Regularized Robust Estimation of Mean and Covariance Matrix under Heavy Tails and Outliers

Daniel P. Palomar

Department of Electronic and Computer Engineering
The Hong Kong University of Science and Technology

2014 Workshop on Complex Systems Modeling and Estimation Challenges in Big Data (CSM-2014)
31 July 2014

Acknowledgment

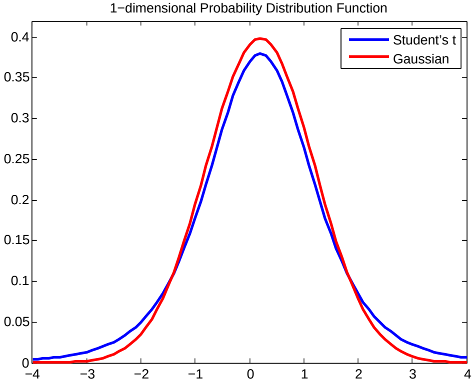
This is a joint work with

- Ying Sun, Ph.D. student, Dept. ECE, HKUST.
- Prabhu Babu, Postdoc, Dept. ECE, HKUST.

Outline

Basic Problem

- Task: estimate mean and covariance matrix from data $\{\mathbf{x}_i\}$.
- Difficulties: outlier corrupted observation (heavy-tailed underlying distribution).



Sample Average

- A straight-forward solution

$$\mu = \mathrm{E}_f(\mathbf{x}) \quad \mathbf{R} = \mathrm{E}_f(\mathbf{x}-\mu)(\mathbf{x}-\mu)^T$$
$$\Downarrow f \leftarrow f_N$$
$$\hat{\mu} = \frac{1}{N}\sum_{i=1}^{N}\mathbf{x}_i \quad \hat{\mathbf{R}} = \frac{1}{N}\sum_{i=1}^{N}(\mathbf{x}_i-\hat{\mu})(\mathbf{x}_i-\hat{\mu})^T.$$

- Works well for i.i.d. Gaussian distributed data.

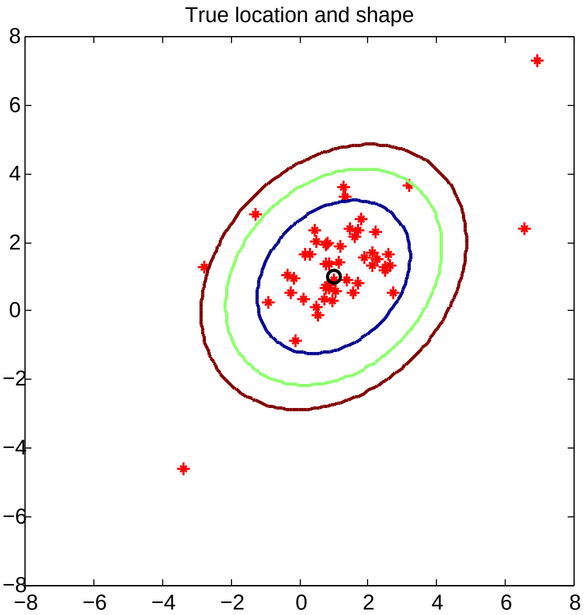
Influence of Outliers

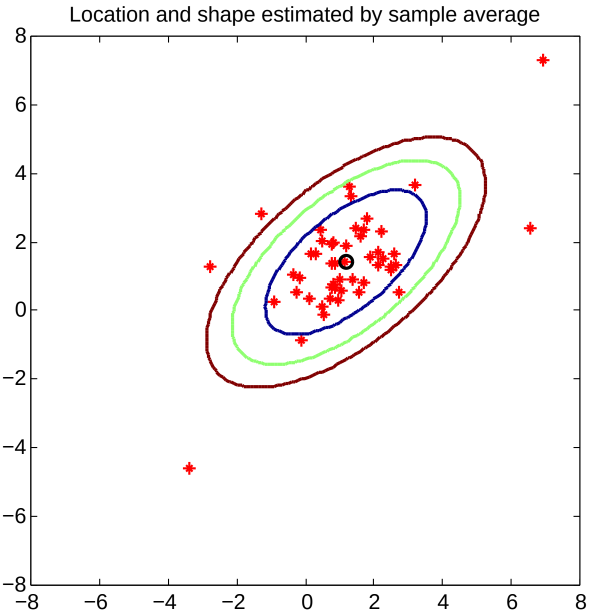
- What if the data is corrupted?
- A real-life example: Kalman filter lost track of the spacecraft during an Apollo mission because of outlier observation (caused by system noise).

