Raninen, Elias; Ollila, Esa

Optimal Pooling of Covariance Matrix Estimates Across Multiple Classes

Published in:
2018 IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP 2018 - Proceedings

DOI:
10.1109/ICASSP.2018.8461327

Published: 10/09/2018

Document Version
Peer-reviewed accepted author manuscript, also known as Final accepted manuscript or Post-print

Please cite the original version:
Raninen, E., & Ollila, E. (2018). Optimal Pooling of Covariance Matrix Estimates Across Multiple Classes. In *2018 IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP 2018 - Proceedings* (Vol. 2018-April, pp. 4224-4228). Article 8461327 (Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing). IEEE. https://doi.org/10.1109/ICASSP.2018.8461327

OPTIMAL POOLING OF COVARIANCE MATRIX ESTIMATES ACROSS MULTIPLE CLASSES

Elias Raninen and Esa Ollila*

Aalto University, Dept. of Signal Processing and Acoustics, P.O. Box 15400, FI-00076 Aalto, Finland

ABSTRACT

The paper considers the problem of estimating the covariance matrices of multiple classes in a low sample support condition, where the data dimensionality is comparable to, or larger than, the sample sizes of the available data sets. In such conditions, a common approach is to shrink the class sample covariance matrices (SCMs) towards the pooled SCM. The success of this approach hinges upon the ability to choose the optimal regularization parameter. Typically, a common regularization level is shared among the classes and determined via a procedure based on cross-validation. We use class-specific regularization levels since this enables the derivation of the optimal regularization parameter for each class in terms of the minimum mean squared error (MMSE). The optimal parameters depend on the true unknown class population covariances. Consistent estimators of the parameters can, however, be easily constructed under the assumption that the class populations follow (unspecified) elliptically symmetric distributions. We demonstrate the performance of the proposed method via a simulation study as well as via an application to discriminant analysis using both synthetic and real data sets.

***Index Terms*—** Covariance matrix estimation, regularization, elliptical distribution, classification.

1. INTRODUCTION

The problem of estimating the covariance matrices of K classes appears in classification problems as well as in graphical models, where in the latter the inverse of the covariance matrix gives information about the conditional dependence structure of the graph [1]. In the context of classification, the covariance matrices determine the shape of the data within each class in the feature space. In low sample support conditions, where the data dimensionality p is comparable to, or larger than, the sample sizes n_k available from each population, conventional SCMs are susceptible to high variance and may not even be positive definite and hence invertible. Therefore, regularized sample covariance matrix (RSCM) estimators are needed. Their usefulness has been validated in many challenging data analysis problems [2] [3] [4]. The level of regularization in the RSCM is determined by the regularization parameter, which is typically chosen via cross-validation. However, procedures based on cross-validation can be computationally prohibitive, especially when p is large. Hence, simple methods that do not require excessive tuning of hyperparameters are more desirable, especially in applications where the aforementioned conditions are predominant, such as in remote sensing [5].

We present an analytical method for choosing the regularization levels of a particular form of the RSCM, where the individual class SCMs are shrunk towards the pooled SCM. This type of regularization often improves the estimation accuracy when there is low sample support and when the population covariance matrices share a common structure. In classification, this type of regularization is used in regularized discriminant analysis [2] (RDA), in which a common regularization level for the classes is determined via cross-validation. Class-specific regularization has been used in a slightly different RSCM formulation in [6], where the parameters were likewise chosen using a method based on cross-validation. Our approach is different in that we derive analytical expressions for the optimal regularization parameters of the classes. Even though the derived expressions depend on the true unknown population covariance matrices, consistent estimates can easily be constructed in a similar fashion as in [7], which studied the single covariance matrix estimation problem. We only need to assume that the class populations follow (possibly different unspecified) elliptical distributions with finite 4th order moments. In the conducted simulations, the proposed method performed not only better than the method based on cross-validation, but also enjoys the advantages of low computational complexity and ease of implementation.

