Supplementary materials for "Optimal shrinkage covariance matrix estimation under random sampling from elliptical distributions"

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

***Abstract*—We evaluate the performance of the regularized sample covariance matrix estimators (RSCM) proposed in [1] in a real data classification problem. Specifically, the proposed estimators are applied to a regularized discriminant analysis (RDA) framework in the classification of phoneme data. The notations are adapted from [1].**

I. Classification based on regularized discriminant analysis

Suppose there are K different p-variate populations with covariance matrices $\boldsymbol{\Sigma}_k \in \mathbb{S}_{++}^{p\times p}$ and a mean vectors $\boldsymbol{\mu}_k \in \mathbb{R}^p$, $k = 1, \ldots, K$. The problem is to classify an observation $\mathbf{x} \in \mathbb{R}^p$ to one of the populations. We assume no knowledge of the class prior probabilities. In quadratic discriminant analysis (QDA) classification, a new observation $\mathbf{x}$ is assigned to class $\hat{k}$ by the rule

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

Commonly, $\boldsymbol{\mu}_k$ and $\boldsymbol{\Sigma}_k$ are estimated by the sample mean vectors $\bar{\mathbf{x}}_k$ and the SCMs $\mathbf{S}_k$ computed from the *training dataset* $\mathbf{X} = (\mathbf{x}_1 \cdots \mathbf{x}_n)$, which consists of n_k observations from each class $k = 1, \ldots, K$ and where $n = n_1 + \cdots + n_k$ denotes the total sample size. In linear discriminant analysis (LDA), one assumes that the class covariance matrices are equal, so $\boldsymbol{\Sigma} = \boldsymbol{\Sigma}_k$ for each $k = 1, \ldots, K$. Then, the unknown common covariance matrix is estimated by the *pooled* SCM defined as

$$\mathbf{S}_{\text{pool}} = \sum_{k=1}^{K} \frac{n_k - 1}{n - K}\, \mathbf{S}_k.$$

The benefit of LDA over QDA is that it can also be applied in the case when $n_k < p$ (for any k) as long as $n = \sum_k n_k > p$. Whereas, in this case, QDA is no longer applicable since the SCM $\mathbf{S}_k$ is not invertible for $n_k < p$. LDA can be viewed as a regularized form of QDA since it decreases the variance of $\mathbf{S}_k$ by using the pooled SCM. In small-sample settings, LDA often has superior performance over QDA.

The performance of LDA and QDA classification rules are highly dependent on the accuracy of the used covariance matrix estimates. Thus, in order to reduce the misclassification rate in low sample support cases, a popular approach is to use RSCM estimators instead of class sample covariance matrices; see e.g., [2]. RSCMs can be applied to LDA and QDA regardless of what the available sample sizes n_k of the classes are. Here, we use a regularized version of QDA and LDA, where we estimate the means by $\bar{\mathbf{x}}_k$, but use Ell1-RSCM, Ell2-RSCM, or LW-RSCM in place of the unknown covariance matrices $\boldsymbol{\Sigma}_k$ in QDA and $\boldsymbol{\Sigma}$ in case of LDA. This approach is referred to as regularized discriminant analysis (RDA) [2].

We compute the misclassification rates of LDA and QDA and different RDA methods for the phoneme dataset [3]. The original data consists of short speech frames of 32 msec duration (512 samples with at a 16kHz sampling rate) and each frame represents one of the following phonemes, "aa", "ao", "dcl", "iy", or "sh" with the number of occurrencies 695, 1022, 757, 1163, and 872, respectively. The full data set consists of 4509 speech frames spoken by 50 different male speakers. The data used for classification consists of the log-periodograms of the speech frames measured at $p = 256$ distinct frequencies. The goal is to classify the spoken phonemes.

In the simulations, we randomly split the dataset into a training set and test set with the ratio 1:12. Then the sizes of the training sets were close to or smaller than the dimension p as this is the regime where regularization is needed the most. The number of occurrences of each of the aforementioned $K = 5$ phonemes in the training set were then 53, 79, 58, 89, and 67, respectively, while the remaining dataset was used as a test set. The full length of the training data was $N = \sum_k n_k = 346 > p = 256$, and thus, the conventional LDA could be applied. QDA, on the other hand, could not be applied since $n_k < p$. The misclassification rates were calculated by classifying the observations from the test set using the classification rule estimated from the training set. We report the corresponding misclassification rates based on 50 repetitions of random splits of the full data set into test sets and training sets.

The boxplots of the test misclassification rates given in Figure 1 compare the conventional LDA with regularized QDA and regularized LDA. Here we also compare the performance of the Ell-RSCM estimators to an estimator that presumes Gaussianity ($\kappa = 0$) and uses the shrinkage parameter estimate $\hat{\beta}_o^{\text{Gau}}$ (as specified in [1, eq. (14)] and the estimate of the sphericity $\hat{\gamma}^{\text{Ell2}}$ in place of the unknown γ.

Several conclusions can be drawn from Figure 1. First, the regularized LDA rules that used Ell-RSCM or LW-RSCM outperformed the LDA with a significant margin: the median test errors of the regularized LDA (resp. regularized QDA) methods based on Ell1-, Ell2-, and LW-RSCM were 9.96%, 10.57%, and 10.62%. (resp. 12.86%, 14.36%, and 15.21%) which may be compared with the 16.8% median error rate of the conventional LDA. Second, the overall performance of the

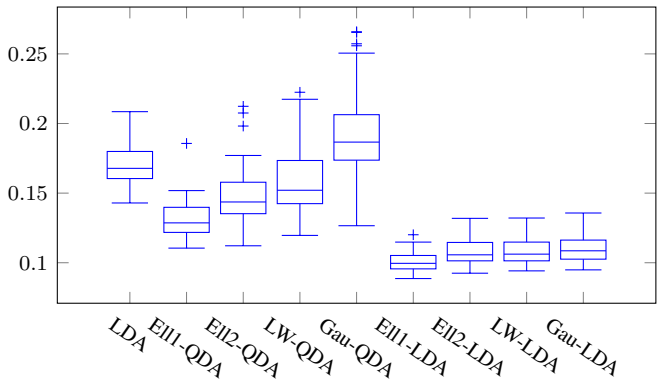

Fig. 1. Phoneme data: Box plots of the test misclassification rates of the conventional LDA compared with the regularized QDA and LDA methods based on different RSCM estimators.

regularized LDA methods was better than the performance of the regularized QDA methods. Third, in all cases, both Ell1-RSCM and Ell2-RSCM outperformed LW-RSCM, and again, Ell1-RSCM had the best performance among all methods. Fourth, we notice that the Gau-RSCM estimator which presumes Gaussianity (and thus uses $\kappa = 0$) is not able to perform better than the other RSCM estimators. In fact, Gau-RSCM had the worst performance among all methods when applied to the QDA rule. This illustrates the fact that the Gaussianity assumption is a poor approximation of reality for many real data analysis problems.

The Matlab script to reproduce Figure 1 is available in the MATLAB toolbox available at http://users.spa.aalto.fi/esollila/regscm/. To have a quick access to the provided demo examples in the toolbox type `demo RegularizedSCM`.

REFERENCES

[1] E. Ollila and E. Raninen, "Optimal shrinkage covariance matrix estimation under random sampling from elliptical distributions," *arXiv preprint arXiv:1808.10188*, 2018.

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

[3] T. Hastie, R. Tibshirani, and J. Friedman, *The elements of statistical learning*. New York: Springer, 2001.