Linear pooling of sample covariance matrices

Esa Ollila

Department of Signal Processing and Acoustics
Aalto University, Finland
http://users.spa.aalto.fi/esollila/

DATAIA Seminar, Dec. 18th, 2020

Aalto University

Joint work with

Elias Raninen
Aalto University

David E. Tyler
Rutgers University

Menu

Multiple covariance matrices problem

- We are given independent p-variate measurements on K classes:

$$\mathbf{x}_{1,1}, \ldots, \mathbf{x}_{n_1,1}, \quad \ldots, \quad \mathbf{x}_{1,K}, \ldots, \mathbf{x}_{n_K,K}$$

- Need to estimate the covariance matrices of the classes:

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

  where $\boldsymbol{\mu}_k = \mathbb{E}[\mathbf{x}_{i,k}]$, for $k = 1, \ldots, K$.
- Each $\boldsymbol{\Sigma}_k \in \mathbb{S}_{++}^{p \times p}$ ($\in$ set of positive definite matrices)
- Common estimate of $\boldsymbol{\Sigma}_k$ is the sample covariance matrix (SCM):

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

  for $k = 1, \ldots, K$.

Multiple covariance matrices problem (cont'd)

- If one assumes equal covariance matrices ($\boldsymbol{\Sigma}_k \equiv \boldsymbol{\Sigma}$)

  ... one may estimate $\boldsymbol{\Sigma}$ via the pooled SCM:

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

  where $n = n_1 + n_2 + \cdots + n_K$.

- Challenges:
  1. High-dimensionality (possibly $p > n_k \ \forall k$)
  2. K large (e.g., multiple classes, and each class has subclasses).
  3. Non-gaussian data.
- Common solution is to use regularized (shrinkage) estimators.

Regularized SCM

Regularized SCM (RSCM) estimator:

$$\mathbf{S}_k(\alpha, \beta) = \beta \mathbf{S}_k + \alpha \mathbf{T}_k,$$

where

- $\mathbf{T}_k \succeq 0$ is some fixed shrinkage *target matrix*
- $\alpha \geq 0$, $\beta \geq 0$ are *weights* (different for each k)

- Weights are optimized by minimizing criterions such as
  1. *Mean squared error* $\mathbb{E}\left[\|\mathbf{S}_k(\alpha, \beta) - \mathbf{\Sigma}_k\|_{\mathrm{F}}^2\right]$
  2. *Metric* $D(\mathbf{S}_k(\alpha, \beta), \mathbf{\Sigma}_k)$ such as Frobenius, Kullback-Leiber, Riemannian distance, ...
  3. *Cross validation*

  or using *Bayesian approaches* or *expected likelihood* approach.

Regularized SCM (cont'd)

$$\mathbf{S}_k(\alpha, \beta) = \beta \mathbf{S}_k + \alpha \mathbf{T}_k.$$

But what target $\mathbf{T}_k$ to use?

1. $\mathbf{T}_k = \mathbf{I}$. [DLS10, Col15]
2. $\mathbf{T}_k = \frac{\mathrm{tr}(\mathbf{S}_k)}{p} \mathbf{I}$ and $\alpha = 1 - \beta \in [0, 1]$. [LW04b, CWEH10, OR19]
3. $\mathbf{T}_k = \mathbf{S}_{\text{pool}}$ and and $\alpha = 1 - \beta$. [Fri89, RO18]
4. Highly structured $\mathbf{T}_k$:
   - Single (market-)factor matrix [LW03]
   - Constant correlation matrix [LW04a]
   - Knowledge aided (KA-)STAP matrix [SLZG08].
   - Generalized banded matrices [LZZ17]

Regularized SCM (cont'd)

$$\mathbf{S}_k(\alpha, \beta) = \beta \mathbf{S}_k + \alpha \mathbf{T}_k.$$

But what target $\mathbf{T}_k$ to use?

1. $\mathbf{T}_k = \mathbf{I}$. [DLS10, Col15]
2. $\mathbf{T}_k = \dfrac{\mathrm{tr}(\mathbf{S}_k)}{p}\mathbf{I}$ and $\alpha = 1 - \beta \in [0, 1]$. [LW04b, CWEH10, OR19]
3. $\mathbf{T}_k = \mathbf{S}_{\text{pool}}$ and and $\alpha = 1 - \beta$. [Fri89, RO18]
4. Highly structured $\mathbf{T}_k$:
   - Single (market-)factor matrix [LW03]
   - Constant correlation matrix [LW04a]
   - Knowledge aided (KA-)STAP matrix [SLZG08].
   - Generalized banded matrices [LZZ17]

Regularized SCM (cont'd)

$$\mathbf{S}_k(\alpha, \beta) = \beta \mathbf{S}_k + \alpha \mathbf{T}_k.$$

But what target $\mathbf{T}_k$ to use?

1. $\mathbf{T}_k = \mathbf{I}$. [DLS10, Col15]
2. $\mathbf{T}_k = \frac{\mathrm{tr}(\mathbf{S}_k)}{p}\mathbf{I}$ and $\alpha = 1 - \beta \in [0, 1]$. [LW04b, CWEH10, OR19]
3. $\mathbf{T}_k = \mathbf{S}_{\text{pool}}$ and and $\alpha = 1 - \beta$. [Fri89, RO18]
4. Highly structured $\mathbf{T}_k$:
   - Single (market-)factor matrix [LW03]
   - Constant correlation matrix [LW04a]
   - Knowledge aided (KA-)STAP matrix [SLZG08].
   - Generalized banded matrices [LZZ17]

Regularized SCM (cont'd)

$$\mathbf{S}_k(\alpha,\beta) = \beta\mathbf{S}_k + \alpha\mathbf{T}_k.$$

But what target $\mathbf{T}_k$ to use?

1. $\mathbf{T}_k = \mathbf{I}$. [DLS10, Col15]
2. $\mathbf{T}_k = \frac{\mathrm{tr}(\mathbf{S}_k)}{p}\mathbf{I}$ and $\alpha = 1-\beta \in [0,1]$. [LW04b, CWEH10, OR19]
3. $\mathbf{T}_k = \mathbf{S}_{\text{pool}}$ and and $\alpha = 1-\beta$. [Fri89, RO18]
4. Highly structured $\mathbf{T}_k$:
   - Single (market-)factor matrix [LW03]
   - Constant correlation matrix [LW04a]
   - Knowledge aided (KA-)STAP matrix [SLZG08].
   - Generalized banded matrices [LZZ17]

Double shrinkage SCM

- Step 1: $\hat{\mathbf{\Sigma}}_k(\beta) = \beta \mathbf{S}_k + (1-\beta)\mathbf{S}_{\text{pool}}, \quad \beta \in [0,1]$

  Shrink each $\mathbf{S}_k$ towards $\mathbf{S}_{\text{pool}}$ to get $\hat{\mathbf{\Sigma}}_k(\beta)$.

- Step 2: $\hat{\mathbf{\Sigma}}_k(\alpha,\beta) = \alpha \hat{\mathbf{\Sigma}}_k(\beta) + (1-\alpha)\frac{\text{tr}(\hat{\mathbf{\Sigma}}_k(\beta))}{p}\mathbf{I}, \quad \alpha \in [0,1].$

  Then regularize $\hat{\mathbf{\Sigma}}_k(\beta)$ further towards the scaled identity matrix to ensure positive definiteness (even for $p > \sum_i n_i$).

- [Fri89] used same α and β for each k, and leave-one-out cross validation for choosing them.
- [RO20] uses different α, β for each k and data-adaptive tuning for parameter selection.

This work

- Define

$$\mathbf{S}(\mathbf{a}) = \sum_{i=1}^{K} a_i \mathbf{S}_i, \quad a_i \geq 0 \ \forall i = 1, \ldots, K$$

or, $$\mathbf{S}(\mathbf{a}) = a_{K+1}\mathbf{I} + \sum_{i=1}^{K} a_i \mathbf{S}_i, \quad a_i \geq 0 \ \forall i = 1, \ldots, K+1$$

- Find weights that minimizes the (total) MSE:

$$\mathbf{a}_k^\star = \arg \min_{(a_i) \geq 0} \mathbb{E}\big[\|\mathbf{S}(\mathbf{a}) - \mathbf{\Sigma}_k\|_{\mathrm{F}}^2\big] \quad \forall k = 1, \ldots, K,$$

$$\Leftrightarrow \mathbf{A}^\star = (\mathbf{a}_1^\star \ \cdots \mathbf{a}_K^\star) = \arg \min_{(a_{ij}) \geq 0} \sum_{k=1}^{K} \mathbb{E}\big[\|\mathbf{S}(\mathbf{a}_k) - \mathbf{\Sigma}_k\|_{\mathrm{F}}^2\big].$$

- Ideally, use $\hat{\mathbf{\Sigma}}_k^\star = \mathbf{S}(\mathbf{a}_k^\star)$. .

