	37.3 (28.6)
LOOCV2	1.1 (0.6)	1.4 (0.2)	2.1 (0.1)	3.0 (0.1)	7.6 (0.7)
LIN1	5.3 (0.5)	4.8 (0.3)	4.6 (0.3)	4.7 (0.3)	19.4 (1.1)
LIN2	0.9 (0.5)	1.3 (0.2)	2.1 (0.1)	3.0 (0.1)	7.3 (0.6)
POLY	0.9 (0.3)	1.3 (0.1)	2.1 (0.1)	3.0 (0.1)	7.2 (0.5)
POLYs	0.8 (0.1)	1.3 (0.1)	2.1 (0.1)	3.0 (0.1)	7.1 (0.3)
POLY-Ave	1.0 (0.5)	1.4 (0.3)	2.1 (0.1)	3.1 (0.1)	7.7 (0.7)
POLYs-Ave	1.0 (0.3)	1.4 (0.2)	2.1 (0.1)	3.1 (0.1)	7.6 (0.5)
Setup B. Compound symmetry					
SCM	14.7 (11.4)	3.6 (3.3)	1.5 (1.0)	0.8 (0.8)	20.6 (11.9)
POOL	10.5 (4.9)	1.8 (1.3)	0.5 (0.3)	0.7 (0.4)	13.5 (6.1)
PPOOL	6.6 (3.4)	1.4 (1.0)	0.5 (0.3)	0.6 (0.4)	9.0 (4.0)
Ell-RSCM	5.7 (1.3)	2.9 (1.1)	1.4 (0.7)	0.7 (0.5)	10.7 (1.9)
LOOCV1	1.7 (4.0)	1.0 (1.8)	0.7 (0.8)	0.7 (0.7)	4.1 (4.5)
LOOCV2	1.5 (3.0)	0.9 (1.2)	0.8 (0.7)	0.7 (0.6)	3.8 (3.3)
LIN1	1.6 (2.2)	0.8 (0.9)	0.6 (0.5)	0.6 (0.4)	3.6 (2.4)
LIN2	1.3 (2.1)	0.7 (0.9)	0.6 (0.5)	0.6 (0.4)	3.2 (2.4)
POLY	1.3 (1.7)	0.7 (0.7)	0.6 (0.4)	0.6 (0.4)	3.2 (2.0)
POLYs	1.3 (1.7)	0.7 (0.7)	0.6 (0.4)	0.6 (0.4)	3.1 (2.0)
POLY-Ave	3.3 (2.2)	0.5 (0.4)	0.8 (0.4)	1.4 (0.5)	6.0 (1.9)
POLYs-Ave	3.3 (2.2)	0.5 (0.4)	0.8 (0.4)	1.4 (0.5)	6.0 (1.9)
Setup C. Mixed					
SCM	12.1 (1.9)	14.7 (6.1)	8.6 (1.5)	10.3 (4.1)	45.6 (7.7)
POOL	6.3 (0.9)	6.3 (0.9)	4.5 (0.5)	4.5 (0.5)	21.5 (2.0)
PPOOL	5.4 (0.5)	5.7 (0.7)	3.8 (0.4)	4.0 (0.5)	19.0 (1.5)
Ell-RSCM	3.7 (0.1)	3.9 (0.2)	3.9 (0.5)	4.1 (0.6)	15.5 (0.8)
LOOCV1	4.9 (0.6)	5.6 (2.4)	4.5 (0.5)	5.0 (1.8)	20.0 (2.9)
LOOCV2	2.9 (0.1)	2.9 (0.1)	2.8 (0.4)	2.8 (0.4)	11.4 (0.8)
LIN1	3.6 (0.2)	3.6 (0.2)	3.0 (0.4)	3.1 (0.4)	13.4 (0.8)
LIN2	2.9 (0.1)	2.9 (0.2)	2.8 (0.4)	2.8 (0.4)	11.4 (0.8)
POLY	3.3 (0.1)	3.4 (0.2)	3.4 (0.5)	3.5 (0.5)	13.7 (0.8)
POLYs	3.3 (0.1)	3.4 (0.2)	3.4 (0.5)	3.5 (0.5)	13.7 (0.8)
POLY-Ave	3.3 (0.1)	3.5 (0.4)	3.4 (0.4)	3.6 (0.5)	13.9 (0.8)
POLYs-Ave	3.3 (0.1)	3.5 (0.4)	3.4 (0.4)	3.6 (0.5)	13.9 (0.8)
Setup D. Randomized					
SCM	20.6 (72.9)	19.2 (39.7)	19.6 (48.7)	23.0 (115.1)	82.4 (149.3)
POOL	41.7 (110.9)	37.9 (102.3)	38.0 (86.3)	41.5 (88.0)	159.1 (254.9)
PPOOL	10.1 (16.0)	9.9 (17.5)	9.9 (17.6)	10.7 (23.6)	40.6 (37.1)
Ell-RSCM	2.1 (1.8)	2.1 (1.7)	2.1 (1.8)	2.1 (3.6)	8.4 (4.8)
LOOCV1	3.6 (39.5)	2.7 (4.7)	3.0 (13.2)	3.2 (16.2)	12.5 (45.2)
LOOCV2	1.7 (9.7)	1.5 (1.7)	1.5 (1.7)	1.6 (2.6)	6.3 (10.4)
LIN1	3.0 (3.0)	2.9 (2.3)	3.0 (2.5)	3.1 (4.1)	12.0 (6.0)
LIN2	1.5 (2.3)	1.4 (1.4)	1.5 (1.7)	1.5 (3.6)	6.0 (4.9)
POLY	1.7 (1.6)	1.6 (1.5)	1.7 (1.6)	1.7 (3.5)	6.6 (4.7)
POLYs	1.7 (1.6)	1.6 (1.5)	1.7 (1.6)	1.7 (3.6)	6.6 (4.9)
POLY-Ave	6.2 (16.3)	5.4 (8.7)	5.8 (10.8)	6.2 (12.1)	23.5 (22.6)
POLYs-Ave	6.2 (15.4)	5.4 (8.7)	5.7 (11.1)	6.3 (24.4)	23.6 (33.6)

