Regularized M-estimators of the Covariance Matrix

Esa Ollila

Aalto University, Department of Signal Processing and Acoustics, Finland
esa.ollila@aalto.fi http://signal.hut.fi/~esollila/

Summer School, Rüdesheim, Sep 22, 2016

Contents

Part A Regularized M-estimators of covariance:

- M-estimation and geodesic (g-)convexity
- Regularization via g-convex penalties
- Penalty (tuning) parameter selection

Part B Applications

- Regularized discriminant analysis
- Radar detection
- Optimal portfolio selection

Covariance estimation problem

- $\mathbf{x}$: p-variate (centered) random vector
- $\mathbf{x}_1, \ldots, \mathbf{x}_n$ i.i.d. realizations of $\mathbf{x}$, collected in data matrix

$$\mathbf{X} = \begin{pmatrix} \mathbf{x}_1^\top \\ \vdots \\ \mathbf{x}_n^\top \end{pmatrix}$$

- Problem: Find an estimate $\hat{\mathbf{\Sigma}} = \hat{\mathbf{\Sigma}}(\{\mathbf{x}_i\}_{i=1}^n)$ of the positive definite covariance matrix

$$\mathbf{\Sigma} = \mathbb{E}[\mathbf{x}\mathbf{x}^\top] \in \mathcal{S}(p)$$

- Solution: Maximum likelihood, M-estimation.

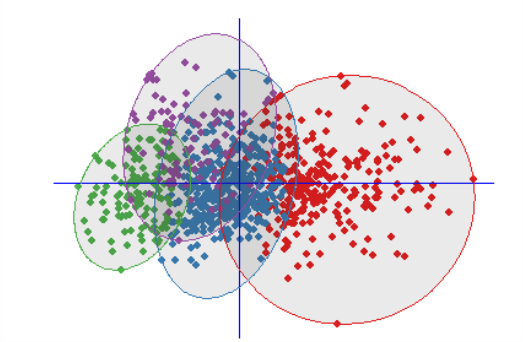
Conventional estimate: the sample covariance matrix (SCM)

$$\mathbf{S} = \frac{1}{n}\mathbf{X}^\top\mathbf{X} = \frac{1}{n}\sum_{i=1}^n \mathbf{x}_i\mathbf{x}_i^\top$$

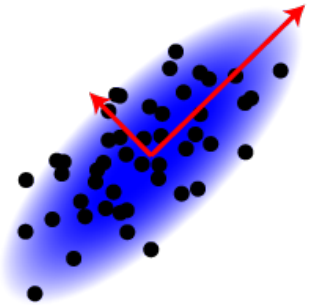
Why covariance estimation?

Portfolio selection

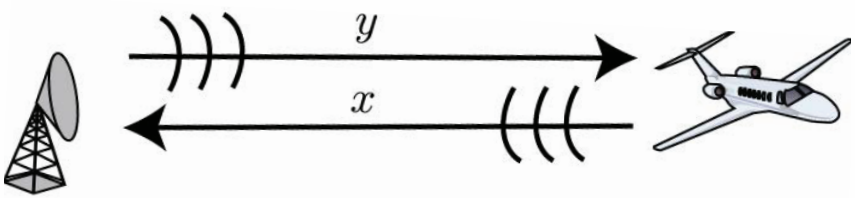
Discriminant Analysis



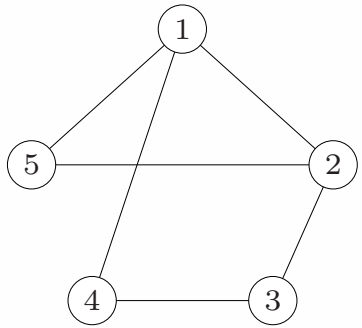
PCA



Radar detection



Graphical models



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

Covariance estimation challenges

1. Insufficient sample support (ISS) case: $p > n$.
   $\Longrightarrow$ Estimate of $\mathbf{\Sigma}^{-1}$ can not be computed!
2. Low sample support (LSS) (i.e., p of the same magnitude as n)
   $\Longrightarrow$ $\hat{\mathbf{\Sigma}}$ is estimated with a lot of error.
3. Outliers or heavy-tailed non-Gaussian data
   $\Longrightarrow$ $\hat{\mathbf{\Sigma}}$ is completely corrupted.

Problem 1 & 2 = Sparse data
$\Rightarrow$ regularization (this talk)
$\Rightarrow$ RMT (Frederic's talk)

Problem 3
$\Rightarrow$ robust estimation

Why robustness?

1. Outliers difficult to glean from high-dimensional data sets
2. Impulsive measurement environments (e.g., fMRI)
3. SCM is vulnerable to outliers and inefficient under non-Gaussianity
4. Most robust estimators can not be computed in $p > n$ cases

Contents

Acknowledgement

- To my co-authors:

David E. Tyler
Rutgers University

Ami Wiesel
Hebrew U. Jerusalem

Ilya Soloveychik
Hebrew U. Jerusalem

- and many inspiring people working in this field:

  Frederic Pascal, Teng Zhang, Lutz Dümbgen, Romain Couillet, Matthew R. McKay, Yuri Abramovich, Olivier Besson, Maria Greco, Fulvio Gini, Daniel Palomar, ...

Menu

I. Ad-hoc shrinkage SCM-s of multiple samples

II. ML- and M-estimators of scatter matrix

III. Geodesic convexity

IV. Regularized M-estimators

V. Penalized estimation of multiple covariances

VI. Estimation of the regularization parameter

VII. Applications

Multiple covariance estimation problem

- We are given K groups of elliptically distributed measurements,

$$\mathbf{x}_{11}, \ldots, \mathbf{x}_{1n_1}, \quad \ldots, \quad \mathbf{x}_{K1}, \ldots, \mathbf{x}_{Kn_K}$$

- Each group $\mathbf{X}_k = (\mathbf{x}_{k1} \ \cdots \ \mathbf{x}_{kn_k})$ containing n_k p-dimensional observation vectors, and

$$N = \sum_{i=1}^{K} n_k = \text{ total sample size}$$

$$\pi_k = \frac{n_k}{N} = \text{ relative sample size of the } k\text{-th group}$$

- Sample populations follow elliptical distributions, $\mathcal{E}_p(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k, g_k)$, with different scatter matrices $\boldsymbol{\Sigma}_k$ possessing mutual structure or a **joint center** $\boldsymbol{\Sigma} \Rightarrow$ need to estimate both $\{\boldsymbol{\Sigma}_k\}_{k=1}^K$ and $\boldsymbol{\Sigma}$.
- We assume that the symmetry center $\boldsymbol{\mu}_k$ of populations is known or that the data sets are *centered*.

Ad-hoc regularization approach

- Gaussian MLE-s of $\mathbf{\Sigma}_1, \ldots, \mathbf{\Sigma}_K$ are the SCM-s $\mathbf{S}_1, \ldots, \mathbf{S}_K$
- If n_k small relative to p, common assumption is $\mathbf{\Sigma}_1 = \ldots = \mathbf{\Sigma}_K$ which is estimated by pooled SCM

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

- Rather than assume the population covariance matrices are all equal (*hard modeling*), simply shrink them towards equality (*soft modeling*):

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

  e.g., as in [Friedman, 1989], where $\beta \in (0,1)$ is a regularization parameter, commonly chosen by cross-validation.
- If the the total sample size N is also small relative to dimension p, then Friedman recommends also shrinking the pooled SCM $\mathbf{S}$ towards $\propto \mathbf{I}$.

Regularized covariance matrices

Q1 Can the Ad-Hoc method be improved or some theory/formalism put behind it?

Q2 Robustness and resistance, e.g., non-Gaussian models and outliers.

Q3 Methods other then convex combinations?

Q4 Shrinkage towards other models?

- E.g., proportional covariance matrices instead of common covariance matrices?
- Other types of shrinkage to the structure?

Q1: Some formalism to the Ad-Hoc method

- Gaussian ML cost function ($-2\times$ neg. log-likelihood) for the kth class:

$$\mathcal{L}_{\mathrm{G},k}(\boldsymbol{\Sigma}_k) = \mathrm{Tr}(\boldsymbol{\Sigma}_k^{-1}\mathbf{S}_k) - \log|\boldsymbol{\Sigma}_k^{-1}|$$

  has a unique minimizer at $\hat{\boldsymbol{\Sigma}}_k = \mathbf{S}_k$ (= SCM of the kth sample).

- Penalized objective function: Add a penalty term and solve

$$\min_{\boldsymbol{\Sigma}_k \in \mathcal{S}(p)} \left\{ \mathcal{L}_{\mathrm{G},k}(\boldsymbol{\Sigma}_k) + \lambda\, d(\boldsymbol{\Sigma}_k, \hat{\boldsymbol{\Sigma}}) \right\}, \quad k = 1, \ldots K,$$

  where
  - $\lambda > 0$ is a penalty/regularization parameter
  - $d(\mathbf{A}, \mathbf{B}) : \mathcal{S}(p) \times \mathcal{S}(p) \to \mathbb{R}_0^+$ is a penalty/distance function minimized whenever $\mathbf{A} = \mathbf{B}$

  Idea: Penalty shrinks $\hat{\boldsymbol{\Sigma}}_k$ towards a fixed **shrinkage target matrix** $\hat{\boldsymbol{\Sigma}} \in \mathcal{S}(p)$, the amount of shrinkage depends on the magnitude of λ

Q1: Some formalism to the Ad-Hoc method

- The information theoretic Kullback-Leibler (KL) divergence [Cover and Thomas, 2012], distance from $\mathcal{N}_p(\mathbf{0}, \mathbf{A})$ to $\mathcal{N}_p(\mathbf{0}, \mathbf{B})$, is

$$d_{\mathrm{KL}}(\mathbf{A}, \mathbf{B}) = \mathrm{Tr}(\mathbf{A}^{-1}\mathbf{B}) - \log|\mathbf{A}^{-1}\mathbf{B}| - p.$$

  As is well known, it verifies $d_{\mathrm{KL}}(\mathbf{A}, \mathbf{B}) \geq 0$ and $= 0$ for $\mathbf{A} = \mathbf{B}$.

- Using $d_{\mathrm{KL}}(\boldsymbol{\Sigma}_k, \hat{\boldsymbol{\Sigma}})$ as the penalty, the optimization problem $\mathcal{L}_{\mathrm{G},k}(\boldsymbol{\Sigma}_k) + \lambda\, d_{\mathrm{KL}}(\boldsymbol{\Sigma}_k, \hat{\boldsymbol{\Sigma}})$ possesses a unique solution given by

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

  where $\beta = (1+\lambda)^{-1} \in (0,1)$ and $k = 1, \ldots, K$.

- This gives Friedman's Ad-Hoc shrinkage SCM estimators when the shrinkage target matrix $\hat{\boldsymbol{\Sigma}}$ is the pooled SCM $\mathbf{S}$

Q1: Some formalism to the Ad-Hoc method

- The information theoretic Kullback-Leibler (KL) divergence [Cover and Thomas, 2012], distance from $\mathcal{N}_p(\mathbf{0}, \mathbf{A})$ to $\mathcal{N}_p(\mathbf{0}, \mathbf{B})$, is

$$d_{\mathrm{KL}}(\mathbf{A}, \mathbf{B}) = \mathrm{Tr}(\mathbf{A}^{-1}\mathbf{B}) - \log|\mathbf{A}^{-1}\mathbf{B}| - p.$$

  As is well known, it verifies $d_{\mathrm{KL}}(\mathbf{A}, \mathbf{B}) \geq 0$ and $= 0$ for $\mathbf{A} = \mathbf{B}$.

- Using $d_{\mathrm{KL}}(\boldsymbol{\Sigma}_k, \hat{\boldsymbol{\Sigma}})$ as the penalty, the optimization problem $\mathcal{L}_{\mathrm{G},k}(\boldsymbol{\Sigma}_k) + \lambda\, d_{\mathrm{KL}}(\boldsymbol{\Sigma}_k, \hat{\boldsymbol{\Sigma}})$ possesses a unique solution given by

$$\mathbf{S}_k(\beta) = \beta\mathbf{S}_k + (1-\beta)\mathbf{S}\,, \quad k = 1, \ldots, K,$$

  where $\beta = (1+\lambda)^{-1} \in (0, 1)$ and $k = 1, \ldots, K$.

- This gives Friedman's Ad-Hoc shrinkage SCM estimators when the shrinkage target matrix $\hat{\boldsymbol{\Sigma}}$ is the pooled SCM $\mathbf{S}$

Discussion

Note: The Gaussian likelihood $\mathcal{L}_{\mathrm{G},k}(\boldsymbol{\Sigma}_k)$ is convex in $\boldsymbol{\Sigma}_k^{-1}$ and so is $d_{\mathrm{KL}}(\boldsymbol{\Sigma}_k, \hat{\boldsymbol{\Sigma}})$.

Comments

- Other (non-Gaussian) ML cost functions $\mathcal{L}_k(\boldsymbol{\Sigma})$ are commonly not convex in $\boldsymbol{\Sigma}^{-1}$
- Swapping the order $d_{\mathrm{KL}}(\boldsymbol{\Sigma}_k, \hat{\boldsymbol{\Sigma}})$ to $d_{\mathrm{KL}}(\hat{\boldsymbol{\Sigma}}, \boldsymbol{\Sigma}_k)$ gives a distance function that is non-convex in $\boldsymbol{\Sigma}_k^{-1}$.

Problems

- The penalized optimization program, $\mathcal{L}_{\mathrm{G},k}(\boldsymbol{\Sigma}_k) + \lambda\, d_{\mathrm{KL}}(\boldsymbol{\Sigma}_k, \hat{\boldsymbol{\Sigma}})$, does not seem to generalize to using other distance functions or other non-Gaussian cost functions.
- KL-distance $d_{\mathrm{KL}}(\boldsymbol{\Sigma}_k, \boldsymbol{\Sigma})$ is not so useful when the assumption is $\boldsymbol{\Sigma}_k \propto \boldsymbol{\Sigma}$, i.e., proportional covariance matrices.

How about a robust Ad-hoc method?

- **Plug-In Robust Estimators**: Let $\hat{\boldsymbol{\Sigma}}_k$ and $\hat{\boldsymbol{\Sigma}}$ represent robust estimates of scatter (covariance) matrix for the kth class and the pooled data respectively.
- Then a robust version of Friedman's approach is given by

$$\hat{\boldsymbol{\Sigma}}_k(\beta) = \beta\hat{\boldsymbol{\Sigma}}_k + (1-\beta)\hat{\boldsymbol{\Sigma}}, \quad k = 1, \ldots, K$$

  where $\beta \in (0, 1)$.
- **Problems**: This approach fails since many robust estimators of scatter, e.g. M, S, MM, MCD, etc., are not defined or do not vary much from the sample covariance when the data is sparse.

Our approach in this series of lectures

- **Regularization** via jointly g-convex distance functions
- **Robust M-estimation** (robust loss fnc downweights outliers)

Menu

I. Ad-hoc shrinkage SCM-s of multiple samples

II. ML- and M-estimators of scatter matrix

III. Geodesic convexity

IV. Regularized M-estimators

V. Penalized estimation of multiple covariances

VI. Estimation of the regularization parameter

VII. Applications

References


R. A. Maronna (1976).
Robust M-estimators of multivariate location and scatter.
Ann. Stat., 5(1):51–67.


D. E. Tyler (1987).
A distribution-free M-estimator of multivariate scatter.
Ann. Stat., 15(1):234–251.


E. Ollila D. E. Tyler V. Koivunen, and H. V. Poor (2012).
Complex elliptically symmetric distributions: survey, new results and applications.
IEEE Trans. Signal Processing, 60(11):5597 – 5625.

The cone of positive definite matrices $\mathcal{S}(p)$

- A square matrix $\mathbf{A}$ is positive definite, denoted $\mathbf{A} \succ 0$, if it is symmetric and satisfies

$$\mathbf{x}^\top \mathbf{A}\mathbf{x} > 0 \quad \forall \mathbf{x} \neq \mathbf{0}$$

  Positive semidefinite ($\mathbf{A} \succeq 0$): $\mathbf{x}^\top \mathbf{A}\mathbf{x} \geq 0 \quad \forall \mathbf{x}$.
- $\mathbf{A} \succ 0$ ($\succeq 0$) if and only if its eigenvalues are positive (non-negative)
- Eigenvalue decomposition (EVD) of $\mathbf{A} \succ 0$;

$$\mathbf{A} = \mathbf{E}\Lambda\mathbf{E}^\top = \sum_{i=1}^{p} \lambda_i \mathbf{e}_i \mathbf{e}_i^\top$$

  $\Lambda = \mathrm{diag}(\lambda_1, \ldots, \lambda_p)$ of positive eigenvalues

  $\mathbf{E} = \begin{pmatrix} \mathbf{e}_1 & \cdots & \mathbf{e}_p \end{pmatrix}$ orthonormal eigenvectors ($\mathbf{E}^\top \mathbf{E} = \mathbf{I}$) as columns
- $\mathcal{S}(p) :=$ the set of all $p \times p$ positive definite matrices.

Elliptically symmetric (ES) distribution

$\mathbf{x} \sim \mathcal{E}_p(\mathbf{0}, \boldsymbol{\Sigma}, g)$: p.d.f. is

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

- $\boldsymbol{\Sigma} \in \mathcal{S}(p)$, unknown positive definite $p \times p$ scatter matrix parameter.
- $g : \mathbb{R}_0^+ \to \mathbb{R}^+$, fixed density generator.

When the covariance matrix exists: $\mathbb{E}[\mathbf{x}\mathbf{x}^\top] = \boldsymbol{\Sigma}$.

Example: Normal distribution $\mathcal{N}_p(\mathbf{0}, \boldsymbol{\Sigma})$ has p.d.f.

$$f(\mathbf{x}) = \pi^{-p/2} |\boldsymbol{\Sigma}|^{-1/2} \exp\Big(-\frac{1}{2} \mathbf{x}^\top \boldsymbol{\Sigma}^{-1} \mathbf{x} \Big).$$

Elliptical distribution with $g(t) = \exp(-t/2)$.

The p-variate t-distribution with ν degrees of freedom

$\mathbf{x} \sim t_{p,\nu}(\mathbf{0}, \boldsymbol{\Sigma})$: pdf is

$$f(\mathbf{x}|\boldsymbol{\Sigma}) \propto |\boldsymbol{\Sigma}|^{-1}\left(1 + \mathbf{x}^\top \boldsymbol{\Sigma}^{-1}\mathbf{x}/\nu\right)^{-(p+\nu)/2}, \quad \nu > 0$$

so the density generator is $g_\nu(t) = (1 + t/\nu)^{-(p+\nu)/2}$

- $\nu > 0$ is the degrees of freedom parameter:
  - $\nu = 1$ is called the complex Cauchy distribution
  - $\nu \to \infty$ yields the p-variate normal distribution
  - finite 2nd-order moments for $\nu > 2$
- Stochastic decomposition:

$$\mathbf{x} =_d \sqrt{\tau}\mathbf{n}, \quad \tau^{-1} \sim \mathrm{Gam}(\nu/2, 2/\nu), \quad \mathbf{n} \sim \mathcal{N}_p(\mathbf{0}, \boldsymbol{\Sigma}), \quad \tau \perp\!\!\!\perp \mathbf{n}.$$

  This also provides a straightforward approach to generate a random sample from $t_{p,\nu}(\mathbf{0}, \boldsymbol{\Sigma})$.

The maximum likelihood estimator (MLE)

- $\{\mathbf{x}_i\} \overset{iid}{\sim} \mathcal{E}_p(\mathbf{0}, \mathbf{\Sigma}, g)$, where $n > p$.
- The MLE $\hat{\mathbf{\Sigma}} \in \mathcal{S}(p)$ minimizes the $(-2/n) \times$ log-likelihood fnc

$$\mathcal{L}(\mathbf{\Sigma}) = \frac{1}{n}\sum_{i=1}^{n} -2\ln f(\mathbf{x}_i|\mathbf{\Sigma}) = \frac{1}{n}\sum_{i=1}^{n} \rho(\mathbf{x}_i^\top \mathbf{\Sigma}^{-1}\mathbf{x}_i) - \ln|\mathbf{\Sigma}^{-1}|$$

  where $\rho(t) = -2\ln g(t)$ is the loss function.
- Critical points are solutions to estimating equations

$$\hat{\mathbf{\Sigma}} = \frac{1}{n}\sum_{i=1}^{n} u(\mathbf{x}_i^\top \hat{\mathbf{\Sigma}}^{-1}\mathbf{x}_i)\mathbf{x}_i\mathbf{x}_i^\top$$

  where $u(t) = \rho'(t)$ is the weight function.
- MLE = "an adaptively weighted sample covariance matrix"

Gaussian MLE

- P.d.f: $f(\mathbf{x}) \propto |\boldsymbol{\Sigma}|^{-1/2} \exp\left(-\frac{1}{2}\mathbf{x}^\top \boldsymbol{\Sigma}^{-1}\mathbf{x}\right)$ with density generator $g(t) = \exp(-t/2)$.
- The Gaussian loss function is $\rho_{\mathrm{G}}(t) = -2\ln g(t) = t$ and hence the $-(2/n)\times$ log-likelihood function is

$$\begin{aligned}
\mathcal{L}_{\mathrm{G}}(\boldsymbol{\Sigma}) &= \frac{1}{n}\sum_{i=1}^{n} \rho_{\mathrm{G}}(\mathbf{x}_i^\top \boldsymbol{\Sigma}^{-1}\mathbf{x}_i) - \ln|\boldsymbol{\Sigma}^{-1}| \\
&= \frac{1}{n}\sum_{i=1}^{n} \underbrace{\mathbf{x}_i^\top \boldsymbol{\Sigma}^{-1}\mathbf{x}_i}_{=\mathrm{Tr}\left(\boldsymbol{\Sigma}^{-1}\mathbf{x}_i\mathbf{x}_i^\top\right)} - \ln|\boldsymbol{\Sigma}^{-1}| \\
&= \mathrm{Tr}\left(\boldsymbol{\Sigma}^{-1}\mathbf{S}\right) - \ln|\boldsymbol{\Sigma}^{-1}|
\end{aligned}$$

where $\mathbf{S} = \frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i\mathbf{x}_i^\top$ is the SCM.

Gaussian MLE (cont'd)

- Taking the matrix derivative w.r.t $\boldsymbol{\Sigma}^{-1}$ yields

$$\frac{\partial}{\partial \boldsymbol{\Sigma}^{-1}} \mathcal{L}_{\mathrm{G}}(\boldsymbol{\Sigma}) = \frac{\partial}{\partial \boldsymbol{\Sigma}^{-1}} \left\{ \sum_{i=1}^{n} \mathrm{Tr}\big(\boldsymbol{\Sigma}^{-1}\mathbf{S}\big) - \ln |\boldsymbol{\Sigma}^{-1}| \right\}$$
$$= \mathbf{S} - \boldsymbol{\Sigma}$$

  where we used that $\frac{\partial}{\partial \mathbf{X}} \mathrm{Tr}(\mathbf{X}\mathbf{A}) = \mathbf{A}^\top$ and $\frac{\partial}{\partial \mathbf{X}} \ln |\mathbf{X}| = (\mathbf{X}^{-1})^\top$.

- Setting the derivative to $\mathbf{0}$ gives $\hat{\boldsymbol{\Sigma}} = \mathbf{S}$ as the critical point.

Q: Is the found solution, i.e., SCM $\mathbf{S}$, also the MLE?

- **Yes**. We have learned this fact *suprisingly recently* from [Watson, 1962] elegant proof.
- Other way to prove this is by showing that $\mathcal{L}_{\mathrm{G}}(\boldsymbol{\Sigma})$ is convex in $\boldsymbol{\Sigma}^{-1}$ (Note: $\mathcal{L}_{\mathrm{G}}(\boldsymbol{\Sigma})$ is non-convex in $\boldsymbol{\Sigma}$).
- Or: later we show that $\mathcal{L}_{\mathrm{G}}(\boldsymbol{\Sigma})$ is g-convex both in $\boldsymbol{\Sigma}$ and $\boldsymbol{\Sigma}^{-1}$.

Gaussian MLE (cont'd)

- Taking the matrix derivative w.r.t $\boldsymbol{\Sigma}^{-1}$ yields

$$\frac{\partial}{\partial \boldsymbol{\Sigma}^{-1}} \mathcal{L}_{\mathrm{G}}(\boldsymbol{\Sigma}) = \frac{\partial}{\partial \boldsymbol{\Sigma}^{-1}} \left\{ \sum_{i=1}^{n} \mathrm{Tr}\big(\boldsymbol{\Sigma}^{-1}\mathbf{S}\big) - \ln |\boldsymbol{\Sigma}^{-1}| \right\}$$
$$= \mathbf{S} - \boldsymbol{\Sigma}$$

  where we used that $\frac{\partial}{\partial \mathbf{X}} \mathrm{Tr}(\mathbf{X}\mathbf{A}) = \mathbf{A}^\top$ and $\frac{\partial}{\partial \mathbf{X}} \ln |\mathbf{X}| = (\mathbf{X}^{-1})^\top$.

- Setting the derivative to $\mathbf{0}$ gives $\hat{\boldsymbol{\Sigma}} = \mathbf{S}$ as the critical point.

Q: Is the found solution, i.e., SCM $\mathbf{S}$, also the MLE?

- **Yes**. We have learned this fact *suprisingly recently* from [Watson, 1962] elegant proof.
- Other way to prove this is by showing that $\mathcal{L}_{\mathrm{G}}(\boldsymbol{\Sigma})$ is convex in $\boldsymbol{\Sigma}^{-1}$ (Note: $\mathcal{L}_{\mathrm{G}}(\boldsymbol{\Sigma})$ is non-convex in $\boldsymbol{\Sigma}$).
- Or: later we show that $\mathcal{L}_{\mathrm{G}}(\boldsymbol{\Sigma})$ is g-convex both in $\boldsymbol{\Sigma}$ and $\boldsymbol{\Sigma}^{-1}$.

Gaussian MLE (cont'd)

- Taking the matrix derivative w.r.t $\boldsymbol{\Sigma}^{-1}$ yields

$$\frac{\partial}{\partial \boldsymbol{\Sigma}^{-1}} \mathcal{L}_{\mathrm{G}}(\boldsymbol{\Sigma}) = \frac{\partial}{\partial \boldsymbol{\Sigma}^{-1}} \left\{ \sum_{i=1}^{n} \mathrm{Tr}\big(\boldsymbol{\Sigma}^{-1}\mathbf{S}\big) - \ln |\boldsymbol{\Sigma}^{-1}| \right\}$$
$$= \mathbf{S} - \boldsymbol{\Sigma}$$

  where we used that $\frac{\partial}{\partial \mathbf{X}} \mathrm{Tr}(\mathbf{X}\mathbf{A}) = \mathbf{A}^\top$ and $\frac{\partial}{\partial \mathbf{X}} \ln |\mathbf{X}| = (\mathbf{X}^{-1})^\top$.

- Setting the derivative to $\mathbf{0}$ gives $\hat{\boldsymbol{\Sigma}} = \mathbf{S}$ as the critical point.

Q: Is the found solution, i.e., SCM $\mathbf{S}$, also the MLE?

- **Yes**. We have learned this fact *suprisingly recently* from [Watson, 1962] elegant proof.
- Other way to prove this is by showing that $\mathcal{L}_{\mathrm{G}}(\boldsymbol{\Sigma})$ is convex in $\boldsymbol{\Sigma}^{-1}$ (Note: $\mathcal{L}_{\mathrm{G}}(\boldsymbol{\Sigma})$ is non-convex in $\boldsymbol{\Sigma}$).
- Or: later we show that $\mathcal{L}_{\mathrm{G}}(\boldsymbol{\Sigma})$ is g-convex both in $\boldsymbol{\Sigma}$ and $\boldsymbol{\Sigma}^{-1}$.

Gaussian MLE (cont'd)

- Taking the matrix derivative w.r.t $\boldsymbol{\Sigma}^{-1}$ yields

$$\frac{\partial}{\partial \boldsymbol{\Sigma}^{-1}} \mathcal{L}_{\mathrm{G}}(\boldsymbol{\Sigma}) = \frac{\partial}{\partial \boldsymbol{\Sigma}^{-1}} \left\{ \sum_{i=1}^{n} \mathrm{Tr}\big(\boldsymbol{\Sigma}^{-1}\mathbf{S}\big) - \ln |\boldsymbol{\Sigma}^{-1}| \right\}$$
$$= \mathbf{S} - \boldsymbol{\Sigma}$$

  where we used that $\frac{\partial}{\partial \mathbf{X}} \mathrm{Tr}(\mathbf{X}\mathbf{A}) = \mathbf{A}^\top$ and $\frac{\partial}{\partial \mathbf{X}} \ln |\mathbf{X}| = (\mathbf{X}^{-1})^\top$.

- Setting the derivative to $\mathbf{0}$ gives $\hat{\boldsymbol{\Sigma}} = \mathbf{S}$ as the critical point.

Q: Is the found solution, i.e., SCM $\mathbf{S}$, also the MLE?

- **Yes**. We have learned this fact *suprisingly recently* from [Watson, 1962] elegant proof.
- Other way to prove this is by showing that $\mathcal{L}_{\mathrm{G}}(\boldsymbol{\Sigma})$ is convex in $\boldsymbol{\Sigma}^{-1}$ (Note: $\mathcal{L}_{\mathrm{G}}(\boldsymbol{\Sigma})$ is non-convex in $\boldsymbol{\Sigma}$).
- Or: later we show that $\mathcal{L}_{\mathrm{G}}(\boldsymbol{\Sigma})$ is g-convex both in $\boldsymbol{\Sigma}$ and $\boldsymbol{\Sigma}^{-1}$.

M-estimators of scatter matrix

$$\hat{\mathbf{\Sigma}} = \frac{1}{n}\sum_{i=1}^{n} u(\mathbf{x}_i^\top \hat{\mathbf{\Sigma}}^{-1}\mathbf{x}_i)\mathbf{x}_i\mathbf{x}_i^\top$$

[Maronna, 1976]

- Among the first proposals for robust covariance matrix estimators

Generalizations of ML-estimators:

- $u(t) = \rho'(t)$ non-neg., continuous and non-increasing.
  (admits more general ρ fnc's)
- $\psi(t) = tu(t)$ strictly increasing $\Rightarrow$ unique solution
- Not too much data lies in some sub-space $\Rightarrow$ solution exists

Popular choices of $u(t)$: Huber's, t-likelihood, Tyler's. ...

Huber's M-estimator

- [Maronna, 1976] defined it as an M-estimator with weight fnc

$$u_{\mathrm{H}}(t;c) = \begin{cases} 1/b, & \text{for } t \leqslant c^2 \\ c^2/(tb), & \text{for } t > c^2 \end{cases}$$

where $c > 0$ is a tuning constant, chosen by the user,and b is a scaling factor used to obtain Fisher consistency at $\mathcal{N}_p(\mathbf{0}, \boldsymbol{\Sigma})$:

$$b = F_{\chi^2_{p+2}}(c^2) + c^2(1 - F_{\chi^2_p}(c^2))/p.$$

Choose c^2 as qth upper quantile of χ^2_p: $c^2 = F^{-1}_{\chi^2_p}(q)$.