The remainder of this paper is organized as follows. Section 2 introduces the proposed estimator and an analytical expression for the optimal class-specific regularization parameter is derived. In Section 3, we construct an estimator of the optimal regularization parameter under the assumption of (unspecified) elliptical populations. In Section 4, we assess the mean squared error (MSE) performance of the proposed method via a simulation study. In Section 5, the method is applied to discriminant analysis, where the classification performance is evaluated using both synthetic and real data examples. Section 6 concludes the paper.

*The research of E. Raninen was supported by the Academy of Finland grant No. 298118.

2. OPTIMAL REGULARIZATION PARAMETER

Consider observing independent and identically distributed (i.i.d.) p-dimensional samples from K different classes (or populations), each class having n_k number of samples. Let $\{\mathbf{x}_{k,i}\}_{i=1}^{n_k}$ denote the data set of the kth class. The scatter of the data in the p-dimensional feature space is characterized by the covariance matrices

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

where $\boldsymbol{\mu}_k = \mathbb{E}[\mathbf{x}_k]$ and $\mathbf{x}_k$ denotes a random vector from the kth class. Let $\overline{\mathbf{x}}_k = (1/n_k)\sum_{i=1}^{n_k}\mathbf{x}_{k,i}$ and

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

denote the sample mean vector and the (unbiased) SCM of the kth class. We are interested in an RSCM defined for class k as

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

where $\beta_k \in [0,1]$, and $\mathbf{S}$ denotes the pooled SCM, i.e., $\mathbf{S} = \sum_{k=1}^{K}\pi_k\mathbf{S}_k$, where $\pi_k = n_k/(\sum_{j=1}^{K} n_j)$. For ease of notation, we will omit the subscript k from β_k hereafter. Our goal is to determine the optimal regularization parameter $\beta^\star$ for class k which minimizes the MSE between the regularized estimator (1) and the true population covariance matrix,

$$\beta^\star = \underset{\beta\in[0,1]}{\arg\min}\ \mathbb{E}\big[\|\hat{\mathbf{\Sigma}}_k(\beta) - \mathbf{\Sigma}_k\|_{\mathrm{F}}^2\big], \tag{2}$$

where $\|\cdot\|_{\mathrm{F}}$ denotes the Frobenius matrix norm, i.e., $\|\mathbf{A}\|_{\mathrm{F}}^2 = \mathrm{tr}\left(\mathbf{A}^\top\mathbf{A}\right)$ for any square matrix $\mathbf{A}$.

Next, we derive the optimal regularization parameter. Write $L_k(\beta) = \mathbb{E}\big[\|\hat{\mathbf{\Sigma}}_k(\beta) - \mathbf{\Sigma}_k\|_{\mathrm{F}}^2\big]$. Then note that

$$\begin{aligned} L_k(\beta) &= \mathbb{E}\left[\|\beta\mathbf{S}_k + (1-\beta)\mathbf{S} - \mathbf{\Sigma}_k\|_{\mathrm{F}}^2\right] \\ &= \beta^2\mathbb{E}\left[\|\mathbf{S}_k - \mathbf{S}\|_{\mathrm{F}}^2\right] + \mathbb{E}\left[\|\mathbf{\Sigma}_k - \mathbf{S}\|_{\mathrm{F}}^2\right] \\ &\quad - 2\beta\mathbb{E}\left[\mathrm{tr}\left((\mathbf{S}_k - \mathbf{S})(\mathbf{\Sigma}_k - \mathbf{S})\right)\right]. \end{aligned} \tag{3}$$

The second derivative $L_k''(\beta)$ is positive whenever $\mathbf{S}_k \neq \mathbf{S}$. Thus under this assumption, the loss function is strictly convex and the optimal regularization parameter value can be found by solving $L_k'(\beta^\star) = 0$, which yields

$$\beta^\star = \frac{\mathbb{E}\left[\mathrm{tr}\left((\mathbf{S}_k - \mathbf{S})(\mathbf{\Sigma}_k - \mathbf{S})\right)\right]}{\mathbb{E}\left[\|\mathbf{S}_k - \mathbf{S}\|_{\mathrm{F}}^2\right]}. \tag{4}$$