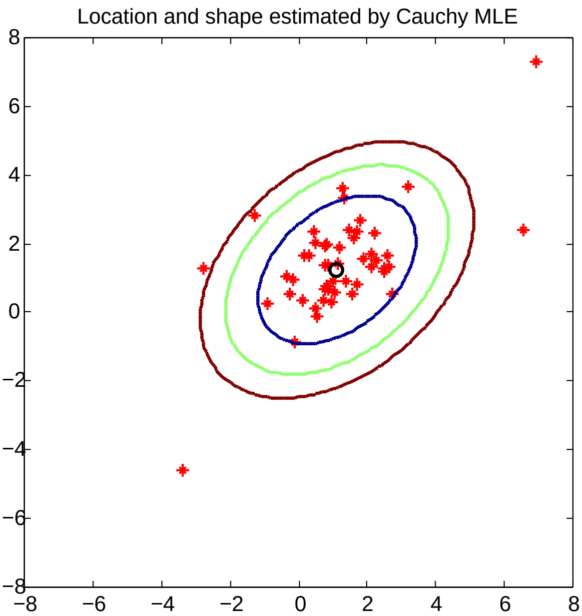
Example 1: Symmetrically Distributed Outliers

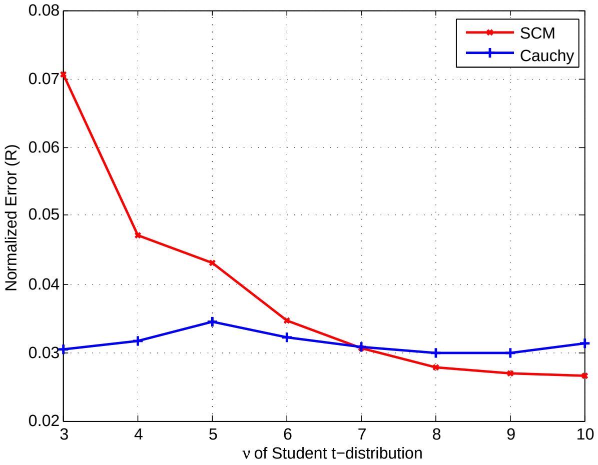
$$\mathbf{x} \sim \text{HeavyTail}\,(\mathbf{1}, \mathbf{R})$$

$$\mathbf{R} = \begin{bmatrix} 1 & 0.5 \\ 0.5 & 1 \end{bmatrix}$$


True location and shape
8
6
4
2
0
−2
−4
−6
−8
−8
−6
−4
−2
0
2
4
6
8


Location and shape estimated by sample average
8
6
4
2
0
−2
−4
−6
−8
−8
−6
−4
−2
0
2
4
6
8


Location and shape estimated by Cauchy MLE
8
6
4
2
0
−2
−4
−6
−8
−8
−6
−4
−2
0
2
4
6
8


SCM
Cauchy
Normalized Error (R)
ν of Student t−distribution
0.08
0.07
0.06
0.05
0.04
0.03
0.02
3
4
5
6
7
8
9
10