and a slightly lower NMSE than the proposed methods. In setup D (randomized case), LIN2 and LOOCV2 had slightly lower NMSE than POLY and POLYs. However, LOOCV2 had over two times the standard deviation compared to LIN2, POLY, and
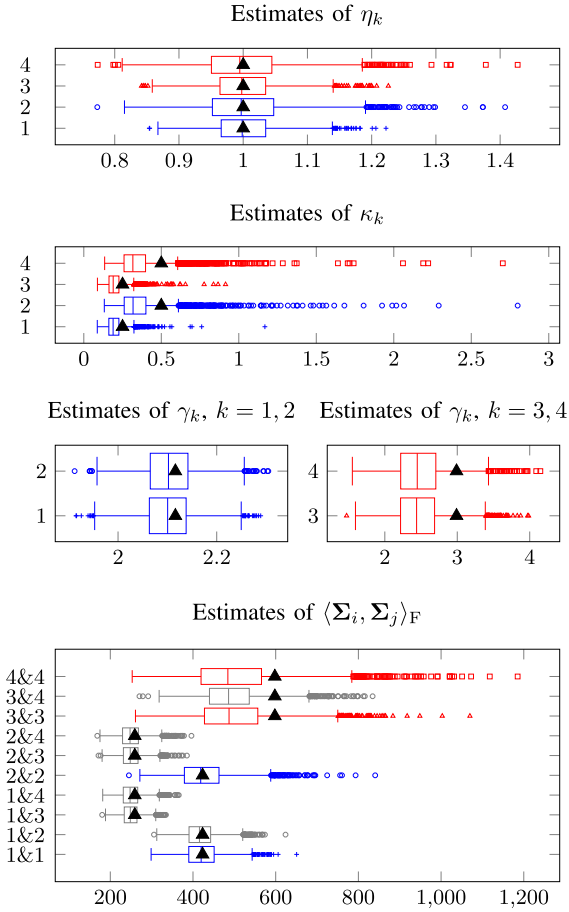
Fig. 3.   Boxplots of the estimates of the parameters η_k, κ_k, γ_k, and the inner products $\langle \boldsymbol{\Sigma}_i, \boldsymbol{\Sigma}_j \rangle_F$ for setup C. The black triangles (▲) denote the true values.

POLYs. In summary, when the covariance matrices had a similar structure (setup A and setup B), the proposed POLY and POLYs methods performed best, whereas when the covariance matrices had a different structure LOOCV2 and LIN2 had a slight edge. The averaged versions of the proposed methods, POLY-Ave and POLYs-Ave, performed well, when the covariance matrices had similar structure (setup A and setup B), whereas in setup C and setup D, their performance degraded compared to the methods using class-specific tuning parameters. Lastly, the performances of POLY and POLYs were almost identical as well as the performances of POLY-Ave and POLYs-Ave.

Fig. 3 shows the boxplots of the estimated statistical parameters η_k, κ_k, γ_k, and the inner products $\langle \boldsymbol{\Sigma}_i, \boldsymbol{\Sigma}_j \rangle_F$ of setup C, where classes 1 and 2 have an AR(1) covariance structure and classes 3 and 4 have a CS structure. The median of the estimated scales η_k coincide with the true values (shown as black triangles (▲) in the plot). The elliptical kurtosis κ_k was more difficult to estimate for heavier tailed distributions (classes 2 and 4 with $\nu = 8$ degrees of freedom) than lighter tailed distributions (classes 1 and 3 with $\nu = 12$ degrees of freedom). The sphericity γ_k was well estimated for classes 1 and 2, which both had the AR(1) covariance structure. Regarding classes 3 and 4, which had the CS covariance structure, the sphericity

TABLE II
SPECIFICATIONS OF THE DATA SETS

Data set	dimension p	classes K	samples n_k
Sonar	59	2	$n_1 = 111$, $n_2 = 97$
Vowel	9	11	$n_k = 90$ for all k
Ionosphere	32	2	$n_1 = 126$, $n_2 = 225$

estimates were noticeably biased. This was most likely due to the sphericity estimate not being asymptotically unbiased for the covariance matrix with CS structure [22]. The inner products were well estimated except for those pairs for which both covariance matrices had a CS structure, i.e., $\langle \boldsymbol{\Sigma}_i, \boldsymbol{\Sigma}_j \rangle_F$, for $(i, j) = \{(3, 3), (3, 4), (4, 4)\}$.