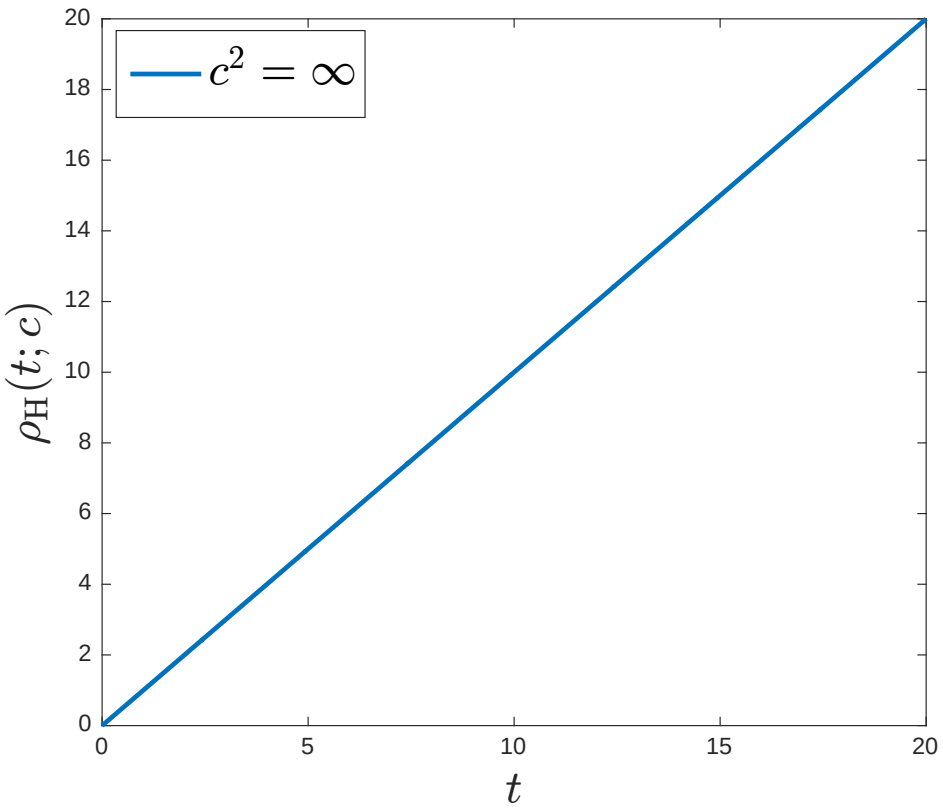
- It is also an MLE with loss function [Ollila et al., 2016]:

$$\rho_{\mathrm{H}}(t;c) = \begin{cases} t/b & \text{for } t \leqslant c^2, \\ (c^2/b)\big(\log(t/c^2) + 1\big) & \text{for } t > c^2. \end{cases}$$

Note: a Gaussian distribution in the middle, but have tails that die down at an inverse polynomial rate. Naturally, $u_{\mathrm{H}}(t;c) = \rho'_{\mathrm{H}}(t;c)$.

Huber's M-estimator

$$\underset{\mathbf{\Sigma}\in\mathcal{S}(p)}{\text{minimize}} \; \frac{1}{n}\sum_{i=1}^{n} \rho_{\text{H}}\Big(\underbrace{\mathbf{x}_i^\top \mathbf{\Sigma}^{-1}\mathbf{x}_i}_{=t}; c \Big) - \ln|\mathbf{\Sigma}^{-1}|,$$



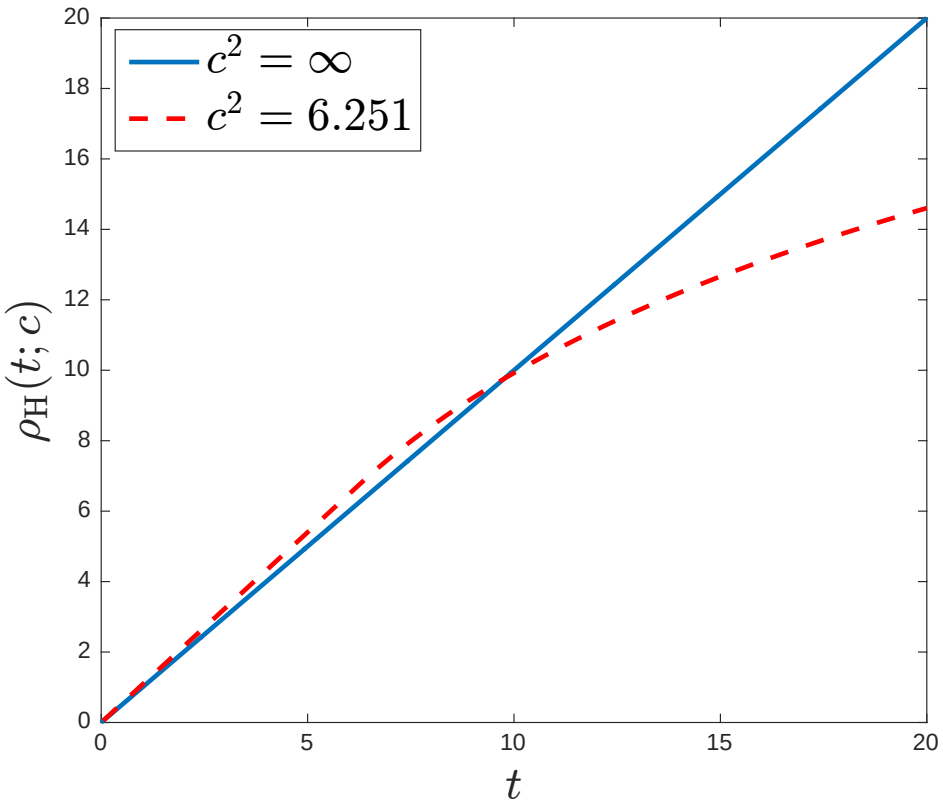
Dimension $p = 3$ and

$$c^2 = F^{-1}_{\chi^2_p}(q)$$

- $q = 1.0 : \; c^2 = \infty$
- $q = 0.9 : \; c^2 = 6.251$
- $q = 0.5 : \; c^2 = 2.366$
- $q = 0.1 : \; c^2 = 0.584$

Huber's M-estimator

$$\underset{\mathbf{\Sigma}\in\mathcal{S}(p)}{\text{minimize}} \; \frac{1}{n}\sum_{i=1}^{n} \rho_{\text{H}}\Big(\underbrace{\mathbf{x}_i^\top \mathbf{\Sigma}^{-1}\mathbf{x}_i}_{=t}; c \Big) - \ln|\mathbf{\Sigma}^{-1}|,$$



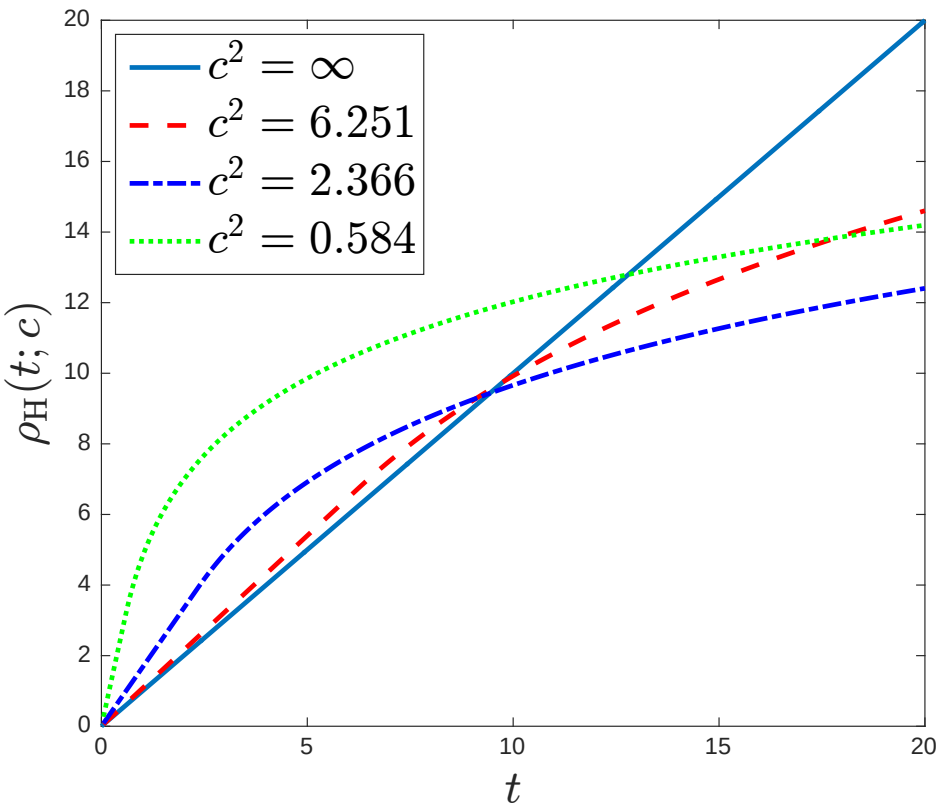
Dimension $p = 3$ and

$$c^2 = F^{-1}_{\chi^2_p}(q)$$

- $q = 1.0 : \; c^2 = \infty$
- $q = 0.9 : \; c^2 = 6.251$
- $q = 0.5 : \; c^2 = 2.366$
- $q = 0.1 : \; c^2 = 0.584$

Huber's M-estimator

$$\underset{\mathbf{\Sigma}\in\mathcal{S}(p)}{\text{minimize}} \; \frac{1}{n}\sum_{i=1}^{n} \rho_{\text{H}}\Big(\underbrace{\mathbf{x}_i^\top \mathbf{\Sigma}^{-1}\mathbf{x}_i}_{=t}; c \Big) - \ln|\mathbf{\Sigma}^{-1}|,$$



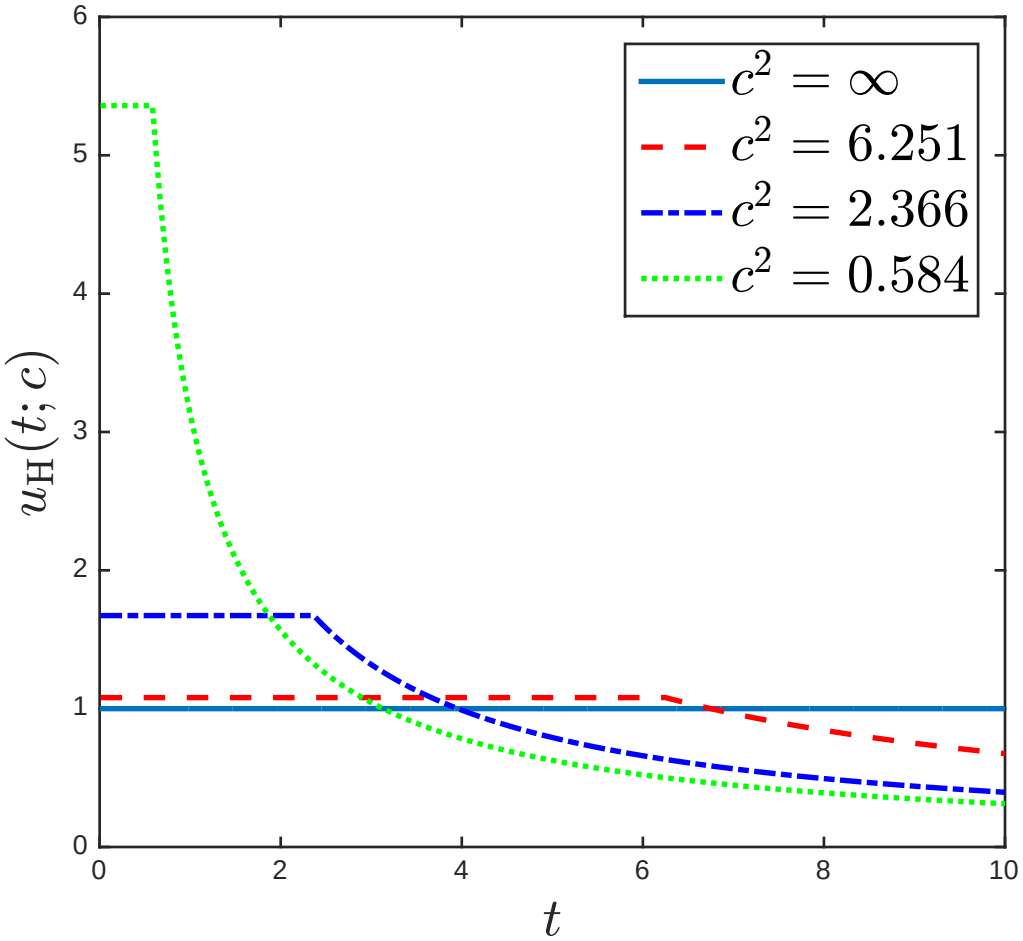
Dimension $p = 3$ and

$$c^2 = F^{-1}_{\chi^2_p}(q)$$

- $q = 1.0 : \; c^2 = \infty$
- $q = 0.9 : \; c^2 = 6.251$
- $q = 0.5 : \; c^2 = 2.366$
- $q = 0.1 : \; c^2 = 0.584$

Huber's M-estimator

solves $$\hat{\boldsymbol{\Sigma}} = \frac{1}{n}\sum_{i=1}^{n} u_{\mathrm{H}}\big(\underbrace{\mathbf{x}_i^\top \hat{\boldsymbol{\Sigma}}^{-1}\mathbf{x}_i}_{=t}; c\big)\mathbf{x}_i\mathbf{x}_i^\top$$



Dimension $p = 3$ and

$$c^2 = F^{-1}_{\chi^2_p}(q)$$

- $q = 1.0 : \; c^2 = \infty$
- $q = 0.9 : \; c^2 = 6.251$
- $q = 0.5 : \; c^2 = 2.366$
- $q = 0.1 : \; c^2 = 0.584$

The MLE of $t_{p,\nu}$-distribution

- Density generator of $t_{p,\nu}(\mathbf{0}, \mathbf{\Sigma})$ is $g_\nu(t) = (1 + t/\nu)^{-\frac{1}{2}(\nu+p)}$.
- M-estimator $\hat{\mathbf{\Sigma}}$ based on the respective loss and weight function

$$\rho_\nu(t) = -2 \ln g_\nu(t) = (\nu + p) \ln(1 + t/\nu)$$
$$u_\nu(t) = \rho'_\nu(t) = \frac{\nu + p}{\nu + t}$$

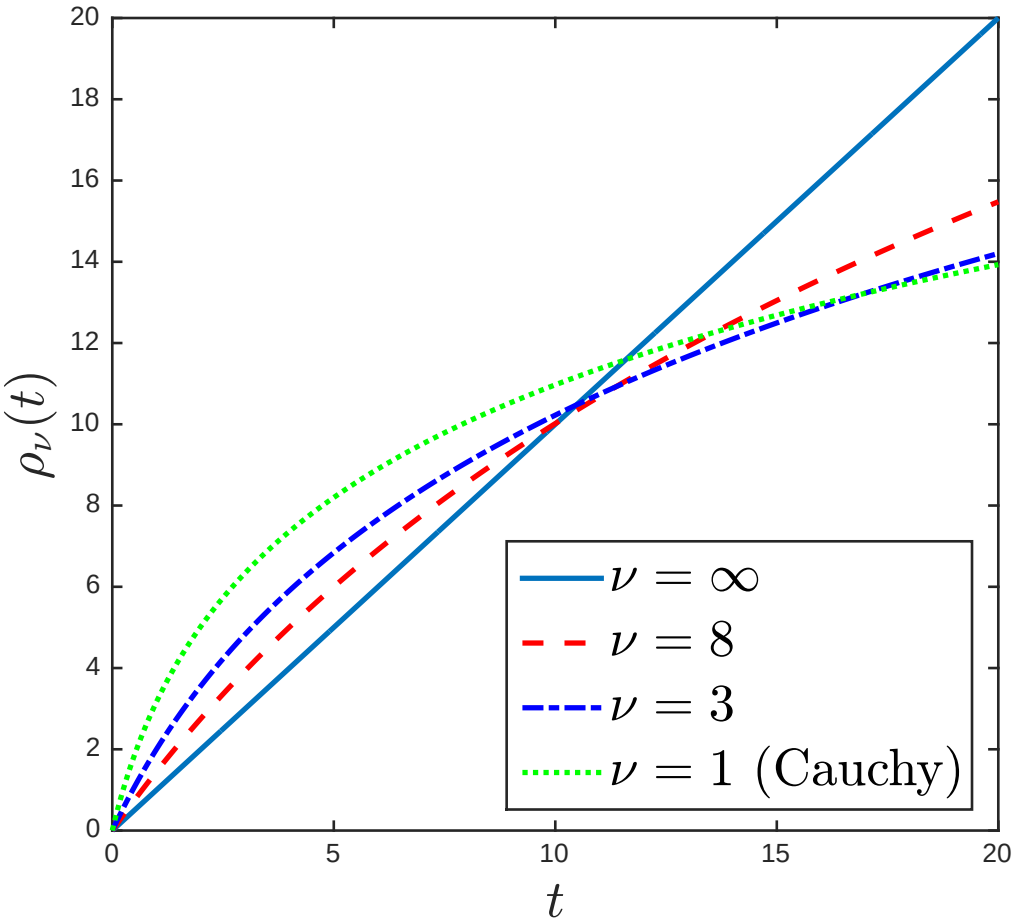
  is referred to as $t_\nu M$-estimator.
- Naturally an MLE when $\{\mathbf{x}_i\}_{i=1}^n \overset{iid}{\sim} t_{p,\nu}(\mathbf{0}, \mathbf{\Sigma})$.
- To obtain a Fisher consistent $t_\nu M$-estimator $\hat{\mathbf{\Sigma}}$ at $\mathcal{N}_p(\mathbf{0}, \mathbf{\Sigma})$, use a scaled loss function:

$$\rho^*_\nu(t) = \frac{1}{b}\rho_\nu(t), \quad b = \{(\nu + p)/p\}\mathbb{E}[\chi^2_p/(\nu + \chi^2_p)].$$

  and respective scaled weight function $u^*_\nu(t) = u_\nu(t)/b$,

$t_\nu M$-estimator

$$\underset{\mathbf{\Sigma}\in\mathcal{S}(p)}{\text{minimize}} \ \frac{1}{n}\sum_{i=1}^{n}\rho_\nu\Big(\underbrace{\mathbf{x}_i^\top\mathbf{\Sigma}^{-1}\mathbf{x}_i}_{=t}\Big) - \ln|\mathbf{\Sigma}^{-1}|,$$

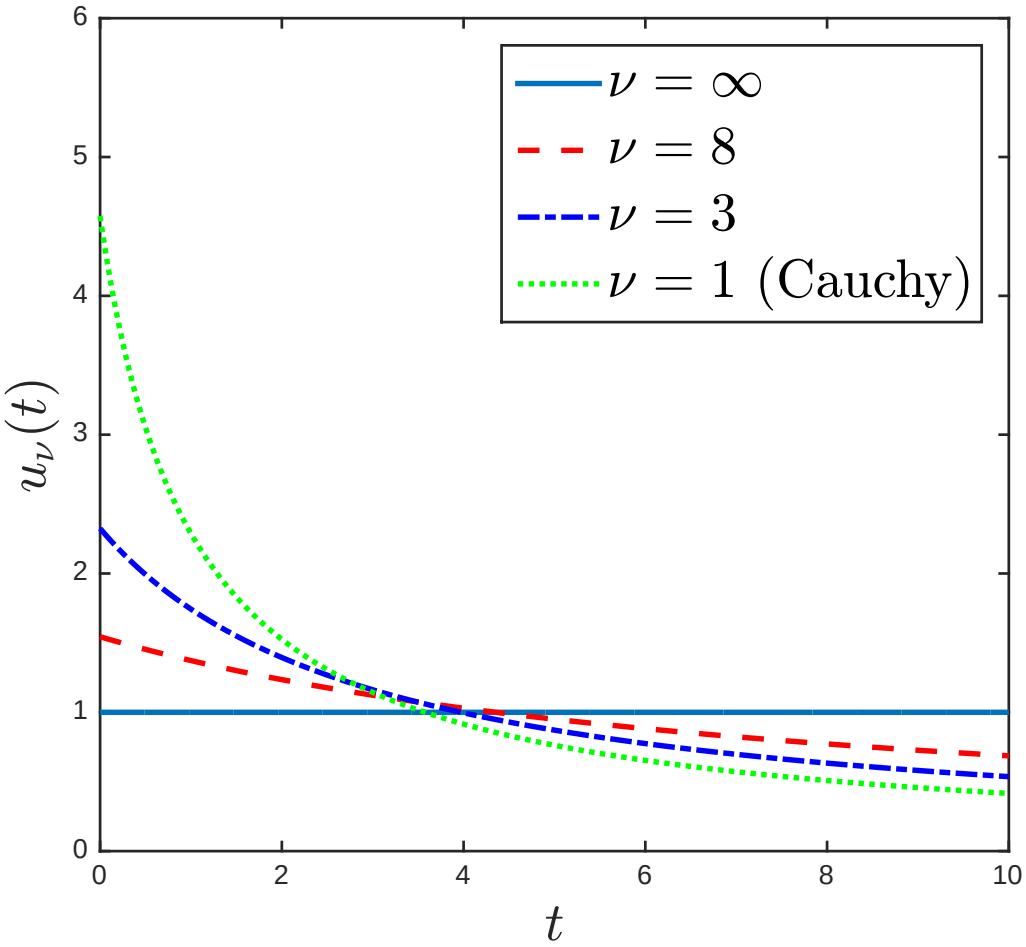


Dimension $p = 3$ and

$$\rho_\nu(t) = (\nu+p)\ln(1+t/\nu)$$

$t_\nu M$-estimator

solves $\hat{\mathbf{\Sigma}} = \frac{1}{n}\sum_{i=1}^{n} u_\nu\big(\underbrace{\mathbf{x}_i^\top \hat{\mathbf{\Sigma}}^{-1}\mathbf{x}_i}_{=t}\big)\mathbf{x}_i\mathbf{x}_i^\top$

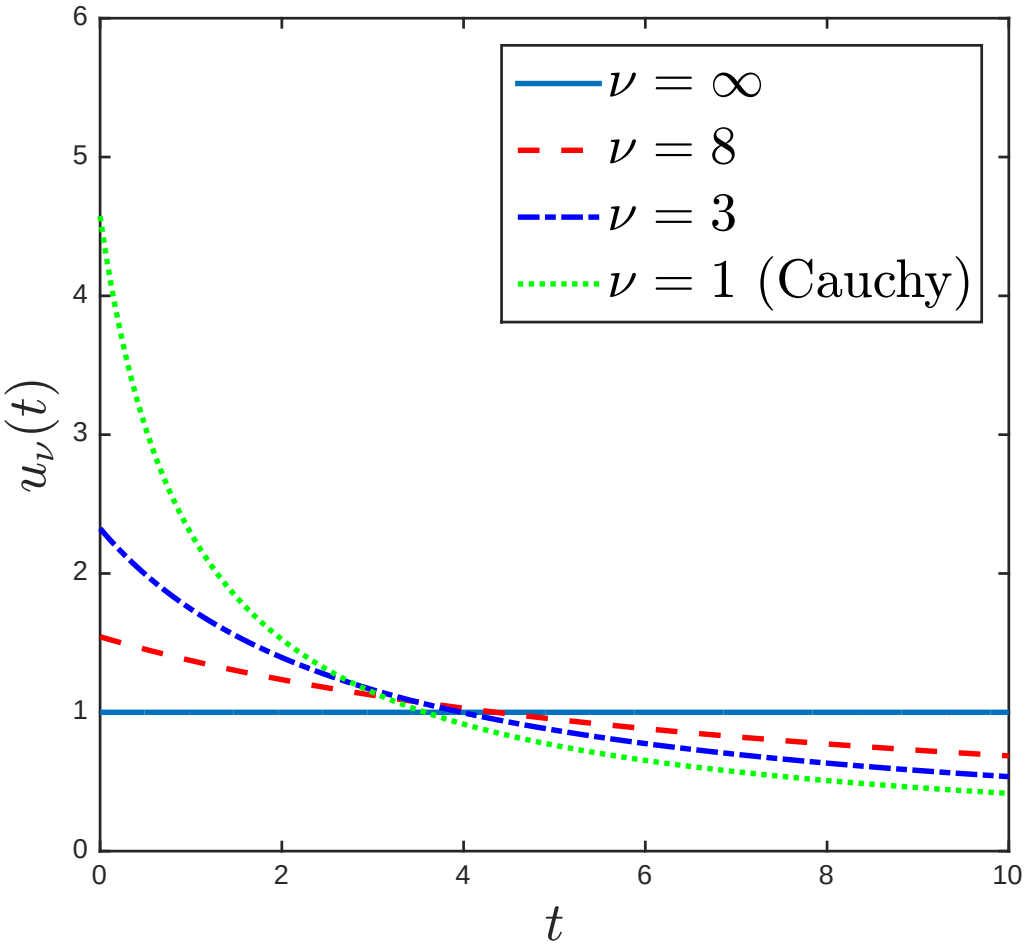


Dimension $p = 3$ and

$$u_\nu(t) = \frac{p+\nu}{\nu+t}$$

$t_\nu M$-estimator

solves $\hat{\mathbf{\Sigma}} = \frac{1}{n}\sum_{i=1}^{n} \frac{(p+\nu)\mathbf{x}_i\mathbf{x}_i^\top}{\nu + \mathbf{x}_i^\top \hat{\mathbf{\Sigma}}^{-1}\mathbf{x}_i}$



Dimension $p = 3$ and

$$u_\nu(t) = \frac{p+\nu}{\nu + t}$$

Tyler's M-estimator

- Distribution-free M-estimator (under elliptical distributions) proposed in [Tyler, 1987].
- Defined as a solution to

$$\hat{\mathbf{\Sigma}} = \frac{p}{n} \sum_{i=1}^{n} \frac{\mathbf{x}_i \mathbf{x}_i^\top}{\mathbf{x}_i^\top \hat{\mathbf{\Sigma}}^{-1} \mathbf{x}_i}$$

  $\Rightarrow$ so an M-estimator with Tyler's weight fnc $u(t) = \rho'(t) = p/t$
- Now it is also known that $\hat{\mathbf{\Sigma}} \in \mathcal{S}(p)$ minimizes the cost fnc

$$\mathcal{L}_{\mathrm{T}}(\mathbf{\Sigma}) = \frac{1}{n} \sum_{i=1}^{n} \underbrace{p \ln(\mathbf{x}_i^\top \mathbf{\Sigma}^{-1} \mathbf{x}_i)}_{\rho(t) = p \ln t} - \ln |\mathbf{\Sigma}^{-1}|$$

  Note: not an MLE for any elliptical density, so $\rho(t) \neq -2 \ln g(t)$!
- Not convex in $\mathbf{\Sigma}$! ... or in $\mathbf{\Sigma}^{-1}$
- Maronna's/Huber's conditions does not apply.

Tyler's M-estimator

- Distribution-free M-estimator (under elliptical distributions) proposed in [Tyler, 1987].
- Defined as a solution to

$$\hat{\mathbf{\Sigma}} = \frac{p}{n} \sum_{i=1}^{n} \frac{\mathbf{x}_i \mathbf{x}_i^\top}{\mathbf{x}_i^\top \hat{\mathbf{\Sigma}}^{-1} \mathbf{x}_i}$$

  $\Rightarrow$ so an M-estimator with Tyler's weight fnc $u(t) = \rho'(t) = p/t$
- Now it is also known that $\hat{\mathbf{\Sigma}} \in \mathcal{S}(p)$ minimizes the cost fnc

$$\mathcal{L}_{\mathrm{T}}(\mathbf{\Sigma}) = \frac{1}{n} \sum_{i=1}^{n} \underbrace{p \ln(\mathbf{x}_i^\top \mathbf{\Sigma}^{-1} \mathbf{x}_i)}_{\rho(t) = p \ln t} - \ln |\mathbf{\Sigma}^{-1}|$$

  Note: not an MLE for any elliptical density, so $\rho(t) \neq -2 \ln g(t)$!
- Not convex in $\mathbf{\Sigma}$! ... or in $\mathbf{\Sigma}^{-1}$
- Maronna's/Huber's conditions does not apply.

Tyler's M-estimator

- Distribution-free M-estimator (under elliptical distributions) proposed in [Tyler, 1987].
- Defined as a solution to

$$\hat{\mathbf{\Sigma}} = \frac{p}{n} \sum_{i=1}^{n} \frac{\mathbf{x}_i \mathbf{x}_i^\top}{\mathbf{x}_i^\top \hat{\mathbf{\Sigma}}^{-1} \mathbf{x}_i}$$

  $\Rightarrow$ so an M-estimator with Tyler's weight fnc $u(t) = \rho'(t) = p/t$
- Now it is also known that $\hat{\mathbf{\Sigma}} \in \mathcal{S}(p)$ minimizes the cost fnc

$$\mathcal{L}_{\mathrm{T}}(\mathbf{\Sigma}) = \frac{1}{n} \sum_{i=1}^{n} \underbrace{p \ln(\mathbf{x}_i^\top \mathbf{\Sigma}^{-1} \mathbf{x}_i)}_{\rho(t) = p \ln t} - \ln |\mathbf{\Sigma}^{-1}|$$

  Note: not an MLE for any elliptical density, so $\rho(t) \neq -2 \ln g(t)$!
- Not convex in $\mathbf{\Sigma}$! ... or in $\mathbf{\Sigma}^{-1}$
- Maronna's/Huber's conditions does not apply.

Tyler's M-estimator, cont'd

$$\hat{\mathbf{\Sigma}} = \frac{p}{n}\sum_{i=1}^{n}\frac{\mathbf{x}_i\mathbf{x}_i^\top}{\mathbf{x}_i^\top\hat{\mathbf{\Sigma}}^{-1}\mathbf{x}_i}$$

Comments:

1. Limiting case of Huber's M-estimator when $c \to 0$ and of $t_\nu M$-estimator when $\nu \to 0$.
2. Minimum is a unique *up to a positive scalar*: if $\hat{\mathbf{\Sigma}}$ is a minimum then so is $b\hat{\mathbf{\Sigma}}$ for any $b > 0$

$\Rightarrow$ $\hat{\mathbf{\Sigma}}$ is a shape matrix estimator. We may choose a solution which verifies $|\hat{\mathbf{\Sigma}}| = 1$.

3. A Fisher consistent estimator at $\mathcal{N}_p(\mathbf{0}, \mathbf{\Sigma})$ can be obtained by scaling any minimum $\hat{\mathbf{\Sigma}}$ by

$$b = \text{Median}\{\mathbf{x}_i^\top\hat{\mathbf{\Sigma}}^{-1}\mathbf{x}_i; i = 1, \ldots, n\}/\text{Median}(\chi^2_p).$$

This scaling is utilized in discriminant analysis later on.

Tyler's M-estimator, cont'd

$$\hat{\mathbf{\Sigma}} = \frac{p}{n}\sum_{i=1}^{n}\frac{\mathbf{x}_i\mathbf{x}_i^\top}{\mathbf{x}_i^\top\hat{\mathbf{\Sigma}}^{-1}\mathbf{x}_i}$$

Comments:

1. Limiting case of Huber's M-estimator when $c \to 0$ and of $t_\nu M$-estimator when $\nu \to 0$.
2. Minimum is a unique *up to a positive scalar*: if $\hat{\mathbf{\Sigma}}$ is a minimum then so is $b\hat{\mathbf{\Sigma}}$ for any $b > 0$

$\Rightarrow$ $\hat{\mathbf{\Sigma}}$ is a shape matrix estimator. We may choose a solution which verifies $|\hat{\mathbf{\Sigma}}| = 1$.

3. A Fisher consistent estimator at $\mathcal{N}_p(\mathbf{0}, \mathbf{\Sigma})$ can be obtained by scaling any minimum $\hat{\mathbf{\Sigma}}$ by

$$b = \text{Median}\{\mathbf{x}_i^\top\hat{\mathbf{\Sigma}}^{-1}\mathbf{x}_i; i = 1, \ldots, n\}/\text{Median}(\chi^2_p).$$

This scaling is utilized in discriminant analysis later on.

Tyler's M-estimator, cont'd

$$c\hat{\mathbf{\Sigma}} = \frac{p}{n}\sum_{i=1}^{n}\frac{\mathbf{x}_i\mathbf{x}_i^\top}{\mathbf{x}_i^\top(c\hat{\mathbf{\Sigma}})^{-1}\mathbf{x}_i}$$

Comments:

1. Limiting case of Huber's M-estimator when $c \to 0$ and of $t_\nu M$-estimator when $\nu \to 0$.
2. Minimum is a unique *up to a positive scalar*: if $\hat{\mathbf{\Sigma}}$ is a minimum then so is $b\hat{\mathbf{\Sigma}}$ for any $b > 0$

$\Rightarrow$ $\hat{\mathbf{\Sigma}}$ is a shape matrix estimator. We may choose a solution which verifies $|\hat{\mathbf{\Sigma}}| = 1$.