This work

- Define

$$\mathbf{S}(\mathbf{a}) = \sum_{i=1}^{K} a_i \mathbf{S}_i, \quad a_i \geq 0 \ \forall i = 1, \ldots, K$$

or, $$\mathbf{S}(\mathbf{a}) = a_{K+1}\mathbf{I} + \sum_{i=1}^{K} a_i \mathbf{S}_i, \quad a_i \geq 0 \ \forall i = 1, \ldots, K+1$$

- Find weights that minimizes the (total) MSE:

$$\mathbf{a}_k^\star = \arg \min_{(a_i) \geq 0} \mathbb{E}\big[\|\mathbf{S}(\mathbf{a}) - \boldsymbol{\Sigma}_k\|_{\mathrm{F}}^2\big] \quad \forall k = 1, \ldots, K,$$

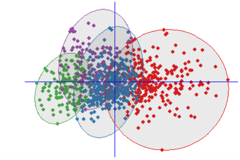
$$\Leftrightarrow \mathbf{A}^\star = (\mathbf{a}_1^\star \ \cdots \mathbf{a}_K^\star) = \arg \min_{(a_{ij}) \geq 0} \sum_{k=1}^{K} \mathbb{E}\big[\|\mathbf{S}(\mathbf{a}_k) - \boldsymbol{\Sigma}_k\|_{\mathrm{F}}^2\big].$$

- Ideally, use $\hat{\boldsymbol{\Sigma}}_k^\star = \mathbf{S}(\mathbf{a}_k^\star)$. In practise, $\hat{\boldsymbol{\Sigma}}_k = \mathbf{S}(\hat{\mathbf{a}}_k)$ (where $\hat{\mathbf{a}}_k \approx \mathbf{a}_k^\star$).

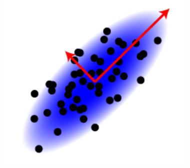
Why covariance estimation?

Portfolio selection

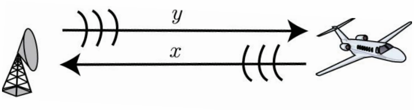
Classification/Clustering



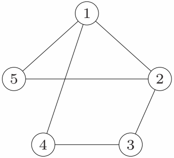
PCA



Radar detection



Graphical models



$$\Sigma^{-1} = \begin{bmatrix} \bullet & \bullet & 0 & \bullet & \bullet \\ \bullet & \bullet & \bullet & 0 & \bullet \\ 0 & \bullet & \bullet & \bullet & 0 \\ \bullet & 0 & \bullet & \bullet & 0 \\ \bullet & \bullet & 0 & 0 & \bullet \end{bmatrix}$$

Why covariance estimation (con'd)?

Pedestrian detection [TPM08, JHS+13]

Feature vector :

$$\mathbf{z}(x,y) = \left(x, y, |I_x|, |I_y|, \sqrt{I_x^2 + I_y^2}, |I_{xx}|, |I_{yy}|, \arctan(I_x|/|I_y|)\right)^\top,$$

where x, y are the pixel coordinates, I_x, I_y the 1^{st} intensity derivatives, ...

(a) Orig. image (b) $|I_x|$

Covariance descriptor of a region R:

$$\mathbf{S}_R = \frac{1}{|R|-1} \sum_{(x,y)\in R} (\mathbf{z}(x,y) - \bar{\mathbf{z}})(\mathbf{z}(x,y) - \bar{\mathbf{z}})^\top$$

where $\bar{\mathbf{z}} = \frac{1}{|R|} \sum_{(x,y)\in R} \mathbf{z}(x,y)$

$\mathbf{S}_R$-s are used as features for an ML algorithm.
See [MRO20] for a review.

Menu

LINPOOL estimator

- Denote the scaled MSE of the k^{th} SCM $\mathbf{S}_k$ by

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

- Define matrices

$$\Delta = \mathrm{diag}(\delta_1, \ldots, \delta_K) \quad \text{and} \qquad \mathbf{C} = (c_{ij}) = \left(\frac{\mathrm{tr}(\boldsymbol{\Sigma}_i \boldsymbol{\Sigma}_j)}{p}\right) \in \mathbb{R}^{K \times K}.$$

- **Theorem:** The MSE of $\mathbf{S}(\mathbf{a}) = \sum_{i=1}^{K} a_i \mathbf{S}_i$ is given by

$$\frac{1}{2p}\mathbb{E}\big[\|\mathbf{S}(\mathbf{a}) - \boldsymbol{\Sigma}_k\|_{\mathrm{F}}^2\big] = \frac{1}{2}\mathbf{a}^\top(\Delta + \mathbf{C})\mathbf{a} - \mathbf{c}_k^\top \mathbf{a} \qquad (+\text{const})$$

and it is a strictly convex quadratic function in $\mathbf{a} \in \mathbb{R}^K$.

LINPOOL estimator (cont'd)

1. Construct estimates (more on this later):

$$\hat{\Delta} = p^{-1}\,\mathrm{diag}(\widehat{\mathrm{MSE}}(\mathbf{S}_1), \ldots, \widehat{\mathrm{MSE}}(\mathbf{S}_K))$$

$$\hat{\mathbf{C}} = (\hat{\mathbf{c}}_1 \ \cdots \hat{\mathbf{c}}_K) = \left(\frac{\mathrm{tr}(\widehat{\boldsymbol{\Sigma}_i \boldsymbol{\Sigma}_j})}{p} \right) \in \mathbb{R}^{K \times K}$$

2. Solve the (unconstrained) strictly convex *quadratic programming (QP)* problem:

$$\begin{aligned} \hat{\mathbf{a}}_k &= \underset{\mathbf{a} \in \mathbb{R}^K}{\arg\min} \ \ \frac{1}{2}\mathbf{a}^\top(\hat{\Delta} + \hat{\mathbf{C}})\mathbf{a} - \hat{\mathbf{c}}_k^\top \mathbf{a} \\ &= (\hat{\Delta} + \hat{\mathbf{C}})^{-1}\hat{\mathbf{c}}_k \end{aligned}$$

3. If any $\hat{a}_{kj} < 0$, then solve

$$\hat{\mathbf{a}}_k = \begin{array}{ll} \text{minimize} & \frac{1}{2}\mathbf{a}^\top(\hat{\Delta} + \hat{\mathbf{C}})\mathbf{a} - \hat{\mathbf{c}}_k^\top \mathbf{a} \\ \text{subject to} & \mathbf{a} \geq \mathbf{0}. \end{array}$$

4. Output: $\hat{\boldsymbol{\Sigma}}_k = \mathbf{S}(\hat{\mathbf{a}}_k)$, where $\mathbf{S}(\mathbf{a}) = \sum_{i=1}^K a_i \mathbf{S}_i$. $\qquad (k = 1, \ldots, K)$

Example 1: single class ($K = 1$) case

- In the single class case, we just need to find shrinkage parameter

$$a_1^\star = \arg\min_{a \in \mathbb{R}} \ \mathbb{E}\big[\|a\mathbf{S}_1 - \boldsymbol{\Sigma}_1\|_{\mathrm{F}}^2\big] = (\delta_1 + c_{11})^{-1} c_{11}$$

$$= \frac{\mathrm{tr}(\boldsymbol{\Sigma}_1^2)}{\mathrm{MSE}(\mathbf{S}_1) + \mathrm{tr}(\boldsymbol{\Sigma}_1^2)} \in (0, 1)$$

- One can show that $\hat{\boldsymbol{\Sigma}}_1^\star = a_1^\star \mathbf{S}_1$ verifies: $\mathrm{MSE}(\hat{\boldsymbol{\Sigma}}_1^\star) = \hat{a}_1^\star \cdot \mathrm{MSE}(\mathbf{S}_1)$.

⇒ Since $0 < a_1^\star < 1$, $\hat{\boldsymbol{\Sigma}}_1^\star = a_1^\star \mathbf{S}_1$ is *always more efficient* than $\mathbf{S}_1$.