It can be seen that if $\mathbf{S}_k \approx \mathbf{\Sigma}_k$, i.e., the SCM is close to the true covariance matrix, then $\beta^\star \approx 1$, and the estimator gives all the weight to the SCM. By expanding the expressions in the numerator and denominator of (4), we get

$$\beta^\star = \frac{\mathrm{tr}\left(\mathbf{\Sigma}_k^2\right) - \mathbb{E}\left[\mathrm{tr}\left(\mathbf{S}_k\mathbf{S}\right)\right] - \mathrm{tr}\left(\mathbf{\Sigma}\mathbf{\Sigma}_k\right) + \mathbb{E}\left[\mathrm{tr}\left(\mathbf{S}^2\right)\right]}{\mathbb{E}\left[\mathrm{tr}\left(\mathbf{S}_k^2\right)\right] - 2\mathbb{E}\left[\mathrm{tr}\left(\mathbf{S}_k\mathbf{S}\right)\right] + \mathbb{E}\left[\mathrm{tr}\left(\mathbf{S}^2\right)\right]}, \tag{5}$$

where $\mathbf{\Sigma} = \mathbb{E}\left[\mathbf{S}\right] = \sum_{k=1}^{K}\pi_k\mathbf{\Sigma}_k$,

$$\begin{aligned} \mathbb{E}\left[\mathrm{tr}\left(\mathbf{S}^2\right)\right] &= \textstyle\sum_j \pi_j^2\mathbb{E}\left[\mathrm{tr}\left(\mathbf{S}_j^2\right)\right] + \sum_{i\neq j}\pi_i\pi_j\mathrm{tr}\left(\mathbf{\Sigma}_i\mathbf{\Sigma}_j\right), \\ \mathbb{E}\left[\mathrm{tr}\left(\mathbf{S}_k\mathbf{S}\right)\right] &= \pi_k\mathbb{E}\left[\mathrm{tr}\left(\mathbf{S}_k^2\right)\right] + \textstyle\sum_{j=1,j\neq k}^{K}\pi_j\mathrm{tr}\left(\mathbf{\Sigma}_k\mathbf{\Sigma}_j\right), \end{aligned}$$

and $\mathrm{tr}\left(\mathbf{\Sigma}_k\mathbf{\Sigma}\right) = \pi_k\mathrm{tr}\left(\mathbf{\Sigma}_k^2\right) + \sum_{j=1,j\neq k}^{K}\pi_j\mathrm{tr}\left(\mathbf{\Sigma}_k\mathbf{\Sigma}_j\right)$. With some further algebra, $\beta^\star$ in (5) becomes

$$\beta^\star = \frac{(1-\pi_k)\mathrm{tr}\left(\mathbf{\Sigma}_k^2\right) - \pi_k\mathbb{E}\left[\mathrm{tr}\left(\mathbf{S}_k^2\right)\right] + \delta_k}{(1-2\pi_k)\mathbb{E}\left[\mathrm{tr}\left(\mathbf{S}_k^2\right)\right] + \delta_k}, \tag{6}$$

where

$$\begin{aligned} \delta_k = &\textstyle\sum_j \pi_j^2\mathbb{E}\left[\mathrm{tr}\left(\mathbf{S}_j^2\right)\right] - 2\sum_{j=1,j\neq k}^{K}\pi_j\mathrm{tr}\left(\mathbf{\Sigma}_k\mathbf{\Sigma}_j\right) \\ &+ \textstyle\sum_{i\neq j}\pi_i\pi_j\mathrm{tr}\left(\mathbf{\Sigma}_i\mathbf{\Sigma}_j\right). \end{aligned}$$

The value of the MSE at the optimum is then

$$\begin{aligned} L_k(\beta^\star) = &(\beta^\star)^2\mathbb{E}\left[\mathrm{tr}\left(\mathbf{S}_k^2\right)\right] + (1-\beta^\star)^2\mathbb{E}\left[\mathrm{tr}\left(\mathbf{S}^2\right)\right] \\ &+ (1-2\beta^\star)\mathrm{tr}\left(\mathbf{\Sigma}_k^2\right) + 2\beta^\star(1-\beta^\star)\mathbb{E}\left[\mathrm{tr}\left(\mathbf{S}_k\mathbf{S}\right)\right] \\ &- 2(1-\beta^\star)\mathrm{tr}\left(\mathbf{\Sigma}\mathbf{\Sigma}_k\right), \end{aligned} \tag{7}$$

which follows from a straightforward calculation of (3).