3. A Fisher consistent estimator at $\mathcal{N}_p(\mathbf{0}, \mathbf{\Sigma})$ can be obtained by scaling any minimum $\hat{\mathbf{\Sigma}}$ by

$$b = \text{Median}\{\mathbf{x}_i^\top \hat{\mathbf{\Sigma}}^{-1}\mathbf{x}_i; i = 1, \ldots, n\}/\text{Median}(\chi^2_p).$$

This scaling is utilized in discriminant analysis later on.

Tyler's M-estimator, cont'd

$$\hat{\mathbf{\Sigma}} = \frac{p}{n}\sum_{i=1}^{n}\frac{\mathbf{x}_i\mathbf{x}_i^\top}{\mathbf{x}_i^\top\hat{\mathbf{\Sigma}}^{-1}\mathbf{x}_i}$$

Comments:

1. Limiting case of Huber's M-estimator when $c \to 0$ and of $t_\nu M$-estimator when $\nu \to 0$.
2. Minimum is a unique *up to a positive scalar*: if $\hat{\mathbf{\Sigma}}$ is a minimum then so is $b\hat{\mathbf{\Sigma}}$ for any $b > 0$

$\Rightarrow$ $\hat{\mathbf{\Sigma}}$ is a shape matrix estimator. We may choose a solution which verifies $|\hat{\mathbf{\Sigma}}| = 1$.

3. A Fisher consistent estimator at $\mathcal{N}_p(\mathbf{0}, \mathbf{\Sigma})$ can be obtained by scaling any minimum $\hat{\mathbf{\Sigma}}$ by

$$b = \text{Median}\{\mathbf{x}_i^\top\hat{\mathbf{\Sigma}}^{-1}\mathbf{x}_i; i = 1, \ldots, n\}/\text{Median}(\chi^2_p).$$

This scaling is utilized in discriminant analysis later on.

Tyler's M-estimator as MLE

Tyler's M-estimator $\hat{\mathbf{\Sigma}}$ is an MLE of $\mathbf{\Sigma}$ in the following 3 cases

Case I [Kent, 1997]

$\{\mathbf{x}_i\}_{i=1}^n$ i.i.d. from an angular central Gaussian (ACG) distribution, $\mathbf{x} \sim \mathcal{AN}_p(\mathbf{0}, \mathbf{\Sigma})$, whose p.d.f. is of the form

$$f(\mathbf{x}) \propto |\mathbf{\Sigma}|^{-1/2}(\mathbf{x}^\top \mathbf{\Sigma}^{-1}\mathbf{x})^{-p/2}, \quad \|\mathbf{x}\| = 1$$

(Note: $\mathbf{\Sigma}$ is identifiable up to a scale).

- Stochastic decomposition:

$$\mathbf{x} =_d \frac{\mathbf{n}}{\|\mathbf{n}\|}, \quad \mathbf{n} \sim \mathcal{N}_p(\mathbf{0}, \mathbf{\Sigma})$$

- The term "angular central Gaussian" is a slight misnomer: $\mathcal{N}_p(\mathbf{0}, \mathbf{\Sigma})$ can be replaced by *any* $\mathcal{E}_p(\mathbf{0}, \mathbf{\Sigma}, g)$.

Tyler's M-estimator as MLE

Tyler's M-estimator $\hat{\mathbf{\Sigma}}$ is an MLE of $\mathbf{\Sigma}$ in the following 3 cases

Case I [Kent, 1997]

$\{\mathbf{x}_i\}_{i=1}^n$ i.i.d. from an angular central Gaussian (ACG) distribution, $\mathbf{x} \sim \mathcal{AN}_p(\mathbf{0}, \mathbf{\Sigma})$, whose p.d.f. is of the form

$$f(\mathbf{x}) \propto |\mathbf{\Sigma}|^{-1/2}(\mathbf{x}^\top \mathbf{\Sigma}^{-1}\mathbf{x})^{-p/2}, \quad \|\mathbf{x}\| = 1$$

(Note: $\mathbf{\Sigma}$ is identifiable up to a scale).

- Stochastic decomposition:

$$\mathbf{x} =_d \frac{\mathbf{n}}{\|\mathbf{n}\|}, \quad \mathbf{n} \sim \mathcal{E}_p(\mathbf{0}, \mathbf{\Sigma}, g)$$

- The term "angular central Gaussian" is a slight misnomer: $\mathcal{N}_p(\mathbf{0}, \mathbf{\Sigma})$ can be replaced by *any* $\mathcal{E}_p(\mathbf{0}, \mathbf{\Sigma}, g)$.

Case II [Gini and Greco, 2002, Conte et al., 2002]

- $\{\mathbf{x}_i\}_{i=1}^n$ independent random vectors with Gaussian distributions of proportional covariance matrices, so

$$\mathbf{x}_i \sim \mathcal{N}_p(\mathbf{0}, \eta_i \boldsymbol{\Sigma}), \quad \eta_i > 0, \quad \mathrm{Tr}(\boldsymbol{\Sigma}) = p$$

where $\{\eta_i\}_{i=1}^n$ and $\boldsymbol{\Sigma} \in \mathcal{S}_p$ (with $\mathrm{Tr}(\boldsymbol{\Sigma}) = p$) are the unknown parameters.

Case III [Ollila and Tyler, 2012] is a generalization of Case II

- $\{\mathbf{x}_i\}_{i=1}^n$ independent from (possibly different) elliptical distributions of proportional covariance matrices, so

$$\mathbf{x}_i \sim \mathcal{E}_p(\mathbf{0}, \eta_i \boldsymbol{\Sigma}, g_i), \quad \eta_i > 0, \quad \mathrm{Tr}(\boldsymbol{\Sigma}) = p$$

- **Note**: density generator (need not be specified or known) and can be different for each $\mathbf{x}_i$, e.g., $\mathbf{x}_1 \sim \mathcal{N}_p(\mathbf{0}, \eta_1 \boldsymbol{\Sigma})$, $\mathbf{x}_2 \sim t_{p,\nu}(\mathbf{0}, \eta_2 \boldsymbol{\Sigma})$, ...

Algorithm

Given an arbitrary initial start $\mathbf{\Sigma}_0 \in \mathcal{S}(p)$, the iterations

$$\hat{\mathbf{\Sigma}}_{k+1} = \frac{1}{n}\sum_{i=1}^{n} u(\mathbf{x}_i^\top \hat{\mathbf{\Sigma}}_k^{-1}\mathbf{x}_i)\mathbf{x}_i\mathbf{x}_i^\top \quad \text{for } k = 0, 1, \ldots$$

converge to the solution

$$\hat{\mathbf{\Sigma}} = \frac{1}{n}\sum_{i=1}^{n} u(\mathbf{x}_i^\top \hat{\mathbf{\Sigma}}^{-1}\mathbf{x}_i)\mathbf{x}_i\mathbf{x}_i^\top$$

under some mild regularity conditions [$\psi(t) = tu(t)$ strictly increasing, not too much data lies in some sub-space]. See details in [Ollila et al., 2012].

Matlab implementation: $t_\nu M$-estimator

```
function C = tMest(X, v)
[n, p] = size(X);
MAX_ITER = 1000; % Max number of iteration
EPS = 1.0e-5;       % Iteration accuracy
C0 = X'*X/n;        % SCM as initial estimate
invC0 = C0 \ eye(p);
iter=1;
while (iter<MAX_ITER)
      t = sum(X*invC0.*X,2);
      u = (v+p)./(v+t);
      C = X'*((X.*repmat(u,1,p)))/n;
      d = norm(eye(p)-invC0*C,Inf);
      if (d<=EPS),
         break;
      end
      invC0 = C \ eye(p);
      iter = iter+1;
end
```

Effect of an outlier

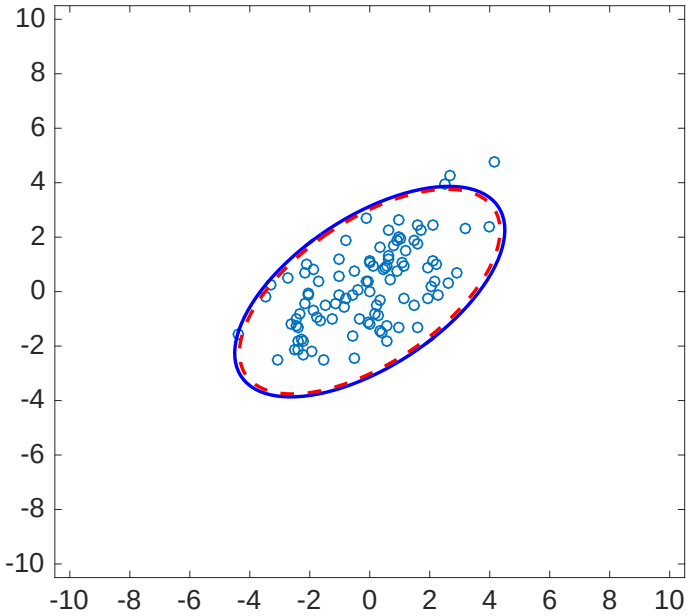
$\{\mathbf{x}_i\}_{i=1}^n \overset{iid}{\sim} \mathcal{N}_2(\mathbf{0}, \mathbf{\Sigma})$, $n = 100$ and $\mathbf{\Sigma} = \mathbf{E}\mathbf{\Lambda}\mathbf{E}^\top$, where

$$\mathbf{E} = \begin{pmatrix} \cos(\frac{\pi}{4}) & -\sin(\frac{\pi}{4}) \\ \sin(\frac{\pi}{4}) & \cos(\frac{\pi}{4}) \end{pmatrix}, \quad \mathbf{\Lambda} = \mathrm{diag}(5, 1)$$

Plot the 95% tolerance ellipses

$$\{x \in \mathbb{R}^2 : \mathbf{x}^\top \hat{\mathbf{\Sigma}}^{-1} \mathbf{x} = F^{-1}_{\chi^2_2}(0.95)\}$$

based on an estimated covariance matrix $\hat{\mathbf{\Sigma}}$.



- A single outlier in *south-east*, gradually increasing in magnitude
- t_3M -estimator (blue ellipse), computed using `C= tMest(X,3)`
- Sample covariance matrix (red ellipse)

Menu

I. Ad-hoc shrinkage SCM-s of multiple samples

II. ML- and M-estimators of scatter matrix

III. Geodesic convexity

- Geodesic
- g-convex functions

IV. Regularized M-estimators

V. Penalized estimation of multiple covariances

VI. Estimation of the regularization parameter

VII. Applications

References


Wiesel, A. (2012a).
Geodesic convexity and covariance estimation.
IEEE Trans. Signal Process., 60(12):6182–6189.


Zhang, T., Wiesel, A., and Greco, M. S. (2013).
Multivariate generalized Gaussian distribution: Convexity and graphical models.
IEEE Trans. Signal Process., 61(16):4141–4148.


Bhatia, R. (2009).
Positive definite matrices.
Princeton University Press.

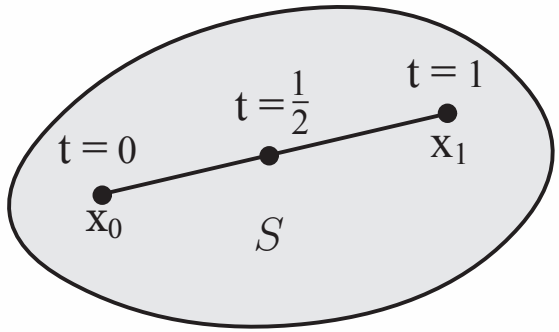
From Euclidean convexity to Riemannian convexity

A set S is convex ...

... if $\forall \mathbf{x}_0, \mathbf{x}_1 \in S$ and $t \in [0,1]$:

$$(1-t)\mathbf{x}_0 + t\mathbf{x}_1 \in S.$$

convex



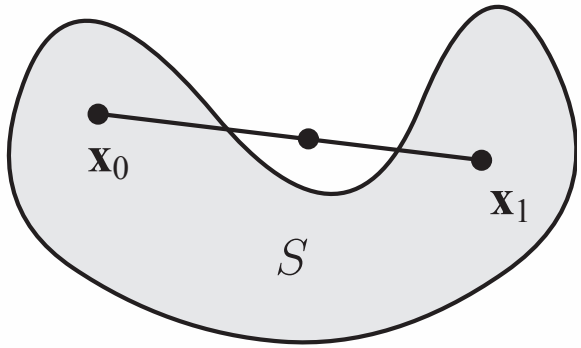
From Euclidean convexity to Riemannian convexity

A set S is convex ...

... if $\forall \mathbf{x}_0, \mathbf{x}_1 \in S$ and $t \in [0,1]$:

$$(1-t)\mathbf{x}_0 + t\mathbf{x}_1 \in S.$$

nonconvex



From Euclidean convexity to Riemannian convexity

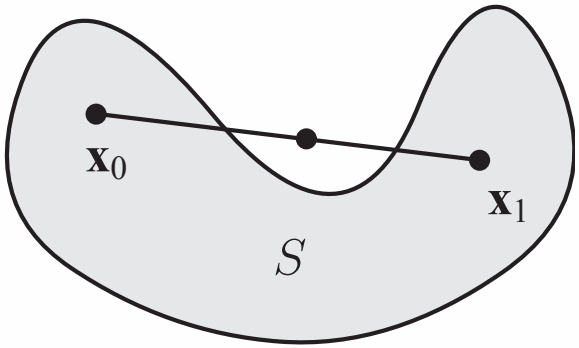
A set S is convex ...

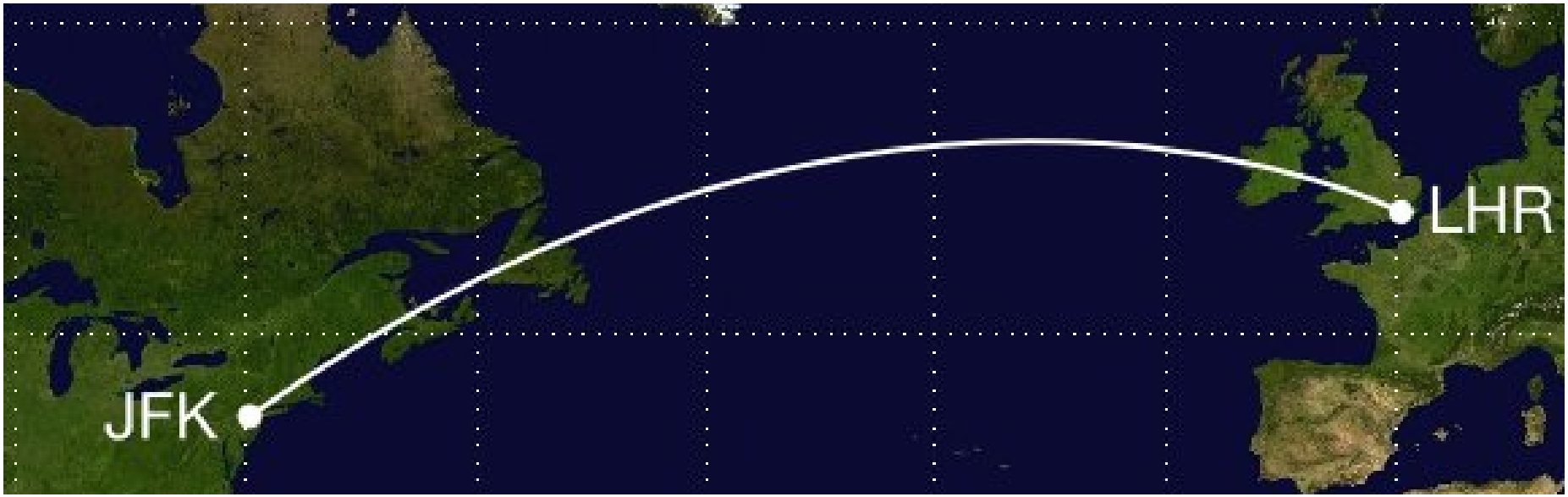
... if $\forall \mathbf{x}_0, \mathbf{x}_1 \in S$ and $t \in [0,1]$:

$$(1-t)\mathbf{x}_0 + t\mathbf{x}_1 \in S.$$

... if together with $\mathbf{x}_0$ and $\mathbf{x}_1$, it contains the shortest path (goedesic) connecting them

nonconvex



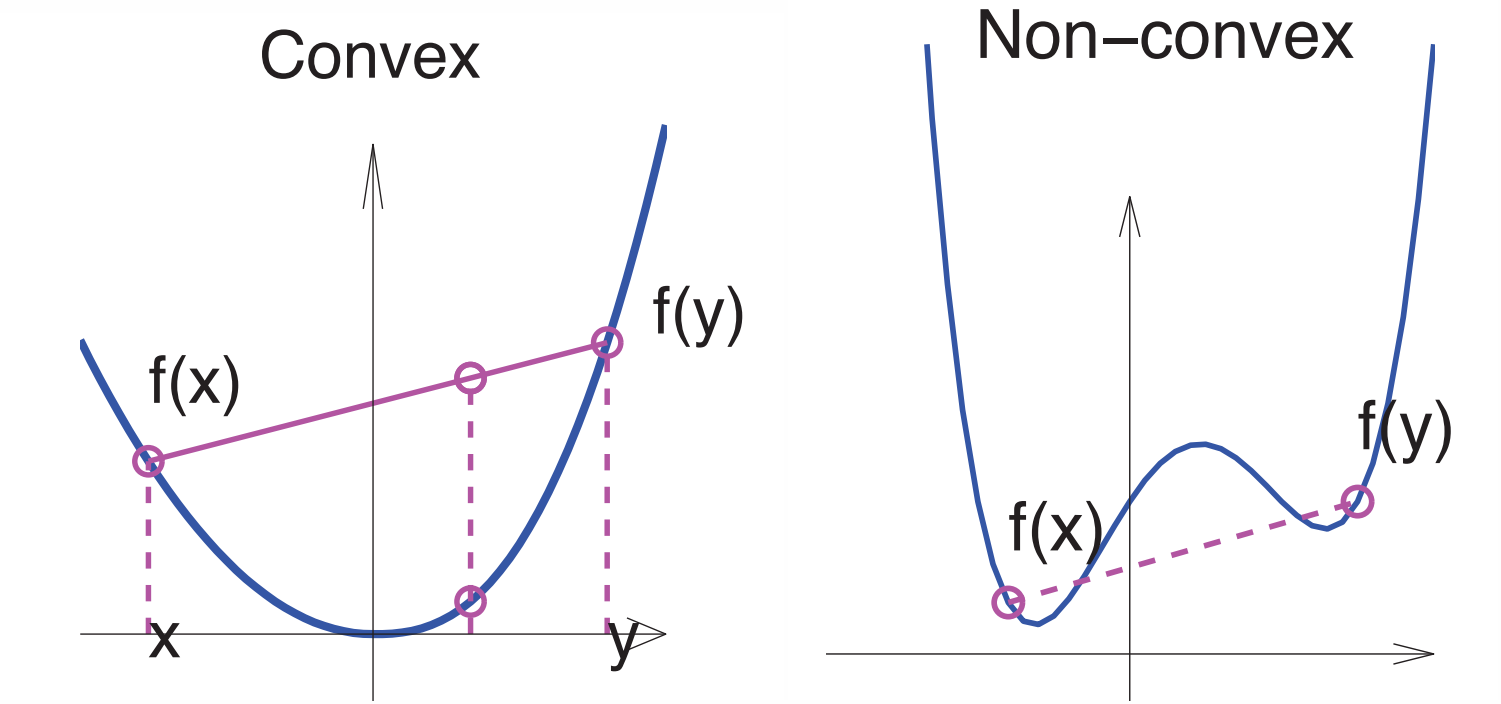


Euclidean convexity

Convex function in $\mathbf{x} \in S$, where S is a convex set

$$f\big((1-t)\mathbf{x}_0 + t\mathbf{x}_1\big) \leq (1-t)f(\mathbf{x}_0) + tf(\mathbf{x}_1), \quad \forall t \in [0,1]$$

strictly convex: strict inequality $(<)$ holds $\forall t \in (0,1)$. concave: $-f$ is convex.



Convex function f: the line segment between any two points on the graph of the function f lies above the graph.

Basic examples

Convex:

- x^2 (strictly) , $|x|$, $1/x$ $(x > 0)$
- e^x, $x \log x$
- log-sum-exp function: $\log(\sum_{i=1}^n e^{x_i})$.
- $\|\mathbf{x}\|_p$ (any norm, so $p \geq 1$)
- $\mathbf{x}^\top \mathbf{\Sigma} \mathbf{x}$ $(\mathbf{\Sigma} \in \mathcal{S}(p))$
- $\mathbf{a}^\top \mathbf{x} + b$
- $\lambda_{\max}(\mathbf{\Sigma})$ $(\mathbf{\Sigma} \in \mathcal{S}(p))$

Concave:

- $x^{1/2}$, $|x|^p$ $(0 \leq p \leq 1)$
- $\log x$
- log determinant: $\log |\mathbf{\Sigma}|$ $(\mathbf{\Sigma} \in \mathcal{S}(p))$
- $\min(x_1, \ldots, x_n)$
- $\lambda_{\min}(\mathbf{\Sigma})$ $(\mathbf{\Sigma} \in \mathbf{S}(p))$
- $\mathbf{a}^\top \mathbf{x} + b$

Geodesic convexity in $p = 1$ variable

convex function in $x \in \mathbb{R}$:

$$f\Big(\underbrace{(1-t)x_0 + tx_1}_{\text{line}}\Big) \leq (1-t)f(x_0) + tf(x_1)$$

g-convex function in $\sigma^2 \in \mathbb{R}_0^+$:

$$\rho\Big(\underbrace{\left(\sigma_0^2\right)^{(1-t)}\left(\sigma_1^2\right)^t}_{\text{geodesic}}\Big) \leq (1-t)\rho(\sigma_0^2) + t\rho(\sigma_1^2)$$

- Convex in $x = \log \sigma^2$ w.r.t. $(1-t)x_0 + tx_1$ is equivalent to g-convex in σ^2 w.r.t. $\sigma_t^2 = \left(\sigma_0^2\right)^{(1-t)}\left(\sigma_1^2\right)^t$.
- But for $\Sigma \in \mathcal{S}(p)$, $p \neq 1$, the solution is **not** a simple change of variables.

Geodesic convexity in $p = 1$ variable

convex function in $x \in \mathbb{R}$: $f(x) = \rho(e^x)$, $x = \log(\sigma^2)$

$$f\Big(\underbrace{(1-t)x_0 + tx_1}_{\text{line}}\Big) \leq (1-t)f(x_0) + tf(x_1)$$

g-convex function in $\sigma^2 \in \mathbb{R}_0^+$: $\rho(\sigma^2) = f(\log \sigma^2)$, $\sigma^2 = e^x$

$$\rho\Big(\underbrace{\left(\sigma_0^2\right)^{(1-t)}\left(\sigma_1^2\right)^t}_{\text{geodesic}}\Big) \leq (1-t)\rho(\sigma_0^2) + t\rho(\sigma_1^2)$$

- Convex in $x = \log \sigma^2$ w.r.t. $(1-t)x_0 + tx_1$ is equivalent to g-convex in σ^2 w.r.t. $\sigma_t^2 = \left(\sigma_0^2\right)^{(1-t)}\left(\sigma_1^2\right)^t$.
- But for $\Sigma \in \mathcal{S}(p)$, $p \neq 1$, the solution is **not** a simple change of variables.

Geodesic convexity in $p = 1$ variable

convex function in $x \in \mathbb{R}$:

$$f\big(\underbrace{(1-t)x_0 + tx_1}_{\text{line}}\big) \leq (1-t)f(x_0) + tf(x_1)$$

g-convex function in $\sigma^2 \in \mathbb{R}_0^+$:

$$\rho\big(\underbrace{(\sigma_0^2)^{(1-t)}(\sigma_1^2)^t}_{\text{geodesic}}\big) \leq (1-t)\rho(\sigma_0^2) + t\rho(\sigma_1^2)$$

- Convex in $x = \log \sigma^2$ w.r.t. $(1-t)x_0 + tx_1$ is equivalent to g-convex in σ^2 w.r.t. $\sigma_t^2 = (\sigma_0^2)^{(1-t)}(\sigma_1^2)^t$.
- But for $\mathbf{\Sigma} \in \mathcal{S}(p)$, $p \neq 1$, the solution is **not** a simple change of variables.

Geodesic ($g-$)convexity

On the Riemannian manifold of positive definite matrices, the

geodesic (shortest) path from $\boldsymbol{\Sigma}_0 \in \mathcal{S}(p)$ to $\boldsymbol{\Sigma}_1 \in \mathcal{S}(p)$ is

$$\boldsymbol{\Sigma}_t = \boldsymbol{\Sigma}_0^{1/2} \left(\boldsymbol{\Sigma}_0^{-1/2} \boldsymbol{\Sigma}_1 \boldsymbol{\Sigma}_0^{-1/2} \right)^t \boldsymbol{\Sigma}_0^{1/2} \text{ for } t \in [0,1].$$

where $\boldsymbol{\Sigma}_t \in \mathcal{S}(p)$ for $0 \leq t \leq 1 \Rightarrow \mathcal{S}(p)$ forms a g-convex set (= all geodesic paths $\boldsymbol{\Sigma}_t$ lie in $\mathcal{S}(p)$).

- Main idea: change the parametric path going from $\boldsymbol{\Sigma}_0$ to $\boldsymbol{\Sigma}_1$.
- Midpoint of the path, $\boldsymbol{\Sigma}_{1/2}$:= Riemannian (geometric) mean between $\boldsymbol{\Sigma}_0$ and $\boldsymbol{\Sigma}_1$.
- For $p = 1$, the path is $\sigma_t^2 = (\sigma_0^2)^{1-t}(\sigma_1^2)^t$ and the midpoint is the geometric mean

$$\sigma_{1/2}^2 = \sqrt{\sigma_0^2 \sigma_1^2} = \exp\left\{ \frac{1}{2} \left[\ln(\sigma_0^2) + \ln(\sigma_1^2) \right] \right\}$$

Riemannian manifold

- Geodesics: informally, shortest paths on a manifold (surface)
- Space of symmetric matrices equipped with inner product

$$\langle \mathbf{A}, \mathbf{B} \rangle = \mathrm{Tr}(\mathbf{AB}) = \mathrm{vec}(\mathbf{A})^\top \mathrm{vec}(\mathbf{B})$$

  and associated Frobenius norm $\|\cdot\|_F = \sqrt{\langle \cdot, \cdot \rangle}$ is a Euclidean space of dimension $p(p+1)/2$.
- Instead, view covariance matrices as elements of a Riemannian manifold
- Endow $\mathcal{S}(p)$ with the **Riemannian metric**
  - local inner product $\langle \mathbf{A}, \mathbf{B} \rangle_{\mathbf{\Sigma}}$ on the tangent space of symmetric matrices

$$\begin{aligned} \langle \mathbf{A}, \mathbf{B} \rangle_{\mathbf{\Sigma}} &= \langle \mathbf{\Sigma}^{-1/2} \mathbf{A} \mathbf{\Sigma}^{-1/2}, \mathbf{\Sigma}^{-1/2} \mathbf{B} \mathbf{\Sigma}^{-1/2} \rangle \\ &= \mathrm{Tr}(\mathbf{A}\mathbf{\Sigma}^{-1}\mathbf{B}\mathbf{\Sigma}^{-1}) = \mathrm{vec}(\mathbf{A})^\top (\mathbf{\Sigma}^{-1} \otimes \mathbf{\Sigma}^{-1}) \mathrm{vec}(\mathbf{B}) \end{aligned}$$

- Geodesic path $\mathbf{\Sigma}_t$ is the shortest path from $\mathbf{\Sigma}_0$ to $\mathbf{\Sigma}_1$.

Geodesically (g-)convex function

A function $h : \mathcal{S}(p) \to \mathbb{R}$ is g-convex function if

$$h(\mathbf{\Sigma}_t) \leq (1-t)\ h(\mathbf{\Sigma}_0) + t\ h(\mathbf{\Sigma}_1) \text{ for } t \in (0,1).$$

If the inequality is strict, then h is strictly g-convex.

Note: Def. of convexity of $h(\mathbf{\Sigma})$ remains the same, i.e., w.r.t. to given path $\mathbf{\Sigma}_t$. Now geodesic instead of Euclidean path.

$$\boxed{g\text{-convexity} = \text{convexity w.r.t. geodesic paths}}$$

Local is Global

1. any local minimum of $h(\mathbf{\Sigma})$ over $\mathcal{S}(p)$ is a global minimum.
2. If h is strictly g-convex and a minimum is in $\mathcal{S}(p)$, then it is a unique minimum.
3. g-convex + g-convex = g-convex

Useful results on g-convexity: my personal top 3

(not in particular order)

$$\boldsymbol{\Sigma}_t = \boldsymbol{\Sigma}_0^{1/2}\left(\boldsymbol{\Sigma}_0^{-1/2}\boldsymbol{\Sigma}_1\boldsymbol{\Sigma}_0^{-1/2}\right)^t\boldsymbol{\Sigma}_0^{1/2}$$

1. Joint diagonalization formulation

The geodesic path can be written equivalently as

$$\boldsymbol{\Sigma}_t = \mathbf{E}\mathbf{D}^t\mathbf{E}^\top, \quad t \in [0,1],$$

where $\boldsymbol{\Sigma}_0 = \mathbf{E}\mathbf{E}^\top$ and $\boldsymbol{\Sigma}_1 = \mathbf{E}\mathbf{D}\mathbf{E}^\top$ by joint diagonalization.

- $\mathbf{E}$ is a *nonsingular* square matrix: row vectors of $\mathbf{E}^{-1}$ are the eigenvectors of $\boldsymbol{\Sigma}_0^{-1}\boldsymbol{\Sigma}_1$
- $\mathbf{D}$ is a diagonal matrix: diagonal elements are the eigenvalues of

$$\boldsymbol{\Sigma}_0^{-1}\boldsymbol{\Sigma}_1 \quad \text{or} \quad \boldsymbol{\Sigma}_0^{-1/2}\boldsymbol{\Sigma}_1\boldsymbol{\Sigma}_0^{-1/2}.$$

Useful results on g-convexity: my personal top 3

$$\boldsymbol{\Sigma}_t = \boldsymbol{\Sigma}_0^{1/2}\left(\boldsymbol{\Sigma}_0^{-1/2}\boldsymbol{\Sigma}_1\boldsymbol{\Sigma}_0^{-1/2}\right)^t\boldsymbol{\Sigma}_0^{1/2}$$

2. Convexity w.r.t. t

A continuous function f on a g-convex set $\mathcal{M}$ is g-convex if $f(\boldsymbol{\Sigma}_t)$ is classically convex in $t \in [0,1]$

3. Midpoint convexity

A continuous function on f on a g-convex set $\mathcal{M}$ is g-convex if

$$f(\boldsymbol{\Sigma}_{1/2}) \leq \frac{1}{2}\{f(\boldsymbol{\Sigma}_0) + f(\boldsymbol{\Sigma}_1)\}$$

for any $\boldsymbol{\Sigma}_0, \boldsymbol{\Sigma}_1 \in \mathcal{M}$.

For more results, see [Wiesel and Zhang, 2015]

Some geodesically (g-)convex functions

1. if $h(\mathbf{\Sigma})$ is g-convex in $\mathbf{\Sigma}$, then it is g-convex in $\mathbf{\Sigma}^{-1}$.

scalar case: if $h(x)$ is convex in $x = \log(\sigma^2) \in \mathbb{R}$, then it is convex in $-x = \log(\sigma^{-2}) = -\log(\sigma^2)$.

2. $\pm \log |\mathbf{\Sigma}|$ is g-convex. (i.e., log-determinant is g-linear function)

scalar case: the scalar g-linear function is the logarithm.

