OPTIMAL POOLING OF COVARIANCE MATRIX ESTIMATES ACROSS MULTIPLE CLASSES

Elias Raninen and Esa Ollila at the Department of Signal Processing and Acoustics, Aalto University, Finland
Contact: elias.raninen@aalto.fi and esa.ollila@aalto.fi

Aalto University

PROBLEM FORMULATION

- Consider data from K distinct classes (populations).
- Let $\{\mathbf{x}_{k,i}\}_{i=1}^{n_k}$ denote the data set of the kth class.
- Our aim is to estimate the $p \times p$ covariance matrices,

$$\boldsymbol{\Sigma}_k = \mathbb{E}[(\mathbf{x}_k - \mathbb{E}[\mathbf{x}_k])(\mathbf{x}_k - \mathbb{E}[\mathbf{x}_k])^\top], \quad k = 1, \dots, K,$$

where $\mathbf{x}_k$ denotes a random vector from the kth class.
- The sample covariance matrix (SCM) of class k is

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

where $\bar{\mathbf{x}}_k = \frac{1}{n_k} \sum_{i=1}^{n_k} \mathbf{x}_{k,i}$.
- If $p \approx n_k$ or $p > n_k$, regularization of the SCM is needed to reduce the variance and to ensure positive definiteness.
- A natural regularization target is the pooled SCM.

We are interested in a regularized SCM for class k:

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

where $\beta \in [0, 1]$, and the regularization target $\mathbf{S}$ is the pooled (average) SCM:

$$\mathbf{S} = \sum_{k=1}^{K} \pi_k \mathbf{S}_k, \quad \text{where} \quad \pi_k = \frac{n_k}{\sum_{j=1}^{K} n_j}.$$

Goal: determine the optimal regularization level,

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

Solution:

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

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

We need to estimate:
$\mathrm{tr}\left(\boldsymbol{\Sigma}_i \boldsymbol{\Sigma}_j\right)$, $i \neq j$, $\mathbb{E}\left[\mathrm{tr}\left(\mathbf{S}_k^2\right)\right]$, and $\mathrm{tr}(\boldsymbol{\Sigma}_k^2)$.

ESTIMATION OF PARAMETERS

- Assume $\{\mathbf{x}_{i,k}\}_{k=1}^{K}, \forall k$, are from (unspecified) elliptical distributions with finite 4th order moments.
- A consistent estimate of $\mathrm{tr}\left(\boldsymbol{\Sigma}_i \boldsymbol{\Sigma}_j\right)$, $i \neq j$, is $\mathrm{tr}\left(\mathbf{S}_i \mathbf{S}_j\right)$.
- By using Corollary 1 from [1], one can show that

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

where $\tau_1 = (n_k - 1)^{-1} + \kappa_k / n_k$ and $\tau_2 = \kappa_k / n_k$.
  - The *elliptical kurtosis*, $\kappa_k = (1/3) \cdot \{\text{excess kurtosis}\}$, is estimated by the average elliptical sample kurtosis of the variables.
  - The *scale*, $\eta_k = \mathrm{tr}\left(\boldsymbol{\Sigma}_k\right)/p$, is estimated by $\hat{\eta}_k = \mathrm{tr}\left(\mathbf{S}_k\right)/p$.
  - The *sphericity*, $\gamma_k = p\mathrm{tr}\left(\boldsymbol{\Sigma}_k^2\right)/\mathrm{tr}\left(\boldsymbol{\Sigma}_k\right)^2$, is estimated by [2]

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

where the *sample sign covariance matrix* is

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

and $\hat{\boldsymbol{\mu}}_k = \arg\min_{\boldsymbol{\mu}} \sum_{i=1}^{n_k} \|\mathbf{x}_{k,i} - \boldsymbol{\mu}\|$.
- An estimate of $\mathrm{tr}\left(\boldsymbol{\Sigma}_k^2\right)$ is obtained by $p\hat{\gamma}_{\mathrm{sgn},k}\hat{\eta}_k^2$.
- As the final estimate of $\beta_k^\star$, we use $\max\{0, \min\{1, \hat{\beta}_k\}\}$.
- We estimate $\hat{\beta}_k$ for each class k, and denote the method by **Prop 1**.