The estimated tuning parameters are shown in Fig. 2 for the setup C. The figure depicts the theoretical NMSE of the estimator as a function of the tuning parameters. The optimal tuning parameter pair $(\alpha^\star, \beta^\star)$ is denoted by the black triangle (▲). The small gray dots correspond to the estimated tuning parameters. The first 400 estimates from the 4000 Monte Carlo trials are shown. The mean of the estimated tuning parameters, denoted by the blue square (■), was very close to the optimal value, although a slight bias could be observed. The estimated tuning parameters were clustered tightly around the optimal point, especially for classes 1 and 2 whose covariance matrices had an AR(1) structure. Regarding classes 3 and 4, whose covariance matrices had the CS structure, there was slightly more spread in the estimates.

B. Discriminant Analysis

We evaluated the performance of the proposed method **POLY-Ave** in discriminant analysis classification. The implementation was written in R (programming language) using three different real data sets: Sonar, Vowel, and Ionosphere, which were obtained from [36] via the R package *mlbench* [37]. The features in the data sets, which were constant for some of the classes were removed. Specifically, from the Vowel data set, we removed the first feature, and from the Ionosphere data set, we removed the first two features. The final specifications of the data sets are given in Table II. In our implementation, we used the package *SpatialNP* [38] for computing the SSCM more efficiently, and the package *tictoc* [39] for recording the computation times of the methods.

In the simulations, we randomly selected a proportion of the samples as training data for estimating the sample means and RSCMs. The remaining data was used as test data to estimate the classification accuracy using the classification rule given in (3). The training set size is given as the number of training samples divided by the total number of samples in the data set. So, the training set size is given as a number between zero and one, where, e.g., 0.3 corresponds to the case where 30% of the data is used as training data and 70% of the data is used as test data to estimate the classification accuracy. The results were averaged over 10 independent Monte Carlo repetitions for each training set size.
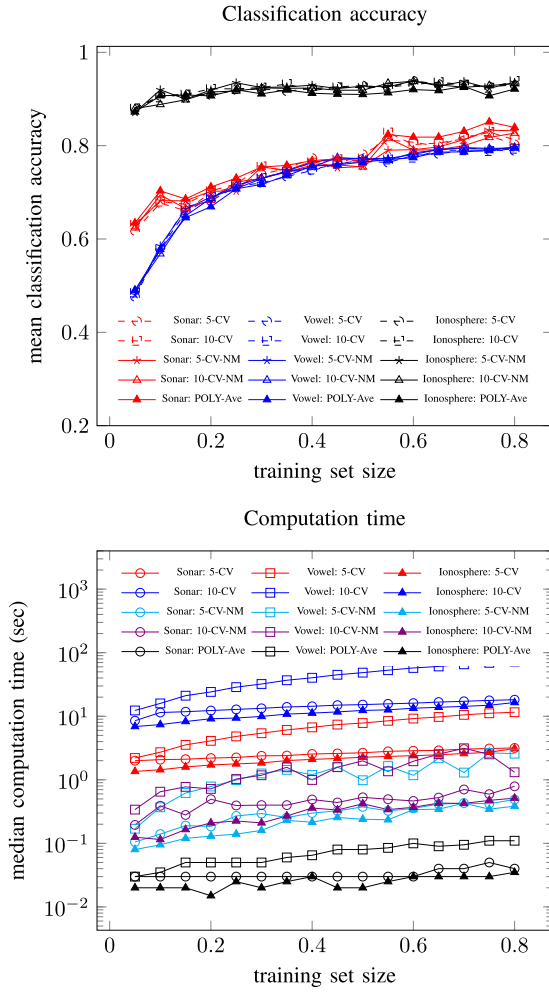
Fig. 4.   Classification accuracy and computation time of the different methods for choosing the tuning parameters.

Fig. 5.   The estimated tuning parameters for the data sets.

For comparison to our proposed method, we included two different methods for implementing cross-validation. The first method chooses the tuning parameters from a two-dimensional grid of candidate parameter values as the minimizer of the 5-fold or 10-fold cross-validated mis-classification rate. In 10-fold cross-validation the two-dimensional grid of tuning parameters scans all the pairs in the set $\{0, 0.125, 0.25, 0.375, 0.5, 0.625, 0.75, 0.875, 1\} \ni \alpha, \beta$. Correspondingly, in 5-fold cross-validation the set is $\{0, 0.25, 0.5, 0.75, 1\} \ni \alpha, \beta$. This method is denoted by **5-CV** and **10-CV** and was implemented using the packages: *caret* [40] and *KlaR* [28], [29]. The second cross-validation method is also based on selecting the tuning parameters that minimize the cross-validated misclassification rate. However, instead of using a predefined two-dimensional grid, the second method uses a Nelder-Mead-algorithm included in the R package *KlaR* [28], [29]. This method was implemented using both 5-fold and 10-fold cross-validation and is denoted by **5-CV-NM** and **10-CV-NM**, respectively.