3. $\mathbf{a}^\top \mathbf{\Sigma}^{\pm 1} \mathbf{a}$ is strictly g-convex ($\mathbf{a} \neq 0$).
4. $\log |\sum_{i=1}^n \mathbf{H}_i \mathbf{\Sigma}^{\pm 1} \mathbf{H}_i|$ is g-convex.

scalar case: log-sum-exp function is convex.

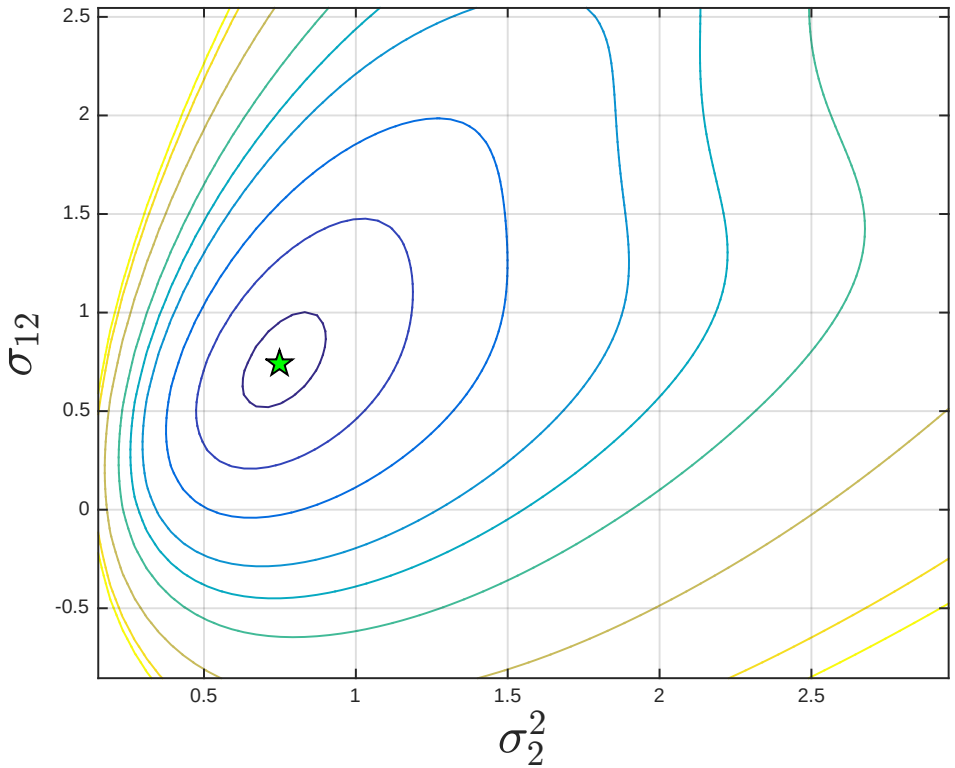
5. if $f(\mathbf{\Sigma})$ is g-convex, then $f(\mathbf{\Sigma}_1 \otimes \mathbf{\Sigma}_2)$ is jointly g-convex.

Example: Tyler's M-estimator of shape

Let's minimize Tyler's cost function $\mathcal{L}_{\mathrm{T}}(\boldsymbol{\Sigma}) = \mathcal{L}_{\mathrm{T}}(\sigma_2^2, \sigma_{12})$ over g-convex set of 2×2 shape matrices:

$$\mathcal{M}(2) = \{\boldsymbol{\Sigma} \in \mathcal{S}(2) : \det(\boldsymbol{\Sigma}) = 1\}$$
$$= \left\{ \begin{pmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{pmatrix} \; : \; \sigma_2^2 > 0, \sigma_{12} \in \mathbb{R}, \sigma_1^2 = \frac{1 + \sigma_{12}^2}{\sigma_2^2} \right\}$$

We generated a Gaussian sample of length $n = 15$ with $\sigma_2^2 = \sigma_{12} = 1$.



$$\min_{\boldsymbol{\Sigma} \in \mathcal{M}(2)} \underbrace{\sum_{i=1}^{n} \ln(\mathbf{x}_i^\top \boldsymbol{\Sigma}^{-1} \mathbf{x}_i)}_{= \mathcal{L}_{\mathrm{T}}(\boldsymbol{\Sigma})}$$

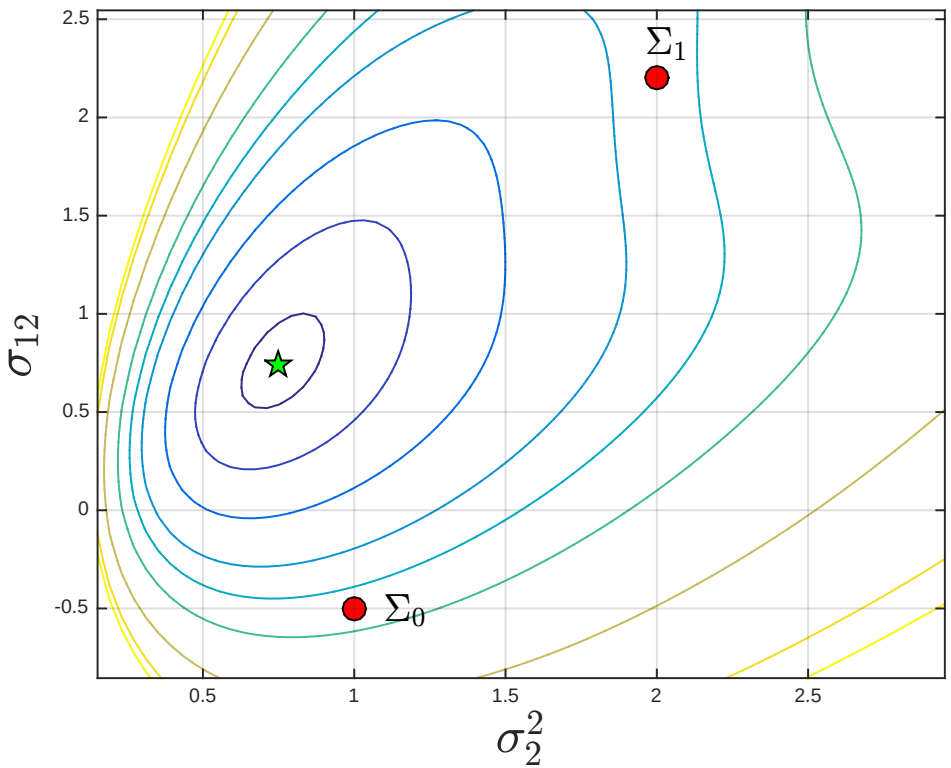
Contours of $\mathcal{L}_{\mathrm{T}}(\boldsymbol{\Sigma})$ and the solution $\hat{\boldsymbol{\Sigma}}$.

Example: Tyler's M-estimator of shape

Let's minimize Tyler's cost function $\mathcal{L}_{\mathrm{T}}(\boldsymbol{\Sigma}) = \mathcal{L}_{\mathrm{T}}(\sigma_2^2, \sigma_{12})$ over g-convex set of 2×2 shape matrices:

$$\mathcal{M}(2) = \{\boldsymbol{\Sigma} \in \mathcal{S}(2) : \det(\boldsymbol{\Sigma}) = 1\}$$
$$= \left\{ \begin{pmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{pmatrix} \; : \; \sigma_2^2 > 0, \sigma_{12} \in \mathbb{R}, \sigma_1^2 = \frac{1 + \sigma_{12}^2}{\sigma_2^2} \right\}$$

We generated a Gaussian sample of length $n = 15$ with $\sigma_2^2 = \sigma_{12} = 1$.



$$\min_{\boldsymbol{\Sigma} \in \mathcal{M}(2)} \underbrace{\sum_{i=1}^{n} \ln(\mathbf{x}_i^\top \boldsymbol{\Sigma}^{-1} \mathbf{x}_i)}_{= \mathcal{L}_{\mathrm{T}}(\boldsymbol{\Sigma})}$$

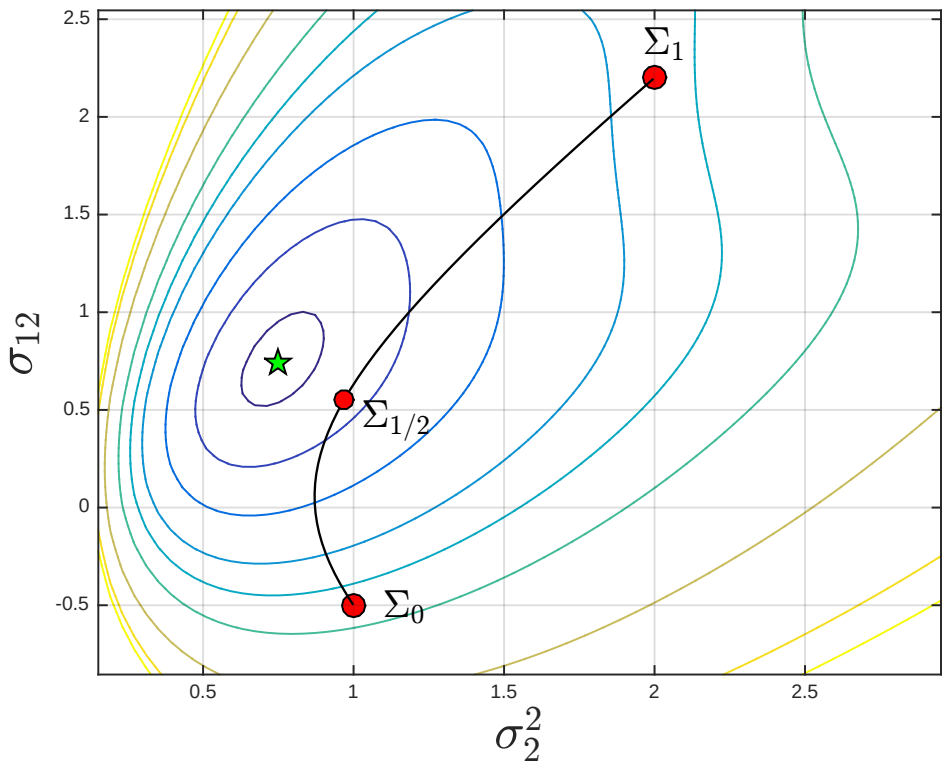
Consider two points $\boldsymbol{\Sigma}_0$ and $\boldsymbol{\Sigma}_1$ of $\mathcal{M}$.

Example: Tyler's M-estimator of shape

Let's minimize Tyler's cost function $\mathcal{L}_{\mathrm{T}}(\boldsymbol{\Sigma}) = \mathcal{L}_{\mathrm{T}}(\sigma_2^2, \sigma_{12})$ over g-convex set of 2×2 shape matrices:

$$\mathcal{M}(2) = \{\boldsymbol{\Sigma} \in \mathcal{S}(2) : \det(\boldsymbol{\Sigma}) = 1\}$$
$$= \left\{ \begin{pmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{pmatrix} \; : \; \sigma_2^2 > 0, \sigma_{12} \in \mathbb{R}, \sigma_1^2 = \frac{1 + \sigma_{12}^2}{\sigma_2^2} \right\}$$

We generated a Gaussian sample of length $n = 15$ with $\sigma_2^2 = \sigma_{12} = 1$.



$$\min_{\boldsymbol{\Sigma} \in \mathcal{M}(2)} \underbrace{\sum_{i=1}^{n} \ln(\mathbf{x}_i^\top \boldsymbol{\Sigma}^{-1} \mathbf{x}_i)}_{= \mathcal{L}_{\mathrm{T}}(\boldsymbol{\Sigma})}$$

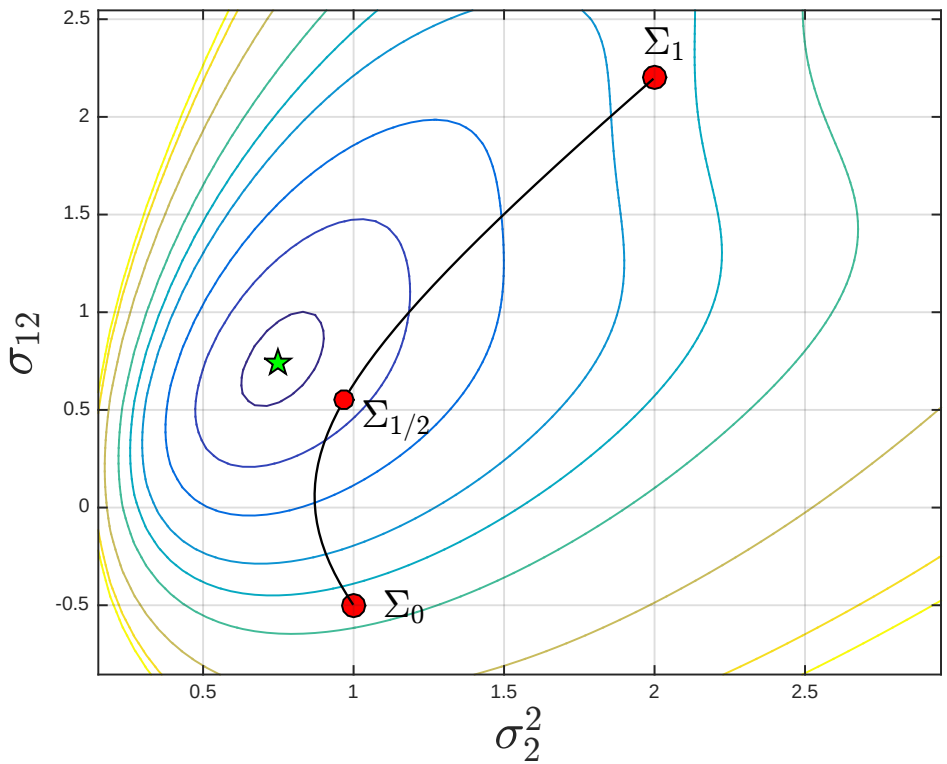
Their geodesic path $\boldsymbol{\Sigma}_t$ and midpoint $\boldsymbol{\Sigma}_{1/2}$

Example: Tyler's M-estimator of shape

Let's minimize Tyler's cost function $\mathcal{L}_{\mathrm{T}}(\boldsymbol{\Sigma}) = \mathcal{L}_{\mathrm{T}}(\sigma_2^2, \sigma_{12})$ over g-convex set of 2×2 shape matrices:

$$\mathcal{M}(2) = \{\boldsymbol{\Sigma} \in \mathcal{S}(2) : \det(\boldsymbol{\Sigma}) = 1\}$$
$$= \left\{ \begin{pmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{pmatrix} \; : \; \sigma_2^2 > 0, \sigma_{12} \in \mathbb{R}, \sigma_1^2 = \frac{1 + \sigma_{12}^2}{\sigma_2^2} \right\}$$

We generated a Gaussian sample of length $n = 15$ with $\sigma_2^2 = \sigma_{12} = 1$.



$$\min_{\boldsymbol{\Sigma} \in \mathcal{M}(2)} \underbrace{\sum_{i=1}^{n} \ln(\mathbf{x}_i^\top \boldsymbol{\Sigma}^{-1} \mathbf{x}_i)}_{= \mathcal{L}_{\mathrm{T}}(\boldsymbol{\Sigma})}$$

By utilizing the proper (Riemannian) metric, Tyler's cost fnc *is convex*.

Examples of g-convex sets

g-convex set $\mathcal{M}$ = all geodescic paths $\mathbf{\Sigma}_t$ lie in the set, where

$$\mathbf{\Sigma}_t = \mathbf{\Sigma}_0^{1/2}\left(\mathbf{\Sigma}_0^{-1/2}\mathbf{\Sigma}_1\mathbf{\Sigma}_0^{-1/2}\right)^t\mathbf{\Sigma}_0^{1/2} \text{ for } t \in [0,1].$$

and $\mathbf{\Sigma}_0$ and $\mathbf{\Sigma}_1$ are in $\mathcal{M}$.

1. The set of PDS matrices: $\mathcal{M} = \mathcal{S}_p$
2. The set of PDS shape matrices: $\mathcal{M} = \{\mathbf{\Sigma} \in \mathcal{S}_p : \det(\mathbf{\Sigma}) = 1\}$
3. The set of PDS block diagonal matrices.
4. Kronenecker model $\mathbf{\Sigma} = \mathbf{\Sigma}_1 \otimes \mathbf{\Sigma}_2$
5. Complex circular symmetric model:

$$\mathbf{\Sigma} = \begin{pmatrix} \mathbf{\Sigma}_1 & \mathbf{\Sigma}_2 \\ -\mathbf{\Sigma}_2 & \mathbf{\Sigma}_1 \end{pmatrix}$$

6. PDS circulant matrices, e.g., $[\mathbf{\Sigma}]_{ij} = \rho^{|i-j|}$, $\rho \in (0,1)$.

Menu

I. Ad-hoc shrinkage SCM-s of multiple samples

II. ML- and M-estimators of scatter matrix

III. Geodesic convexity

IV. Regularized M-estimators

- Shrinkage towards an identity matrix
- Shrinkage towards a target matrix

V. Penalized estimation of multiple covariances

VI. Estimation of the regularization parameter

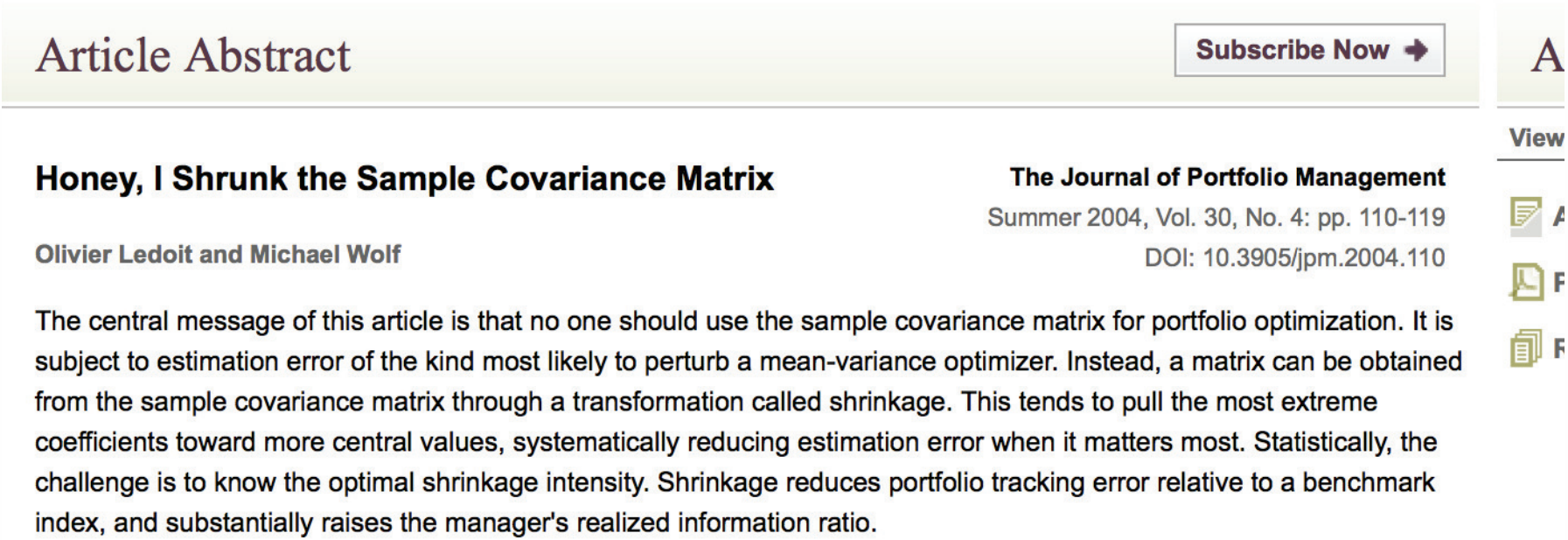
VII. Applications

References

Ollila, E. and Tyler, D. E. (2014).
Regularized M-estimators of scatter matrix.
IEEE Trans. Signal Process., 62(22):6059–6070.

Ollila, E., Soloveychik, I., Tyler, D. E. and Wiesel, A. (2016).
Simultaneous penalized M-estimation of covariance matrices using geodesically convex optimization
Journal of Multivariate Analysis (under review), Cite as:
arXiv:1608.08126 [stat.ME]
`http://arxiv.org/abs/1608.08126`

Honey, I shrunk the M-estimator of covariance matrix



Article Abstract

Subscribe Now

Honey, I Shrunk the Sample Covariance Matrix

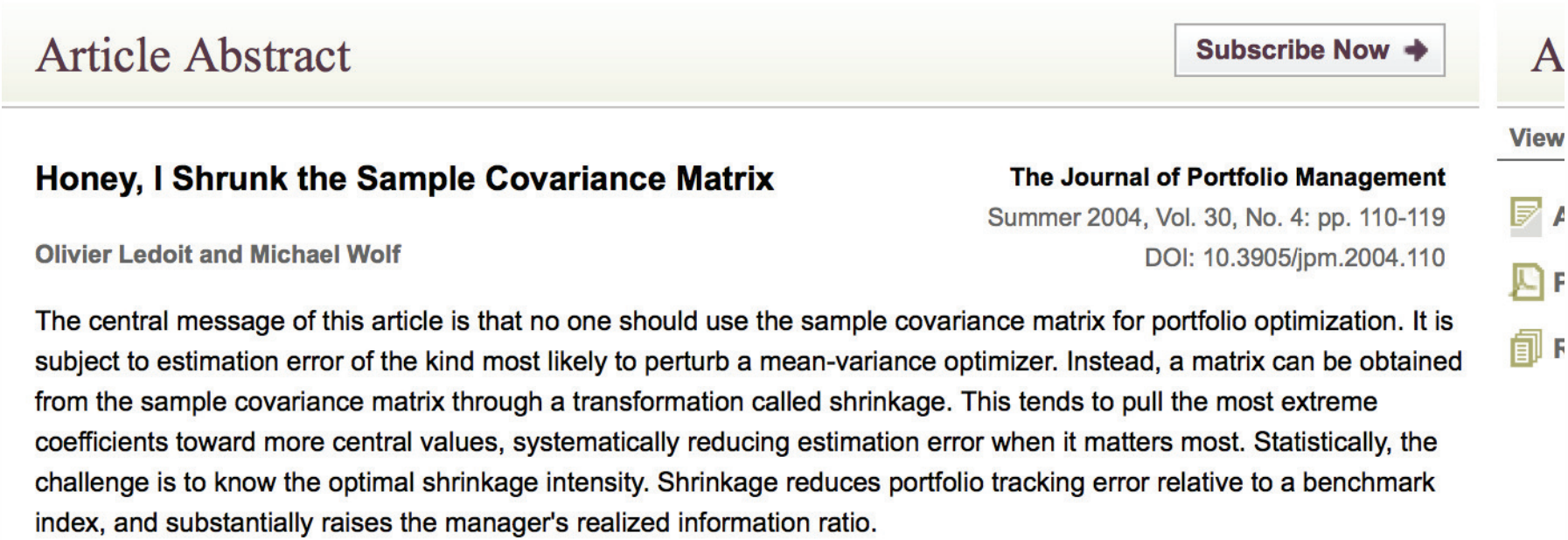
Olivier Ledoit and Michael Wolf

The Journal of Portfolio Management
Summer 2004, Vol. 30, No. 4: pp. 110-119
DOI: 10.3905/jpm.2004.110

The central message of this article is that no one should use the sample covariance matrix for portfolio optimization. It is subject to estimation error of the kind most likely to perturb a mean-variance optimizer. Instead, a matrix can be obtained from the sample covariance matrix through a transformation called shrinkage. This tends to pull the most extreme coefficients toward more central values, systematically reducing estimation error when it matters most. Statistically, the challenge is to know the optimal shrinkage intensity. Shrinkage reduces portfolio tracking error relative to a benchmark index, and substantially raises the manager's realized information ratio.

- $\Rightarrow$ The popular shrinkage SCM estimator (a.k.a Ledoit-Wolf estimator) is a member in the large class of regularized M-estimators of [Ollila and Tyler, 2014] that are presented in this section.
- $\Rightarrow$ **Our message**: for non-Gaussian data or in the presence of outliers, you will do MUCH BETTER by shrinking a robust M-estimator of covariance!

Honey, I shrunk the M-estimator of covariance matrix



Article Abstract

Subscribe Now

Honey, I Shrunk the Sample Covariance Matrix

Olivier Ledoit and Michael Wolf

The Journal of Portfolio Management
Summer 2004, Vol. 30, No. 4: pp. 110-119
DOI: 10.3905/jpm.2004.110

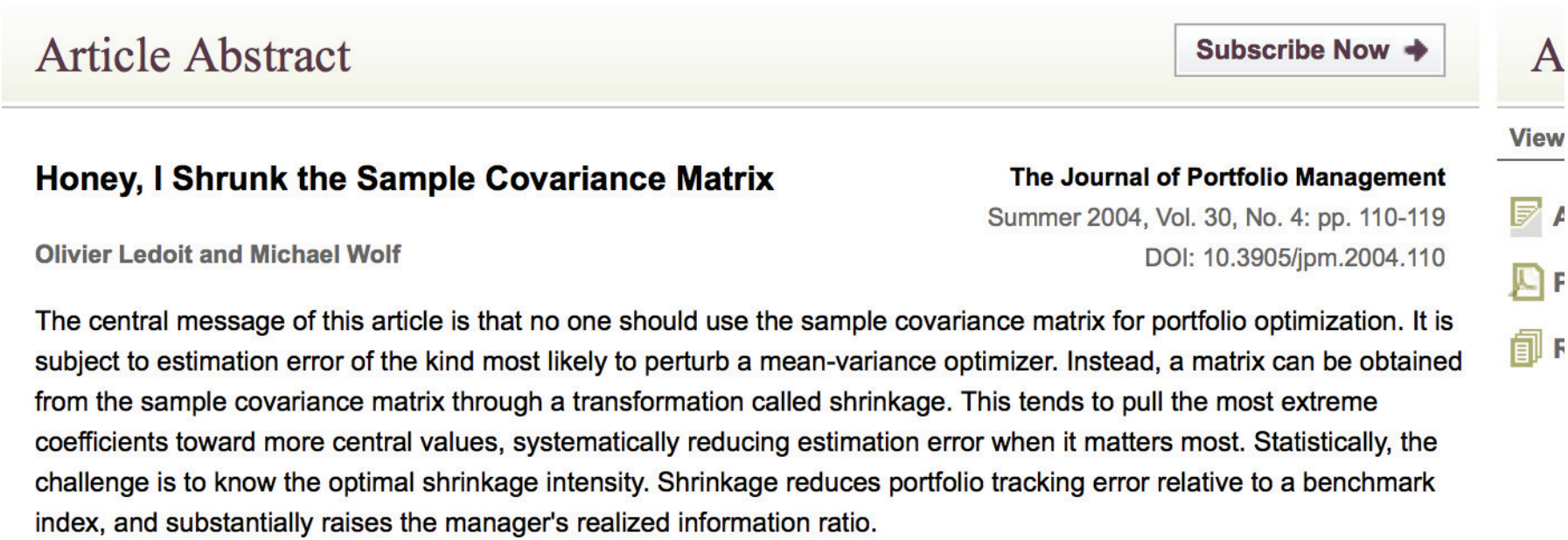
The central message of this article is that no one should use the sample covariance matrix for portfolio optimization. It is subject to estimation error of the kind most likely to perturb a mean-variance optimizer. Instead, a matrix can be obtained from the sample covariance matrix through a transformation called shrinkage. This tends to pull the most extreme coefficients toward more central values, systematically reducing estimation error when it matters most. Statistically, the challenge is to know the optimal shrinkage intensity. Shrinkage reduces portfolio tracking error relative to a benchmark index, and substantially raises the manager's realized information ratio.

View

$\Rightarrow$ The popular shrinkage SCM estimator (a.k.a Ledoit-Wolf estimator) is a member in the large class of regularized M-estimators of [Ollila and Tyler, 2014] that are presented in this section.

$\Rightarrow$ **Our message**: for non-Gaussian data or in the presence of outliers, you will do MUCH BETTER by shrinking a robust M-estimator of covariance!

Honey, I shrunk the M-estimator of covariance matrix



Article Abstract

Subscribe Now

Honey, I Shrunk the Sample Covariance Matrix

Olivier Ledoit and Michael Wolf

The Journal of Portfolio Management
Summer 2004, Vol. 30, No. 4: pp. 110-119
DOI: 10.3905/jpm.2004.110

The central message of this article is that no one should use the sample covariance matrix for portfolio optimization. It is subject to estimation error of the kind most likely to perturb a mean-variance optimizer. Instead, a matrix can be obtained from the sample covariance matrix through a transformation called shrinkage. This tends to pull the most extreme coefficients toward more central values, systematically reducing estimation error when it matters most. Statistically, the challenge is to know the optimal shrinkage intensity. Shrinkage reduces portfolio tracking error relative to a benchmark index, and substantially raises the manager's realized information ratio.

$\Rightarrow$ The popular shrinkage SCM estimator (a.k.a Ledoit-Wolf estimator) is a member in the large class of regularized M-estimators of [Ollila and Tyler, 2014] that are presented in this section.

$\Rightarrow$ **Our message**: for non-Gaussian data or in the presence of outliers, you will do MUCH BETTER by shrinking a robust M-estimator of covariance!

Regularized M-estimators of scatter matrix

Penalized cost function:

$$\mathcal{L}_\alpha(\mathbf{\Sigma}) = \frac{1}{n}\sum_{i=1}^{n} \rho(\mathbf{x}_i^\top \mathbf{\Sigma}^{-1}\mathbf{x}_i) - \ln|\mathbf{\Sigma}^{-1}| + \alpha\mathcal{P}(\mathbf{\Sigma})$$

where $\alpha \geq 0$ is a fixed regularization parameter.
Q: Existence, Uniqueness, computation?

Our penalty function

$$\mathcal{P}(\mathbf{\Sigma}) = \mathrm{Tr}(\mathbf{\Sigma}^{-1})$$

- Q: Why is this penalty useful or what is its effect?
- Q: Any other penalties ?

Regularized M-estimators of scatter matrix

Penalized cost function:

$$\mathcal{L}_\alpha(\mathbf{\Sigma}) = \frac{1}{n}\sum_{i=1}^{n} \rho(\mathbf{x}_i^\top \mathbf{\Sigma}^{-1}\mathbf{x}_i) - \ln|\mathbf{\Sigma}^{-1}| + \alpha\mathcal{P}(\mathbf{\Sigma})$$

where $\alpha \geq 0$ is a fixed regularization parameter.
Q: Existence, Uniqueness, computation?

Our penalty function

$$\mathcal{P}(\mathbf{\Sigma}) = \mathrm{Tr}(\mathbf{\Sigma}^{-1})$$

- **Q**: Why is this penalty useful or what is its effect?
- **Q**: Any other penalties ?

Regularized M-estimators of scatter matrix

Penalized cost function:

$$\mathcal{L}_\alpha(\mathbf{\Sigma}) = \frac{1}{n}\sum_{i=1}^{n} \rho(\mathbf{x}_i^\top \mathbf{\Sigma}^{-1}\mathbf{x}_i) - \ln|\mathbf{\Sigma}^{-1}| + \alpha\mathcal{P}(\mathbf{\Sigma})$$

where $\alpha \geq 0$ is a fixed regularization parameter.
Q: Existence, Uniqueness, computation?

Our penalty function *pulls* Σ *away from singularity*

$$\mathcal{P}(\mathbf{\Sigma}) = \mathrm{Tr}(\mathbf{\Sigma}^{-1})$$

- **Q**: Why is this penalty useful or what is its effect?
- **Q**: Any other penalties ?

Condition 1. [Zhang et al., 2013, Ollila and Tyler, 2014]

- $\rho(t)$ is nondecreasing and continuous for $0 < t < \infty$.
- $\rho(t)$ is g-convex (i.e., $\rho(e^x)$ is convex in $-\infty < x < \infty$)

Are the loss functions presented earlier g-convex?

- Gaussian loss function $\rho_{\mathrm{G}}(t) = t$

  as $f(x) = \rho_{\mathrm{G}}(e^x) = e^x$ is strictly convex in x

- Tyler's loss function $\rho_{\mathrm{T}}(t) = p \ln t$

  as $f(x) = \rho_{\mathrm{T}}(e^x) = px$ is g-linear.