The optimal regularization parameter depends on the unknown true covariance matrices. Hence, this parameter must be estimated, which forms the topic of the next section.

3. ESTIMATING THE OPTIMAL REGULARIZATION PARAMETER

An estimate $\hat{\beta}$ of the optimal regularization parameter $\beta^\star$ of class k in (6) is obtained by the estimates of $\mathrm{tr}(\mathbf{\Sigma}_k^2)$, $\mathbb{E}\left[\mathrm{tr}\left(\mathbf{S}_k^2\right)\right]$, and $\mathrm{tr}\left(\mathbf{\Sigma}_i\mathbf{\Sigma}_j\right)$, where $i \neq j$. The last expression can simply be estimated with $\mathrm{tr}\left(\mathbf{S}_i\mathbf{S}_j\right)$, which follows from independence, i.e., $\mathbb{E}\left[\mathrm{tr}\left(\mathbf{S}_i\mathbf{S}_j\right)\right] = \mathrm{tr}\left(\mathbb{E}\left[\mathbf{S}_i\right]\mathbb{E}\left[\mathbf{S}_j\right]\right) = \mathrm{tr}\left(\mathbf{\Sigma}_i\mathbf{\Sigma}_j\right)$.

We now assume that the random sample of the kth class, $\{\mathbf{x}_{k,i}\}_{i=1}^{n_k}$, is from an elliptically symmetric distribution with mean vector $\boldsymbol{\mu}_k$ and covariance matrix $\mathbf{\Sigma}_k$, and that it possesses finite 4th order moments. Then the probability density function (p.d.f.) of $\mathbf{x}_{k,i}$ is of the form

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

where $g_k : [0,\infty) \to [0,\infty)$ is a fixed function, called the *density generator*, which is independent of $\mathbf{x}$, $\boldsymbol{\mu}_k$ and $\mathbf{\Sigma}_k$, and C_k is a normalizing constant ensuring that $f_k(\mathbf{x})$ integrates to 1. We denote this case as $\mathbf{x}_{k,i} \sim \mathcal{E}_p(\boldsymbol{\mu}_k, \mathbf{\Sigma}_k, g_k)$. The functional form of the density generator $g_k(\cdot)$ determines the elliptical distribution. For example, the multivariate normal (MVN) distribution is obtained when $g_k(t) = \exp(-t/2)$. In our derivations, we do not have to assume that the density generators are the same for each class or even specify the elliptical populations; e.g., $g_k(t)$ can be the generator of the

multivariate t-distribution, MVN, or any other elliptical distribution with finite 4th order moments. Under the assumption that $\mathbf{x}_{k,i} \sim \mathcal{E}_p(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k, g_k)$ and by using the results from [8], one can show that

$$\mathbb{E}\left[\mathrm{tr}\left(\mathbf{S}_k^2\right)\right] = p\eta_k^2\left\{\tau_1(p+\gamma_k) + (\tau_2+1)\gamma_k\right\}, \quad (8)$$

where $\tau_1 = (n_k-1)^{-1} + \kappa_k/n_k$, $\tau_2 = \kappa_k/n_k$, $\eta_k = \mathrm{tr}\left(\boldsymbol{\Sigma}_k\right)/p$, $\gamma_k = p\mathrm{tr}\left(\boldsymbol{\Sigma}_k^2\right)/\mathrm{tr}\left(\boldsymbol{\Sigma}_k\right)^2$, and κ_k is the elliptical kurtosis parameter of the kth population,

$$\kappa_k = (1/3)\cdot \mathrm{kurt}(x_j), \quad (9)$$

where x_j denotes any element of $\mathbf{x}_{k,i} \sim \mathcal{E}_p(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k, g_k)$ and $\mathrm{kurt}(x)$ denotes the kurtosis of a random variable x.