Fig. 4 depicts the mean classification accuracy on the test set as a function of the training set size as well as the median
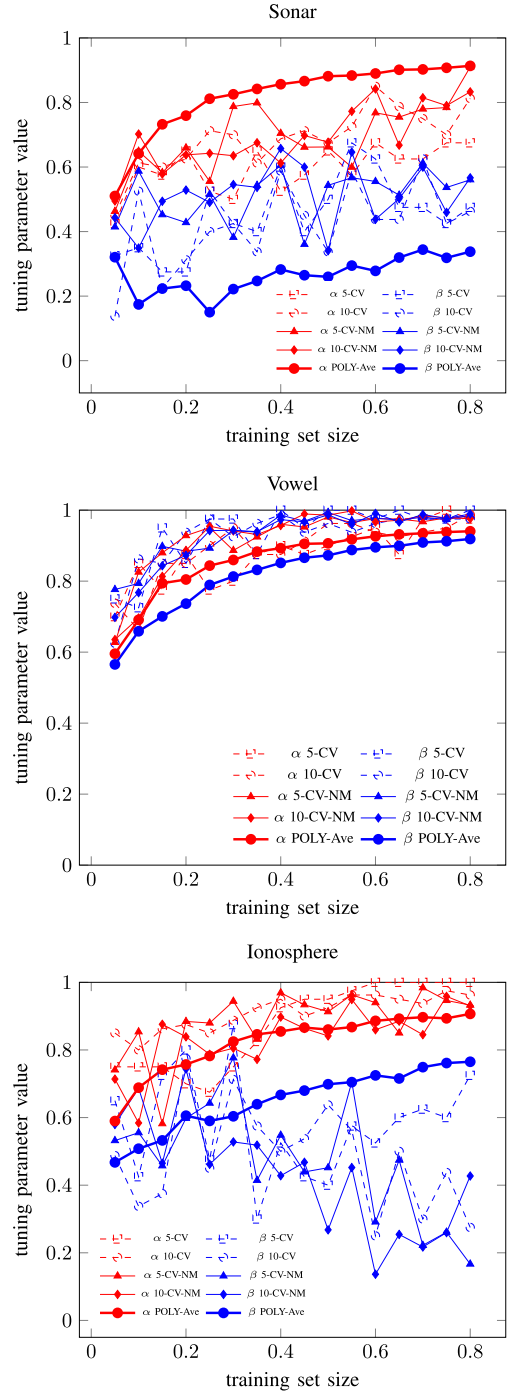
computation time. It can be seen that all of the tested methods gave comparable classification performance. Depending on the data set and training set size, there were only minor differences. The computation time was, however, substantially longer for the cross-validation based methods and it tended to increase along with the training data size faster than for the proposed method. The chosen tuning parameter values are shown in Fig. 5. It can be seen that for the cross-validation based methods the tuning parameter values varied a lot with different training set sizes. The

tuning parameters remained much more stable for our proposed method.

VII. Conclusion

Two regularized sample covariance matrix estimators specified in (4) and in (9) for multiclass problems were considered in this work and their theoretically optimal (in terms of MSE) class-specific tuning parameters were derived in Theorem 1 and Theorem 3, respectively. Since the optimal tuning parameters depend on unknown scalar parameters in (10), a method for their estimation was proposed. The usefulness of the method was supported by the conducted numerical simulations, which demonstrated very good MSE performance when compared to similar methods. By averaging the class-specific tuning parameters, the method was applied for choosing the tuning parameters in an RDA classification framework. The proposed approach had comparable classification performance to cross-validation on each of the three tested real data sets, but had significantly faster computation time. The codes for the proposed methods in Matlab, R, and Python programming languages are available at https://github.com/EliasRaninen.

Appendix A
The MSE and Optimal Tuning Parameters

A Notation and Useful Identities

Let $\mathbf{A}$ and $\mathbf{B}$ be symmetric positive semidefinite matrices of same size. Using the notation

$$\mathbf{I_A} = \frac{\mathrm{tr}(\mathbf{A})}{p}\mathbf{I} \quad \text{and} \quad \mathbf{A^I} = \mathbf{A} - \mathbf{I_A},$$

we have the following identities:

- $\mathrm{tr}(\mathbf{I_A}) = \mathrm{tr}(\mathbf{A})$
- $\mathrm{tr}(\mathbf{A^I}) = 0$
- $\mathbf{I_{A+B}} = \mathbf{I_A} + \mathbf{I_B}$
- $(\mathbf{A} + \mathbf{B})^\mathbf{I} = \mathbf{A^I} + \mathbf{B^I}$
- $\langle \mathbf{I_A}, \mathbf{I_B} \rangle_\mathrm{F} = \langle \mathbf{I_A}, \mathbf{B} \rangle_\mathrm{F} = \langle \mathbf{A}, \mathbf{I_B} \rangle_\mathrm{F} = p^{-1}\mathrm{tr}(\mathbf{A})\mathrm{tr}(\mathbf{B})$
- $\langle \mathbf{A^I}, \mathbf{I_B} \rangle_\mathrm{F} = 0$
- $\langle \mathbf{A^I}, \mathbf{B^I} \rangle_\mathrm{F} = \langle \mathbf{A}, \mathbf{B} \rangle_\mathrm{F} - \langle \mathbf{I_A}, \mathbf{I_B} \rangle_\mathrm{F}$
- $\langle \mathbf{A^I}, \cdot \rangle_\mathrm{F} = \langle \mathbf{A}, \cdot \rangle_\mathrm{F} - \langle \mathbf{I_A}, \cdot \rangle_\mathrm{F}$
- $\|\mathbf{A^I}\|_\mathrm{F}^2 = \|\mathbf{A}\|_\mathrm{F}^2 - \|\mathbf{I_A}\|_\mathrm{F}^2$.