- Gaussian data: $(n_1 - 1)\mathbf{S}_1 \sim \mathcal{W}_p(n-1, \boldsymbol{\Sigma}_1)$, so

$$\mathrm{MSE}(\mathbf{S}_1) = \frac{1}{n_1 - 1}\big(\mathrm{tr}(\boldsymbol{\Sigma}_1)^2 + \mathrm{tr}(\boldsymbol{\Sigma}_1^2)\big) \Rightarrow a_1^\star = \frac{n_1 - 1}{n_1 + \gamma/p}$$

where $\gamma = p\,\mathrm{tr}(\boldsymbol{\Sigma}_1^2)/\,\mathrm{tr}(\boldsymbol{\Sigma}_1)^2 \in [1, p]$ is a *measure of sphericity*.

⇒ LINPOOL estimator (for Gaussian data) is $\hat{\boldsymbol{\Sigma}}_1 = \dfrac{n_1 - 1}{n_1 + \gamma/p}\mathbf{S}_1$.

Example 1: single class ($K = 1$) case

- In the single class case, we just need to find shrinkage parameter

$$a_1^\star = \arg\min_{a \in \mathbb{R}} \ \mathbb{E}\big[\|a\mathbf{S}_1 - \mathbf{\Sigma}_1\|_{\mathrm{F}}^2\big] = (\delta_1 + c_{11})^{-1} c_{11}$$
$$= \frac{\operatorname{tr}(\mathbf{\Sigma}_1^2)}{\operatorname{MSE}(\mathbf{S}_1) + \operatorname{tr}(\mathbf{\Sigma}_1^2)} \in (0, 1)$$

- One can show that $\hat{\mathbf{\Sigma}}_1^\star = a_1^\star \mathbf{S}_1$ verifies: $\operatorname{MSE}(\hat{\mathbf{\Sigma}}_1^\star) = \hat{a}_1^\star \cdot \operatorname{MSE}(\mathbf{S}_1)$.

⇒ Since $0 < a_1^\star < 1$, $\hat{\mathbf{\Sigma}}_1^\star = a_1^\star \mathbf{S}_1$ is *always more efficient* than $\mathbf{S}_1$.

- Gaussian data: $(n_1 - 1)\mathbf{S}_1 \sim \mathcal{W}_p(n - 1, \mathbf{\Sigma}_1)$, so

$$\operatorname{MSE}(\mathbf{S}_1) = \frac{1}{n_1 - 1}(\operatorname{tr}(\mathbf{\Sigma}_1)^2 + \operatorname{tr}(\mathbf{\Sigma}_1^2)) \Rightarrow a_1^\star = \frac{n_1 - 1}{n_1 + \gamma/p}$$

where $\gamma = p \operatorname{tr}(\mathbf{\Sigma}_1^2) / \operatorname{tr}(\mathbf{\Sigma}_1)^2 \in [1, p]$ is a *measure of sphericity*.

⇒ LINPOOL estimator (for Gaussian data) is $\hat{\mathbf{\Sigma}}_1 = \dfrac{n_1 - 1}{n_1 + \gamma/p}\mathbf{S}_1$.

Example 1: single class ($K = 1$) case

- In the single class case, we just need to find shrinkage parameter

$$a_1^\star = \arg\min_{a \in \mathbb{R}} \ \mathbb{E}\left[\|a\mathbf{S}_1 - \mathbf{\Sigma}_1\|_{\mathrm{F}}^2\right] = (\delta_1 + c_{11})^{-1} c_{11}$$
$$= \frac{\operatorname{tr}(\mathbf{\Sigma}_1^2)}{\operatorname{MSE}(\mathbf{S}_1) + \operatorname{tr}(\mathbf{\Sigma}_1^2)} \in (0,1)$$

- One can show that $\hat{\mathbf{\Sigma}}_1^\star = a_1^\star \mathbf{S}_1$ verifies: $\operatorname{MSE}(\hat{\mathbf{\Sigma}}_1^\star) = \hat{a}_1^\star \cdot \operatorname{MSE}(\mathbf{S}_1)$.

$\Rightarrow$ Since $0 < a_1^\star < 1$, $\hat{\mathbf{\Sigma}}_1^\star = a_1^\star \mathbf{S}_1$ is *always more efficient* than $\mathbf{S}_1$.

- Gaussian data: $(n_1 - 1)\mathbf{S}_1 \sim \mathcal{W}_p(n - 1, \mathbf{\Sigma}_1)$, so

$$\operatorname{MSE}(\mathbf{S}_1) = \frac{1}{n_1 - 1}\left(\operatorname{tr}(\mathbf{\Sigma}_1)^2 + \operatorname{tr}(\mathbf{\Sigma}_1^2)\right) \Rightarrow a_1^\star = \frac{n_1 - 1}{n_1 + \gamma/p}$$

where $\gamma = p \operatorname{tr}(\mathbf{\Sigma}_1^2)/\operatorname{tr}(\mathbf{\Sigma}_1)^2 \in [1, p]$ is a *measure of sphericity*.

$\Rightarrow$ LINPOOL estimator (for Gaussian data) is $\hat{\mathbf{\Sigma}}_1 = \dfrac{n_1 - 1}{n_1 + \hat{\gamma}/p}\mathbf{S}_1$.

Example 1: single class ($K = 1$) case

- In the single class case, we just need to find shrinkage parameter

$$a_1^\star = \arg\min_{a \in \mathbb{R}} \mathbb{E}\left[\|a\mathbf{S}_1 - \mathbf{\Sigma}_1\|_{\mathrm{F}}^2\right] = (\delta_1 + c_{11})^{-1} c_{11}$$
$$= \frac{\operatorname{tr}(\mathbf{\Sigma}_1^2)}{\operatorname{MSE}(\mathbf{S}_1) + \operatorname{tr}(\mathbf{\Sigma}_1^2)} \in (0,1)$$

- One can show that $\hat{\mathbf{\Sigma}}_1^\star = a_1^\star \mathbf{S}_1$ verifies: $\operatorname{MSE}(\hat{\mathbf{\Sigma}}_1^\star) = \hat{a}_1^\star \cdot \operatorname{MSE}(\mathbf{S}_1)$.

$\Rightarrow$ Since $0 < a_1^\star < 1$, $\hat{\mathbf{\Sigma}}_1^\star = a_1^\star \mathbf{S}_1$ is *always more efficient* than $\mathbf{S}_1$.

- Gaussian data: $(n_1 - 1)\mathbf{S}_1 \sim \mathcal{W}_p(n - 1, \mathbf{\Sigma}_1)$, so

$$\operatorname{MSE}(\mathbf{S}_1) = \frac{1}{n_1 - 1}(\operatorname{tr}(\mathbf{\Sigma}_1)^2 + \operatorname{tr}(\mathbf{\Sigma}_1^2)) \Rightarrow a_1^\star = \frac{n_1 - 1}{n_1 + \gamma/p}$$

where $\gamma = p \operatorname{tr}(\mathbf{\Sigma}_1^2)/\operatorname{tr}(\mathbf{\Sigma}_1)^2 \in [1, p]$ is a *measure of sphericity*.

$\Rightarrow$ LINPOOL estimator (for Gaussian data) is $\hat{\mathbf{\Sigma}}_1 = \frac{n_1 - 1}{n_1 + \hat{\gamma}/p}\mathbf{S}_1$.

Example 1: single class ($K = 1$) case

- In the single class case, we just need to find shrinkage parameter

$$a_1^\star = \underset{a \in \mathbb{R}}{\arg\min} \ \mathbb{E}\left[\|a\mathbf{S}_1 - \mathbf{\Sigma}_1\|_{\mathrm{F}}^2\right] = (\delta_1 + c_{11})^{-1} c_{11}$$
$$= \frac{\operatorname{tr}(\mathbf{\Sigma}_1^2)}{\mathrm{MSE}(\mathbf{S}_1) + \operatorname{tr}(\mathbf{\Sigma}_1^2)} \in (0,1)$$

- One can show that $\hat{\mathbf{\Sigma}}_1^\star = a_1^\star \mathbf{S}_1$ verifies: $\mathrm{MSE}(\hat{\mathbf{\Sigma}}_1^\star) = \hat{a}_1^\star \cdot \mathrm{MSE}(\mathbf{S}_1)$.

$\Rightarrow$ Since $0 < a_1^\star < 1$, $\hat{\mathbf{\Sigma}}_1^\star = a_1^\star \mathbf{S}_1$ is *always more efficient* than $\mathbf{S}_1$.