A consistent estimate $\hat{\beta}$ of $\beta^\star$ for the kth class is obtained given consistent estimates of η_k, γ_k and κ_k. An obvious estimate of η_k is $\hat{\eta}_k = \mathrm{tr}\left(\mathbf{S}_k\right)/p$. For γ_k, we used the estimator

$$\hat{\gamma}_{\mathrm{sgn},k} = p\mathrm{tr}\left(\mathbf{S}_{\mathrm{sgn},k}^2\right) - \frac{p}{n_k}, \quad (10)$$

which uses the *sample sign covariance matrix* [9],

$$\mathbf{S}_{\mathrm{sgn},k} = \frac{1}{n_k}\sum_{i=1}^{n_k}\frac{(\mathbf{x}_{k,i}-\hat{\boldsymbol{\mu}}_k)(\mathbf{x}_{k,i}-\hat{\boldsymbol{\mu}}_k)^\top}{\|\mathbf{x}_{k,i}-\hat{\boldsymbol{\mu}}_k\|^2}, \quad (11)$$

where $\hat{\boldsymbol{\mu}}_k = \arg\min_{\boldsymbol{\mu}} \sum_{i=1}^{n_k}\|\mathbf{x}_{k,i}-\boldsymbol{\mu}\|$ is the spatial sample median [10] of the kth data set. It was shown in [11] that $\hat{\gamma}_{\mathrm{sgn},k}$ is a consistent estimator of γ_k under the random matrix theory regime. Their proof considered the centered case ($\boldsymbol{\mu}_k = 0$), where the centering by the spatial sample median was not required in (11). We have not yet extended this result to the case of unknown center, however, the estimator (10) performed well in our simulations. As in [7], a consistent estimator of κ_k is

$$\hat{\kappa}_k = \max\left\{-\tfrac{2}{p+2}, \tfrac{1}{3p}\sum_{j=1}^{p}\hat{q}_{k,j}\right\}, \quad (12)$$

where $\hat{q}_{k,j}$ is the sample kurtosis of the jth variable of the kth class, i.e., $\hat{q}_{k,j} = m_{k,j}^{(4)}/(m_{k,j}^{(2)})^2 - 3$, where $m_{k,j}^{(l)} = (1/n_k)\sum_{i=1}^{n_k}((\mathbf{x}_{k,i})_j - (\overline{\mathbf{x}}_k)_j)^l$ denotes the lth order sample moment of the jth variable of the kth class.

As the final estimate of the optimal regularization parameter, we used $\max\{0, \min\{1, \hat{\beta}\}\}$ since $\beta^\star$ needs to be within $[0,1]$.

4. SIMULATION STUDY

We now illustrate the performance of the proposed method via a simulation study. We generated training data of dimension $p = 20$ comprising $K = 4$ classes and $n = \sum_k n_k = 100$ samples. The data was generated from a Student's t_ν-distribution with $\nu = 10$ degrees of freedom. The first class, $k = 1$, had zero mean and the subsequent classes, $k = 2$, 3, and 4, had means in orthogonal directions such that $\|\boldsymbol{\mu}_k\| = 1 + k$. We simulated three different set-ups:

Table 1. The empirical NMSE $\tilde{L}_k$ for the covariance estimates in the set-ups 1 to 3 (from top to down). The corresponding standard deviations are shown in parenthesis.