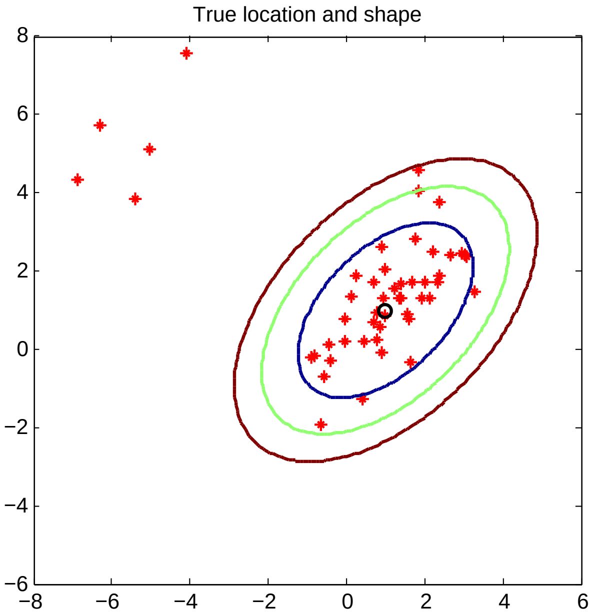
Influence of Outliers

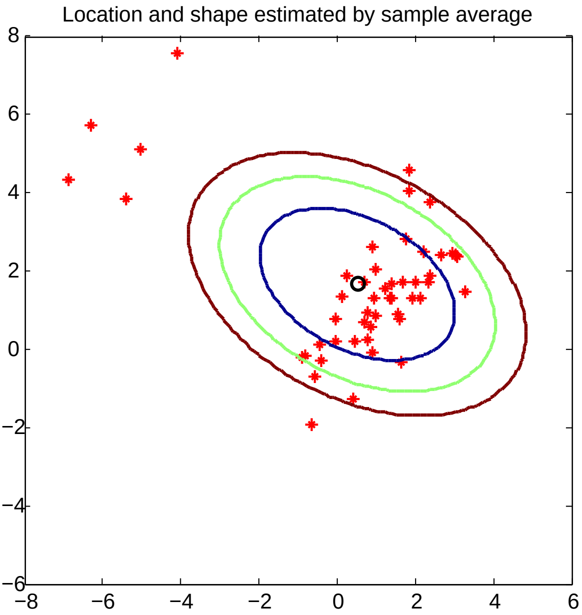
- What if the data is corrupted?

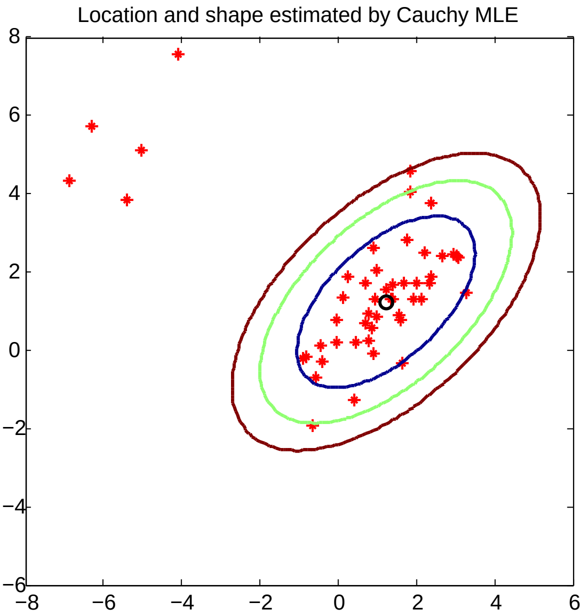
Example 2: Asymmetrically Distributed Outliers

$$\mathbf{x} \sim 0.9\mathcal{N}(\mathbf{1}, \mathbf{R}) + 0.1\mathcal{N}(\boldsymbol{\mu}, \mathbf{R})$$

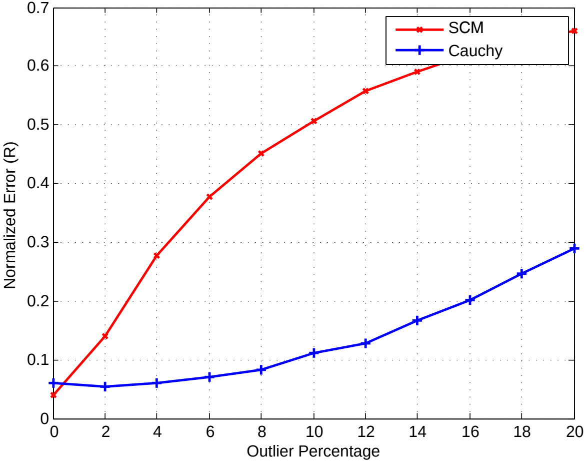
$$\boldsymbol{\mu} = \begin{bmatrix} 5 \\ -5 \end{bmatrix} \quad \mathbf{R} = \begin{bmatrix} 1 & 0.5 \\ 0.5 & 1 \end{bmatrix}$$


True location and shape
8
6
4
2
0
−2
−4
−6
−8
−6
−4
−2
0
2
4
6


Location and shape estimated by sample average
8
6
4
2
0
−2
−4
−6
−8
−6
−4
−2
0
2
4
6



Location and shape estimated by Cauchy MLE
8
6
4
2
0
−2
−4
−6
−8
−6
−4
−2
0
2
4
6



SCM
Cauchy
Normalized Error (R)
Outlier Percentage
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0
2
4
6
8
10
12
14
16
18
20