- Gaussian data: $(n_1 - 1)\mathbf{S}_1 \sim \mathcal{W}_p(n-1, \mathbf{\Sigma}_1)$, so

$$\mathrm{MSE}(\mathbf{S}_1) = \frac{1}{n_1 - 1}(\operatorname{tr}(\mathbf{\Sigma}_1)^2 + \operatorname{tr}(\mathbf{\Sigma}_1^2)) \Rightarrow a_1^\star = \frac{n_1 - 1}{n_1 + \gamma/p}$$

where $\gamma = p \operatorname{tr}(\mathbf{\Sigma}_1^2)/\operatorname{tr}(\mathbf{\Sigma}_1)^2 \in [1, p]$ is a *measure of sphericity*.

$\Rightarrow$ LINPOOL estimator (for Gaussian data) is $\hat{\mathbf{\Sigma}}_1 = \dfrac{n_1 - 1}{n_1 + \hat{\gamma}/p}\mathbf{S}_1$.

Examples: equal covariance matrices $\Sigma_k \equiv \Sigma \ \forall k$

- In this case, $\mathbf{C} = c\mathbf{1}\mathbf{1}^\top$ with $c = \mathrm{tr}(\mathbf{\Sigma}^2)/p$ and

$$\mathbf{a}_k^\star = (\Delta + \mathbf{C})^{-1}\mathbf{c}_k = c(\Delta + c\mathbf{1}\mathbf{1}^\top)^{-1}\mathbf{1}$$

$$\Rightarrow a_{jk}^\star = \frac{\mathrm{MSE}(\mathbf{S}_j)^{-1}}{\|\mathbf{\Sigma}\|^{-2} + a}, \quad a = \sum_{i=1}^{K} \mathrm{MSE}(\mathbf{S}_i)^{-1}.$$

- Remarks:
  1. $a_{jk}^\star > 0$ and $a_{jk}^\star \propto \mathrm{MSE}(\mathbf{S}_j)^{-1}$.
  2. $\mathbf{a}_1^\star = \cdots = \mathbf{a}_K^\star \Rightarrow \hat{\mathbf{\Sigma}}^\star = \sum_{j=1}^{K} a_{jk}^\star \mathbf{S}_j$.
  3. If $\mathrm{MSE}(\mathbf{S}_j)$ is large, then the weight for summand $\mathbf{S}_j$ is small.
  4. Gaussian data: $(n_j - 1)\mathbf{S}_j \sim \mathcal{W}(n_j - 1, \mathbf{\Sigma}) \ \forall j$, so

$$\mathrm{MSE}(\mathbf{S}_j) = \frac{1}{n_j - 1}(\mathrm{tr}(\mathbf{\Sigma})^2 + \mathrm{tr}(\mathbf{\Sigma}^2)) \Rightarrow a_{jk}^\star = \frac{n_j - 1}{n + 1 - K + \gamma/p}.$$

Compare against $\mathbf{S}_{\text{pool}} = \sum_{j=1}^{K} \frac{n_j}{n}\mathbf{S}_j$ (where $n = \sum_i n_i$).

Examples: equal covariance matrices $\Sigma_k \equiv \Sigma \ \forall k$

- In this case, $\mathbf{C} = c\mathbf{1}\mathbf{1}^\top$ with $c = \mathrm{tr}(\mathbf{\Sigma}^2)/p$ and

$$\mathbf{a}_k^\star = (\Delta + \mathbf{C})^{-1}\mathbf{c}_k = c(\Delta + c\mathbf{1}\mathbf{1}^\top)^{-1}\mathbf{1}$$

$$\Rightarrow a_{jk}^\star = \frac{\mathrm{MSE}(\mathbf{S}_j)^{-1}}{\|\mathbf{\Sigma}\|^{-2} + a}, \quad a = \sum_{i=1}^{K} \mathrm{MSE}(\mathbf{S}_i)^{-1}.$$

- Remarks:
  1. $a_{jk}^\star > 0$ and $a_{jk}^\star \propto \mathrm{MSE}(\mathbf{S}_j)^{-1}$.
  2. $\mathbf{a}_1^\star = \cdots = \mathbf{a}_K^\star \Rightarrow \hat{\mathbf{\Sigma}}^\star = \sum_{j=1}^K a_{jk}^\star \mathbf{S}_j$.
  3. If $\mathrm{MSE}(\mathbf{S}_j)$ is large, then the weight for summand $\mathbf{S}_j$ is small.
  4. Gaussian data: $(n_j - 1)\mathbf{S}_j \sim \mathcal{W}(n_j - 1, \mathbf{\Sigma}) \ \forall j$, so

$$\mathrm{MSE}(\mathbf{S}_j) = \frac{1}{n_j - 1}(\mathrm{tr}(\mathbf{\Sigma})^2 + \mathrm{tr}(\mathbf{\Sigma}^2)) \Rightarrow a_{jk}^\star = \frac{n_j - 1}{n + 1 - K + \gamma/p}.$$

Compare against $\mathbf{S}_{\text{pool}} = \sum_{j=1}^K \frac{n_j}{n}\mathbf{S}_j$ (where $n = \sum_i n_i$).

LINPOOL estimator with identity shrinkage

- Is $\hat{\boldsymbol{\Sigma}}_k = \sum_{j=1}^{K} \hat{a}_{jk}\mathbf{S}_j$, $\hat{a}_{jk} \geq 0$, positive definite ($\hat{\boldsymbol{\Sigma}}_k \succ 0$)?
- To account for this, we add $\mathbf{I}$ as an additional summand:

$$\mathbf{S}(\mathbf{a}) = a_{K+1}\mathbf{I} + \sum_{i=1}^{K} a_i \mathbf{S}_i,$$

where $a_i \geq 0$, $i = 1, \ldots, K$, $a_{K+1} > 0$ and $\mathbf{a} = (a_1, \ldots, a_K, a_{K+1})^\top$.

- The solution is found identically, since now the MSE is

$$\frac{1}{2p}\mathbb{E}\big[\|\mathbf{S}(\mathbf{a}) - \boldsymbol{\Sigma}_k\|_{\mathrm{F}}^2\big] = \frac{1}{2}\mathbf{a}^\top(\tilde{\boldsymbol{\Delta}} + \tilde{\mathbf{C}})\mathbf{a} - \tilde{\mathbf{c}}_k^\top \mathbf{a},$$

where

$$\tilde{\mathbf{C}} = \begin{pmatrix} \mathbf{C} & \boldsymbol{\eta} \\ \boldsymbol{\eta}^\top & 1 \end{pmatrix} \quad \text{and} \quad \tilde{\boldsymbol{\Delta}} = \begin{pmatrix} \boldsymbol{\Delta} & \mathbf{0} \\ \mathbf{0}^\top & 0 \end{pmatrix}$$

where $\boldsymbol{\eta} = (p^{-1}\operatorname{tr}(\boldsymbol{\Sigma}_1), \ldots, p^{-1}\operatorname{tr}(\boldsymbol{\Sigma}_K))^\top$.

LINPOOL estimator with identity shrinkage

- Is $\hat{\boldsymbol{\Sigma}}_k = \sum_{j=1}^{K} \hat{a}_{jk}\mathbf{S}_j$, $\hat{a}_{jk} \geq 0$, positive definite ($\hat{\boldsymbol{\Sigma}}_k \succ 0$)?
- To account for this, we add $\mathbf{I}$ as an additional summand:

$$\mathbf{S}(\mathbf{a}) = a_{K+1}\mathbf{I} + \sum_{i=1}^{K} a_i \mathbf{S}_i,$$

where $a_i \geq 0$, $i = 1, \ldots, K$, $a_{K+1} > 0$ and $\mathbf{a} = (a_1, \ldots, a_K, a_{K+1})^\top$.

- The solution is found identically, since now the MSE is

$$\frac{1}{2p}\mathbb{E}\left[\|\mathbf{S}(\mathbf{a}) - \boldsymbol{\Sigma}_k\|_{\mathrm{F}}^2\right] = \frac{1}{2}\mathbf{a}^\top(\tilde{\Delta} + \tilde{\mathbf{C}})\mathbf{a} - \tilde{\mathbf{c}}_k^\top \mathbf{a},$$

where

$$\tilde{\mathbf{C}} = \begin{pmatrix} \mathbf{C} & \boldsymbol{\eta} \\ \boldsymbol{\eta}^\top & 1 \end{pmatrix} \quad \text{and} \quad \tilde{\Delta} = \begin{pmatrix} \Delta & \mathbf{0} \\ \mathbf{0}^\top & 0 \end{pmatrix}$$

where $\boldsymbol{\eta} = (p^{-1}\operatorname{tr}(\boldsymbol{\Sigma}_1), \ldots, p^{-1}\operatorname{tr}(\boldsymbol{\Sigma}_K))^\top$.