	$\tilde{L}_1$	$\tilde{L}_2$	$\tilde{L}_3$	$\tilde{L}_4$	Sum
Oracle	0.99 (0.41)	0.52 (0.16)	0.27 (0.06)	0.28 (0.03)	2.06 (0.48)
Prop 1	0.98 (0.38)	0.50 (0.15)	0.28 (0.06)	0.29 (0.03)	2.04 (0.45)
Pool	4.47 (1.15)	0.62 (0.21)	0.27 (0.06)	0.28 (0.03)	5.63 (1.42)
SCM	1.13 (0.48)	1.12 (0.41)	1.20 (0.67)	1.18 (0.53)	4.63 (1.07)
Oracle	2.13 (1.09)	0.70 (0.29)	0.32 (0.12)	0.24 (0.05)	3.40 (1.23)
Prop 1	2.07 (0.97)	0.67 (0.20)	0.31 (0.08)	0.24 (0.05)	3.29 (1.06)
Pool	6.80 (1.85)	0.96 (0.37)	0.32 (0.13)	0.24 (0.05)	8.32 (2.38)
SCM	2.89 (1.63)	1.42 (0.84)	0.92 (0.35)	0.73 (0.41)	5.95 (1.94)
Oracle	1.25 (0.71)	0.93 (0.40)	0.40 (0.17)	0.24 (0.09)	2.81 (0.90)
Prop 1	1.18 (0.60)	0.88 (0.30)	0.38 (0.12)	0.24 (0.07)	2.68 (0.72)
Pool	6.25 (1.77)	2.04 (0.73)	0.43 (0.23)	0.29 (0.05)	9.01 (2.71)
SCM	1.50 (1.05)	1.32 (0.75)	0.86 (0.34)	0.39 (0.31)	4.07 (1.36)

1. The true covariance matrices were $\boldsymbol{\Sigma}_k = k\mathbf{I}$, and the sample sizes were $n_k = 25$ for all k.
2. The true covariance matrices were $\boldsymbol{\Sigma}_k = k\mathbf{I}$, and the sample sizes were $n_k = 10\cdot k$.
3. The true covariance matrices were generated such that the ijth entry of the covariance matrix was $(\boldsymbol{\Sigma}_k)_{ij} = k\rho_k^{|i-j|}$, where $\rho_1 = -0.6$, $\rho_2 = -0.2$, $\rho_3 = 0.2$, and $\rho_4 = 0.6$, and the sample sizes were $n_k = 10\cdot k$.

We report the empirical normalized MSE (NMSE), $\tilde{L}_k(\hat{\boldsymbol{\Sigma}}_k) = \mathrm{Ave}\|\hat{\boldsymbol{\Sigma}}_k - \boldsymbol{\Sigma}_k\|_{\mathrm{F}}^2/\|\boldsymbol{\Sigma}_k\|_{\mathrm{F}}^2$, of our proposed estimator (Prop 1) and the oracle estimator (Oracle), which uses the true values of η_k, γ_k, κ_k, and $\boldsymbol{\Sigma}_k$ in (6) and (8). The NMSEs of the SCM and the pooled SCM (Pool) are also shown. The results were averaged over 300 Monte-Carlo (MC) trials and are given in Table 1. The proposed method (Prop 1) provided not only a significant improvement both over the SCM and the pooled SCM by yielding the smallest NMSEs and standard deviations, but was also somewhat robust to estimation errors in the sense that it was able to perform at the level of the oracle estimator.

5. APPLICATION TO DISCRIMINANT ANALYSIS

In linear and quadratic discriminant analysis, one uses a discriminant rule which assigns any new observation $\mathbf{x}$ to class $\hat{k}$, such that it minimizes the quadratic discriminant function

$$\hat{k} = \arg\min_k\ (\mathbf{x}-\bar{\mathbf{x}}_k)^\top\hat{\boldsymbol{\Sigma}}_k^{-1}(\mathbf{x}-\bar{\mathbf{x}}_k) + \log|\hat{\boldsymbol{\Sigma}}_k|,$$

where $\hat{\boldsymbol{\Sigma}}_k$ denotes an estimator of $\boldsymbol{\Sigma}_k$. In quadratic discriminant analysis (QDA), one uses $\hat{\boldsymbol{\Sigma}}_k = \mathbf{S}_k$ for all k, whereas in linear discriminant analysis (LDA), one uses $\hat{\boldsymbol{\Sigma}}_k = \mathbf{S}$ for all k. In RDA [2], $\hat{\boldsymbol{\Sigma}}_k(\beta_k)$ of (1), using $\beta_k = \beta$ for all k, is further regularized towards a scaled identity matrix by

$$\hat{\boldsymbol{\Sigma}}_k(\alpha,\beta) = \alpha\hat{\boldsymbol{\Sigma}}_k(\beta) + (1-\alpha)\big(\mathrm{tr}(\hat{\boldsymbol{\Sigma}}_k(\beta))/p\big)\mathbf{I}, \quad (13)$$