Because of linearity, we also have $\mathbb{E}[\mathbf{I_A}] = \mathbf{I}_{\mathbb{E}[\mathbf{A}]}$ and $\mathbb{E}[\mathbf{A^I}] = \mathbb{E}[\mathbf{A}]^\mathbf{I}$. Regarding expressions involving the pooled SCM $\mathbf{S}$, we have

- $\|\mathbf{S}\|_\mathrm{F}^2 = \sum_i \pi_i^2 \|\mathbf{S}_i\|_\mathrm{F}^2 + \sum_{i \neq j} \pi_i \pi_j \langle \mathbf{S}_i, \mathbf{S}_j \rangle_\mathrm{F}$
- $\|\mathbf{I_S}\|_\mathrm{F}^2 = \sum_i \pi_i^2 \|\mathbf{I}_{\mathbf{S}_i}\|_\mathrm{F}^2 + \sum_{i \neq j} \pi_i \pi_j \langle \mathbf{I}_{\mathbf{S}_i}, \mathbf{I}_{\mathbf{S}_j} \rangle_\mathrm{F}$
- $\langle \cdot, \mathbf{S} \rangle_\mathrm{F} = \sum_j \pi_j \langle \cdot, \mathbf{S}_j \rangle_\mathrm{F}$
- $\langle \cdot, \mathbf{I_S} \rangle_\mathrm{F} = \sum_j \pi_j \langle \cdot, \mathbf{I}_{\mathbf{S}_j} \rangle_\mathrm{F}$.

B MSE of the Estimator

The estimator $\hat{\mathbf{\Sigma}}_k(\alpha, \beta)$ in (4) can be rewritten as

$$\hat{\mathbf{\Sigma}}_k(\alpha, \beta) = \alpha\beta(\mathbf{S}_k - \mathbf{S})^\mathbf{I} + \alpha\mathbf{S^I} + \beta\mathbf{I}_{(\mathbf{S}_k - \mathbf{S})} + \mathbf{I_S}.$$

The squared error $\|\hat{\mathbf{\Sigma}}_k(\alpha, \beta) - \mathbf{\Sigma}_k\|_\mathrm{F}^2$ is

$$\left\| \alpha\beta(\mathbf{S}_k - \mathbf{S})^\mathbf{I} + \alpha\mathbf{S^I} + \beta\mathbf{I}_{(\mathbf{S}_k - \mathbf{S})} + \mathbf{I_S} - \mathbf{\Sigma}_k \right\|_\mathrm{F}^2$$

$$= \alpha^2\beta^2\tilde{C}_{22} + \alpha^2\beta\tilde{C}_{21} + \alpha\beta^2\tilde{C}_{12} + \alpha^2\tilde{C}_{20} + \beta^2\tilde{C}_{02}$$

$$+ \alpha\beta\tilde{C}_{11} + \alpha\tilde{C}_{10} + \beta\tilde{C}_{01} + \tilde{C}_{00},$$

where

$$\begin{aligned}
\tilde{C}_{22} &= \|\mathbf{S}_k^\mathbf{I} - \mathbf{S^I}\|_\mathrm{F}^2 & \tilde{C}_{21} &= 2\langle \mathbf{S}_k^\mathbf{I} - \mathbf{S^I}, \mathbf{S^I} \rangle_\mathrm{F} \\
\tilde{C}_{12} &= 0 & \tilde{C}_{20} &= \|\mathbf{S^I}\|_\mathrm{F}^2 \\
\tilde{C}_{02} &= \|\mathbf{I}_{\mathbf{S}_k} - \mathbf{I_S}\|_\mathrm{F}^2 & \tilde{C}_{11} &= -2\langle \mathbf{S}_k^\mathbf{I} - \mathbf{S^I}, \mathbf{\Sigma}_k \rangle_\mathrm{F} \\
\tilde{C}_{10} &= -2\langle \mathbf{S^I}, \mathbf{\Sigma}_k \rangle_\mathrm{F} & \tilde{C}_{01} &= 2\langle \mathbf{I}_{\mathbf{S}_k} - \mathbf{I_S}, \mathbf{I_S} - \mathbf{\Sigma}_k \rangle_\mathrm{F} \\
\tilde{C}_{00} &= \|\mathbf{I_S} - \mathbf{\Sigma}_k\|_\mathrm{F}^2.
\end{aligned}$$

Taking the expectation of the squared error gives the MSE, with the coefficients $C_{mn} = \mathbb{E}[\tilde{C}_{mn}]$, where $m, n \in \{0, 1, 2\}$.