- t_ν loss function $\rho_\nu(t) = (\nu + p) \ln(\nu + t)$

  as $f(x) = \rho_\nu(e^x) = (\nu + p)(\log \nu + x)$

- Huber's loss function $\rho_{\mathrm{H}}(t; c)$

  as it is a hybrid of Gaussian and Tyler's loss functions

Condition 1. [Zhang et al., 2013, Ollila and Tyler, 2014]

- $\rho(t)$ is nondecreasing and continuous for $0 < t < \infty$.
- $\rho(t)$ is g-convex (i.e., $\rho(e^x)$ is convex in $-\infty < x < \infty$)

Are the loss functions presented earlier g-convex?

- Gaussian loss function $\rho_{\mathrm{G}}(t) = t$ is strictly g-convex

  as $f(x) = \rho_{\mathrm{G}}(e^x) = e^x$ is strictly convex in x
- Tyler's loss function $\rho_{\mathrm{T}}(t) = p \ln t$

  as $f(x) = \rho_{\mathrm{T}}(e^x) = px$ is g-linear.
- t_ν loss function $\rho_\nu(t) = (\nu + p) \ln(\nu + t)$

  as $f(x) = \rho_\nu(e^x) = (\nu + p)(\log \nu + x)$
- Huber's loss function $\rho_{\mathrm{H}}(t; c)$

  as it is a hybrid of Gaussian and Tyler's loss functions

Condition 1. [Zhang et al., 2013, Ollila and Tyler, 2014]

- $\rho(t)$ is nondecreasing and continuous for $0 < t < \infty$.
- $\rho(t)$ is g-convex (i.e., $\rho(e^x)$ is convex in $-\infty < x < \infty$)

Are the loss functions presented earlier g-convex?

- Gaussian loss function $\rho_{\mathrm{G}}(t) = t$ is strictly g-convex

  as $f(x) = \rho_{\mathrm{G}}(e^x) = e^x$ is strictly convex in x
- Tyler's loss function $\rho_{\mathrm{T}}(t) = p \ln t$ is g-convex

  as $f(x) = \rho_{\mathrm{T}}(e^x) = px$ is g-linear.
- t_ν loss function $\rho_\nu(t) = (\nu + p) \ln(\nu + t)$

  as $f(x) = \rho_\nu(e^x) = (\nu + p)(\log \nu + x)$
- Huber's loss function $\rho_{\mathrm{H}}(t; c)$

  as it is a hybrid of Gaussian and Tyler's loss functions

Condition 1. [Zhang et al., 2013, Ollila and Tyler, 2014]

- $\rho(t)$ is nondecreasing and continuous for $0 < t < \infty$.
- $\rho(t)$ is g-convex (i.e., $\rho(e^x)$ is convex in $-\infty < x < \infty$)

Are the loss functions presented earlier g-convex?

- Gaussian loss function $\rho_{\mathrm{G}}(t) = t$ is strictly g-convex

  as $f(x) = \rho_{\mathrm{G}}(e^x) = e^x$ is strictly convex in x
- Tyler's loss function $\rho_{\mathrm{T}}(t) = p \ln t$ is g-convex

  as $f(x) = \rho_{\mathrm{T}}(e^x) = px$ is g-linear.
- t_ν loss function $\rho_\nu(t) = (\nu + p) \ln(\nu + t)$ is g-convex

  as $f(x) = \rho_\nu(e^x) = (\nu + p)(\log \nu + x)$
- Huber's loss function $\rho_{\mathrm{H}}(t; c)$ is g-convex

  as it is a hybrid of Gaussian and Tyler's loss functions

Main results

$$\mathcal{L}_\alpha(\mathbf{\Sigma}) = \frac{1}{n}\sum_{i=1}^{n}\rho(\mathbf{x}_i^\top\mathbf{\Sigma}^{-1}\mathbf{x}_i) - \ln|\mathbf{\Sigma}^{-1}| + \alpha\mathrm{Tr}(\mathbf{\Sigma}^{-1}),\ \alpha > 0$$

Result 1 [Ollila and Tyler, 2014]

Assume $\rho(t)$ satisfies Condition 1.

(a) Uniqueness: $\mathcal{L}_\alpha(\mathbf{\Sigma})$ is strictly g-convex in $\mathbf{\Sigma} \in \mathcal{S}(p)$

(b) Existence: If $\rho(t)$ is bounded below, then the solution to $\mathcal{L}_\alpha(\mathbf{\Sigma})$ allways exists and is unique.

(c) Furthermore, if $\rho(t)$ is also differentiable, then the minimum corresponds to the unique solution of the regularized M-estimating equation:

$$\hat{\mathbf{\Sigma}} = \frac{1}{n}\sum_{i=1}^{n} u(\mathbf{x}_i^\top\hat{\mathbf{\Sigma}}^{-1}\mathbf{x}_i)\mathbf{x}_i\mathbf{x}_i^\top + \alpha\mathbf{I}$$

Main results (cont'd)

Result 1 implies

- $u(t)$ need not be nonincreasing
- Unlike the non-regularized case, no conditions on the data are needed!

  $\rightarrow$ breakdown point is $= 1$.

Result 1(d) [Ollila and Tyler, 2014, Theorem 2]

Suppose $\rho(t)$ is continuously differentiable, satisfies Condition 1 and that $u(t) = \rho'(t)$ is non-increasing, Then the Fixed-point (FP) algorithm

$$\hat{\mathbf{\Sigma}}_{k+1} = \frac{1}{n}\sum_{i=1}^{n} u(\mathbf{x}_i^\top \hat{\mathbf{\Sigma}}_k^{-1}\mathbf{x}_i)\mathbf{x}_i\mathbf{x}_i^\top + \alpha\mathbf{I}$$

converges to the solution of regularized M-estimating equation given in Result 1(c).

Matlab implementation of regularized $t_\nu M$-estimator

Regularized $t_\nu M$-estimator is a solution to

$$\hat{\mathbf{\Sigma}} = \frac{1}{n}\sum_{i=1}^{n} u_\nu(\mathbf{x}_i^\top \hat{\mathbf{\Sigma}}_k^{-1}\mathbf{x}_i)\mathbf{x}_i\mathbf{x}_i^\top + \alpha\mathbf{I}$$

for $\alpha, \beta > 0$.

- Define a new function `C = regtMest(X, v,al,be)`
- Copy-paste the code from Slide 36.
- You *only need to change one line*:

  replace

  ```
  C = X'*((X.*repmat(u,1,p)))/n;
  ```

  to

  ```
  C = be*X'*((X.*repmat(u,1,p)))/n + al*eye(p);
  ```

Tuning the $\rho(t)$ function

- Result 1 is general and allows us to tune the $\rho(t)$ function
- For a given ρ-function, a class of tuned ρ-functions are defined as

$$\rho_\beta(t) = \beta\rho(t) \quad \text{for } \beta > 0.$$

  where β represents *additional tuning constant* which can be used to tune the estimator towards some desirable property.
- Using $\rho_\beta(t) = \beta\rho(t)$, our optimization program is

$$\mathcal{L}_{\alpha,\beta}(\boldsymbol{\Sigma}) = \beta\frac{1}{n}\sum_{i=1}^{n}\rho(\mathbf{x}_i^\top\boldsymbol{\Sigma}^{-1}\mathbf{x}_i) - \ln|\boldsymbol{\Sigma}^{-1}| + \alpha\mathrm{Tr}(\boldsymbol{\Sigma}^{-1})$$

- The solution verifies

$$\hat{\boldsymbol{\Sigma}} = \beta\frac{1}{n}\sum_{i=1}^{n}u(\mathbf{x}_i^\top\hat{\boldsymbol{\Sigma}}^{-1}\mathbf{x}_i)\mathbf{x}_i\mathbf{x}_i^\top + \alpha\mathbf{I}$$

- **Special cases**: $\alpha = 1-\beta$ or $\beta = (1-\alpha)$.

Tuning the $\rho(t)$ function

- Result 1 is general and allows us to tune the $\rho(t)$ function
- For a given ρ-function, a class of tuned ρ-functions are defined as

$$\rho_\beta(t) = \beta\rho(t) \quad \text{for } \beta > 0.$$

where β represents *additional tuning constant* which can be used to tune the estimator towards some desirable property.

- Using $\rho_\beta(t) = \beta\rho(t)$, our optimization program is

$$\mathcal{L}_{\alpha,\beta}(\boldsymbol{\Sigma}) = \beta\frac{1}{n}\sum_{i=1}^{n}\rho(\mathbf{x}_i^\top\boldsymbol{\Sigma}^{-1}\mathbf{x}_i) - \ln|\boldsymbol{\Sigma}^{-1}| + \alpha\mathrm{Tr}(\boldsymbol{\Sigma}^{-1})$$

- The solution verifies

$$\hat{\boldsymbol{\Sigma}} = \beta\frac{1}{n}\sum_{i=1}^{n}u(\mathbf{x}_i^\top\hat{\boldsymbol{\Sigma}}^{-1}\mathbf{x}_i)\mathbf{x}_i\mathbf{x}_i^\top + \alpha\mathbf{I}$$

- **Special cases**: $\alpha = 1-\beta$ or $\beta = (1-\alpha)$.

A class of regularized SCM's

- Let use *tuned* Gaussian cost fnc $\rho(t) = \beta t$, where $\beta > 0$ is a fixed tuning parameter.
- The penalized cost fnc is then

$$\mathcal{L}_{\alpha,\beta}(\mathbf{\Sigma}) = \mathrm{Tr}\big\{(\beta\mathbf{S} + \alpha\mathbf{I})\mathbf{\Sigma}^{-1}\big\} - \ln|\mathbf{\Sigma}^{-1}|$$

  where $\mathbf{S}$ denotes the SCM.
- Due to ▸ Result 1, its unique minimizer $\hat{\mathbf{\Sigma}}$ is

$$\mathbf{S}_{\alpha,\beta} = \beta\mathbf{S} + \alpha\mathbf{I}$$

  which corresponds to [Ledoit and Wolf, 2004] shrinkage estimator.
- **Note**: Ledoit and Wolf did not show that $\mathbf{S}_{\alpha,\beta}$ solves an penalized Gaussian optimization program.

A class of regularized Tyler's M-estimators

- Let use *tuned* Tyler's cost fnc $\rho(t) = p\beta \log t$ for fixed $0 < \beta < 1$.
- The penalized Tyler's cost fnc is

$$\mathcal{L}_{\alpha,\beta}(\mathbf{\Sigma}) = \beta \frac{p}{n} \sum_{i=1}^{n} \log(\mathbf{x}_i^\top \mathbf{\Sigma}^{-1} \mathbf{x}_i) - \ln |\mathbf{\Sigma}^{-1}| + \alpha \mathrm{Tr}(\mathbf{\Sigma}^{-1}),$$

- The weight fnc is $u(t) = p\beta/t$, so the regularized M-estimating eq. is

$$\hat{\mathbf{\Sigma}} = \beta \frac{p}{n} \sum_{i=1}^{n} \frac{\mathbf{x}_i \mathbf{x}_i^\top}{\mathbf{x}_i^\top \hat{\mathbf{\Sigma}}^{-1} \mathbf{x}_i} + \alpha \mathbf{I}$$

- We commonly use $\alpha = 1 - \beta$, $\beta \in (0, 1]$.
- Target $\mathcal{L}_{\alpha,\beta}(\mathbf{\Sigma})$ is g-convex in $\mathbf{\Sigma}$, but ρ is not bounded below $\Rightarrow$ ▸ Result 1(b), for existence does not hold.
- Conditions for existence needs to be considered separately for Tyler's M-estimator;

A class of regularized Tyler's M-estimators

- Let use *tuned* Tyler's cost fnc $\rho(t) = p\beta \log t$ for fixed $0 < \beta < 1$.
- The penalized Tyler's cost fnc is

$$\mathcal{L}_{\alpha,\beta}(\mathbf{\Sigma}) = \beta \frac{p}{n} \sum_{i=1}^{n} \log(\mathbf{x}_i^\top \mathbf{\Sigma}^{-1} \mathbf{x}_i) - \ln |\mathbf{\Sigma}^{-1}| + \alpha \mathrm{Tr}(\mathbf{\Sigma}^{-1}),$$

- The weight fnc is $u(t) = p\beta/t$, so the regularized M-estimating eq. is

$$\hat{\mathbf{\Sigma}} = \beta \frac{p}{n} \sum_{i=1}^{n} \frac{\mathbf{x}_i \mathbf{x}_i^\top}{\mathbf{x}_i^\top \hat{\mathbf{\Sigma}}^{-1} \mathbf{x}_i} + (1-\beta)\mathbf{I}$$

- We commonly use $\alpha = 1 - \beta$, $\beta \in (0, 1]$.
- Target $\mathcal{L}_{\alpha,\beta}(\mathbf{\Sigma})$ is g-convex in $\mathbf{\Sigma}$, but ρ is not bounded below
  $\Rightarrow$ ▸ Result 1(b), for existence does not hold.
- Conditions for existence needs to be considered separately for Tyler's M-estimator;

A class of regularized Tyler's M-estimators

- Let use *tuned* Tyler's cost fnc $\rho(t) = p\beta \log t$ for fixed $0 < \beta < 1$.
- The penalized Tyler's cost fnc is

$$\mathcal{L}_{\alpha,\beta}(\mathbf{\Sigma}) = \beta \frac{p}{n} \sum_{i=1}^{n} \log(\mathbf{x}_i^\top \mathbf{\Sigma}^{-1} \mathbf{x}_i) - \ln |\mathbf{\Sigma}^{-1}| + \alpha \mathrm{Tr}(\mathbf{\Sigma}^{-1}),$$

- The weight fnc is $u(t) = p\beta/t$, so the regularized M-estimating eq. is

$$\hat{\mathbf{\Sigma}} = \beta \frac{p}{n} \sum_{i=1}^{n} \frac{\mathbf{x}_i \mathbf{x}_i^\top}{\mathbf{x}_i^\top \hat{\mathbf{\Sigma}}^{-1} \mathbf{x}_i} + (1-\beta)\mathbf{I}$$

- We commonly use $\alpha = 1 - \beta$, $\beta \in (0, 1]$.
- Target $\mathcal{L}_{\alpha,\beta}(\mathbf{\Sigma})$ is g-convex in $\mathbf{\Sigma}$, but ρ is not bounded below
  $\Rightarrow$ ▸ Result 1(b), for existence does not hold.
- Conditions for existence needs to be considered separately for Tyler's M-estimator;

- **(Sufficient) Condition A.** For any subspace $\mathcal{V}$ of $\mathbb{R}^p$, $1 \leq \dim(\mathcal{V}) < p$, the inequality

$$\frac{\#\{\mathbf{x}_i \in \mathcal{V}\}}{n} < \frac{\dim(\mathcal{V})}{p\beta}$$

  holds. [**(Necessary) Condition B:** As earlier but with inequality.]
- Cond A implies $\beta < n/p$ whenever the sample is in "general position" (e.g., when sampling from a continuous distribution)

Result 2 [Ollila and Tyler, 2014]

Consider tuned Tyler's cost $\rho_\beta(t) = p\beta \ln t$ and $\alpha > 0$, $0 \leq \beta < 1$. If Condition A holds, then $\mathcal{L}_{\alpha,\beta}(\mathbf{\Sigma})$ has a unique minimum $\hat{\mathbf{\Sigma}}$ in $\mathcal{S}(p)$, the minimum being obtained at the unique solution to

$$\hat{\mathbf{\Sigma}} = \beta \frac{p}{n} \sum_{i=1}^{n} \frac{\mathbf{x}_i \mathbf{x}_i^\top}{\mathbf{x}_i^\top \hat{\mathbf{\Sigma}}^{-1} \mathbf{x}_i} + \alpha \mathbf{I},$$

Similar result found independently in [Pascal et al., 2014, Sun et al., 2014].

- For **fixed** $0 < \beta < 1$, consider two different values α_1 and α_2, and let $\hat{\mathbf{\Sigma}}_1$ and $\hat{\mathbf{\Sigma}}_2$ represent the respective regularized Tyler's M-estimators.
- It then follows that

$$\hat{\mathbf{\Sigma}}_1 = \frac{\alpha_1}{\alpha_2} \cdot \hat{\mathbf{\Sigma}}_2$$

  $\Rightarrow$ for any fixed $0 < \beta < 1$, the regularized Tyler's M-estimators are *proportional to one another* as α varies.
- Consequently, when the main interest is on estimation of the covariance matrix *up to a scale*, one may set w.lo.g.

$$\alpha = 1 - \beta \quad \text{[or equivalently } \beta = 1 - \alpha \text{]}.$$

  In these cases, it holds that $\mathrm{Tr}(\hat{\mathbf{\Sigma}}^{-1}) = p$.

Related approach for regularizing Tyler's M-estimator

- A related (but *different*) regularized Tyler's M-estimator was proposed by [Abramovich and Spencer, 2007] as the limit of the algorithm

$$\boldsymbol{\Sigma}_{k+1} \leftarrow (1-\alpha)\frac{p}{n}\sum_{i=1}^{n}\frac{\mathbf{x}_i\mathbf{x}_i^\top}{\mathbf{x}_i^\top\mathbf{V}_k^{-1}\mathbf{x}_i} + \alpha\mathbf{I}$$

$$\mathbf{V}_{k+1} \leftarrow p\boldsymbol{\Sigma}_{k+1}/\mathrm{Tr}(\boldsymbol{\Sigma}_{k+1}),$$

  where $\alpha \in (0,1]$ is a fixed regularization parameter.
- [Chen et al., 2011] proved that the recursive algorithm above converges to a unique solution regardless of the initialization. [Convergence means convergence in $\mathbf{V}_k$ and not necessarily in $\boldsymbol{\Sigma}_k$.]
- **Note 1:** Diagonally loaded version of the fixed-point algorithm for Tyler's M-estimator. Hence we refer to it by DL-FP.
- **Note 2:** DL-FP was not shown to be a solution to any penalized form of Tyler's cost function.

Comments

$$\mathcal{L}_\alpha(\boldsymbol{\Sigma}) = \underbrace{\frac{1}{n}\sum_{i=1}^{n} \rho(\mathbf{x}_i^\top \boldsymbol{\Sigma}^{-1}\mathbf{x}_i) - \ln|\boldsymbol{\Sigma}^{-1}|}_{=\mathcal{L}(\boldsymbol{\Sigma}),\ \text{M(L) cost function}} + \alpha\mathcal{P}(\boldsymbol{\Sigma})$$

- ℓ_1-penalization $\mathcal{P}(\boldsymbol{\Sigma}) = \|\boldsymbol{\Sigma}^{-1}\|_1$ is convex but not g-convex, so GLASSO [Friedman et al., 2008] does not fit our framework.
- $\mathcal{L}_\alpha(\boldsymbol{\Sigma})$ is strictly g-convex if either $\mathcal{L}(\boldsymbol{\Sigma})$ or $\mathcal{P}(\boldsymbol{\Sigma})$ are.
- Penalties $\mathcal{P}(\boldsymbol{\Sigma})$, other than $\mathcal{P}(\boldsymbol{\Sigma}) = \mathrm{Tr}(\boldsymbol{\Sigma}^{-1})$, that shrink towards a target matrix are considered in the next section.

Shrinkage towards a target matrix

- Fixed shrinkage target matrix $\mathbf{T} \in \mathcal{S}(p)$
- Define penalized M-estimator of scatter matrix as solution to

$$\min_{\mathbf{\Sigma} \in \mathcal{S}(p)} \left\{ \mathcal{L}(\mathbf{\Sigma}) + \lambda \, d(\mathbf{\Sigma}, \mathbf{T}) \right\},$$

or equivalently,

$$\min_{\mathbf{\Sigma} \in \mathcal{S}(p)} \left\{ \beta \mathcal{L}(\mathbf{\Sigma}) + (1-\beta) d(\mathbf{\Sigma}, \mathbf{T}) \right\}, \qquad \text{where } \lambda = \frac{1-\beta}{\beta}$$

where

- $\lambda > 0$ or $\beta \in (0, 1]$ is a regularization/penalty parameter
- $d(\mathbf{A}, \mathbf{B}) : \mathcal{S}(p) \times \mathcal{S}(p) \to \mathbb{R}_0^+$ is penalty/distance fnc.

Distance $d(\mathbf{\Sigma}, \mathbf{T})$ is used to enforce similarity of $\mathbf{\Sigma}$ to target $\mathbf{T}$ and β controls the amount of shrinkage of solution $\hat{\mathbf{\Sigma}}$ towards $\mathbf{T}$.

Properties of the penalty (distance) function

D1 $d(\mathbf{A}, \mathbf{B}) = 0$ if $\mathbf{A} = \mathbf{B}$,

D2 $d(\mathbf{A}, \mathbf{B})$ is jointly g-convex

D3 *symmetry*: $d(\mathbf{A}, \mathbf{B}) = d(\mathbf{B}, \mathbf{A})$.

D4 *affine invariance* $d(\mathbf{A}, \mathbf{B}) = d(\mathbf{C}\mathbf{A}\mathbf{C}^\top, \mathbf{C}\mathbf{B}\mathbf{C}^\top)$, $\forall$ nonsingular $\mathbf{C}$

D5 *scale invariance*: $d(c_1\mathbf{A}, c_2\mathbf{B}) = d(\mathbf{A}, \mathbf{B})$ for $c_1, c_2 > 0$,

Comments:

- D3-D5 are considered optional properties
- Property D5 is needed for shape matrix estimators (e.g. Tyler's). It is also important if $\mathbf{\Sigma}_k$-s share a common shape matrix only.

Note: Each distance $d(\mathbf{\Sigma}_k, \mathbf{\Sigma})$ *induce a notion of mean (or center).*

$\Rightarrow$ one might expect that a judicious choice of $d(\cdot, \cdot)$ should induce a natural notion of the mean of pos. def. matrices.

Properties of the penalty (distance) function

D1 $d(\mathbf{A}, \mathbf{B}) = 0$ if $\mathbf{A} = \mathbf{B}$,

D2 $d(\mathbf{A}, \mathbf{B})$ is jointly g-convex

D3 *symmetry*: $d(\mathbf{A}, \mathbf{B}) = d(\mathbf{B}, \mathbf{A})$.

D4 *affine invariance* $d(\mathbf{A}, \mathbf{B}) = d(\mathbf{C}\mathbf{A}\mathbf{C}^\top, \mathbf{C}\mathbf{B}\mathbf{C}^\top)$, $\forall$ nonsingular $\mathbf{C}$

D5 *scale invariance*: $d(c_1\mathbf{A}, c_2\mathbf{B}) = d(\mathbf{A}, \mathbf{B})$ for $c_1, c_2 > 0$,

Comments:

- D3-D5 are considered optional properties
- Property D5 is needed for shape matrix estimators (e.g. Tyler's). It is also important if $\mathbf{\Sigma}_k$-s share a common shape matrix only.

Note: Each distance $d(\mathbf{\Sigma}_k, \mathbf{\Sigma})$ *induce a notion of mean (or center).*

$\Rightarrow$ one might expect that a judicious choice of $d(\cdot, \cdot)$ should induce a natural notion of the mean of pos. def. matrices.

The induced mean or center

- Let $\{\boldsymbol{\Sigma}_k\}_{k=1}^K$ be given matrices in $\mathcal{S}(p)$
- Let weights $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_K)$, $\sum_{k=1}^K \pi_k = 1$, be given.

Then

$$\boldsymbol{\Sigma}(\boldsymbol{\pi}) = \arg\min_{\boldsymbol{\Sigma} \in \mathcal{S}(p)} \sum_{i=1}^{K} \pi_k \, d(\boldsymbol{\Sigma}_k, \boldsymbol{\Sigma}),$$

is a weighted mean associated with distance (penalty) d.

- **Q**: What is a natural mean of positive definite matrices?
- If $p = 1$, so we have $\sigma_1^2, \ldots, \sigma_K^2 > 0$, we could consider
  - a geometric mean $\sigma^2 = (\sigma_1^2 \cdots \sigma_K^2)^{1/K}$
- Note: For a pair σ_0^2, σ_1^2, the geometric mean is the midpoint of the geodesic $\sigma_t^2 = (\sigma_0^2)^{1-t}(\sigma_1^2)^t$.

The induced mean or center

- Let $\{\mathbf{\Sigma}_k\}_{k=1}^K$ be given matrices in $\mathcal{S}(p)$
- Let weights $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_K)$, $\sum_{k=1}^K \pi_k = 1$, be given.

Then

$$\mathbf{\Sigma}(\boldsymbol{\pi}) = \underset{\mathbf{\Sigma} \in \mathcal{S}(p)}{\arg\min} \sum_{i=1}^{K} \pi_k \, d(\mathbf{\Sigma}_k, \mathbf{\Sigma}),$$

is a weighted mean associated with distance (penalty) d.

- **Q**: What is a natural mean of positive definite matrices?
- If $p = 1$, so we have $\sigma_1^2, \ldots, \sigma_K^2 > 0$, we could consider
  - arithmetic mean $\sigma^2 = \frac{1}{K} \sum_{k=1}^K \sigma_k^2$.
  - geometric mean $\sigma^2 = \left(\sigma_1^2 \cdots \sigma_K^2\right)^{1/K}$
  - harmonic mean $\sigma^2 = \left(\frac{1}{K} \sum_{k=1}^K (\sigma_k^2)^{-1}\right)^{-1}$
- **Note:** For a pair σ_0^2, σ_1^2, the geometric mean is the midpoint of the geodesic $\sigma_t^2 = (\sigma_0^2)^{1-t} (\sigma_1^2)^t$.

The induced mean or center

- Let $\{\boldsymbol{\Sigma}_k\}_{k=1}^K$ be given matrices in $\mathcal{S}(p)$
- Let weights $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_K)$, $\sum_{k=1}^K \pi_k = 1$, be given.

Then

$$\boldsymbol{\Sigma}(\boldsymbol{\pi}) = \underset{\boldsymbol{\Sigma} \in \mathcal{S}(p)}{\arg\min} \sum_{i=1}^K \pi_k \, d(\boldsymbol{\Sigma}_k, \boldsymbol{\Sigma}),$$

is a weighted mean associated with distance (penalty) d.

- **Q**: What is a natural mean of positive definite matrices?
- If $p = 1$, so we have $\sigma_1^2, \ldots, \sigma_K^2 > 0$, we could consider
  - arithmetic mean $\sigma^2 = \frac{1}{K} \sum_{k=1}^K \sigma_k^2$.
  - geometric mean $\sigma^2 = \left(\sigma_1^2 \cdots \sigma_K^2\right)^{1/K}$
  - harmonic mean $\sigma^2 = \left(\frac{1}{K} \sum_{k=1}^K (\sigma_k^2)^{-1}\right)^{-1}$
- **Note:** For a pair σ_0^2, σ_1^2, the geometric mean is the midpoint of the geodesic $\sigma_t^2 = (\sigma_0^2)^{1-t} (\sigma_1^2)^t$.

The induced mean or center

- Let $\{\boldsymbol{\Sigma}_k\}_{k=1}^K$ be given matrices in $\mathcal{S}(p)$
- Let weights $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_K)$, $\sum_{k=1}^K \pi_k = 1$, be given.

Then

$$\boldsymbol{\Sigma}(\boldsymbol{\pi}) = \underset{\boldsymbol{\Sigma} \in \mathcal{S}(p)}{\arg\min} \sum_{i=1}^K \pi_k \, d(\boldsymbol{\Sigma}_k, \boldsymbol{\Sigma}),$$

is a weighted mean associated with distance (penalty) d.

- **Q**: What is a natural mean of positive definite matrices?
- If $p = 1$, so we have $\sigma_1^2, \ldots, \sigma_K^2 > 0$, we could consider
  - arithmetic mean $\sigma^2 = \frac{1}{K} \sum_{k=1}^K \sigma_k^2$.
  - geometric mean $\sigma^2 = \left(\sigma_1^2 \cdots \sigma_K^2\right)^{1/K}$
  - harmonic mean $\sigma^2 = \left(\frac{1}{K} \sum_{k=1}^K (\sigma_k^2)^{-1}\right)^{-1}$
- **Note:** For a pair σ_0^2, σ_1^2, the geometric mean is the midpoint of the geodesic $\sigma_t^2 = (\sigma_0^2)^{1-t} (\sigma_1^2)^t$.

The induced mean or center

- Let $\{\boldsymbol{\Sigma}_k\}_{k=1}^K$ be given matrices in $\mathcal{S}(p)$
- Let weights $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_K)$, $\sum_{k=1}^K \pi_k = 1$, be given.

Then

$$\boldsymbol{\Sigma}(\boldsymbol{\pi}) = \underset{\boldsymbol{\Sigma} \in \mathcal{S}(p)}{\arg\min} \sum_{i=1}^K \pi_k \, d(\boldsymbol{\Sigma}_k, \boldsymbol{\Sigma}),$$

is a weighted mean associated with distance (penalty) d.

- **Q**: What is a natural mean of positive definite matrices?
- If $p = 1$, so we have $\sigma_1^2, \ldots, \sigma_K^2 > 0$, we could consider
  - arithmetic mean $\sigma^2 = \frac{1}{K} \sum_{k=1}^K \sigma_k^2$.
  - geometric mean $\sigma^2 = \left(\sigma_1^2 \cdots \sigma_K^2\right)^{1/K}$
  - harmonic mean $\sigma^2 = \left(\frac{1}{K} \sum_{k=1}^K (\sigma_k^2)^{-1}\right)^{-1}$
- **Note:** For a pair σ_0^2, σ_1^2, the geometric mean is the midpoint of the geodesic $\sigma_t^2 = (\sigma_0^2)^{1-t} (\sigma_1^2)^t$.

So for $p > 1$ what penalties could one use?

- Frobenius distance

$$d_{\mathrm{F}}(\boldsymbol{\Sigma}_k, \boldsymbol{\Sigma}) = \left\{ \mathrm{Tr}[(\boldsymbol{\Sigma}_k - \boldsymbol{\Sigma})^2] \right\}^{1/2}$$

gives the standard weighted arithmetic mean $\boldsymbol{\Sigma}_{\mathrm{F}}(\boldsymbol{\pi}) = \sum_{k=1}^{K} \pi_k \boldsymbol{\Sigma}_k$.

. . . but not g-convex!

✓ Riemannian distance $d_{\mathrm{R}}(\mathbf{A}, \mathbf{B})$

✓ Kullback-Leibler (KL) divergence $d_{\mathrm{KL}}(\mathbf{A}, \mathbf{B})$

✓ Ellipticity distance $d_{\mathrm{E}}(\mathbf{A}, \mathbf{B})$

Note: there are also some other distances that are jointly g-convex, and hence fit our framework, e.g., S-divergence of [Sra, 2011].

So for $p > 1$ what penalties could one use?

✗ ~~Frobenius distance~~

$$d_{\mathrm{F}}(\boldsymbol{\Sigma}_k, \boldsymbol{\Sigma}) = \left\{ \mathrm{Tr}[(\boldsymbol{\Sigma}_k - \boldsymbol{\Sigma})^2] \right\}^{1/2}$$

gives the standard weighted arithmetic mean $\boldsymbol{\Sigma}_{\mathrm{F}}(\boldsymbol{\pi}) = \sum_{k=1}^{K} \pi_k \boldsymbol{\Sigma}_k$.

. . . but not g-convex!

✓ Riemannian distance $d_{\mathrm{R}}(\mathbf{A}, \mathbf{B})$

✓ Kullback-Leibler (KL) divergence $d_{\mathrm{KL}}(\mathbf{A}, \mathbf{B})$

✓ Ellipticity distance $d_{\mathrm{E}}(\mathbf{A}, \mathbf{B})$

Note: there are also some other distances that are jointly g-convex, and hence fit our framework, e.g., S-divergence of [Sra, 2011].

So for $p > 1$ what penalties could one use?