LINPOOL estimator with identity shrinkage

- Is $\hat{\boldsymbol{\Sigma}}_k = \sum_{j=1}^{K} \hat{a}_{jk}\mathbf{S}_j$, $\hat{a}_{jk} \geq 0$, positive definite ($\hat{\boldsymbol{\Sigma}}_k \succ 0$)?
- To account for this, we add $\mathbf{I}$ as an additional summand:

$$\mathbf{S}(\mathbf{a}) = a_{K+1}\mathbf{I} + \sum_{i=1}^{K} a_i \mathbf{S}_i,$$

  where $a_i \geq 0$, $i = 1, \ldots, K$, $a_{K+1} > 0$ and $\mathbf{a} = (a_1, \ldots, a_K, a_{K+1})^\top$.
- The solution is found identically, since now the MSE is

$$\frac{1}{2p}\mathbb{E}\big[\|\mathbf{S}(\mathbf{a}) - \boldsymbol{\Sigma}_k\|_{\mathrm{F}}^2\big] = \frac{1}{2}\mathbf{a}^\top(\tilde{\boldsymbol{\Delta}} + \tilde{\mathbf{C}})\mathbf{a} - \tilde{\mathbf{c}}_k^\top \mathbf{a},$$

  where

$$\tilde{\mathbf{C}} = \begin{pmatrix} \mathbf{C} & \boldsymbol{\eta} \\ \boldsymbol{\eta}^\top & 1 \end{pmatrix} \quad \text{and} \quad \tilde{\boldsymbol{\Delta}} = \begin{pmatrix} \boldsymbol{\Delta} & \mathbf{0} \\ \mathbf{0}^\top & 0 \end{pmatrix}$$

  where $\boldsymbol{\eta} = (p^{-1}\operatorname{tr}(\boldsymbol{\Sigma}_1), \ldots, p^{-1}\operatorname{tr}(\boldsymbol{\Sigma}_K))^\top$.

Menu

LINPOOL estimator with convex combination

- Recall that LINPOOL estimator is $\hat{\mathbf{\Sigma}}_k = \sum_{j=1}^{K} \hat{a}_{jk}\mathbf{S}_j$.
- A natural modification is to require that the weights sum to 1:

$$\mathbf{1}^\top \hat{\mathbf{a}}_k = \sum_{j=1}^{K} \hat{a}_{jk} = 1$$

- *Note:* such constraint presumes that the true covariance matrices share similar scale $(\mathrm{tr}(\mathbf{\Sigma}_i) \approx \mathrm{tr}(\mathbf{\Sigma}_j))$
- This results in the following QP problem:

$$\begin{array}{ll} \text{minimize} & \frac{1}{2}\mathbf{a}^\top(\mathbf{\Delta} + \mathbf{C})\mathbf{a} - \mathbf{c}_k^\top \mathbf{a} \\ \text{subject to} & \mathbf{a} \geq \mathbf{0} \\ & \mathbf{1}^\top \mathbf{a} = 1. \end{array}$$

LINPOOL estimator using SDP

- Write $\mathbf{B} = \mathbf{C} + \Delta$. Then note that

$$\frac{1}{2p}\mathbb{E}\big[\|\mathbf{S}(\mathbf{a}) - \boldsymbol{\Sigma}_k\|_{\mathrm{F}}^2\big] = \frac{1}{2}\mathbf{a}^\top \mathbf{B}\mathbf{a} - \mathbf{c}_k^\top \mathbf{a} \qquad (+\,\text{const})$$
$$= \frac{1}{2}(\mathbf{a} - \mathbf{B}^{-1}\mathbf{c}_k)^\top \mathbf{B}(\mathbf{a} - \mathbf{B}^{-1}\mathbf{c}_k) \qquad (+\,\text{const}).$$

- It is possible to minimize the MSE under the constraint $\mathbf{S}(\mathbf{a}) \succeq 0$ by solving following *semidefinite program (SDP)*:

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

- *Note:* When $\mathbf{a}_k^\star = \mathbf{B}^{-1}\mathbf{c}_k$ has positive elements, then it is also the solution to SDP (and the constrained QP) problem.

LINPOOL estimator for multitarget problems

- Single class ($K = 1$) problem, in which we estimate the covariance matrix $\mathbf{\Sigma}_1$ from data $\mathcal{X}_1 = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$.
- Let $\mathbf{S}_1$ denote the SCM based on the data $\mathcal{X}_1$ and $\{\mathbf{T}_m\}_{m=1}^M$, $\mathbf{T}_m \succeq 0$, our set of target matrices.
- Then the **multitarget (MT-)RSCM** is defined as

$$\hat{\mathbf{\Sigma}}_1 = \beta \mathbf{S}_1 + \sum_{m=1}^{M} \alpha_m \mathbf{T}_m.$$

Q: How to determine the optimal weights β and $\{\alpha_m\}_{m=1}^M$?

- Often the target matrices are not fixed, but also based on the data $\mathcal{X}_1$.

  $\Rightarrow$ SCM $\mathbf{S}_1$ can not be considered independent of $\mathbf{T}_i$-s.
- We enhance independence and use LINPOOL estimator to construct a multitarget-style shrinkage estimator.

LINPOOL estimator for multitarget problems

- Single class ($K = 1$) problem, in which we estimate the covariance matrix $\mathbf{\Sigma}_1$ from data $\mathcal{X}_1 = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$.
- Let $\mathbf{S}_1$ denote the SCM based on the data $\mathcal{X}_1$ and $\{\mathbf{T}_m\}_{m=1}^{M}$, $\mathbf{T}_m \succeq 0$, our set of target matrices.
- Then the **multitarget (MT-)RSCM** is defined as

$$\hat{\mathbf{\Sigma}}_1 = \beta \mathbf{S}_1 + \sum_{m=1}^{M} \alpha_m \mathbf{T}_m.$$

Q: How to determine the optimal weights β and $\{\alpha_m\}_{m=1}^{M}$?

- Often the target matrices are not fixed, but also based on the data $\mathcal{X}_1$.

  $\Rightarrow$ SCM $\mathbf{S}_1$ can not be considered independent of $\mathbf{T}_i$-s.
- We enhance independence and use LINPOOL estimator to construct a multitarget-style shrinkage estimator.

The MT-LINPOOL estimator

1. Generate i.i.d. samples $\mathcal{X}_{m+1} \sim \mathcal{N}_p(\mathbf{0}, \mathbf{T}_m)$ for $m = 1, \ldots, M$ each of size L.
2. Compute $\mathbf{S}_1$ from $\mathcal{X}_1$ and $\mathbf{S}_2, \ldots, \mathbf{S}_{M+1}$ from $\mathcal{X}_2, \ldots, \mathcal{X}_{M+1}$.
3. Compute $\hat{\mathbf{C}}$ and $\hat{\Delta}$ based on data sets $\mathcal{X}_1$ and $\{\mathcal{X}_{m+1}\}_{m=1}^{M}$.
4. $\hat{\mathbf{a}} = \arg\min_{\mathbf{a} \geq \mathbf{0}} \ \frac{1}{2}\mathbf{a}^\top(\hat{\Delta} + \hat{\mathbf{C}})\mathbf{a} - \hat{\mathbf{c}}_1^\top \mathbf{a}$
5. $\hat{\mathbf{\Sigma}}_1 = \hat{a}_1\mathbf{S}_1 + \hat{a}_2\mathbf{S}_2 + \ldots + \hat{a}_{M+1}\mathbf{S}_{M+1}$

Note: one may view L as an additional regularization parameter.

Complex-valued case

- Our framework is general: the LINPOOL estimator can be constructed as earlier, but based on SCM-s,

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

  of complex-valued observations $\mathbf{x}_{i,k} \in \mathbb{C}^p$ $(k = 1, \ldots, K)$.
  Note: $(\cdot)^{\mathsf{H}}$ denotes the Hermitian transpose.
- Only estimation of $\mathbf{C}$ and Δ are affected (and this is the topic of the next section).