For example, $C_{20} = \mathbb{E}[\tilde{C}_{20}] = \mathbb{E}[\|\mathbf{S^I}\|_\mathrm{F}^2] = \mathbb{E}[\|\mathbf{S}\|_\mathrm{F}^2] - \mathbb{E}[\|\mathbf{I_S}\|_\mathrm{F}^2]$, where $\mathbb{E}[\|\mathbf{S}\|_\mathrm{F}^2] = \sum_i \pi_i^2 \mathbb{E}[\|\mathbf{S}_i\|_\mathrm{F}^2] + \sum_{i \neq j} \pi_i \pi_j \langle \mathbf{\Sigma}_i, \mathbf{\Sigma}_j \rangle_\mathrm{F}$ and $\mathbb{E}[\|\mathbf{I_S}\|_\mathrm{F}^2] = \sum_i \pi_i^2 \mathbb{E}[\|\mathbf{I}_{\mathbf{S}_i}\|_\mathrm{F}^2] + \sum_{i \neq j} \pi_i \pi_j \langle \mathbf{I}_{\mathbf{\Sigma}_i}, \mathbf{I}_{\mathbf{\Sigma}_j} \rangle_\mathrm{F}$. The estimation of these terms is explained in Section IV.

The optimal tuning parameters can be solved exactly as follows. Let $L_k = \mathrm{MSE}(\hat{\mathbf{\Sigma}}_k)$. The gradient equations of the MSE can be written as

$$\begin{cases} \partial_\alpha L_k = 2\alpha\beta^2 C_{22} + 2\alpha\beta C_{21} + 2\alpha C_{20} + \beta C_{11} + C_{10} \\ \partial_\beta L_k = 2\alpha^2\beta C_{22} + \alpha^2 C_{21} + 2\beta C_{02} + \alpha C_{11} + C_{01}. \end{cases}$$

Set $\partial_\alpha L_k = 0$. Then, solving for α gives

$$\alpha = -\frac{1}{2} \frac{\beta C_{11} + C_{10}}{\beta^2 C_{22} + \beta C_{21} + C_{20}} \tag{11}$$

when $\beta^2 C_{22} + \beta C_{21} + C_{20} \neq 0$, which holds a.s. for continuous distributions (see Appendix A–C). By substituting (11) into the equation $\partial_\beta L_k = 0$, after some algebra, we obtain a rational function. The zeros of the rational function can be computed from the roots of its numerator, which is a quintic (fifth order) polynomial in β. To this end, general polynomial solvers can be used. The critical points of the MSE are obtained by substituting the roots into (11). The optimal tuning parameters then correspond to the critical point, which yields the minimum estimated MSE. If this critical point is not in the feasible set $[0, 1] \times [0, 1]$, the other critical points and the boundaries need to be considered, i.e., the special cases C1–C4 (see Section III). For this purpose the equations (7) and (8) of Theorem 2 can be used.

In practise, it may be simpler to find an approximately optimal tuning parameter pair by forming a two-dimensional grid of (α, β) and choosing the point with the smallest estimated MSE as explained in Subsection V-A. The approximate solution can then further be finetuned by iterating (7) and (8) (see Appendix A–C).

C Alternate Convex Minimization

By considering one of the tuning parameters α or β fixed, the optimization of the remaining one is a convex problem. This is known as biconvexity. To prove this, we show that the second derivatives of the MSE L_k are positive (implying strict convexity). Indeed, we have

$$\begin{cases} \partial_\alpha^2 L_k = 2\beta^2 C_{22} + 2\beta C_{21} + 2C_{20} > 0 \\ \partial_\beta^2 L_k = 2\alpha^2 C_{22} + 2C_{02} > 0. \end{cases}$$

The first equation is an upward opening quadratic function in β. Hence, it is positive if its corresponding discriminant function is negative (there are no real roots). Indeed, the discriminant, $C_{21}^2 - 4C_{22}C_{20}$, is negative, i.e.,

$$(2\,\mathbb{E}[\langle \mathbf{S}_k^{\mathbf{I}} - \mathbf{S}^{\mathbf{I}}, \mathbf{S}^{\mathbf{I}}\rangle_{\mathrm{F}}])^2 - 4\,\mathbb{E}[\|\mathbf{S}_k^{\mathbf{I}} - \mathbf{S}^{\mathbf{I}}\|_{\mathrm{F}}^2]\,\mathbb{E}[\|\mathbf{S}^{\mathbf{I}}\|_{\mathrm{F}}^2] < 0 \text{ a.s.},$$

which follows from the Cauchy-Schwartz inequality. The second equation is also positive since C_{22} and C_{02} are a.s. both positive.

Having concluded that the problem is biconvex, the (unconstrained) solution for β given a fixed value of α is

$$\beta = -\frac{1}{2}\frac{\alpha^2 C_{21} + \alpha C_{11} + C_{01}}{\alpha^2 C_{22} + C_{02}}.$$

The corresponding (unconstrained) solution for α given a fixed value of β was already given in (11). To guarantee that $\alpha, \beta \in [0, 1]$, a projection to the feasible set by the clip function $[\cdot]_0^1$ has to be applied. Hence, we obtain (7) and (8).