✗ ~~Frobenius distance~~

$$d_{\mathrm{F}}(\boldsymbol{\Sigma}_k, \boldsymbol{\Sigma}) = \left\{ \mathrm{Tr}[(\boldsymbol{\Sigma}_k - \boldsymbol{\Sigma})^2] \right\}^{1/2}$$

gives the standard weighted arithmetic mean $\boldsymbol{\Sigma}_{\mathrm{F}}(\boldsymbol{\pi}) = \sum_{k=1}^{K} \pi_k \boldsymbol{\Sigma}_k$.

. . . but not g-convex!

✓ Riemannian distance $d_{\mathrm{R}}(\mathbf{A}, \mathbf{B})$

✓ Kullback-Leibler (KL) divergence $d_{\mathrm{KL}}(\mathbf{A}, \mathbf{B})$

✓ Ellipticity distance $d_{\mathrm{E}}(\mathbf{A}, \mathbf{B})$

Note: there are also some other distances that are jointly g-convex, and hence fit our framework, e.g., S-divergence of [Sra, 2011].

Riemannian distance

- Riemannian distance

$$d_{\mathrm{R}}(\mathbf{A},\mathbf{B}) = \| \log(\mathbf{A}^{-1/2}\mathbf{B}\mathbf{A}^{-1/2})\|_{\mathrm{F}}^2,$$

  is the length of the geodesic curve between $\mathbf{A}$ and $\mathbf{B}$.
- The induced mean, called the Riemannian (or Karcher) mean is a unique solution to [Bhatia, 2009]

$$\sum_{k=1}^{K} \pi_k \log(\boldsymbol{\Sigma}_{\mathrm{R}}^{1/2}\boldsymbol{\Sigma}_k^{-1}\boldsymbol{\Sigma}_{\mathrm{R}}^{1/2}) = \mathbf{0}$$

☹ No closed-form solution: a number of complex numerical approaches have been proposed in the literature.

Kullback-Leibler (KL) divergence

$$d_{\mathrm{KL}}(\mathbf{A},\mathbf{B}) = \mathrm{Tr}(\mathbf{A}^{-1}\mathbf{B}) - \log|\mathbf{A}^{-1}\mathbf{B}| - p$$

- KL-distance verifies $d_{\mathrm{KL}}(\mathbf{A},\mathbf{B}) \geq 0$ and $= 0$ for $\mathbf{A} = \mathbf{B}$.
- utilized as shrinkage penalty in [Sun et al., 2014].

Result 3 [Ollila et al., 2016]

$d_{\mathrm{KL}}(\mathbf{A},\mathbf{B})$ is jointly strictly g-convex and affine invariant and the mean based on it has a unique solution in closed form:

$$\begin{aligned}\boldsymbol{\Sigma}_{\mathrm{I}}(\boldsymbol{\pi}) &= \arg\min_{\boldsymbol{\Sigma}\in\mathcal{S}(p)} \sum_{i=1}^{K} \pi_k \, d_{\mathrm{KL}}(\boldsymbol{\Sigma}_k, \boldsymbol{\Sigma}) \\ &= \left(\sum_{k=1}^{K} \pi_k \boldsymbol{\Sigma}_k^{-1}\right)^{-1},\end{aligned}$$

which is a **weighted harmonic mean** of PDS matrices.

Special case: target matrix $\mathbf{T} = \mathbf{I}$

- If the shrinkage target is $\mathbf{T} = \mathbf{I}$, then the criterion using KL-distance

$$\mathcal{L}_{\mathrm{KL},\beta}(\boldsymbol{\Sigma}) = \beta\mathcal{L}(\boldsymbol{\Sigma}) + (1-\beta)d_{\mathrm{KL}}(\boldsymbol{\Sigma}, \mathbf{I})$$
$$= \beta\left\{\frac{1}{n}\sum_{i=1}^{n}\rho(\mathbf{x}_i^\top\boldsymbol{\Sigma}^{-1}\mathbf{x}_i) - \ln|\boldsymbol{\Sigma}^{-1}|\right\} + (1-\beta)\underbrace{\{\mathrm{Tr}(\boldsymbol{\Sigma}^{-1}) - \ln|\boldsymbol{\Sigma}^{-1}|\}}_{d_{\mathrm{KL}}(\boldsymbol{\Sigma},\mathbf{I})}$$

looks closely similar to the optimization program which we studied earlier:

$$\mathcal{L}_{\alpha,\beta}(\boldsymbol{\Sigma}) = \beta\frac{1}{n}\sum_{i=1}^{n}\rho(\mathbf{x}_i^\top\boldsymbol{\Sigma}^{-1}\mathbf{x}_i) - \ln|\boldsymbol{\Sigma}^{-1}| + \alpha\mathrm{Tr}(\boldsymbol{\Sigma}^{-1}),$$

which utilized the penalty $\mathcal{P}(\boldsymbol{\Sigma}) = \mathrm{Tr}(\boldsymbol{\Sigma}^{-1})$ and a tuned ρ-function $\rho_\beta(t) = \beta\rho(t)$, $\beta > 0$.

Special case: target matrix $\mathbf{T} = \mathbf{I}$ (cont'd)

- Note that

$$\mathcal{L}_{\alpha,\beta}(\boldsymbol{\Sigma}) = \beta \frac{1}{n} \sum_{i=1}^{n} \rho(\mathbf{x}_i^\top \boldsymbol{\Sigma}^{-1} \mathbf{x}_i) - \ln |\boldsymbol{\Sigma}^{-1}| + \alpha \mathrm{Tr}(\boldsymbol{\Sigma}^{-1})$$

$$= \beta \underbrace{\left\{ \frac{1}{n} \sum_{i=1}^{n} \rho(\mathbf{x}_i^\top \boldsymbol{\Sigma}^{-1} \mathbf{x}_i) - \ln |\boldsymbol{\Sigma}^{-1}| \right\}}_{=\mathcal{L}(\boldsymbol{\Sigma})} - (1-\beta) \ln |\boldsymbol{\Sigma}^{-1}| + \alpha \mathrm{Tr}(\boldsymbol{\Sigma}^{-1})$$

- This shows that $\mathcal{L}_{\alpha,\beta}(\boldsymbol{\Sigma}) = \mathcal{L}_{\mathrm{KL},\beta}(\boldsymbol{\Sigma})$ when $\alpha = 1 - \beta$
- Thus results given earlier (e.g. ▸ Result 1(b)) transfer directly to penalization using KL-penalty.

Ellipticity distance

$$d_{\mathrm{E}}(\mathbf{A},\mathbf{B}) = p\log\frac{1}{p}\mathrm{Tr}(\mathbf{A}^{-1}\mathbf{B}) - \log|\mathbf{A}^{-1}\mathbf{B}|$$

- d_{E} is scale invariant. Note: Scale invariance is a useful property for estimators that are scale invariant, e.g., Tyler's M-estimator.
- utilized as shrinkage penalty in [Wiesel, 2012]
- Related to ellipticity factor, $e(\mathbf{\Sigma}) = \frac{1}{p}\mathrm{Tr}(\mathbf{\Sigma})/|\mathbf{\Sigma}|^{1/p}$, the ratio of the arithmetic and geometric means of the eigenvalues of $\mathbf{\Sigma}$.

Result 4 [Ollila et al., 2016]

$d_{\mathrm{E}}(\mathbf{A},\mathbf{B})$ is jointly g-convex and affine and scale invariant. The induced mean is unique (up to a scale) and solves

$$\mathbf{\Sigma}_{\mathrm{E}} = \left(\sum_{k=1}^{K}\pi_k\frac{p\mathbf{\Sigma}_k^{-1}}{\mathrm{Tr}(\mathbf{\Sigma}_k^{-1}\mathbf{\Sigma}_{\mathrm{E}})}\right)^{-1},$$

which is an (implicitly) weighted harmonic mean of normalized $\mathbf{\Sigma}_k$-s.

Critical points

$$\min_{\mathbf{\Sigma}\in\mathcal{S}(p)} \left\{\beta\mathcal{L}(\mathbf{\Sigma}) + (1-\beta)d(\mathbf{\Sigma},\mathbf{T})\right\}, \quad \beta\in(0,1]$$

- Write $\mathcal{P}_0(\mathbf{\Sigma}) = d(\mathbf{\Sigma},\mathbf{T})$ and $\mathcal{P}_0'(\mathbf{\Sigma}) = \partial\mathcal{P}(\mathbf{\Sigma})/\partial\mathbf{\Sigma}^{-1}$.
- The critical points then verify

$$\mathbf{0} = \beta\left\{\frac{1}{n}\sum_{i=1}^{n} u(\mathbf{x}_i^\top\mathbf{\Sigma}^{-1}\mathbf{x}_i)\mathbf{x}_i\mathbf{x}_i^\top - \mathbf{\Sigma}\right\} + (1-\beta)\mathcal{P}_0'(\mathbf{\Sigma})$$

$$\Leftrightarrow \quad \beta\mathbf{\Sigma} = \beta\frac{1}{n}\sum_{i=1}^{n} u(\mathbf{x}_i^\top\mathbf{\Sigma}^{-1}\mathbf{x}_i)\mathbf{x}_i\mathbf{x}_i^\top + (1-\beta)\mathcal{P}_0'(\mathbf{\Sigma})$$

$$\Leftrightarrow \quad \mathbf{\Sigma} = \beta\frac{1}{n}\sum_{i=1}^{n} u(\mathbf{x}_i^\top\mathbf{\Sigma}^{-1}\mathbf{x}_i)\mathbf{x}_i\mathbf{x}_i^\top + (1-\beta)\{\mathcal{P}_0'(\mathbf{\Sigma}) + \mathbf{\Sigma}\}.$$

- For $\mathcal{P}_0(\mathbf{\Sigma}) = d_{\mathrm{KL}}(\mathbf{\Sigma},\mathbf{T}) = \mathrm{Tr}(\mathbf{\Sigma}^{-1}\mathbf{T}) - \log|\mathbf{\Sigma}^{-1}\mathbf{T}| - p$, this gives

$$\mathbf{\Sigma} = \beta\frac{1}{n}\sum_{i=1}^{n} u(\mathbf{x}_i^\top\mathbf{\Sigma}^{-1}\mathbf{x}_i)\mathbf{x}_i\mathbf{x}_i^\top + (1-\beta)\mathbf{T}.$$

Menu

I. Ad-hoc shrinkage SCM-s of multiple samples

II. ML- and M-estimators of scatter matrix

III. Geodesic convexity

IV. Regularized M-estimators

V. Penalized estimation of multiple covariances

VI. Estimation of the regularization parameter

VII. Applications

Reference


Ollila, E., Soloveychik, I., Tyler, D. E. and Wiesel, A. (2016).
Simultaneous penalized M-estimation of covariance matrices using geodesically convex optimization
Journal of Multivariate Analysis (under review), Cite as:
arXiv:1608.08126 [stat.ME]
http://arxiv.org/abs/1608.08126

Multiple covariance estimation problem

- We are given K groups of elliptically distributed measurements,

$$\mathbf{x}_{11}, \ldots, \mathbf{x}_{1n_1}, \quad \ldots, \quad \mathbf{x}_{K1}, \ldots, \mathbf{x}_{Kn_K}$$

- Each group $\mathbf{X}_k = \{\mathbf{x}_{k1}, \ldots, \mathbf{x}_{kn_k}\}$ containing n_k p-dimensional samples, and

$$N = \sum_{i=1}^{K} n_k = \text{ total sample size}$$

$$\pi_k = \frac{n_k}{N} = \text{ relative sample size of the } k\text{-th group}$$

- Sample populations follow elliptical distributions, $\mathcal{E}_p(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k, g_k)$, with different scatter matrices $\boldsymbol{\Sigma}_k$ possessing mutual structure or a **joint center** $\boldsymbol{\Sigma} \Rightarrow$ need to estimate both $\{\boldsymbol{\Sigma}_k\}_{k=1}^K$ and $\boldsymbol{\Sigma}$.
- We assume that symmetry center $\boldsymbol{\mu}_k$ of populations is known or that data is *centered*.

Proposal 1: Regularization towards a pooled center

- A pooled M-estimator of scatter is defined as a minimum of

$$\mathcal{L}(\mathbf{\Sigma}) = \sum_{k=1}^{K} \pi_k \mathcal{L}_k(\mathbf{\Sigma}) = \frac{1}{N} \left\{ \sum_{k=1}^{K} \sum_{i=1}^{n_k} \rho_k(\mathbf{x}_{ki}^\top \mathbf{\Sigma}^{-1} \mathbf{x}_{ki}) \right\} - \log |\mathbf{\Sigma}^{-1}|$$

  over $\mathbf{\Sigma} \in \mathcal{S}(p)$.

- Penalized M-estimators of scatter for the individual groups solve

$$\min_{\mathbf{\Sigma}_k \in \mathcal{S}(p)} \left\{ \beta \mathcal{L}_k(\mathbf{\Sigma}_k) + (1-\beta)\, d(\mathbf{\Sigma}_k, \hat{\mathbf{\Sigma}}) \right\}, \quad k = 1, \ldots K,$$

  where
  - $\beta \in (0, 1]$ is a regularization/penalty parameter
  - $d(\mathbf{A}, \mathbf{B}) : \mathcal{S}(p) \times \mathcal{S}(p) \to \mathbb{R}_0^+$ is penalty/distance fnc.

Distance $d(\mathbf{\Sigma}_k, \hat{\mathbf{\Sigma}})$ enforce similarity of $\mathbf{\Sigma}_k$-s to joint center $\hat{\mathbf{\Sigma}}$ and β controls the amount of shrinkage towards $\hat{\mathbf{\Sigma}}$.

Proposal 2: Joint regularization enforcing similarity among the group scatter matrices

$$\underset{\{\boldsymbol{\Sigma}_k\}_{k=1}^K, \boldsymbol{\Sigma} \in \mathcal{S}(p)}{\text{minimize}} \quad \sum_{k=1}^{K} \pi_k \left\{ \beta \mathcal{L}_k(\boldsymbol{\Sigma}_k) + (1-\beta)\, d(\boldsymbol{\Sigma}_k, \boldsymbol{\Sigma}) \right\}$$

where β is the penalty parameter, $d(\boldsymbol{\Sigma}_k, \boldsymbol{\Sigma})$ is the distance function as before, and

$$\mathcal{L}_k(\boldsymbol{\Sigma}_k) = \frac{1}{n_k} \sum_{i=1}^{n_k} \rho_k(\mathbf{x}_{ki}^\top \boldsymbol{\Sigma}_k^{-1} \mathbf{x}_{ki}) - \log |\boldsymbol{\Sigma}_k^{-1}|$$

is the M(L)-cost fnc for the k-th class and $\rho_k(\cdot)$ is the loss fnc.

'Center' $\boldsymbol{\Sigma}$ can now be viewed as 'average' of $\boldsymbol{\Sigma}_k$-s. Namely, for fixed $\boldsymbol{\Sigma}_k$-s, the minumum $\hat{\boldsymbol{\Sigma}}$ is found by solving

$$\hat{\boldsymbol{\Sigma}}(\boldsymbol{\pi}) = \underset{\boldsymbol{\Sigma} \in \mathcal{S}(p)}{\arg\min} \sum_{i=1}^{K} \pi_k\, d(\boldsymbol{\Sigma}_k, \boldsymbol{\Sigma}),$$

which represents the weighted mean associated with the distance d.

Modifications to Proposals 1 and 2

- Penalty parameter β can be replaced by individual tuning constants β_k, $k = 1, \ldots, K$ for each class.

  Comment: typically one tends to choose small β_k when sample size is small, but this does not seem to be necessary in our framework

- In Proposal 1, if the total sample size N is small (e.g., $N < p$), then one may add a penalty $\mathcal{P}(\mathbf{\Sigma}) = \mathrm{Tr}(\mathbf{\Sigma}^{-1})$ and compute pooled center $\hat{\mathbf{\Sigma}}$ as a pooled regularized M-estimator:

$$\min_{\mathbf{\Sigma}} \sum_{k=1}^{K} \pi_k \mathcal{L}_k(\mathbf{\Sigma}) + \gamma \mathcal{P}(\mathbf{\Sigma})$$

  where $\gamma > 0$ is the (additional) penalty parameter for the center.

- Such a penalty term can be added to Proposal 2 as well.

- We consider the cases that penalty function $d(\mathbf{A}, \mathbf{B})$ is the KL-distance or ellipticity distance.
- Both distances are *affine invariant*, i.e.

$$d(\mathbf{A}, \mathbf{B}) = d(\mathbf{C}\mathbf{A}\mathbf{C}^\top, \mathbf{C}\mathbf{B}\mathbf{C}^\top), \ \forall \text{ nonsingular } \mathbf{C}.$$

  which is Property D4 in Slide
- If D4 holds, the resulting estimators are affine equivariant:

$$\text{if} \quad \mathbf{x}_{ki} \to \mathbf{C}\mathbf{x}_{ki} \ \text{ for all } k = 1, \ldots, K;\ i = 1, \ldots, n_k$$
$$\text{then } \{\mathbf{\Sigma}_1, \ldots, \mathbf{\Sigma}_K, \mathbf{\Sigma}\} \to \{\mathbf{C}\mathbf{\Sigma}_1\mathbf{C}^\top, \ldots, \mathbf{C}\mathbf{\Sigma}_K\mathbf{C}^\top, \mathbf{C}\mathbf{\Sigma}\mathbf{C}^\top\}.$$

Critical points/algorithm using KL-divergence penalty

Problem: $$\min_{\{\boldsymbol{\Sigma}_k\}_{k=1}^K, \boldsymbol{\Sigma}} \quad \sum_{k=1}^K \pi_k \left\{ \beta \mathcal{L}_k(\boldsymbol{\Sigma}_k) + (1-\beta)\, d_{\mathrm{KL}}(\boldsymbol{\Sigma}_k, \boldsymbol{\Sigma}) \right\}$$

Solving

$$\mathbf{0} = \beta \frac{\partial \mathcal{L}_k(\boldsymbol{\Sigma}_k)}{\partial \boldsymbol{\Sigma}_k^{-1}} + (1-\beta) \frac{\partial d_{\mathrm{KL}}(\boldsymbol{\Sigma}_k, \boldsymbol{\Sigma})}{\partial \boldsymbol{\Sigma}_k^{-1}}, \quad k = 1, \ldots, K$$

$$\mathbf{0} = \sum_{k=1}^K \pi_k \frac{\partial d_{\mathrm{KL}}(\boldsymbol{\Sigma}_k, \boldsymbol{\Sigma})}{\partial \boldsymbol{\Sigma}}$$

yields estimating equations

$$\boldsymbol{\Sigma}_k = \beta \frac{1}{n_k} \sum_{i=1}^{n_k} u_k(\mathbf{x}_{ki}^\top \boldsymbol{\Sigma}_k^{-1} \mathbf{x}_{ki}) \mathbf{x}_{ki} \mathbf{x}_{ki}^\top + (1-\beta)\boldsymbol{\Sigma}$$

$$\boldsymbol{\Sigma} = \left(\sum_{k=1}^K \pi_k \boldsymbol{\Sigma}_k^{-1} \right)^{-1}$$

where $u_k(t) = \rho_k'(t)$, $k = 1, \ldots, K$.

Critical points/algorithm using KL-divergence penalty

Problem: $$\min_{\{\boldsymbol{\Sigma}_k\}_{k=1}^K, \boldsymbol{\Sigma}} \quad \sum_{k=1}^K \pi_k \left\{ \beta \mathcal{L}_k(\boldsymbol{\Sigma}_k) + (1-\beta)\, d_{\mathrm{KL}}(\boldsymbol{\Sigma}_k, \boldsymbol{\Sigma}) \right\}$$

Solving

$$\mathbf{0} = \beta \frac{\partial \mathcal{L}_k(\boldsymbol{\Sigma}_k)}{\partial \boldsymbol{\Sigma}_k^{-1}} + (1-\beta) \frac{\partial d_{\mathrm{KL}}(\boldsymbol{\Sigma}_k, \boldsymbol{\Sigma})}{\partial \boldsymbol{\Sigma}_k^{-1}}, \quad k = 1, \ldots, K$$

$$\mathbf{0} = \sum_{k=1}^K \pi_k \frac{\partial d_{\mathrm{KL}}(\boldsymbol{\Sigma}_k, \boldsymbol{\Sigma})}{\partial \boldsymbol{\Sigma}}$$

yields algorithm that updates covariances cyclically from $\boldsymbol{\Sigma}_1, \ldots \boldsymbol{\Sigma}_K$ to $\boldsymbol{\Sigma}$

$$\boldsymbol{\Sigma}_k \leftarrow \beta \frac{1}{n_k} \sum_{i=1}^{n_k} u_k(\mathbf{x}_{ki}^\top \boldsymbol{\Sigma}_k^{-1} \mathbf{x}_{ki}) \mathbf{x}_{ki} \mathbf{x}_{ki}^\top + (1-\beta)\boldsymbol{\Sigma}$$

$$\boldsymbol{\Sigma} \leftarrow \left(\sum_{k=1}^K \pi_k \boldsymbol{\Sigma}_k^{-1} \right)^{-1}$$

where $u_k(t) = \rho_k'(t)$, $k = 1, \ldots, K$.

Critical points/algorithm using ellipticity distance

As for KL-distance, we can easily solve the estimating equations and propose a cyclic algorithm to find the solutions.

Estimating equations

$$\boldsymbol{\Sigma}_k = \beta \frac{1}{n_k} \sum_{i=1}^{n_k} u_k(\mathbf{x}_{ki}^\top \boldsymbol{\Sigma}_k^{-1} \mathbf{x}_{ki}) \mathbf{x}_{ki} \mathbf{x}_{ki}^\top + (1-\beta) \frac{p\boldsymbol{\Sigma}}{\mathrm{Tr}(\boldsymbol{\Sigma}_k^{-1}\boldsymbol{\Sigma})},$$

$$\boldsymbol{\Sigma} = \left(\sum_{k=1}^{K} \pi_k \frac{p\boldsymbol{\Sigma}_k^{-1}}{\mathrm{Tr}(\boldsymbol{\Sigma}_k^{-1}\boldsymbol{\Sigma})} \right)^{-1}$$

where $u_k(t) = \rho_k'(t)$, $k = 1, \ldots, K$.

Critical points/algorithm using ellipticity distance

As for KL-distance, we can easily solve the estimating equations and propose a cyclic algorithm to find the solutions.

Algorithm updates covariances cyclically from $\boldsymbol{\Sigma}_1, \ldots \boldsymbol{\Sigma}_K$ to $\boldsymbol{\Sigma}$

$$\boldsymbol{\Sigma}_k \leftarrow \beta \frac{1}{n_k} \sum_{i=1}^{n_k} u_k(\mathbf{x}_{ki}^\top \boldsymbol{\Sigma}_k^{-1} \mathbf{x}_{ki}) \mathbf{x}_{ki} \mathbf{x}_{ki}^\top + (1-\beta) \frac{p\boldsymbol{\Sigma}}{\mathrm{Tr}(\boldsymbol{\Sigma}_k^{-1}\boldsymbol{\Sigma})},$$

$$\boldsymbol{\Sigma} \leftarrow \left(\sum_{k=1}^{K} \pi_k \frac{p\boldsymbol{\Sigma}_k^{-1}}{\mathrm{Tr}(\boldsymbol{\Sigma}_k^{-1}\boldsymbol{\Sigma})} \right)^{-1}$$

where $u_k(t) = \rho_k'(t)$, $k = 1, \ldots, K$.

Menu

I. Ad-hoc shrinkage SCM-s of multiple samples

II. ML- and M-estimators of scatter matrix

III. Geodesic convexity

IV. Regularized M-estimators

V. Penalized estimation of multiple covariances

VI. Estimation of the regularization parameter
- Cross-validation
- Oracle approach

VII. Applications

Estimation of the regularization parameter

- Recall that the regularized M-estimator introduced in Section V solve

$$\hat{\mathbf{\Sigma}} = \beta \frac{1}{n} \sum_{i=1}^{n} u(\mathbf{x}_i^\top \hat{\mathbf{\Sigma}}^{-1} \mathbf{x}_i) \mathbf{x}_i \mathbf{x}_i^\top + \alpha \mathbf{I}$$

- For simplicity, we often tune only one parameter and set:

$$\beta = (1 - \alpha),\ \alpha \in (0, 1) \quad \text{or} \quad \alpha = (1 - \beta),\ \beta \in (0, 1).$$

- We need a disciplined way of choosing β (or α): it determines the amount of shrinkage towards the target matrix (here the identity matrix)
- Also related to a wider topic of model selection.
- A judicious choice of β is the one that provides an estimate $\hat{\mathbf{\Sigma}}$ that minimizes the mean squared error (or PE).

Approaches:

1. Cross-validation
2. Oracle/Clairvoyant approach
3. Expected likelihood approach
   [Abramovich and Besson, 2013, Besson and Abramovich, 2013]
4. Random matrix theory (Frederic's talk).

Only the first two topics are addressed in this tutorial.

Cross-validation (CV)

- Ideally, we would split our available data into two portions:
  - training set $\{\mathbf{x}_i\}_{i=1}^n$ to estimate $\hat{\mathbf{\Sigma}}$ which then gives the fit $\hat{F}(\mathbf{x})$.
  - test set $\{\mathbf{x}_i^t\}_{i=1}^n$ to validate the model, i.e., to "test" or assess how small is (say) the sum of the fitted values, $\sum_i \hat{F}(\mathbf{x}_i^t)$.
- This may not be plausible (especially if $n \approx p$) and also impractical/inefficient usage of data
- **Note**: The choice of the fit function $\hat{F}(\mathbf{x})$ is not as obvious as it in the regression setting, where the (squared) prediction error is a natural choice.
- Cross validation solves the problem by splitting the data into K folds, fits the data on $K-1$ folds, and evaluates the performance (using fit criterion) on the fold that was left out.

Cross-validation (cont'd)

- Partition $\mathbf{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$ into Q separate sets of similar size

$$I_1 \cup I_2 \cup \cdots \cup I_Q = \{1, \ldots, n\} \equiv [n]$$

- Common choices: $Q = 5, 10$ or $Q = n$ (*leave-one-out CV*).
- Taking qth fold out (all $\mathbf{x}_i$, $i \in I_q$) gives a reduced data set $\mathbf{X}_{-q}$.
- As the fit function, we use the (non-penalized) criterion function

$$\boxed{\hat{F}(\mathbf{x}) = \rho\big(\mathbf{x}^\top \hat{\boldsymbol{\Sigma}}^{-1} \mathbf{x}\big) - \log \big|\hat{\boldsymbol{\Sigma}}^{-1}\big|}$$

Test	Train	Train	Train	Train
Train	Test	Train	Train	Train
Train	Train	Test	Train	Train
Train	Train	Train	Test	Train
Train	Train	Train	Train	Test

Figure: $K = 5$ cross validation

CV procedure [Ollila et al., 2016]

For simplicity, tune only one parameter and set $\alpha = 1 - \beta$.

1. **for** $\beta \in [\beta]$ ($=$ a grid of β values in $(0,1)$) and $q \in \{1, \ldots, Q\}$ **do**
   - Compute regularized M-estimator based on $\mathbf{X}_{-q}$, denoted $\hat{\mathbf{\Sigma}}(\beta, q)$
   - CV fit for β is computed as the sum of fits over the qth folds that were left out:

$$\mathrm{CV}(\beta, q) = \sum_{\tilde{q} \in I_q} \rho\big(\mathbf{x}_{\tilde{q}}^\top \big[\hat{\mathbf{\Sigma}}(\beta, q)\big]^{-1} \mathbf{x}_{\tilde{q}}\big) - (\#I_q) \cdot \log \big|\hat{\mathbf{\Sigma}}(\beta, q)^{-1}\big|$$

   end

2. Compute the average CV fit: $\mathrm{CV}(\beta) = \frac{1}{Q} \sum_{q=1}^{Q} \mathrm{CV}(\beta, q)$, $\forall \beta \in [\beta]$.
3. Select $\hat{\beta}_{\mathrm{CV}} = \arg\min_{\beta \in [\beta]} \mathrm{CV}(\beta)$.
4. Compute $\hat{\mathbf{\Sigma}}$ based on the entire data set $\mathbf{X}$ using $\beta = \hat{\beta}_{\mathrm{CV}}$.

Oracle/Clairvoyant approach

- The regularized M-estimator of scatter is a solution of

$$\hat{\mathbf{\Sigma}} = \beta \frac{1}{n} \sum_{i=1}^{n} u(\mathbf{x}_i^\top \hat{\mathbf{\Sigma}}^{-1} \mathbf{x}_i) \mathbf{x}_i \mathbf{x}_i^\top + \alpha \mathbf{I}$$

  for given penalty parameters $\alpha, \beta > 0$.

- Since $\hat{\mathbf{\Sigma}}$ does not have a closed form expression, deriving an analytic expression for an expected loss (such as the MSE) is virtually impossible.

- **Solution**: use an approximation for $\hat{\Sigma}$.

- Given the true scatter (covariance) matrix $\mathbf{\Sigma}_0$, then

$$\mathbf{\Sigma}_{\alpha,\beta} = (1-\alpha) \frac{1}{n} \sum_{i=1}^{n} u(\mathbf{x}_i^\top \mathbf{\Sigma}_0^{-1} \mathbf{x}_i) \mathbf{x}_i \mathbf{x}_i^\top + \alpha \mathbf{I}$$

  called the clairvoyant estimator, is a closed-form approximation of $\hat{\Sigma}$.

Oracle/Clairvoyant approach

- The regularized M-estimator of scatter is a solution of

$$\hat{\mathbf{\Sigma}} = \beta \frac{1}{n} \sum_{i=1}^{n} u(\mathbf{x}_i^\top \hat{\mathbf{\Sigma}}^{-1} \mathbf{x}_i) \mathbf{x}_i \mathbf{x}_i^\top + \alpha \mathbf{I}$$

  for given penalty parameters $\alpha, \beta > 0$.
- Since $\hat{\mathbf{\Sigma}}$ does not have a closed form expression, deriving an analytic expression for an expected loss (such as the MSE) is virtually impossible.
- **Solution**: use an approximation for $\hat{\mathbf{\Sigma}}$.
- Given the true scatter (covariance) matrix $\mathbf{\Sigma}_0$, then

$$\mathbf{\Sigma}_{\alpha,\beta} = (1-\alpha) \frac{1}{n} \sum_{i=1}^{n} u(\mathbf{x}_i^\top \mathbf{\Sigma}_0^{-1} \mathbf{x}_i) \mathbf{x}_i \mathbf{x}_i^\top + \alpha \mathbf{I}$$

  called the clairvoyant estimator, is a closed-form approximation of $\hat{\mathbf{\Sigma}}$.

Oracle/Clairvoyant approach

- The regularized M-estimator of scatter is a solution of

$$\hat{\mathbf{\Sigma}} = \beta \frac{1}{n} \sum_{i=1}^{n} u(\mathbf{x}_i^\top \hat{\mathbf{\Sigma}}^{-1} \mathbf{x}_i) \mathbf{x}_i \mathbf{x}_i^\top + \alpha \mathbf{I}$$

  for given penalty parameters $\alpha, \beta > 0$.
- Since $\hat{\mathbf{\Sigma}}$ does not have a closed form expression, deriving an analytic expression for an expected loss (such as the MSE) is virtually impossible.
- **Solution**: use an approximation for $\hat{\mathbf{\Sigma}}$.
- Given the true scatter (covariance) matrix $\mathbf{\Sigma}_0$, then

$$\mathbf{\Sigma}_{\alpha,\beta} = (1-\alpha) \frac{1}{n} \sum_{i=1}^{n} u(\mathbf{x}_i^\top \mathbf{\Sigma}_0^{-1} \mathbf{x}_i) \mathbf{x}_i \mathbf{x}_i^\top + \alpha \mathbf{I}$$

  called the clairvoyant estimator, is a closed-form approximation of $\hat{\mathbf{\Sigma}}$.

Oracle/Clairvoyant approach (cont'd)

Let $\mathbf{\Sigma}_0$ denote the true unknown scatter (covariance) matrix. Oracle penalty parameters (α_0, β_0) minimize the expected loss,

$$(\alpha_o, \beta_o) = \arg \min_{\alpha,\beta>0} \mathbb{E}\big[d(\mathbf{\Sigma}_{\alpha,\beta}, \mathbf{\Sigma}_0)\big],$$

for some suitable distance function $d(\mathbf{A}, \mathbf{B})$.