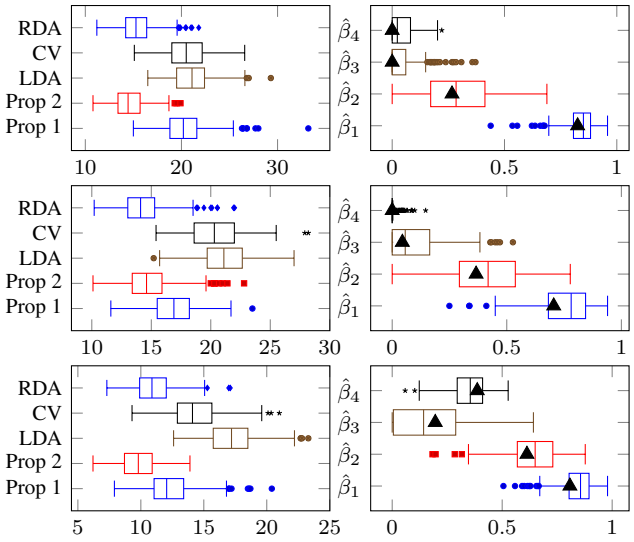

Fig. 1. The misclassification rate $\times 100$ and the corresponding boxplots of $\hat{\beta}_k$ for the set-ups 1, 2 and 3 shown from top to bottom. The black triangles denote the oracle values $\beta_k^\star$.

and $\alpha, \beta \in [0, 1]$ are common over the classes and chosen via cross-validation. Additional regularization via parameter α helps to stabilize the estimates if the total sample size $n = \sum_k n_k$ is small compared to p. We also applied (13) to further shrink the covariance estimates obtained by our method using the approach in [7], thus acquiring the estimates of α_k by

$$\hat{\alpha}_k = \max\left\{0, \frac{T_k}{T_k + \frac{1}{n_k}(\hat{\kappa}_k(2\hat{\gamma}_k + p) + \hat{\gamma}_k + p)}\right\}, \quad (14)$$

where $T_k = \hat{\gamma}_k - 1$, yielding an estimator which we abbreviate Prop 2. Unlike in the single covariance matrix estimation problem of [7], Prop 2 does not have the interpretation of being the MMSE estimator.

In our next example, we used the same simulation set-ups as in Section 4. However, for each MC trial, an additional test data set comprising $10 \cdot n_k$ observations from each class was generated and used for computing the misclassification rate. In addition to our proposed estimators (Prop 1 and 2), the results are reported for LDA, RDA, and CV, of which the last uses only a common cross-validated β in (1) for shrinking the SCMs towards the pooled SCM. Using 10-fold cross-validation, the grid of the regularization parameter values for α and β ranged from 0 to 1 with a step of 0.05. In situations, where equal training errors were obtained with distinct values of α and β, we averaged the performance of all of them. The QDA estimate was omitted, since it does not exist if the SCM is not invertible. The average misclassification rates, over 300 MC trials, are given in Figure 1 along with the boxplots of the estimated regularization parameters $\hat{\beta}_k$. The black triangles in the boxplots denote the oracle values $\beta_k^\star$ computed using the true values of η_k, γ_k, κ_k, and $\mathbf{\Sigma}_k$ in (6) and (8). As shown, the optimal regularization parameters were estimated reasonably well. Also, Prop 1 performed better than CV, whereas Prop 2 performed at the level of RDA, if not slightly better.