Lastly, we make some remarks about the convergence of sequentially iterating (7) and (8). Let $\alpha^{(i)}$ and $\beta^{(i)}$ denote the tuning parameter values at the ith iteration. Denote the MSE function to be minimized as $L_k : X \times Y \to \mathbb{R}$, where $X = [0, 1]$ and $Y = [0, 1]$. Since the MSE L_k is bounded from below and the sequence $\{L_k(\alpha^{(i)}, \beta^{(i)})\}_{i \in \mathbb{N}}$ is monotonically decreasing, it converges [41, Theorem 4.5]. Then, since L_k is continuous, X and Y are closed sets, and both (7) and (8) have unique solutions (due to strict convexity) in the compact set [0,1], by [41, Theorem 4.9], we have $\lim_{i \to \infty} \|(\alpha^{(i+1)}, \beta^{(i+1)}) - (\alpha^{(i)}, \beta^{(i)})\| = 0$. Furthermore, since X and Y are subsets of $\mathbb{R}$, the algorithm is coordinate-wise and the sequence $\{(\alpha^{(i)}, \beta^{(i)})\}_{i \in \mathbb{N}}$ converges [41, pp. 398]. If the accumulation point of the sequence $\{(\alpha^{(i)}, \beta^{(i)})\}_{i \in \mathbb{N}}$ is in the interior of $[0, 1] \times [0, 1]$, then it is a stationary point [41, Corollary 4.10]. However, in theory, the stationary point can be a global minimum, local minimum, or a saddle point. Therefore, it is important to select the starting point for the iterations carefully.

APPENDIX B
STREAMLINED ANALYTICAL ESTIMATOR

Let $\mathbf{T} \in \{\mathbf{S}, \mathbf{S}_k\}$, then the analytical estimator is

$$\tilde{\mathbf{\Sigma}}_k(\alpha, \beta) = \alpha\beta(\mathbf{S}_k - \mathbf{S}) + \alpha(\mathbf{S} - \mathbf{I_T}) + \mathbf{I_T}.$$

The squared error is

$$\|\alpha\beta(\mathbf{S}_k - \mathbf{S}) + \alpha(\mathbf{S} - \mathbf{I_T}) + \mathbf{I_T} - \mathbf{\Sigma}_k\|_{\mathrm{F}}^2$$
$$= \alpha^2\beta^2 \tilde{B}_{22} + \alpha^2\beta \tilde{B}_{21} + \alpha^2 \tilde{B}_{20} + \alpha\beta \tilde{B}_{11} + \alpha \tilde{B}_{10} + \tilde{B}_{00},$$

where

$$\begin{array}{ll}
\tilde{B}_{22} = \|\mathbf{S}_k - \mathbf{S}\|_{\mathrm{F}}^2 & \tilde{B}_{21} = 2\langle \mathbf{S}_k - \mathbf{S}, \mathbf{S} - \mathbf{I_T}\rangle_{\mathrm{F}} \\
\tilde{B}_{20} = \|\mathbf{S} - \mathbf{I_T}\|_{\mathrm{F}}^2 & \tilde{B}_{11} = 2\langle \mathbf{S}_k - \mathbf{S}, \mathbf{I_T} - \mathbf{\Sigma}_k\rangle_{\mathrm{F}} \\
\tilde{B}_{10} = 2\langle \mathbf{S} - \mathbf{I_T}, \mathbf{I_T} - \mathbf{\Sigma}_k\rangle_{\mathrm{F}} & \tilde{B}_{00} = \|\mathbf{I_T} - \mathbf{\Sigma}_k\|_{\mathrm{F}}^2.
\end{array}$$

By taking the expectation, we get the MSE, i.e., $B_{mn} = \mathbb{E}[\tilde{B}_{mn}]$, for $m, n \in \{0, 1, 2\}$. The partial derivatives of the MSE are

$$\begin{cases} \partial_\alpha L_k = 2\alpha\beta^2 B_{22} + 2\alpha\beta B_{21} + 2\alpha B_{20} + \beta B_{11} + B_{10} \\ \partial_\beta L_k = 2\alpha^2\beta B_{22} + \alpha^2 B_{21} + \alpha B_{11}. \end{cases}$$

By setting $\partial_\beta L_k = 0$, it is easy to see that $(\alpha = 0, \beta = -B_{10}/B_{11})$ is a critical point. Note, however, that when $\alpha = 0$, the estimator does not depend on β. Then, if we assume $\alpha \neq 0$, solving for β yields

$$\beta = -\frac{1}{2}\frac{\alpha B_{21} + B_{11}}{\alpha B_{22}}.$$

Substituting this expression into $\partial_\alpha L_k = 0$ and solving for α yields

$$\alpha = \frac{2B_{10}B_{22} - B_{11}B_{21}}{B_{21}^2 - 4B_{20}B_{22}}. \tag{12}$$

Substituting this back to the equation for β gives

$$\beta = \frac{2B_{11}B_{20} - B_{10}B_{21}}{2B_{10}B_{22} - B_{11}B_{21}}. \tag{13}$$

The Hessian matrix is

$$\mathbf{H} = \begin{pmatrix} L_{\alpha\alpha} & L_{\alpha\beta} \\ L_{\beta\alpha} & L_{\beta\beta} \end{pmatrix},$$

where the second partial derivatives are

$$L_{\alpha\alpha} = \partial_\alpha^2 L_k = 2\beta^2 B_{22} + 2\beta B_{21} + 2B_{20}$$
$$L_{\beta\beta} = \partial_\beta^2 L_k = 2\alpha^2 B_{22}$$
$$L_{\beta\alpha} = L_{\alpha\beta} = \partial_\alpha\partial_\beta L_k = 4\alpha\beta B_{22} + 2\alpha B_{21} + B_{11}.$$