- Naturally the found solution (α_o, β_o) will depend on the true scatter matrix $\mathbf{\Sigma}_0$ which is unknown in practise.
- Replace the unknown true $\mathbf{\Sigma}_0$ in α_0 and β_0 with some preliminary estimate or guess $\hat{\mathbf{\Sigma}}_0$

  $\Rightarrow \hat{\alpha}_o = \alpha_o(\hat{\mathbf{\Sigma}}_0)$ and $\hat{\beta} = \beta_0(\hat{\mathbf{\Sigma}}_0)$ are the oracle/clairvoyant estimates.

Oracle approach for the regularized SCM

- Let us first consider the regularized M-estimator using (tuned) Gaussian loss function $\rho(t) = \beta t$ ($u(t) = \beta$)
- This gives the regularized SCM (a.k.a. Ledoit-Wolf estimator) :

$$\mathbf{S}_{\alpha,\beta} = \beta\mathbf{S} + \alpha\mathbf{I},$$

where $\mathbf{S}$ is the unbiased SCM $\mathbf{S} = \frac{1}{n-1}\sum_{i=1}^{n} \mathbf{x}_i\mathbf{x}_i^\top$, so $\mathbb{E}[\mathbf{S}] = \boldsymbol{\Sigma}$

- Clairvoyant (oracle) estimator of $(\alpha, \beta) \in \mathbb{R}_0^+ \times \mathbb{R}^+$ is defined as a minimizer of the mean squared error (MSE),

$$(\alpha_o, \beta_o) = \arg\min_{(\alpha,\beta)} \ \mathbb{E}\Big[\|\boldsymbol{\Sigma}_0 - \mathbf{S}_{\alpha,\beta}\|_{\mathrm{F}}^2\Big].$$

See e.g., [Du et al., 2010].

- The solution will naturally depend on the true unknown covariance matrix $\boldsymbol{\Sigma}_0$.

- The oracle solution is easily found to be

$$\beta_o = \frac{\gamma(\mathbf{\Sigma}_0)}{\gamma(\mathbf{\Sigma}_0) + \varrho(\mathbf{\Sigma}_0)}$$
$$\alpha_o = (1 - \beta_0)\mathrm{Tr}(\mathbf{\Sigma}_0)/p.$$

$$\text{where } \varrho(\mathbf{\Sigma}_0) = \mathbb{E}[\|\mathbf{S} - \mathbf{\Sigma}_0\|_{\mathrm{F}}^2] = \frac{1}{n}\mathbb{E}[\|\mathbf{x}\|^4] - \frac{1}{n}\mathrm{Tr}(\mathbf{\Sigma}_0^2)$$

$$\gamma(\mathbf{\Sigma}) = \mathrm{Tr}(\mathbf{\Sigma}_0^2) - \frac{1}{p}\big\{\mathrm{Tr}(\mathbf{\Sigma}_0)\big\}^2$$

- **Note**: $\varrho(\mathbf{\Sigma}_0)$ is the MSE between $\mathbf{S}$ and $\mathbf{\Sigma}_0$ and $\beta \in [0, 1]$ and $\alpha \geq 0$.
- **Estimation**: replace the unknown $\mathbf{\Sigma}$ by $\mathbf{S}$, and $\mathbb{E}[\|\mathbf{x}\|^4]$ by finite sample average.
- For more advanced approach, see [Ledoit and Wolf, 2004].

- The oracle solution is easily found to be

$$\hat{\beta}_o = \frac{\gamma(\mathbf{S})}{\gamma(\mathbf{S}) + \varrho(\mathbf{S})}$$
$$\hat{\alpha}_o = (1 - \hat{\beta}_0)\mathrm{Tr}(\mathbf{S})/p \quad .$$

$$\text{where } \varrho(\mathbf{S}) \quad = \mathbb{E}[\|\mathbf{S} - \mathbf{\Sigma}_0\|_{\mathrm{F}}^2] = \frac{1}{n^2}\sum_{i=1}^{n}\|\mathbf{x}_i\|^4 - \frac{1}{n}\mathrm{Tr}(\mathbf{S}^2)$$
$$\gamma(\mathbf{S}) \ = \mathrm{Tr}(\mathbf{S}^2) - \frac{1}{p}\big\{\mathrm{Tr}(\mathbf{S})\big\}^2$$

- **Note**: $\varrho(\mathbf{\Sigma}_0)$ is the MSE between $\mathbf{S}$ and $\mathbf{\Sigma}_0$ and $\beta \in [0, 1]$ and $\alpha \geq 0$.
- **Estimation**: replace the unknown $\mathbf{\Sigma}$ by $\mathbf{S}$, and $\mathbb{E}[\|\mathbf{x}\|^4]$ by finite sample average.
- For more advanced approach, see [Ledoit and Wolf, 2004].

- The oracle solution is easily found to be

$$\hat{\beta}_o = \frac{\gamma(\mathbf{S})}{\gamma(\mathbf{S}) + \varrho(\mathbf{S})}$$
$$\hat{\alpha}_o = (1 - \hat{\beta}_0)\mathrm{Tr}(\mathbf{S})/p \quad .$$

$$\text{where } \varrho(\mathbf{S}) \quad = \mathbb{E}[\|\mathbf{S} - \mathbf{\Sigma}_0\|_{\mathrm{F}}^2] = \frac{1}{n^2}\sum_{i=1}^{n}\|\mathbf{x}_i\|^4 - \frac{1}{n}\mathrm{Tr}(\mathbf{S}^2)$$
$$\gamma(\mathbf{S}) \ = \mathrm{Tr}(\mathbf{S}^2) - \frac{1}{p}\big\{\mathrm{Tr}(\mathbf{S})\big\}^2$$

- **Note**: $\varrho(\mathbf{\Sigma}_0)$ is the MSE between $\mathbf{S}$ and $\mathbf{\Sigma}_0$ and $\beta \in [0, 1]$ and $\alpha \geq 0$.
- **Estimation**: replace the unknown $\mathbf{\Sigma}$ by $\mathbf{S}$, and $\mathbb{E}[\|\mathbf{x}\|^4]$ by finite sample average.
- For more advanced approach, see [Ledoit and Wolf, 2004].

Oracle approach for regularized Tyler's M-estimator

- Choose $\beta = 1 - \alpha$, $\alpha \in (0, 1)$.
- The clairvoyant estimator of Tyler's M-estimator $\hat{\mathbf{\Sigma}}$ is

$$\mathbf{\Sigma}_\alpha = (1-\alpha)\frac{p}{n}\sum_{i=1}^{n}\frac{\mathbf{x}_i\mathbf{x}_i^\top}{\mathbf{x}_i^\top\mathbf{\Sigma}_0^{-1}\mathbf{x}_i} + \alpha\mathbf{I}.$$

- **Q:** What distance d to use in minimizing the expected loss?
- **idea**: Given a shape matrix $\mathbf{\Sigma}_0$, verifying $\mathrm{Tr}(\mathbf{\Sigma}_0^{-1}) = p$, choose α so that $\mathbf{\Sigma}_0^{-1}\mathbf{\Sigma}_\alpha$ *is as close as possible* to $c\mathbf{I}$, for some $c > 0$.
- A natural distance that measures similarity in shape:

$$d(\mathbf{\Sigma}_0, \mathbf{\Sigma}_\alpha) = \|\mathbf{\Sigma}_0^{-1}\mathbf{\Sigma}_\alpha - \tfrac{1}{p}\mathrm{Tr}(\mathbf{\Sigma}_0^{-1}\mathbf{\Sigma}_\alpha)\mathbf{I}\|^2$$

- The value $\alpha_o \in (0, 1)$ that minimizes the expected loss $\mathbb{E}[d(\mathbf{\Sigma}_0, \mathbf{\Sigma}_\alpha)]$ is [Ollila and Tyler, 2014, Theorem 5]:

$$\alpha_o = \frac{p - 2 + p\mathrm{Tr}(\mathbf{\Sigma}_0)}{p - 2 + p\mathrm{Tr}(\mathbf{\Sigma}_0) + n(p+2)\{p^{-1}\mathrm{Tr}(\mathbf{\Sigma}_0^{-2}) - 1\}}$$

Oracle approach for regularized Tyler's M-estimator

- Choose $\beta = 1 - \alpha$, $\alpha \in (0, 1)$.
- The clairvoyant estimator of Tyler's M-estimator $\hat{\mathbf{\Sigma}}$ is

$$\mathbf{\Sigma}_\alpha = (1-\alpha)\frac{p}{n}\sum_{i=1}^{n}\frac{\mathbf{x}_i\mathbf{x}_i^\top}{\mathbf{x}_i^\top\mathbf{\Sigma}_0^{-1}\mathbf{x}_i} + \alpha\mathbf{I}.$$

- **Q:** What distance d to use in minimizing the expected loss?
- **idea**: Given a shape matrix $\mathbf{\Sigma}_0$, verifying $\mathrm{Tr}(\mathbf{\Sigma}_0^{-1}) = p$, choose α so that $\mathbf{\Sigma}_0^{-1}\mathbf{\Sigma}_\alpha$ *is as close as possible* to $c\mathbf{I}$, for some $c > 0$.
- A natural distance that measures similarity in shape:

$$d(\mathbf{\Sigma}_0, \mathbf{\Sigma}_\alpha) = \|\mathbf{\Sigma}_0^{-1}\mathbf{\Sigma}_\alpha - \tfrac{1}{p}\mathrm{Tr}(\mathbf{\Sigma}_0^{-1}\mathbf{\Sigma}_\alpha)\mathbf{I}\|^2$$

- The value $\alpha_o \in (0, 1)$ that minimizes the expected loss $\mathbb{E}[d(\mathbf{\Sigma}_0, \mathbf{\Sigma}_\alpha)]$ is [Ollila and Tyler, 2014, Theorem 5]:

$$\alpha_o = \frac{p - 2 + p\mathrm{Tr}(\mathbf{\Sigma}_0)}{p - 2 + p\mathrm{Tr}(\mathbf{\Sigma}_0) + n(p+2)\{p^{-1}\mathrm{Tr}(\mathbf{\Sigma}_0^{-2}) - 1\}}$$

Oracle approach for regularized Tyler's M-estimator

- Choose $\beta = 1 - \alpha$, $\alpha \in (0, 1)$.
- The clairvoyant estimator of Tyler's M-estimator $\hat{\mathbf{\Sigma}}$ is

$$\mathbf{\Sigma}_\alpha = (1-\alpha)\frac{p}{n}\sum_{i=1}^{n}\frac{\mathbf{x}_i\mathbf{x}_i^\top}{\mathbf{x}_i^\top\mathbf{\Sigma}_0^{-1}\mathbf{x}_i} + \alpha\mathbf{I}.$$

- **Q:** What distance d to use in minimizing the expected loss?
- **idea**: Given a shape matrix $\mathbf{\Sigma}_0$, verifying $\mathrm{Tr}(\mathbf{\Sigma}_0^{-1}) = p$, choose α so that $\mathbf{\Sigma}_0^{-1}\mathbf{\Sigma}_\alpha$ *is as close as possible* to $c\mathbf{I}$, for some $c > 0$.
- A natural distance that measures similarity in shape:

$$d(\mathbf{\Sigma}_0, \mathbf{\Sigma}_\alpha) = \|\mathbf{\Sigma}_0^{-1}\mathbf{\Sigma}_\alpha - \tfrac{1}{p}\mathrm{Tr}(\mathbf{\Sigma}_0^{-1}\mathbf{\Sigma}_\alpha)\mathbf{I}\|^2$$

- The value $\alpha_o \in (0, 1)$ that minimizes the expected loss $\mathbb{E}[d(\mathbf{\Sigma}_0, \mathbf{\Sigma}_\alpha)]$ is [Ollila and Tyler, 2014, Theorem 5]:

$$\alpha_o = \frac{p - 2 + p\mathrm{Tr}(\mathbf{\Sigma}_0)}{p - 2 + p\mathrm{Tr}(\mathbf{\Sigma}_0) + n(p+2)\{p^{-1}\mathrm{Tr}(\mathbf{\Sigma}_0^{-2}) - 1\}}$$

Oracle approach for regularized Tyler's M-estimator

- Choose $\beta = 1 - \alpha$, $\alpha \in (0, 1)$.
- The clairvoyant estimator of Tyler's M-estimator $\hat{\mathbf{\Sigma}}$ is

$$\mathbf{\Sigma}_\alpha = (1-\alpha)\frac{p}{n}\sum_{i=1}^{n}\frac{\mathbf{x}_i\mathbf{x}_i^\top}{\mathbf{x}_i^\top\mathbf{\Sigma}_0^{-1}\mathbf{x}_i} + \alpha\mathbf{I}.$$

- **Q:** What distance d to use in minimizing the expected loss?
- **idea**: Given a shape matrix $\mathbf{\Sigma}_0$, verifying $\mathrm{Tr}(\mathbf{\Sigma}_0^{-1}) = p$, choose α so that $\mathbf{\Sigma}_0^{-1}\mathbf{\Sigma}_\alpha$ *is as close as possible* to $c\mathbf{I}$, for some $c > 0$.
- A natural distance that measures similarity in shape:

$$d(\mathbf{\Sigma}_0, \mathbf{\Sigma}_\alpha) = \|\mathbf{\Sigma}_0^{-1}\mathbf{\Sigma}_\alpha - \tfrac{1}{p}\mathrm{Tr}(\mathbf{\Sigma}_0^{-1}\mathbf{\Sigma}_\alpha)\mathbf{I}\|^2$$

- The value $\alpha_o \in (0, 1)$ that minimizes the expected loss $\mathbb{E}[d(\mathbf{\Sigma}_0, \mathbf{\Sigma}_\alpha)]$ is [Ollila and Tyler, 2014, Theorem 5]:

$$\alpha_o = \frac{p - 2 + p\mathrm{Tr}(\mathbf{\Sigma}_0)}{p - 2 + p\mathrm{Tr}(\mathbf{\Sigma}_0) + n(p+2)\{p^{-1}\mathrm{Tr}(\mathbf{\Sigma}_0^{-2}) - 1\}}$$

Oracle approach for regularized Tyler's M-estimator

- Similar expression is derived for complex-valued case also in [Ollila and Tyler, 2014].
- Estimate $\hat{\alpha}_o = \alpha_o(\hat{\mathbf{\Sigma}}_0)$ is obtained by using an estimate $\hat{\mathbf{\Sigma}}_0$ which is
  - Conventional (non-regularized) Tyler's M-estimator normalized so that $\mathrm{Tr}(\hat{\mathbf{\Sigma}}_0^{-1}) = p$ when $n \geq p$
  - Regularized Tyler's M-estimator using $\beta < n/p$ and $\alpha = 1 - \beta$ when $n < p$.

Oracle approach for DL-FP estimator

$$\boldsymbol{\Sigma}_\alpha = (1-\alpha)\frac{p}{n}\sum_{i=1}^{n}\frac{\mathbf{x}_i\mathbf{x}_i^\top}{\mathbf{x}_i^\top\boldsymbol{\Sigma}_0^{-1}\mathbf{x}_i} + \alpha\mathbf{I}.$$

[Chen et al., 2011] proposed an oracle estimator for the tuning parameter of DL-FP estimator ▸ defined in this slide

- Given a shape matrix $\boldsymbol{\Sigma}_0$, verifying $\mathrm{Tr}(\boldsymbol{\Sigma}_0) = p$, find α as

$$\alpha_o = \arg\min_\alpha \mathbb{E}\left[\|\boldsymbol{\Sigma}_0 - \boldsymbol{\Sigma}_\alpha\|_{\mathrm{F}}^2\right]$$

- The obtained oracle estimator is (in the real-valued case):

$$\alpha_o = \frac{p^3 + (p-2)\mathrm{Tr}(\boldsymbol{\Sigma}_0^2)}{\{p^3 + (p-2)\mathrm{Tr}(\boldsymbol{\Sigma}_0^2)\} + n(p+2)(\mathrm{Tr}(\boldsymbol{\Sigma}_0^2) - p)}.$$

- Estimate $\hat{\alpha}_o = \alpha_o(\hat{\boldsymbol{\Sigma}}_0)$ is obtained using trace normalized sample sign covariance matrix

$$\hat{\boldsymbol{\Sigma}}_0 = \frac{p}{n}\sum_{i=1}^{n}\frac{\mathbf{x}_i\mathbf{x}_i^\top}{\|\mathbf{x}_i\|^2}.$$

Numerical Example

- n realizations from a p-variate (complex) Gaussian distribution with covariance matrix

$$[\mathbf{\Sigma}]_{ij} = \rho^{|i-j|}, \quad \rho \in (0,1).$$

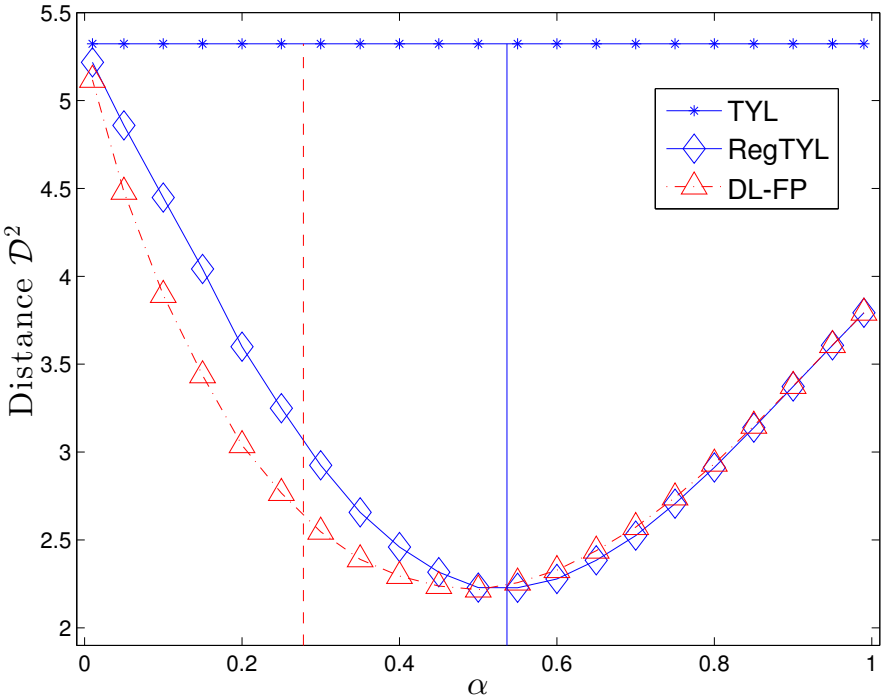
- Distance measure:

$$\mathcal{D}^2 \equiv \mathcal{D}^2(\mathbf{\Sigma}, \hat{\mathbf{\Sigma}}) = \|\{p/\mathrm{Tr}(\mathbf{\Sigma}^{-1}\hat{\mathbf{\Sigma}})\}\,\mathbf{\Sigma}^{-1}\hat{\mathbf{\Sigma}} - \mathbf{I}\|_{\mathrm{F}}^2$$

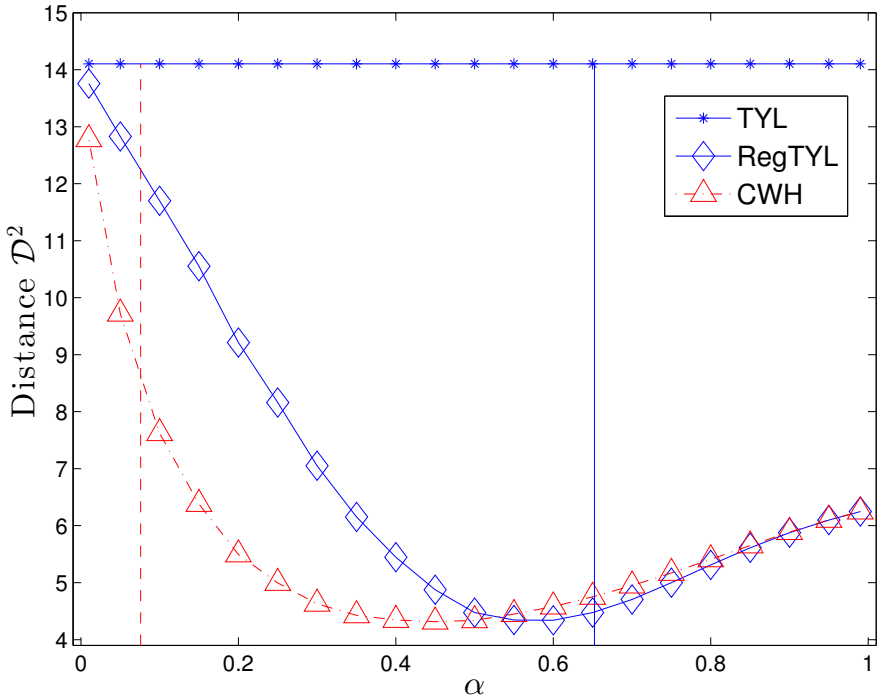
  measures the ability $\hat{\mathbf{\Sigma}}$ to estimate $\mathbf{\Sigma}$ up to a scale.
- Simulation report: Average $\mathcal{D}^2(\mathbf{\Sigma}, \hat{\mathbf{\Sigma}})$ over 1000 MC-trials for both the regularized Tyler M-estimator (using $\beta = 1 - \alpha$) and DL-FP estimator for different values of penalty parameter $\alpha \in (0,1)$.

Numerical Example (cont'd)

$n = 48,\ p = 12$



$\rho = 0.5$



$\rho = 0.8$

- Solid vertical line: oracle value α_o for the regularized Tyler's M-estimator given in Slide 92.
- Dotted vertical line: oracle value α_o of DL-FP given in Slide 94.

Menu

I. Ad-hoc shrinkage SCM-s of multiple samples

II. ML- and M-estimators of scatter matrix

III. Geodesic convexity

IV. Regularized M-estimators

V. Penalized estimation of multiple covariances

VI. Estimation of the regularization parameter

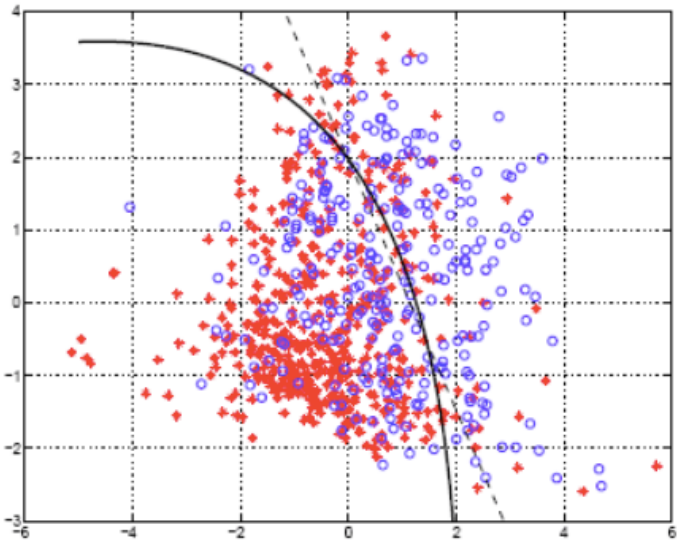
VII. Applications

- Regularized discriminant analysis
- Matched filter detection

Quadratic discriminant analysis (QDA)

QDA assigns $\mathbf{x}$ to a group $\hat{k}$:

$$\hat{k} = \min_{1 \leq k \leq K} \left\{ (\mathbf{x} - \bar{\mathbf{x}}_k)^\top \mathbf{S}_k^{-1} (\mathbf{x} - \bar{\mathbf{x}}_k) + \ln |\mathbf{S}_k| \right\}.$$



where

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

is the SCM of a training data set $\mathbf{X}_k$ from kth population ($k = 1, \ldots, K$).

Assumptions:

- Gaussian populations $\mathcal{N}_p(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$
- Covariance matrices can be *different* for each class $\boldsymbol{\Sigma}_i \neq \boldsymbol{\Sigma}_j \; i \neq j$

Linear discriminant analysis (LDA)

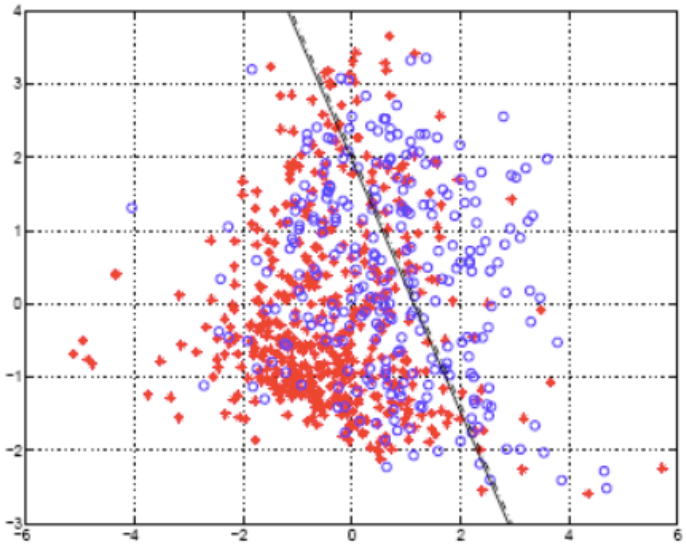
LDA assigns $\mathbf{x}$ to a group $\hat{k}$:

$$\hat{k} = \min_{1 \leq k \leq K} \left\{ (\mathbf{x} - \bar{\mathbf{x}}_k)^\top \mathbf{S}^{-1} (\mathbf{x} - \hat{\boldsymbol{\mu}}_k) \right\}.$$

where

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

is the pooled SCM estimator.



Assumptions:

- Gaussian populations $\mathcal{N}_p(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$
- Covariance matrices are the *same* for each class $\boldsymbol{\Sigma}_i = \boldsymbol{\Sigma}_j \; i \neq j$

Regularized Discriminant Analysis (RDA)

RDA* assigns $\mathbf{x}$ to a group $\hat{k}$:

$$\hat{k} = \min_{1 \leq k \leq K} \left\{ (\mathbf{x} - \hat{\boldsymbol{\mu}}_k)^\top [\hat{\boldsymbol{\Sigma}}_k(\beta)]^{-1} (\mathbf{x} - \hat{\boldsymbol{\mu}}_k) + \ln |\hat{\boldsymbol{\Sigma}}_k(\beta)| \right\}.$$

where $\hat{\boldsymbol{\Sigma}}_k(\beta)$ are the penalized estimators of scatter matrices obtained either using Proposal 1 or Proposal 2.

Interpretation:

- if $\beta \to 1$, we do not shrink towards joint center
  $\Rightarrow$ RDA $\to$ QDA
- if $\beta \to 0$, we shrink towards joint center
  $\Rightarrow$ RDA $\to$ LDA
- $0 < \beta < 1 \Rightarrow$ a compromise between LDA and QDA.

For robust loss fnc-s, we use spatial median as an estimate $\hat{\boldsymbol{\mu}}_k$ of location

* Inspired by Friedman, "Regularized discriminant Analysis", JASA (1989)

Simulation set-up

- We use the same loss function $\rho = \rho_k$ for each K samples
- **Training data**: $\mathbf{X}_k$-s generated from $\mathcal{N}_p(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$ or $t_\nu(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$, $\nu = 2$. These are used to estimate the discriminant rules.
- **Test data**: generated in exactly the same manner (same size $N = 100$) and classified with the discriminant rules thereby yielding an estimate of the misclassification risk.
- RDA rules are computed over a grid of penalty parameters $\beta \in (0, 1)$ and optimal (*smallest*) misclassification risk is reported.
- Prop1(ρ, d) and Prop2(ρ, d) refer to RDA rules based on Proposal 1 and Proposal 2 estimators, respectively, where
  - ρ refers to the used loss fnc: Gaussian, Huber's, Tyler's)
  - d refers to the used distance fnc: (KL or Ellipticity).

Unequal spherical covariance matrices ($\Sigma_k = k\mathbf{I}$)

- Nr of classes is $K = 3$, total sample size $N = \sum_{k=1}^{K} n_k = 100$.
- $(n_1, n_2, n_3) \sim \text{Multin}(N; p_1 = p_2 = \frac{1}{4}, p_3 = \frac{1}{2})$.
- $\boldsymbol{\mu}_1 = \mathbf{0}$ and remaining classes $\boldsymbol{\mu}_k$ have norm equal to $\delta_k = \|\boldsymbol{\mu}_k\| = 3 + k$ in orthogonal directions

Gaussian case: test misclassification errors %

method	$p = 10$	$p = 20$	$p = 30$
Oracle1	$8.8_{(2.6)}$	$6.2_{(2.3)}$	$4.6_{(1.9)}$
Oracle2	$9.8_{(3.1)}$	$7.6_{(2.6)}$	$6.0_{(2.3)}$
QDA	$19.9_{(4.4)}$	–	–
LDA	$17.1_{(3.8)}$	$20.5_{(4.3)}$	$24.0_{(4.9)}$
Prop1(G,KL)	$12.2_{(3.1)}$	$14.6_{(3.5)}$	$17.9_{(4.3)}$
Prop1(H,KL)	$12.4_{(3.2)}$	$14.6_{(3.5)}$	$17.7_{(4.1)}$
Prop1(T,E)	$10.9_{(3.1)}$	$12.1_{(3.3)}$	$16.5_{(3.9)}$
Prop2(G,E)	$10.5_{(3.0)}$	$11.5_{(3.3)}$	$15.9_{(3.8)}$
Prop2(T,E)	$10.9_{(3.1)}$	$12.1_{(3.3)}$	$16.5_{(3.9)}$
Prop2(H,E)	$10.5_{(3.0)}$	$11.6_{(3.3)}$	$15.7_{(3.8)}$
Prop2(H,KL)	$12.3_{(3.2)}$	$14.8_{(3.6)}$	$18.0_{(4.1)}$

standard deviations inside parantheses in subscript

Oracle1 = QDA rule using true $\boldsymbol{\mu}_k$ and $\boldsymbol{\Sigma}_k$.

Oracle2 = QDA rule using true $\boldsymbol{\Sigma}_k$, but estimated $\hat{\boldsymbol{\mu}}_k$.

- = sample means in Gaussian case
- = spatial median in t_2 case

Unequal spherical covariance matrices ($\Sigma_k = k\mathbf{I}$)

- Nr of classes is $K = 3$, total sample size $N = \sum_{k=1}^{K} n_k = 100$.
- $(n_1, n_2, n_3) \sim \text{Multin}(N; p_1 = p_2 = \frac{1}{4}, p_3 = \frac{1}{2})$.
- $\boldsymbol{\mu}_1 = \mathbf{0}$ and remaining classes $\boldsymbol{\mu}_k$ have norm equal to $\delta_k = \|\boldsymbol{\mu}_k\| = 4 + k$ in orthogonal directions

t_2 **case**: test misclassification errors %

Oracle1	$15.7_{(3.8)}$	$18.2_{(3.9)}$	$21.1_{(4.0)}$
Oracle2	$16.2_{(3.5)}$	$19.1_{(4.2)}$	$21.9_{(4.1)}$
QDA	$26.9_{(5.2)}$	–	–
LDA	$21.8_{(4.9)}$	$25.3_{(5.3)}$	$27.7_{(5.3)}$
Prop1(G,KL)	$19.7_{(4.8)}$	$22.7_{(5.2)}$	$24.7_{(5.1)}$
Prop1(H,KL)	$15.5_{(3.7)}$	$17.9_{(4.0)}$	$20.3_{(4.1)}$
Prop1(T,E)	$16.8_{(4.0)}$	$20.4_{(4.3)}$	$23.4_{(4.7)}$
Prop2(G,E)	$22.3_{(5.9)}$	$24.3_{(5.1)}$	$25.9_{(4.8)}$
Prop2(T,E)	$16.8_{(4.0)}$	$20.4_{(4.4)}$	$23.5_{(4.8)}$
Prop2(H,E)	$16.6_{(3.9)}$	$20.2_{(4.4)}$	$23.6_{(4.6)}$
Prop2(H,KL)	$15.5_{(3.7)}$	$17.9_{(4.0)}$	$20.5_{(4.1)}$

standard deviations inside parantheses in subscript

Oracle1 = QDA rule using true $\boldsymbol{\mu}_k$ and $\boldsymbol{\Sigma}_k$.