Menu

Estimation of C and Δ

- We needs to estimate the following parameter matrices:

$$\Delta = p^{-1}\,\mathrm{diag}(\mathbb{E}[\|\mathbf{S}_1 - \mathbf{\Sigma}_1\|_{\mathrm{F}}^2], \dots, \mathbb{E}[\|\mathbf{S}_k - \mathbf{\Sigma}_k\|_{\mathrm{F}}^2])$$
$$\mathbf{C} = (c_{ij}) = \left(\frac{\mathrm{tr}(\mathbf{\Sigma}_i \mathbf{\Sigma}_j)}{p}\right) \in \mathbb{R}^{K \times K}.$$

- We constuct estimates $\hat{\Delta}$ and $\hat{\mathbf{C}}$ under the assumption that the class distributions are (unspecified) elliptical distributions:

$$\{\mathbf{x}_{i,k}\}_{i=1}^{n_k} \overset{iid}{\sim} \mathcal{E}_p(\boldsymbol{\mu}_k, \mathbf{\Sigma}_k, g_k) \quad \text{for each } k$$

(defined on next slide)

Elliptically symmetric (ES) distributions

$\mathbf{x} \sim \mathcal{E}_p(\mathbf{0}, \mathbf{\Sigma}, g)$ when its pdf is [FKN90]

$$f(\mathbf{x}) \propto |\mathbf{\Sigma}|^{-1/2} g\big(\mathbf{x}^\top \mathbf{\Sigma}^{-1} \mathbf{x}\big),$$

where

- $\mathbf{\Sigma} \in \mathbb{S}_{++}^{p \times p}$ is the unknown **covariance matrix**.
- $g : [0, \infty) \to [0, \infty)$ is **density generator**

- We assume that ES distribution has finite 4^{th}-order moments.
- Multivariate normal (MVN) : $g(t) = \exp(-t/2)$
- The ES family also Includes other distributions such as multivariate t (MVT) with $\nu > 2$ d.o.f, generalized Gaussian distribution, etc.
- The (circular) complex elliptically symmetric distributions [OTKP12] can be defined similarly.

Estimate of MSE

We need the following statistics of $\mathbf{x} = (x_1, \ldots, x_p)^\top \sim \mathcal{E}_p(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i, g_i)$:

- *sphericity*: $\gamma_i = \dfrac{p\,\mathrm{tr}(\boldsymbol{\Sigma}_i^2)}{\mathrm{tr}(\boldsymbol{\Sigma}_i)^2} \in [1, p]$
- *scale*: $\eta_i = \dfrac{\mathrm{tr}(\boldsymbol{\Sigma}_i)}{p} > 0$
- *elliptical kurtosis*:
  $$\kappa_i = \begin{cases} \frac{1}{3} \cdot \mathrm{kurt}(x_1), & \text{real case} \\ \frac{1}{2} \cdot \mathrm{kurt}(x_1), & \text{complex case} \end{cases}$$

Lemma: The MSE of SCM $\mathbf{S}_i$ when data is from $\mathcal{E}_p(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i, g_i)$ is

$$\frac{\mathrm{MSE}(\mathbf{S}_i)}{p} = \eta_i^2 \times \begin{cases} \left(\dfrac{1}{n_i - 1} + \dfrac{\kappa_i}{n_i}\right)(p + \gamma_i) + \dfrac{\kappa_i}{n_i}\gamma_i, & \text{real case} \\ \left(\dfrac{1}{n_i - 1} + \dfrac{\kappa_i}{n_i}\right)p + \dfrac{\kappa_i}{n_i}\gamma_i, & \text{complex case} \end{cases}$$

Estimate of MSE

We need the following statistics of $\mathbf{x} = (x_1, \ldots, x_p)^\top \sim \mathcal{E}_p(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i, g_i)$:

- *sphericity*: $\gamma_i = \dfrac{p\,\mathrm{tr}(\boldsymbol{\Sigma}_i^2)}{\mathrm{tr}(\boldsymbol{\Sigma}_i)^2} \in [1, p]$ $\Rightarrow$ next slide
- *scale*: $\eta_i = \dfrac{\mathrm{tr}(\boldsymbol{\Sigma}_i)}{p} > 0 \Rightarrow \hat{\eta}_i = \mathrm{tr}(\mathbf{S}_i)$
- *elliptical kurtosis*:

$$\kappa_i = \begin{cases} \frac{1}{3} \cdot \mathrm{kurt}(x_1), & \text{real case} \\ \frac{1}{2} \cdot \mathrm{kurt}(x_1), & \text{complex case} \end{cases} \Rightarrow \hat{\kappa}_i = \begin{cases} \frac{1}{3} \cdot \widehat{\mathrm{kurt}}(x_1) \\ \frac{1}{2} \cdot \widehat{\mathrm{kurt}}(x_1) \end{cases}$$

Lemma: The MSE of SCM $\mathbf{S}_i$ when data is from $\mathcal{E}_p(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i, g_i)$ is

$$\frac{\mathrm{MSE}(\mathbf{S}_i)}{p} = \eta_i^2 \times \begin{cases} \left(\dfrac{1}{n_i - 1} + \dfrac{\kappa_i}{n_i}\right)(p + \gamma_i) + \dfrac{\kappa_i}{n_i}\gamma_i, & \text{real case} \\ \left(\dfrac{1}{n_i - 1} + \dfrac{\kappa_i}{n_i}\right)p + \dfrac{\kappa_i}{n_i}\gamma_i, & \text{complex case} \end{cases}$$

Estimate of sphericity

- Define a *shape matrix* $\mathbf{\Lambda}_k = p\dfrac{\mathbf{\Sigma}_k}{\mathrm{tr}(\mathbf{\Sigma}_k)}$.

  The sphericity measure can then be expressed as $\gamma_k = \dfrac{\mathrm{tr}(\mathbf{\Lambda}_k^2)}{p}$.

- As an estimator of $\mathbf{\Lambda}_k$, we use

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

$$\hat{\boldsymbol{\mu}}_k = \arg\min_{\boldsymbol{\mu}} \sum_{i=1}^{n_k} \|\mathbf{x}_{i,k} - \boldsymbol{\mu}\| \qquad \text{(spatial median [Bro83])}$$

- $\hat{\mathbf{\Lambda}}_k$ is a scaled ($\times p$) *spatial sign covariance matrix* [VKO00].

Estimate of sphericity

Theorem 2: Under assumption

(A) $\{\mathbf{x}_{i,k}\}_{i=1}^{n_k} \sim \mathcal{E}_p(\mathbf{0}, \boldsymbol{\Sigma}_k, g_k)$ and $\gamma_k = o(p)$ as $p \to \infty$

it holds that

$$\mathbb{E}[\hat{\boldsymbol{\Lambda}}_k] = \boldsymbol{\Lambda}_k + o(\|\boldsymbol{\Lambda}_k\|_{\mathrm{F}}).$$

- It is easy to show that

$$\frac{\mathbb{E}[\mathrm{tr}(\hat{\boldsymbol{\Lambda}}_k^2)]}{p} = \frac{p}{n_k} + \frac{n_k - 1}{n_k} \underbrace{\frac{\mathrm{tr}(\mathbb{E}[\hat{\boldsymbol{\Lambda}}]^2)}{p}}_{\text{Th. 2:}\to\gamma \text{ as } p\to\infty}$$

- Hence

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

  is an *asymptotically unbiased estimator* of γ_k.

.

Estimate of sphericity

Theorem 2: Under assumption

(A) $\{\mathbf{x}_{i,k}\}_{i=1}^{n_k} \sim \mathcal{E}_p(\mathbf{0}, \boldsymbol{\Sigma}_k, g_k)$ and $\gamma_k = o(p)$ as $p \to \infty$

it holds that

$$\mathbb{E}[\hat{\boldsymbol{\Lambda}}_k] = \boldsymbol{\Lambda}_k + o(\|\boldsymbol{\Lambda}_k\|_{\mathrm{F}}).$$

- It is easy to show that

$$\frac{\mathbb{E}[\mathrm{tr}(\hat{\boldsymbol{\Lambda}}_k^2)]}{p} = \frac{p}{n_k} + \frac{n_k - 1}{n_k} \underbrace{\frac{\mathrm{tr}(\mathbb{E}[\hat{\boldsymbol{\Lambda}}]^2)}{p}}_{\text{Th. 2:}\to\gamma \text{ as } p\to\infty}$$

- Hence

$$\hat{\gamma}_k = \frac{n_k}{n_k - 1}\left(\frac{\mathrm{tr}(\hat{\boldsymbol{\Lambda}}_k^2)}{p} - \frac{p}{n_k}\right) - d_k$$

is an *asymptotically unbiased estimator* of γ_k. We also use correction term d_k proposed in [ZPFW14].