The determinant of the Hessian matrix is

$$\det(\mathbf{H}) = L_{\alpha\alpha}L_{\beta\beta} - L_{\alpha\beta}^2.$$

For the critical point $(\alpha = 0, \beta = -B_{10}/B_{11})$, we have $L_{\beta\beta} = 0$ and $L_{\alpha\beta} = B_{11}$. The determinant of the Hessian is then $-B_{11}^2 < 0$. This implies that the eigenvalues of the Hessian matrix have different signs, and hence, the corresponding critical point is a saddle point. Regarding the other critical point defined by (12) and (13), after some manipulations, one has that the determinant of the Hessian is

$$\frac{(2B_{10}B_{22} - B_{11}B_{21})^2}{4B_{20}B_{22} - B_{21}^2},$$

which is positive since by applying the Cauchy-Schwartz inequality

$$0 < B_{21}^2 = (2\,\mathbb{E}[\langle \mathbf{S}_k - \mathbf{S}, \mathbf{S} - \mathbf{I_T}\rangle_{\mathrm{F}}])^2$$
$$< 4\,\mathbb{E}[\|\mathbf{S}_k - \mathbf{S}\|_{\mathrm{F}}^2]\,\mathbb{E}[\|\mathbf{S} - \mathbf{I_T}\|_{\mathrm{F}}^2] = 4B_{22}B_{20},$$

which holds with strict inequalities a.s. for any continuous distribution. Furthermore, as also $L_{\beta\beta} > 0$ a.s. (and hence $L_{\alpha\alpha} > 0$ a.s.), the second critical point is a local minimum.

REFERENCES

[1] M. Bilodeau and D. Brenner, *Theory of Multivariate Statistics*. New York, NY, USA: Springer-Verlag, 1999.
[2] J. H. Friedman, "Regularized discriminant analysis," *J. Amer. Stat. Assoc.*, vol. 84, no. 405, pp. 165–175, 1989.

[3] 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.

[4] 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.

[5] 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, 2005, Art. no. 32.

[6] 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.

[7] 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.

[8] 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.

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

[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] 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.

[13] 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.

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

[15] Y. I. Abramovich and O. Besson, "On the expected likelihood approach for assessment of regularization covariance matrix," *IEEE Signal Process. Lett.*, vol. 22, no. 6, pp. 777–781, Jun. 2015.

[16] 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.

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

[18] 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.

[19] 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.

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

[22] E. Raninen, D. E. Tyler, and E. Ollila, "Linear pooling of sample covariance matrices," 2021, *arXiv: 2008.05854*.

[23] 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.

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

[25] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, 2nd ed. New York, NY, USA: Springer, 2009.

[26] X. Yang, K. Elkhalil, A. Kammoun, T. Y. Al-Naffouri, and M.-S. Alouini, "Regularized discriminant analysis: A large dimensional study," in *Proc. IEEE Int. Symp. Inf. Theory*, 2018, pp. 536–540.

[27] K. Elkhalil, A. Kammoun, R. Couillet, T. Y. Al-Naffouri, and M.-S. Alouini, "A large dimensional study of regularized discriminant analysis," *IEEE Trans. Signal Process.*, vol. 68, pp. 2464–2479, 2020.

[28] C. Weihs, U. Ligges, K. Luebke, and N. Raabe, "klaR analyzing German business cycles," in *Data Analysis and Decision Support*, D. Baier, R. Decker, and L. Schmidt-Thieme, Eds. Berlin, Heidelberg: Springer-Verlag, 2005, pp. 335–343.

[29] C. Roever *et al.*, "klaR: Classification and visualization," R package version 0.6-15, 2020. [Online]. Available: https://CRAN.R-project.org/package=klaR

[30] 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.

[31] K. W. Fang, *Symmetric Multivariate and Related Distributions*. New York, NY, USA: CRC Press, 2018.

[32] G. Frahm, "Generalized elliptical distributions: Theory and applications," Ph.D. dissertation, Fac. Manage. Economics and Social Sciences, Univ. Cologne, 2004.

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

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

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

[36] D. J. Newman, S. Hettich, C. L. Blake, and C. J. Merz, "UCI repository of machine learning databases," 1998. [Online]. Available: https://archive.ics.uci.edu/ml/

[37] F. Leisch and E. Dimitriadou, "MLBENCH: Machine learning benchmark problems," R package version 2.1-3, 2011. [Online]. Available: https://CRAN.R-project.org/package=mlbench

[38] S. Sirkiä, J. Miettinen, K. Nordhausen, H. Oja, and S. Taskinen, "SpatialNP: Multivariate nonparametric methods based on spatial signs and ranks," R package version 1.1-4, 2019. [Online]. Available: https://CRAN.R-project.org/package=SpatialNP

[39] S. Izrailev, "Tictoc: Functions for timing R scripts, as well as implementations of stack and list structures," R package version 1.0, 2014. [Online]. Available: https://CRAN.R-project.org/package=tictoc

[40] M. Kuhn, "Caret: Classification and regression training," R package version 6.0-86, 2020. [Online]. Available: https://CRAN.R-project.org/package=caret

[41] J. Gorski, F. Pfeuffer, and K. Klamroth, "Biconvex sets and optimization with biconvex functions: A survey and extensions," *Math. Methods Operations Res.*, vol. 66, no. 3, pp. 373–407, Dec. 2007.

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.

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, 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, Finland. He is also an Adjunct Professor (Statistics) of Oulu University. He is a 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*.