SIMULATION SET-UPS

1. $\boldsymbol{\Sigma}_k = k\mathbf{I}$.
2. $(\boldsymbol{\Sigma}_k)_{ij} = k\rho_k^{|i-j|}$, where $\rho_1 = -0.6$, $\rho_2 = -0.2$, $\rho_3 = 0.2$, and $\rho_4 = 0.6$.

- $K = 4$, $p = 20$, $n_k = 10k$, and $n = \sum_k n_k = 100$.
- The data was Student's t_ν-distributed with $\nu = 10$.
- $\boldsymbol{\mu}_1 = \mathbf{0}$, and for the classes $k = 2, 3$, and 4, $\|\boldsymbol{\mu}_k\| = 1 + k$ in orthogonal directions.
- 300 Monte-Carlo trials.

MSE PERFORMANCE

The empirical NMSE, $\tilde{L}_k = \mathrm{Ave}\|\hat{\boldsymbol{\Sigma}}_k - \boldsymbol{\Sigma}_k\|_{\mathrm{F}}^2 / \|\boldsymbol{\Sigma}_k\|_{\mathrm{F}}^2$, for the set-ups 1 and 2 (from top to down). LB denotes the lower bound and Oracle uses $\beta_k^\star$ from (1). Standard deviations are in parenthesis.

	$\tilde{L}_1$	$\tilde{L}_2$	$\tilde{L}_3$	$\tilde{L}_4$	Sum
LB	2.04 (0.97)	0.65 (0.20)	0.30 (0.08)	0.24 (0.05)	3.24 (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)
LB	1.17 (0.57)	0.87 (0.29)	0.37 (0.11)	0.22 (0.05)	2.63 (0.68)
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)

APPLICATION IN CLASSIFICATION

- In discriminant analysis, any new observation $\mathbf{x}$ is assigned to class $\hat{k}$ by the rule:

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

- In **RDA** [3], $\hat{\boldsymbol{\Sigma}}_k(\beta)$ is further regularized towards scaled identity by

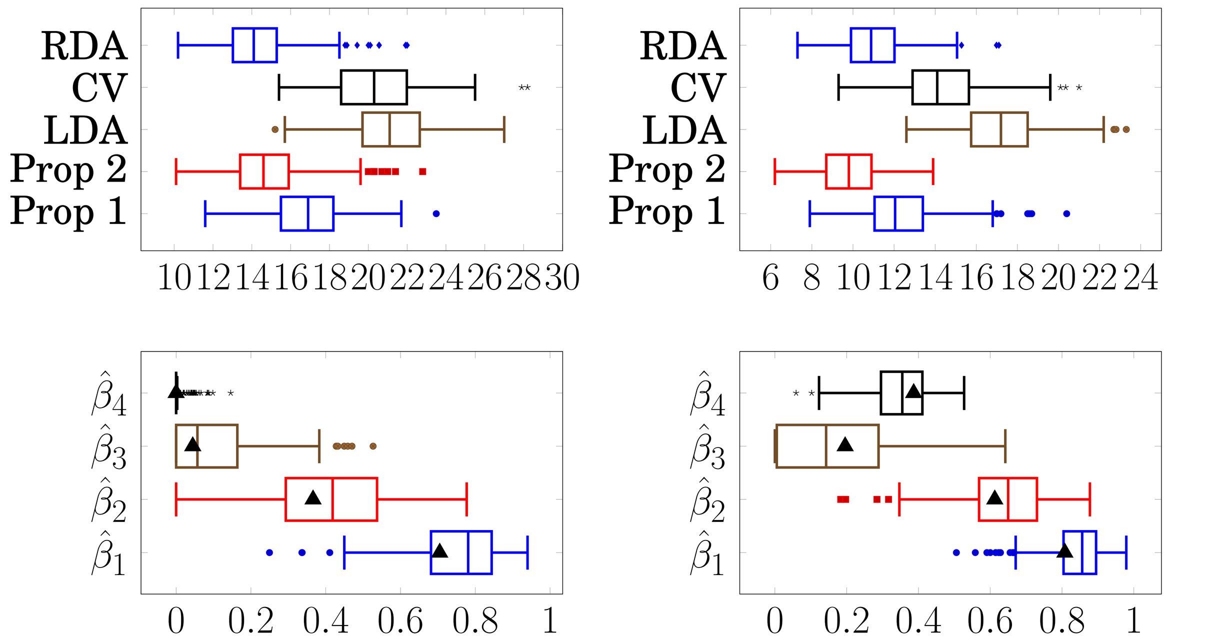
$$\hat{\boldsymbol{\Sigma}}_k(\alpha, \beta) = \alpha\hat{\boldsymbol{\Sigma}}_k(\beta) + (1 - \alpha)\big(\mathrm{tr}(\hat{\boldsymbol{\Sigma}}_k(\beta))/p\big)\mathbf{I}, \qquad (2)$$