Oracle2 = QDA rule using true $\boldsymbol{\Sigma}_k$, but estimated $\hat{\boldsymbol{\mu}}_k$:

- = sample means in Gaussian case
- = spatial median in t_2 case

Comments

- I do not want to bug you with more simulations...
- I just mention that, we can perform *much better* than estimators regularized sample covariance matrices (SCM-s) $\mathbf{S}_k(\beta)$ with shrinkage towards pooled SCM $\mathbf{S}$ (as in Friedman's RDA) even when the clusters follow Gaussian distributions.

Why?

- We use more natural Riemannian geometry and our class of joint regularized estimators is **huge**:

✓ many different g-convex penalty fnc's $d(\mathbf{A}, \mathbf{B})$: Kullback-Leibler , Ellipticity, Riemannian distance, ...

✓ many different g-convex loss fnc's $\rho(t)$: Gaussian, Tyler's, Huber's, ...

✓ robust: good performance under non-Gaussianity or outliers

Comments

- I do not want to bug you with more simulations...
- I just mention that, we can perform *much better* than estimators regularized sample covariance matrices (SCM-s) $\mathbf{S}_k(\beta)$ with shrinkage towards pooled SCM $\mathbf{S}$ (as in Friedman's RDA) even when the clusters follow Gaussian distributions.

Why?

- We use more natural Riemannian geometry and our class of joint regularized estimators is **huge**:

✓ many different g-convex penalty fnc's $d(\mathbf{A}, \mathbf{B})$: Kullback-Leibler , Ellipticity, Riemannian distance, ...

✓ many different g-convex loss fnc's $\rho(t)$: Gaussian, Tyler's, Huber's, ...

✓ robust: good performance under non-Gaussianity or outliers

A short recap on complex numbers/calculus

We consider *complex-valued* measurements. Our results are also valid for the real-valued case.

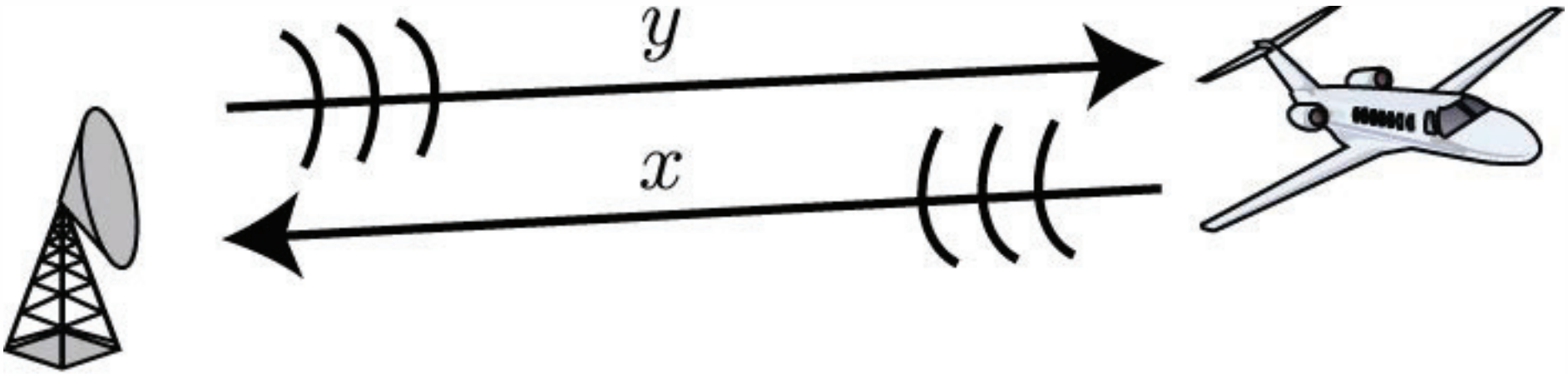
- A complex vector $\mathbf{a} = \mathbf{a}_R + \jmath\mathbf{a}_I$ in complex Euclidean p-space $\mathbb{C}^p$, where $\jmath = \sqrt{-1}$ is the imaginary unit.
- A complex conjugate: $\mathbf{a}^* = \mathbf{a}_R - \jmath\mathbf{a}_I$
- (Hermitian) *norm*: $\|\mathbf{a}\| = \sqrt{\mathbf{a}^\top\mathbf{a}}$, where $(\cdot)^\top = [(\cdot)^*]^\top$ denotes the *Hermitian (conjugate) transpose*.
- Modulus of $a = a_R + \jmath a_I \in \mathbb{C}$ is $|a|^2 = \sqrt{aa^*} = \sqrt{a_R^2 + a_I^2}$.
- A complex matrix $\mathbf{A} = \mathbf{A}_R + \jmath\mathbf{A}_I$ is *Hermitian* if $\mathbf{A}^\top = \mathbf{A}$ which implies that $\mathbf{A}_R$ is symmetric ($\mathbf{A}_R^\top = \mathbf{A}_R$) and $\mathbf{A}_I$ is skew-symmetric ($\mathbf{A}_I^\top = -\mathbf{A}_I$).

Matched filter detection

- Matched filter (MF) is a linear filter that correlates a known "signal" $\mathbf{p}$ with a received (measured) data $\mathbf{x}$ with the purpose of testing

$$\mathcal{H}_0 : \text{signal-absent (noise only)}$$
$$\text{vs } \mathcal{H}_1 : \text{signal-present (signal + noise)} .$$

- Optimal = maximizes the signal to noise ratio (SNR).
- Decide (absence/precense) by comparing the magnitude of the (normalized) matched filter ouput to a *threshold*.
- Threshold is set to attain the desired probability of false alarm).



- The output of a linear filter with filter weights $c_1, \ldots, c_p$ is

$$y = \mathbf{c}^\top \mathbf{x} = \sum_{i=1}^{p} c_i^* z_i$$

where $\mathbf{x}$ denotes the measured p-variate data $\mathbf{x}$.

- The MF filter weights are

$$\mathbf{c} = \frac{\mathbf{\Sigma}^{-1}\mathbf{p}}{(\mathbf{p}^{\mathsf{H}}\mathbf{\Sigma}^{-1}\mathbf{p})^{1/2}}$$

where $\mathbf{p}$ denotes the (known) transmitted p-variate signal.

- In practise, the noise covariance matrix $\mathbf{\Sigma}$ is unknown and is estimated from a set of *secondary data* $\mathbf{x}_1, \ldots, \mathbf{x}_n$ acquired under noise-only ($\mathcal{H}_0$) environment.
- **Problem**: $p > n$ or $n \approx p$.

Radar detection using NMF

$$\boxed{\mathcal{H}_0 : \mathbf{x} = \mathbf{c}} \quad \text{vs.} \quad \boxed{\mathcal{H}_1 : \mathbf{x} = \gamma\mathbf{p} + \mathbf{c}}$$

- $\mathbf{x}$: p-variate received data.
- $\mathbf{p}$: known complex signal vector
- $\mathbf{c}$: noise r.v., $\mathbf{c} \sim \mathcal{E}_p(\mathbf{0}, \mathbf{\Sigma}, g)$. [complex elliptically symmetric random vector]
- $\gamma \in \mathbb{C}$: signal parameter

EX In radar, γ is accounting for both channel propagation effect and target backscattering; $\mathbf{p}$ is the transmitted known radar pulse vector.

The normalized matched filter (NMF) detector

$$\Lambda = \frac{|\mathbf{c}^{\mathsf{H}}\mathbf{x}|^2}{\|\mathbf{\Sigma}^{-1/2}\mathbf{x}\|^2} = \frac{|\mathbf{p}^{\mathsf{H}}\mathbf{\Sigma}^{-1}\mathbf{x}|^2}{(\mathbf{x}^{\mathsf{H}}\mathbf{\Sigma}^{-1}\mathbf{x})(\mathbf{p}^{\mathsf{H}}\mathbf{\Sigma}^{-1}\mathbf{p})} \underset{\mathcal{H}_0}{\overset{\mathcal{H}_1}{\gtrless}} \lambda$$

- **Null distribution**: $\Lambda \sim_{\mathcal{H}_0} \mathrm{Beta}(1, p-1)$ i.e., distribution-free under the class of CES distributed noise.
- **Threshold**: To obtain a desired level P_{FA}, threshold λ can be set as $(1 - P_{\mathrm{FA}})$th quantile of $\mathrm{Beta}(1, p-1)$:

$$P_{\mathrm{FA}} = \Pr(\Lambda > \lambda | \mathcal{H}_0) = (1-\lambda)^{p-1}$$
$$\Rightarrow \quad \lambda = 1 - {P_{\mathrm{FA}}}^{1/(p-1)}.$$

- Adaptive NMF detector $\hat{\Lambda}$ = replace $\mathbf{\Sigma}$ by its estimate $\hat{\mathbf{\Sigma}}$.
- **Problems**: $p > n$ or $n \approx p$. Furthermore, noise (clutter) in radar applications often heavy-tailed non-Gaussian and hence accurate estimation of $\mathbf{\Sigma}$ is crucial and SCM will do a poor job.

$\Rightarrow$ Adaptive NMF detector often does not retain the constant false alarm rate (CFAR) of NMF nor its probability of detection

- **Solution**: Adaptive NMF based on regularized M-estimator

- **Null distribution**: $\Lambda \sim_{\mathcal{H}_0} \mathrm{Beta}(1, p-1)$ i.e., distribution-free under the class of CES distributed noise.
- **Threshold**: To obtain a desired level P_{FA}, threshold λ can be set as $(1 - P_{\mathrm{FA}})$th quantile of $\mathrm{Beta}(1, p-1)$:

$$P_{\mathrm{FA}} = \Pr(\Lambda > \lambda | \mathcal{H}_0) = (1-\lambda)^{p-1}$$
$$\Rightarrow \quad \lambda = 1 - {P_{\mathrm{FA}}}^{1/(p-1)}.$$

- Adaptive NMF detector $\hat{\Lambda}$ = replace $\mathbf{\Sigma}$ by its estimate $\hat{\mathbf{\Sigma}}$.
- **Problems**: $p > n$ or $n \approx p$. Furthermore, noise (clutter) in radar applications often heavy-tailed non-Gaussian and hence accurate estimation of $\mathbf{\Sigma}$ is crucial and SCM will do a poor job.

$\Rightarrow$ Adaptive NMF detector often does not retain the constant false alarm rate (CFAR) of NMF nor its probability of detection

- **Solution**: Adaptive NMF based on regularized M-estimator

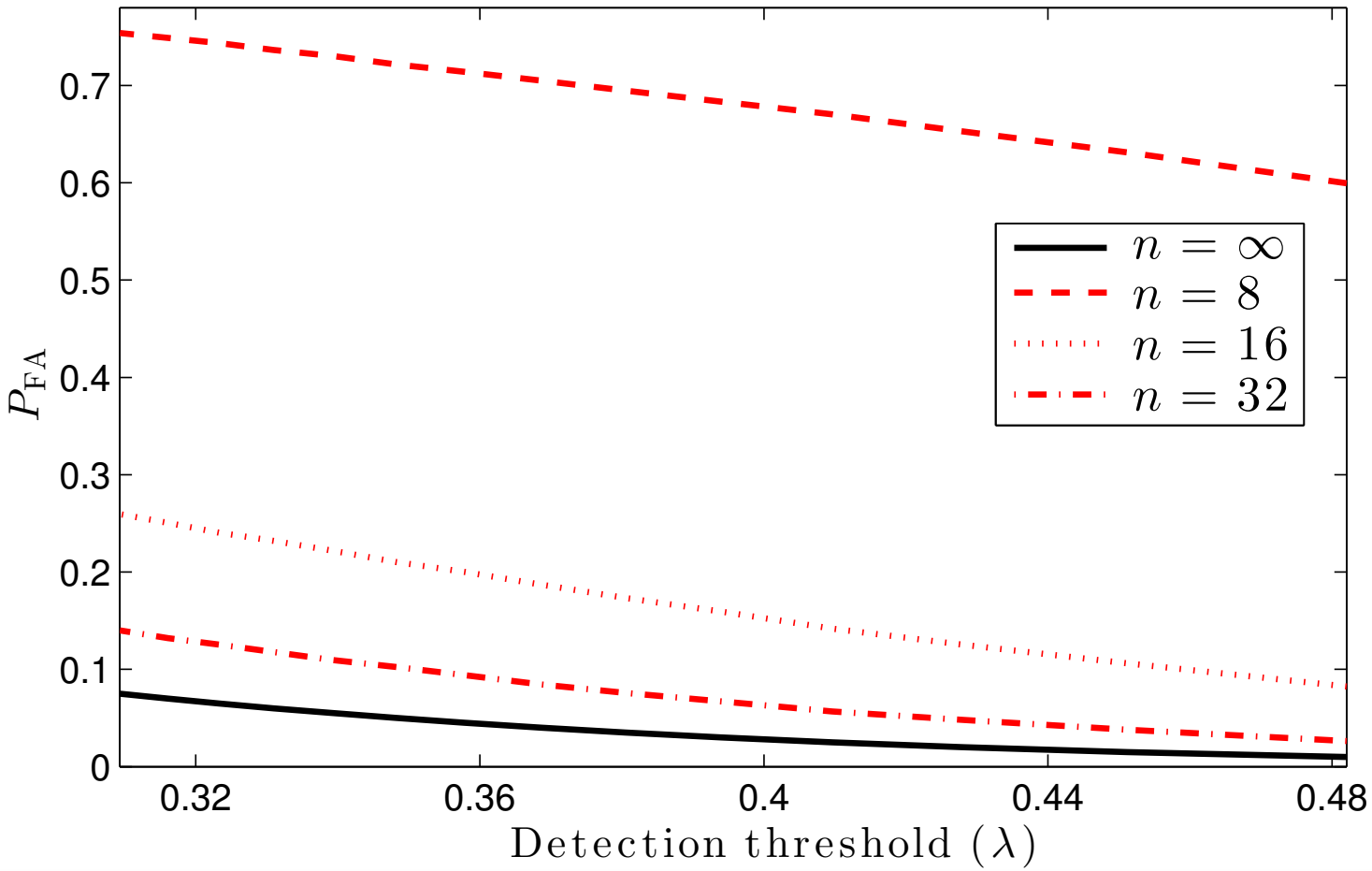
Simulation set-up

Q: How well does the adaptive NMF detector $\hat{\Lambda}$ based on estimated $\boldsymbol{\Sigma}$ maintain the preset PFA?

Simulation set-up: $\mathcal{H}_0$ holds, and $\mathbf{x} = \mathbf{c} \sim \mathbb{C}K_{p,\nu}(\mathbf{0}, \boldsymbol{\Sigma})$

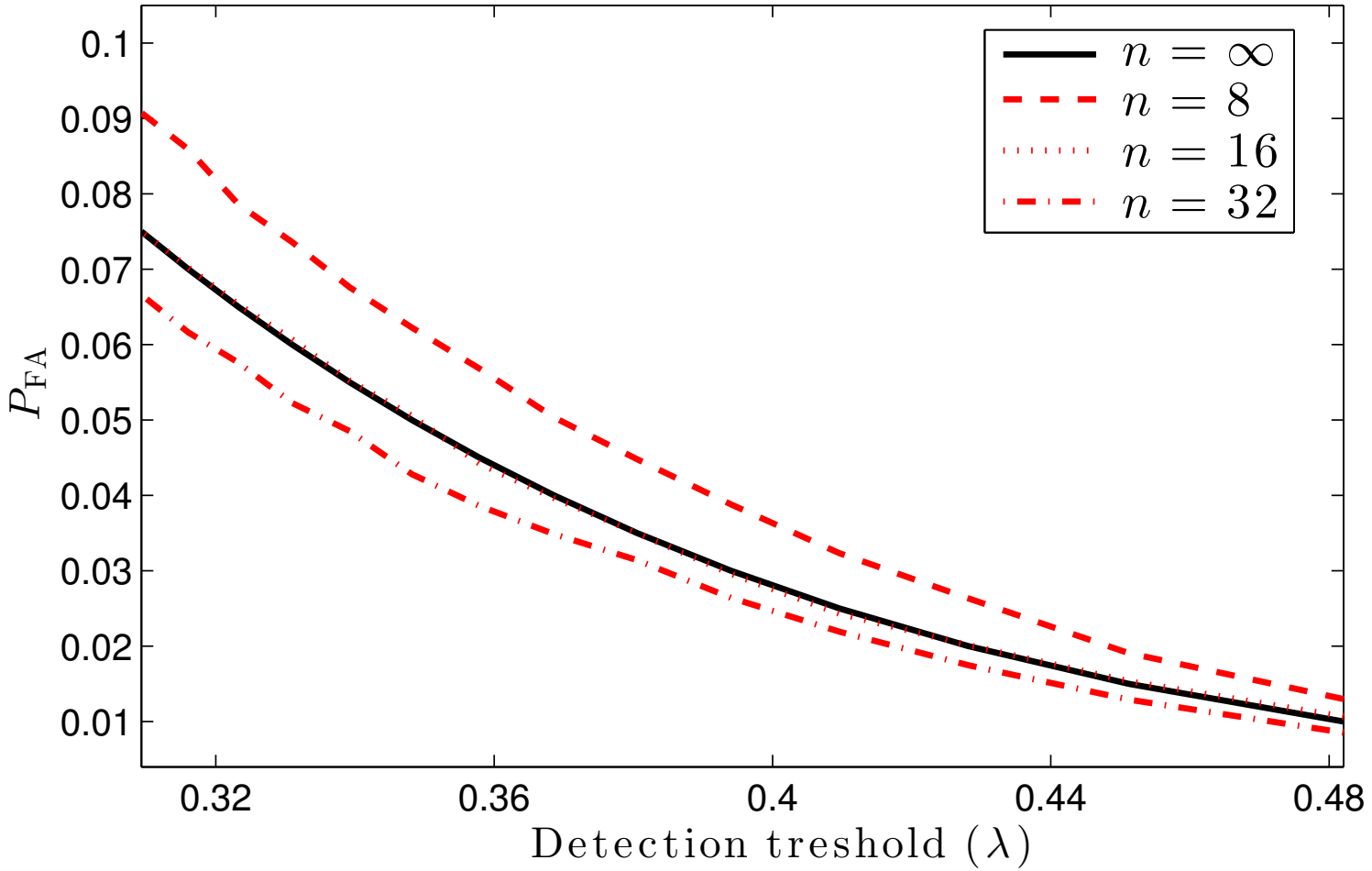
- received data $\mathbf{x}$ (input for the NMF detector)
- secondary data $\mathbf{x}_1, \ldots, \mathbf{x}_n$ (input to estimate $\hat{\boldsymbol{\Sigma}}$).
- $p = 8$ and the shape parameter of the K-distribution is $\nu = 4.5$.

For 10000 trials, we calculated the empirical P_{FA} for fixed λ where the true covariance matrix $\boldsymbol{\Sigma}$ was generated *randomly for each trial*.



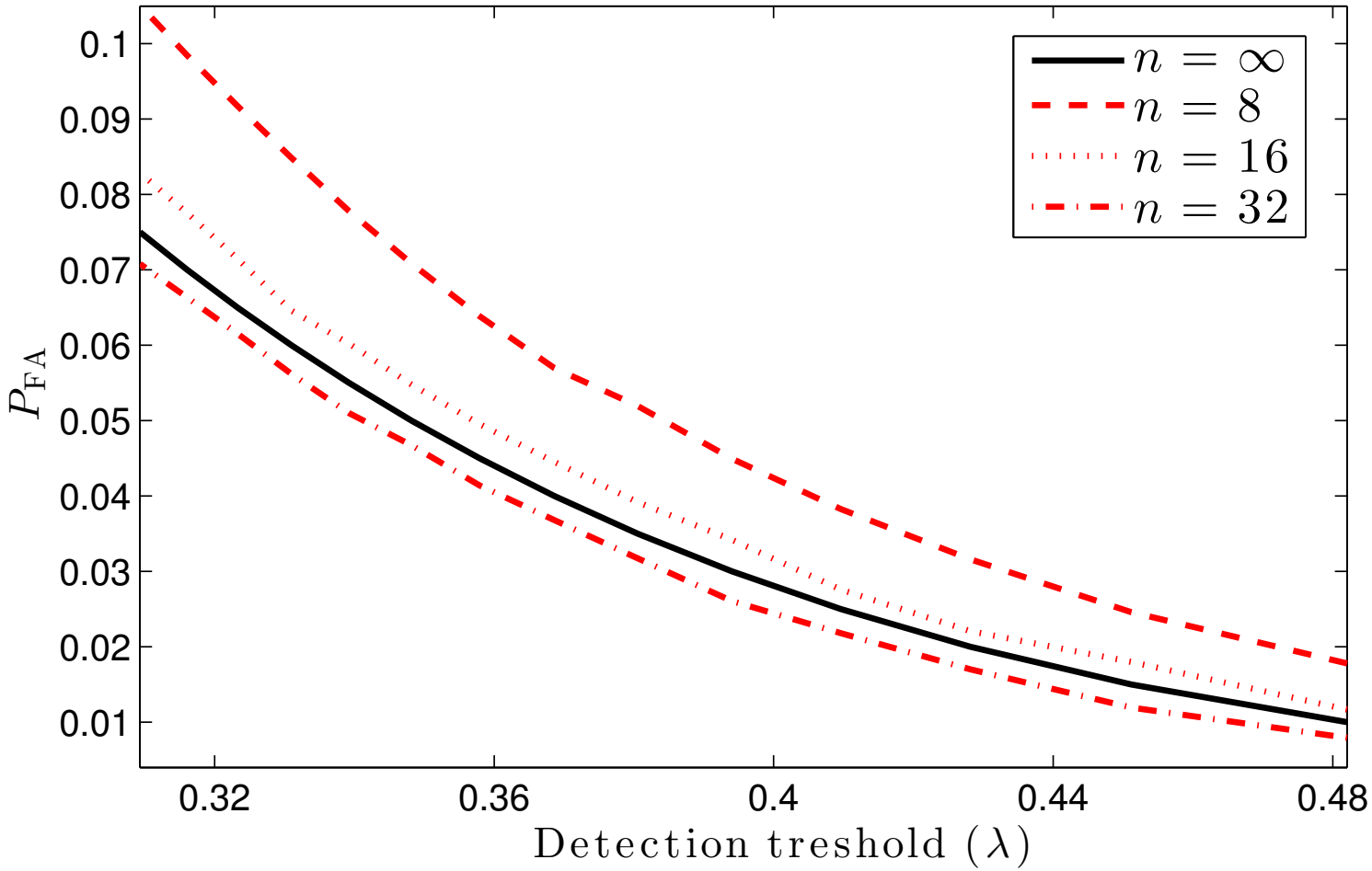
Adaptive NMF detector $\hat{\Lambda}$ using non-regularized Tyler's M-estimator as an estimate for $\mathbf{\Sigma}$.

$n :=$ # of secondary samples $\{\mathbf{x}_i\}_{i=1}^n$ availabe for estimation



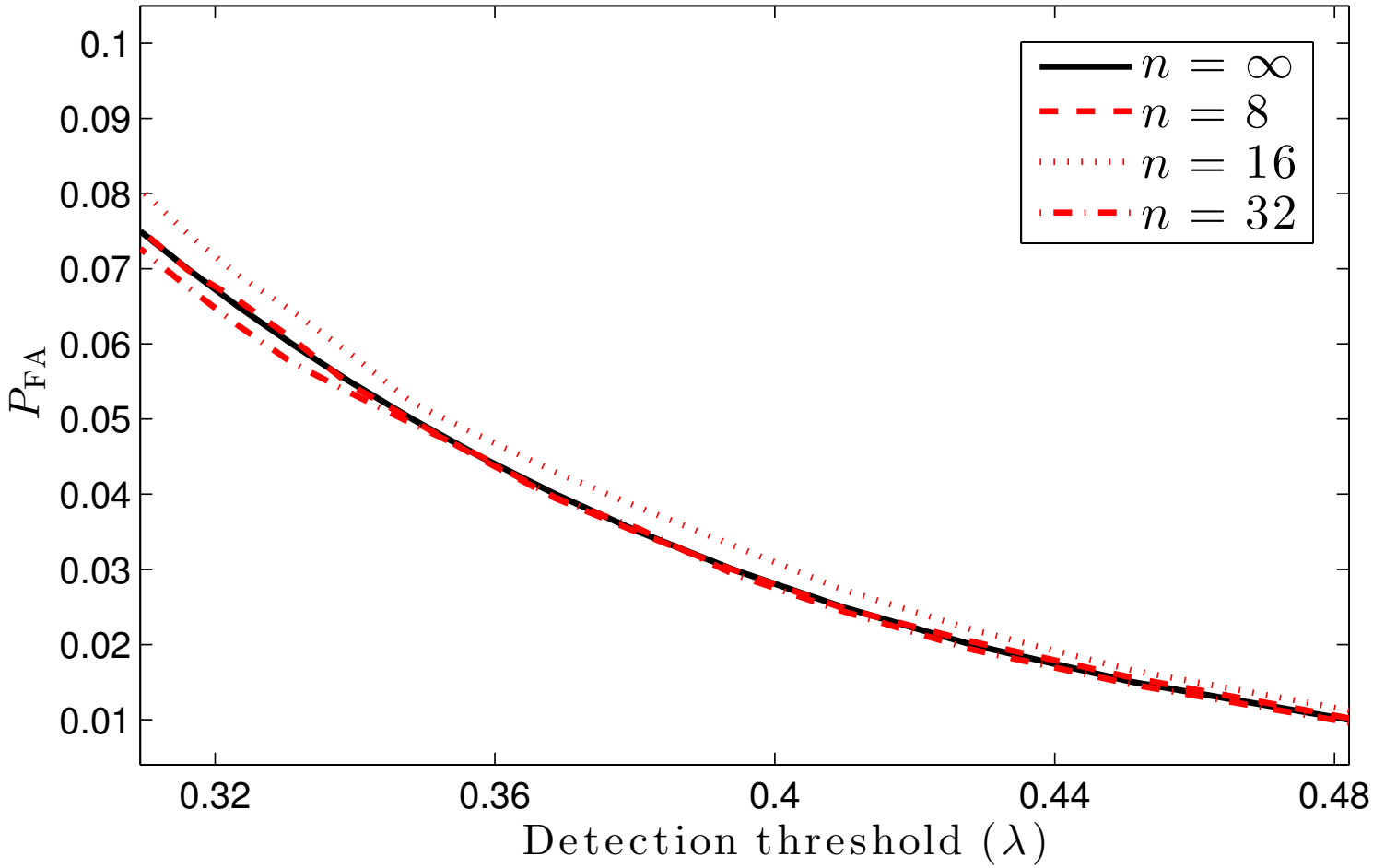
Adaptive NMF detector $\hat{\Lambda}$ using regularized SCM $\mathbf{S}_{\alpha,\beta}$ and plug-in oracle estimator $\hat{\alpha}_o$ and $\hat{\beta}_0$ given in Slide 91.

$n :=$ # of secondary samples $\{\mathbf{x}_i\}_{i=1}^n$ availabe for estimation



Adaptive NMF detector $\hat{\Lambda}$ using DL-FP estimator and plug-in oracle estimator $\hat{\alpha}_0$ given in Slide 94

$n := \#$ of secondary samples $\{\mathbf{x}_i\}_{i=1}^n$ availabe for estimation



Adaptive NMF detector $\hat{\Lambda}$ using regularized Tyler's M-estimator and plug-in oracle estimator $\hat{\alpha}_o$ (and $\hat{\beta}_o = 1 - \hat{\alpha}_0$) given in Slide 92.

$n :=$ # of secondary samples $\{\mathbf{x}_i\}_{i=1}^n$ availabe for estimation

Thank you !!!

References: see the next slides

Abramovich, Y. and Besson, O. (2013).

Regularized covariance matrix estimation in complex elliptically symmetric distributions using the expected likelihood approach-part 1: The over-sampled case.

IEEE Trans. Signal Process., 61(23):5807–5818.

Abramovich, Y. I. and Spencer, N. K. (2007).

Diagonally loaded normalised sample matrix inversion (LNSMI) for outlier-resistant adaptive filtering.

In *Proc. Int'l Conf. on Acoustics, Speech and Signal Processing (ICASSP'07)*, pages 1105–1108.

Besson, O. and Abramovich, Y. (2013).

Regularized covariance matrix estimation in complex elliptically symmetric distributions using the expected likelihood approach-part 2: The under-sampled case.

IEEE Trans. Signal Process., 61(23):5819–5829.

Bhatia, R. (2009).

Positive definite matrices.

Princeton University Press.

Chen, Y., Wiesel, A., and Hero, A. O. (2011).

Robust shrinkage estimation of high-dimensional covariance matrices.
IEEE Trans. Signal Process., 59(9):4097 – 4107.


Conte, E., De Maio, A., and Ricci, G. (2002).
Recursive estimation of the covariance matrix of a compound-Gaussian process and its application to adaptive CFAR detection.
IEEE Trans. Signal Process., 50(8):1908 – 1915.


Cover, T. M. and Thomas, J. A. (2012).
Elements of information theory.
John Wiley & Sons.


Du, L., Li, J., and Stoica, P. (2010).
Fully automatic computation of diagonal loading levels for robust adaptive beamforming.
IEEE Trans. Aerosp. Electron. Syst., 46(1):449–458.


Friedman, J., Hastie, T., and Tibshirani, R. (2008).
Sparse inverse covariance estimation with the graphical lasso.
Biostatistics, 9(3):432–441.


Friedman, J. H. (1989).
Regularized discriminant analysis.
J. Amer. Stat. Assoc., 84(405):165–175.

Gini, F. and Greco, M. (2002).
Covariance matrix estimation for CFAR detection in correlated heavy-tailed clutter.
Signal Processing, 82(12):1847–1859.

Kent, J. T. (1997).
Data analysis for shapes and images.
J. Statist. Plann. Inference, 57(2):181–193.

Ledoit, O. and Wolf, M. (2004).
A well-conditioned estimator for large-dimensional covariance matrices.
J. Mult. Anal., 88:365–411.

Maronna, R. A. (1976).
Robust M-estimators of multivariate location and scatter.
Ann. Stat., 5(1):51–67.

Ollila, E., Soloveychik, I., Tyler, D. E., and Wiesel, A. (2016).
Simultaneous penalized M-estimation of covariance matrices using geodesically convex optimization.
Journal of Multivariate Analysis, submitted.

Ollila, E. and Tyler, D. E. (2012).
Distribution-free detection under complex elliptically symmetric clutter distribution.

In *Proc. IEEE Sensor Array and Multichannel Signal Processing Workshop (SAM'12)*, Hoboken, NJ, USA.

Ollila, E. and Tyler, D. E. (2014).

Regularized M-estimators of scatter matrix.

IEEE Trans. Signal Process., 62(22):6059–6070.

Pascal, F., Chitour, Y., and Quek, Y. (2014).

Generalized robust shrinkage estimator and its application to stap detection problem.

IEEE Trans. Signal Process., 62(21):5640–5651.

Sra, S. (2011).

Positive definite matrices and the S-divergence.

arXiv preprint arXiv:1110.1773.

Sun, Y., Babu, P., and Palomar, D. P. (2014).

Regularized tyler's scatter estimator: Existence, uniqueness, and algorithms.

IEEE Trans. Signal Process., 62(19):5143–5156.

Wiesel, A. (2012).

Unified framework to regularized covariance estimation in scaled gaussian models.

IEEE Trans. Signal Process., 60(1):29–38.


Wiesel, A. and Zhang, T. (2015).
Structured robust covariance estimation.
Foundations and Trends in Signal Processing, 8(3):127–216.


Zhang, T., Wiesel, A., and Greco, M. S. (2013).
Multivariate generalized Gaussian distribution: Convexity and graphical models.
IEEE Trans. Signal Process., 61(16):4141–4148.