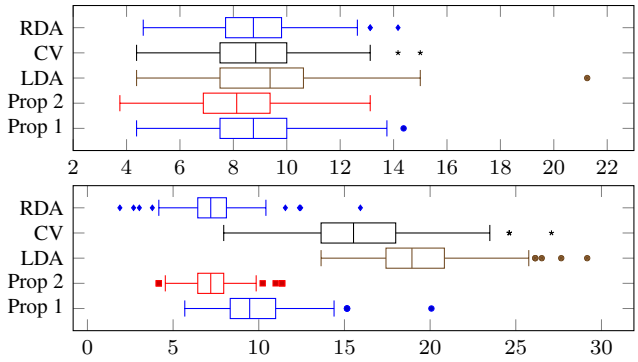

Fig. 2. Misclassification rate $\times 100$ for the glass and ionosphere data shown from top to bottom.

Next, we applied the proposed methods to the following real data examples obtained from the UCI Machine Learning Repository [12]. The **glass data set** had $p = 9$ variables and sample sizes of $n_1 = 51$ (window glass) and $n_2 = 163$ (non-window glass). The **ionosphere data set** had $p = 32$ variables[1] and sample sizes $n_1 = 126$ (bad radar return) and $n_2 = 225$ (good radar return). A fraction 1/4 of the samples from each class were chosen randomly as training data in each MC trial, and the remaining samples were used as test data. This made the training data sizes comparable to the dimension for at least one of the classes. The misclassification rates were averaged over 300 MC trials and are shown in Figure 2. In addition to the proposed estimators, the performances of LDA, CV, and RDA are given. As Figure 2 shows, Prop 1 performed better than CV, whereas Prop 2 and RDA performed equally well.

6. CONCLUSION

We considered joint covariance matrix estimation of multiple classes, where the SCMs of the classes are individually regularized towards the pooled SCM. We derived the optimal class-specific regularization parameters and showed how they could be estimated when the data is considered to be elliptically distributed. The conducted synthetic simulation study showed that the optimal regularization parameters could be estimated with an MSE performance close to the optimal oracle level. When applied to discriminant analysis classification, the conducted synthetic and real data simulations indicated that the performance of the proposed methods are on a par with, and often better than, the computationally more intensive methods based on cross-validation.

[1] Two variables, which were zero for all samples in the classes, were removed from the original data.

7. REFERENCES

[1] Patrick Danaher, Pei Wang, and Daniela M. Witten, "The joint graphical lasso for inverse covariance estimation across multiple classes," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 76, no. 2, pp. 373–397, 2014.

[2] Jerome H. Friedman, "Regularized discriminant analysis," *Journal of the American Statistical Association*, vol. 84, no. 405, pp. 165–175, 1989.

[3] Yaqian Guo, Trevor Hastie, and Robert Tibshirani, "Regularized linear discriminant analysis and its application in microarrays," *Biostatistics*, vol. 8, no. 1, pp. 86–100, 2007.

[4] Trevor Hastie, Robert Tibshirani, and Jerome Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, Springer, 2nd edition, 2009.

[5] Tatyana V. Bandos, Lorenzo Bruzzone, and Gustavo Camps-Valls, "Classification of hyperspectral images with regularized linear discriminant analysis," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 47, no. 3, pp. 862–873, 2009.

[6] Joseph P. Hoffbeck and David A. Landgrebe, "Covariance matrix estimation and classification with limited training data," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 18, no. 7, pp. 763–767, 1996.

[7] Esa Ollila, "Optimal high-dimensional shrinkage covariance estimation for elliptical distributions," in *Proc. 25th European Signal Processing Conference (EUSIPCO 2017)*, Kos, Greece, 2017, pp. 1639–1643.

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

[9] Christophe Croux, Esa Ollila, and Hannu Oja, "Sign and rank covariance matrices: statistical properties and application to principal components analysis," in *Statistical data analysis based on the L1-norm and related methods*, pp. 257–269. Birkhäuser, Basel, 2002.

[10] B. M. Brown, "Statistical uses of the spatial median," *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 45, no. 1, pp. 25–30, 1983.

[11] Teng Zhang and Ami Wiesel, "Automatic diagonal loading for Tyler's robust covariance estimator," in *IEEE Statistical Signal Processing Workshop (SSP 2016)*, 2016, pp. 1–5.

[12] M. Lichman, "UCI machine learning repository," `http://archive.ics.uci.edu/ml`, 2013.