Estimates of $c_{ij} = \mathrm{tr}(\mathbf{\Sigma}_i \mathbf{\Sigma}_j)/p$

- $i = j$: Use $\hat{c}_{ii} = \hat{\eta}_i^2 \hat{\gamma}_i$ as an estimator of

$$c_{ii} = \frac{\mathrm{tr}(\mathbf{\Sigma}_i^2)}{p} = \eta_i^2 \gamma_i, \quad i = 1, \ldots, K.$$

(where $\eta_i = \mathrm{tr}(\mathbf{\Sigma}_i)/p$)

- $i \neq j$: use $\hat{c}_{ij} = \hat{\eta}_i \hat{\eta}_j \, \mathrm{tr}(\hat{\mathbf{\Lambda}}_i \hat{\mathbf{\Lambda}}_j)/p$ as an estimator of

$$c_{ij} = \frac{\mathrm{tr}(\mathbf{\Sigma}_i \mathbf{\Sigma}_j)}{p} = \eta_i \eta_j \frac{\mathrm{tr}(\mathbf{\Lambda}_i \mathbf{\Lambda}_j)}{p}.$$

Menu

A simulation study: set-up

dimension	# of. classes	sample lengths
300	4	$n_k = n \ \forall k$

- $\{\mathbf{x}_{i,k}\}_{i=1}^{n_k} \overset{iid}{\sim} t_{p,\nu}(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$ with degr. of freedom $\nu = 7$.
- $\boldsymbol{\Sigma}_k$ has an AR(1) structure, $(\boldsymbol{\Sigma}_k)_{ij} = \eta_k \varrho^{|i-j|}$, where

  $$\varrho_1 = 0.3, \varrho_2 = 0.4, \varrho_3 = 0.5, \varrho_4 = 0.6$$

  and $\eta_k = \mathrm{tr}(\boldsymbol{\Sigma}_k)/p = k$, $k = 1, \ldots, K$.
- We compute the normalized MSE (NMSE)

  $$\|\hat{\boldsymbol{\Sigma}}_k - \boldsymbol{\Sigma}_k\|_{\mathrm{F}}^2 / \|\boldsymbol{\Sigma}_k\|_{\mathrm{F}}^2$$

  and total NMSE

  $$\sum_{k=1}^{K} \|\hat{\boldsymbol{\Sigma}}_k - \boldsymbol{\Sigma}_k\|_{\mathrm{F}}^2 / \|\boldsymbol{\Sigma}_k\|_{\mathrm{F}}^2$$

  averaged over 1000 MC trials.

A simulation study: estimators

- We use LINPOOL estimator with identity shrinkage:

$$\hat{\boldsymbol{\Sigma}}_k = \sum_{i=1}^{K} \hat{a}_{ik}\mathbf{S}_i + \hat{a}_{(K+1)k}\mathbf{I}.$$

- We compare with the MT-RSCM estimators of the form:

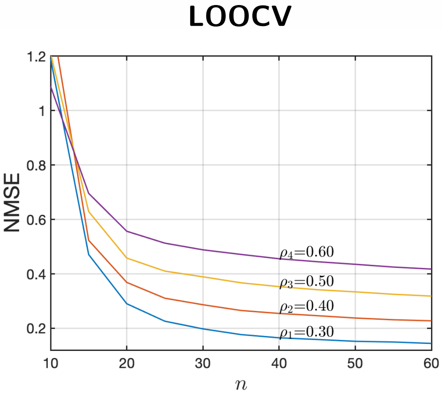
$$\tilde{\boldsymbol{\Sigma}}_k = \sum_{i=1}^{K} \tilde{a}_{ik}\mathbf{T}_i^{(k)} + \tilde{a}_{(K+1)k}\mathbf{S}_k.$$

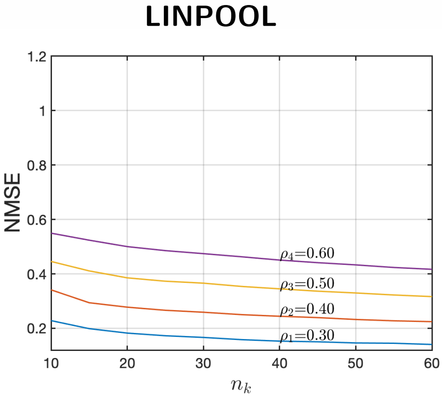
  where $\mathbf{T}_i^{(k)} \succeq 0$ are K target matrices for the k^{th} class.
- As the set of target matrices, we use

$$\{\mathbf{T}_i^{(k)}\}_{i=1}^{K} = \{\mathbf{I}\} \cup \{\mathbf{S}_i\}_{i \in \{1,\dots,K\} \setminus k}.$$

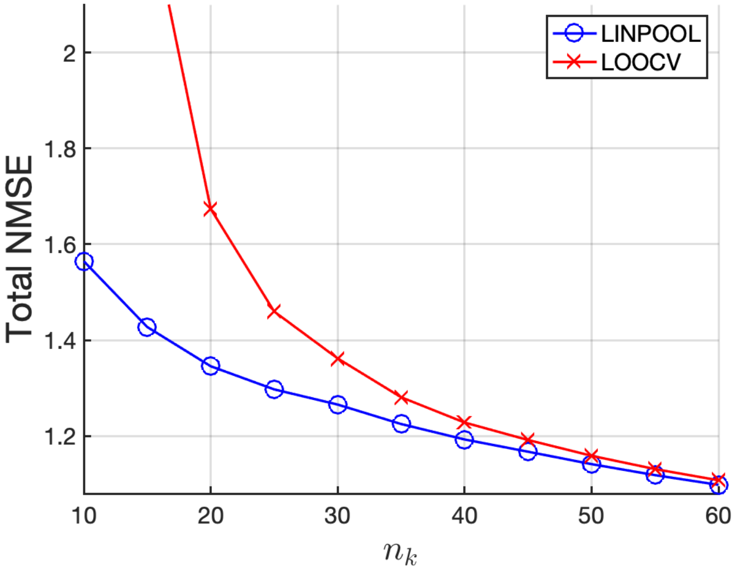
  Hence the MT-estimator equals LINPOOL estimator, except for its choice of weights.
- We use LOOCV [THX+18] method for computing the optimal MT weights.

Results: NMSE

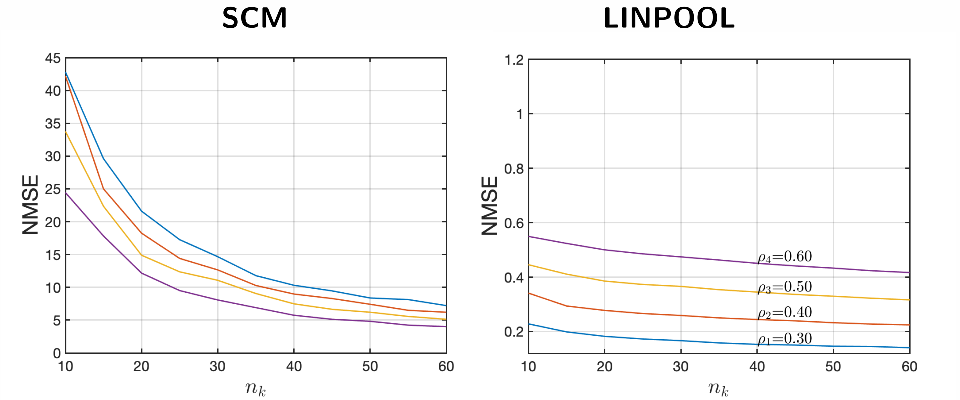




Results: Total NMSE



What about just using plain SCMS-s?



$$\text{NMSE}(\mathbf{S}_k) \approx 100\times\ \text{NMSE}(\hat{\mathbf{\Sigma}}_k)$$

Menu

Basic definitions (cont'd)

- $p :=$ # of stocks in the portfolio
- $w_i :=$ proportion of total wealth allocated to i^{th} asset, verifying

$$1 = \sum_{i=1}^{p} w_i = \mathbf{w}^\top \mathbf{1}.$$

- $\mathbf{r} = (r_1, \ldots, r_p)^\top :=$ net returns of p assets (at some time t).
- Two key statistics of portfolio net return $R = \mathbf{w}^\top \mathbf{r}$ are

$$\text{mean return} \qquad \mu_{\mathbf{w}} = \mathbb{E}[R] = \mathbf{w}^\top \boldsymbol{\mu}$$

$$\text{variance (risk)} \quad \sigma_{\mathbf{w}}^2 = \mathrm{var}(R) = \mathbf{w}^\top \boldsymbol{\Sigma} \mathbf{w}.$$

- Global minimum variance portfolio (GMVP) allocation strategy:

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

$$\Rightarrow \mathbf{w}_o = \frac{\boldsymbol{\Sigma}^{-1}\mathbf{1}}{\mathbf{1}^\top \boldsymbol{\Sigma}^{-1}\mathbf{1}}.$$

- See [FP16] for a great reference on financial engineering.