and $\alpha, \beta \in [0, 1]$ are common across classes and chosen via cross-validation.
- We applied (2) to our estimator by using

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

from [4]. We denote this estimator by **Prop 2**.
- **CV** is the RDA estimator in (2) with fixed $\alpha = 1$.
- **LDA** uses the pooled SCM.
- Note: Prop 1 and Prop 2 are computationally significantly more efficient than CV and RDA since no cross-validation is needed.

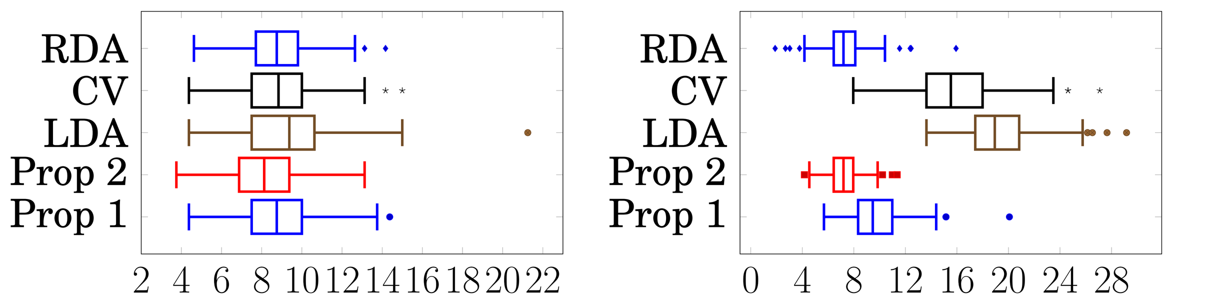
SYNTHETIC DATA EXAMPLES



Boxplots of the misclassification rate $\times 100$ and $\hat{\beta}_k$ for the set-ups 1 (left) and 2 (right). The black triangles denote $\beta_k^\star$.

REAL DATA EXAMPLES

- Glass data set [5]: $p = 9$, $n_1 = 51$ (window glass) and $n_2 = 163$ (non-window glass).
- Ionosphere data set [5]: $p = 32$, $n_1 = 126$ (bad radar return) and $n_2 = 225$ (good radar return).
- A fraction 1/4 of the samples from each class were used as training data.



Boxplots of the misclassification rate $\times 100$ for the glass data (left) and the ionosphere data (right).

REFERENCES

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

[2] Teng Zhang and Ami Wiesel, "Automatic diagonal loading for Tyler's robust covariance estimator," in *SSP*, 2016, pp. 1–5.

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

[4] Esa Ollila, "Optimal high-dimensional shrinkage covariance estimation for elliptical distributions," in *EUSIPCO*, Kos, Greece, 2017, pp. 1639–1643.

[5] "UCI machine learning repository," `http://archive.ics.uci.edu/ml`.

ICASSP'18, Calgary, Alberta, Canada