GMVP stock data analysis

Data sets (daily net returns of daily closing prices)

- $p = 50$ stocks in Hang Seng Index (HSI), 1/2016 - 12/2017.
- $p = 45$ stocks in Hang Seng Index (HSI), 1/2010 - 12/2011.

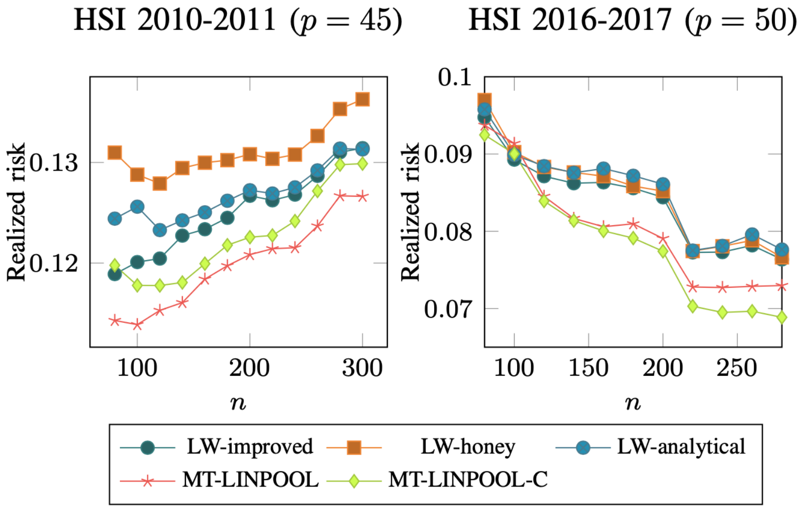
Sliding window method

- At day t, we use the previous n days to estimate $\boldsymbol{\Sigma}$ and $\mathbf{w}$.
- portfolio returns are then computed for the following 20 days.
- Window is shifted 20 trading days forward, new allocations and portfolio returns for another 20 days are computed.

GMVP stock data analysis: methods

- We use **MT-LINPOOL** method with 2 target matrices:
  - The single factor market index model $\mathbf{T}_F$ computed as in [LW03].
  - The constant correlation model $\mathbf{T}_C$ computed as in [LW04a]
  - **MT-LINPOOL-C** is same as earlier method but with constraint that weights $\hat{a}_{ik}$ sum to 1, i.e., $1 = \sum_i \hat{a}_{ik} = 1$.
- We compare against the following methods developed by finance experts (Profs. O. Ledoit and M. Wolf):
  1. **LW-improved** [LW03]: RSCM with shrinkage towards $\mathbf{T}_F$
  2. **LW-honey** [LW04a]: RSCM with shrinkage towards $\mathbf{T}_C$.
  3. **LW-analytical** [LW20]: nonlinear shrinkage of eigenvalues of SCM.

GMVP stock data analysis: results



The proposed MT-LINPOOL approach is able to provide the smallest realised risk results

What's cooking

- The paper is available at ArXiv:

  https://arxiv.org/abs/2008.05854

  Note: we are currently revising the paper, and extension to complex-valued data is not (yet!) available in the ArXiv submission.
- The codes (MATLAB, python) are also available at:

  https://github.com/EliasRaninen
- Also take a look at the related double shrinkage RSCM method:

  https://arxiv.org/abs/2011.04315

  Coupled regularized sample covariance matrix estimator for multiple classes, E. Raninen and E. Ollila.
- Or find out about robust linear shrinkage methods:

  https://doi.org/10.1109/TSP.2020.3043952

  Shrinking the eigenvalues of M-estimators of covariance matrix, E. Ollila, D.P. Palomar, F. Pascal, *TSP* 2020 (early access).

REFERENCES

BM Brown, *Statistical uses of the spatial median*, Journal of the Royal Statistical Society. Series B (Methodological) (1983), 25–30.

Angelo Coluccia, *Regularized covariance matrix estimation via empirical bayes*, IEEE Signal Processing Letters **22** (2015), no. 11, 2127–2131.

Yilun Chen, Ami Wiesel, Yonina C. Eldar, and Alfred O. Hero, *Shrinkage algorithms for MMSE covariance estimation*, IEEE Trans. Signal Process. **58** (2010), no. 10, 5016–5029.

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

Kai-Tai Fang, Samuel Kotz, and Kai-Wang Ng, *Symmetric multivariate and related distributions*, Chapman and hall, London, 1990.

Yiyong Feng and Daniel P Palomar, *A signal processing perspective on financial engineering*, vol. 9, Now Publishers, 2016.

J. H. Friedman, *Regularized discriminant analysis*, J. Amer. Stat. Assoc. **84** (1989), no. 405, 165–175.

Sadeep Jayasumana, Richard Hartley, Mathieu Salzmann, Hongdong Li, and Mehrtash Harandi, *Kernel methods on the riemannian manifold of symmetric* *positive definite matrices*, Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR), 2013, pp. 73–80.

Olivier Ledoit and Michael Wolf, *Improved estimation of the covariance matrix of stock returns with an application to portfolio selection*, Journal of empirical finance **10** (2003), no. 5, 603–621.

O. Ledoit and M. Wolf, *Honey, i shrunk the sample covariance matrix*, The Journal of Portfolio Management **30** (2004), no. 4, 110–119.

Olivier Ledoit and Michael Wolf, *A well-conditioned estimator for large-dimensional covariance matrices*, J. Mult. Anal. **88** (2004), no. 2, 365–411.

______, *Analytical nonlinear shrinkage of large-dimensional covariance matrices*, Annals of Statistics **48** (2020), no. 5, 3043–3065.

Jianbo Li, Jie Zhou, and Bin Zhang, *Estimation of large covariance matrices by shrinking to structured target in normal and non-normal distributions*, IEEE Access **6** (2017), 2158–2169.

Ammar Mian, Elias Raninen, and Esa Ollila, *A comparative study of supervised learning algorithms for symmetric positive definite features*, The 28th European Signal Processing Conference (EUSIPCO'20) (Amsterdam, Netherlands), Aug. 24 – 28 2020, pp. 699–703.

Esa Ollila and Elias Raninen, *Optimal shrinkage covariance matrix estimation under random sampling from elliptical distributions*, IEEE Transactions on Signal Processing **67** (2019), no. 10, 2707–2719.

Esa Ollila, David E. Tyler, Visa Koivunen, and H. Vincent Poor, *Complex elliptically symmetric distributions: survey, new results and applications*, IEEE Trans. Signal Process. **60** (2012), no. 11, 5597–5625.

Elias Raninen and Esa Ollila, *Optimal pooling of covariance matrix estimates across multiple classes*, Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing (ICASSP'18) (Calgary, Canada), April 15–20, 2018, pp. 4224–4228.

Elias Raninen and Esa Ollila, *Coupled regularized sample covariance matrix estimator for multiple classes*, 2020.

Petre Stoica, Jian Li, Xumin Zhu, and Joseph R Guerci, *On using a priori knowledge in space-time adaptive processing*, IEEE transactions on signal processing **56** (2008), no. 6, 2598–2602.

Jun Tong, Rui Hu, Jiangtao Xi, Zhitao Xiao, Qinghua Guo, and Yanguang Yu, *Linear shrinkage estimation of covariance matrices using low-complexity cross-validation*, Signal Processing **148** (2018), 223–233.

Oncel Tuzel, Fatih Porikli, and Peter Meer, *Pedestrian detection via classification on riemannian manifolds*, IEEE Trans. Pattern Anal. Mach. Intell. **30** (2008), no. 10, 1713–1727.

S. Visuri, V. Koivunen, and H. Oja, *Sign and rank covariance matrices*, J. Statist. Plann. Inference **91** (2000), 557–575.

Changliang Zou, Liuhua Peng, Long Feng, and Zhaojun Wang, *Multivariate sign-based high-dimensional tests for sphericity*, Biometrika **101** (2014), no. 1.