More Sophisticated Models

- Factor model:

$$\mathbf{y} = \mu + \mathbf{A}\mathbf{x} + \varepsilon.$$

- Vector ARMA:

$$\left(1 - \sum_{i=1}^{p} \mathbf{\Phi}_i L^i\right)(\mathbf{y}_t - \mu) = \left(1 - \sum_{i=1}^{q} \mathbf{\Theta}_i L^i\right)\mathbf{u}_t.$$

- VECM:

$$\left(1 - \sum_{i=1}^{p} \mathbf{\Gamma}_i L^i\right)\Delta\mathbf{y}_t = \mathbf{\Phi}\mathbf{D}_t + \mathbf{\Pi}\mathbf{y}_{t-1} + \varepsilon_t.$$

- State-space model:

$$\mathbf{y}_t = \mathbf{C}\mathbf{x}_t + \varepsilon_t$$

$$\mathbf{x}_t = \mathbf{A}\mathbf{x}_{t-1} + \mathbf{u}_t.$$

Outline

1. Motivation
2. Robust Covariance Matrix Estimators
   - Introduction
   - Examples
   - Unsolved Problems
3. Robust Mean-Covariance Estimators
   - Introduction
   - Joint Mean-Covariance Estimation for Elliptical Distributions
4. Small Sample Regime
   - Shrinkage Robust Estimator with Known Mean
   - Shrinkage Robust Estimator for Unknown Mean

Warm-up

- Recall the Gaussian distribution

$$f(\mathbf{x}) = C \det(\mathbf{\Sigma})^{-\frac{1}{2}} \exp\left(-\frac{1}{2}\mathbf{x}^T \mathbf{\Sigma}^{-1}\mathbf{x}\right).$$

- Negative log-likelihood function

$$L(\mathbf{\Sigma}) = \frac{N}{2} \log\det(\mathbf{\Sigma}) + \frac{1}{2}\sum_{i=1}^{N} \mathbf{x}_i^T \mathbf{\Sigma}^{-1}\mathbf{x}_i.$$

- Sample covariance matrix

$$\hat{\mathbf{\Sigma}} = \frac{1}{N}\sum_{i=1}^{N} \mathbf{x}_i \mathbf{x}_i^T.$$

M-estimator

- Minimizer of loss function [Mar-Mar-Yoh'06]:

$$L(\mathbf{\Sigma}) = \frac{N}{2}\log\det(\mathbf{\Sigma}) + \sum_{i=1}^{N}\rho\left(\mathbf{x}_i^T\mathbf{\Sigma}^{-1}\mathbf{x}_i\right).$$

- Solution to fixed-point equation:

$$\mathbf{\Sigma} = \frac{1}{N}\sum_{i=1}^{N} w\left(\mathbf{x}_i^T\mathbf{\Sigma}^{-1}\mathbf{x}_i\right)\mathbf{x}_i\mathbf{x}_i^T.$$

- If ρ is differentiable

$$w = \frac{\rho'}{2}.$$

Outline

Sample Covariance Matrix

- SCM can be viewed as:

$$\hat{\mathbf{\Sigma}} = \sum_{i=1}^{N} w_i \mathbf{x}_i \mathbf{x}_i^T$$

with $w_i = \frac{1}{N}, \ \forall i$.

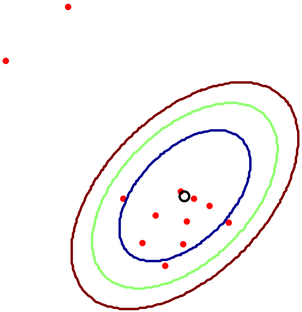
- MLE of a Gaussian distribution with loss function

$$\frac{N}{2} \log\det(\mathbf{\Sigma}) + \frac{1}{2} \sum_{i=1}^{N} \mathbf{x}_i^T \mathbf{\Sigma}^{-1} \mathbf{x}_i.$$

- Why is SCM sensitive to outliers? ☹

Sample Covariance Matrix

- Consider distance
  $$d_i = \sqrt{\mathbf{x}_i^T \boldsymbol{\Sigma}^{-1} \mathbf{x}_i}.$$
- $w_i = \frac{1}{N}$
  normal samples and outliers contribute to $\hat{\boldsymbol{\Sigma}}$ equally.
- Quadratic loss.



Tyler's M-estimator

- Given $f(\mathbf{x}) \rightarrow$ use MLE.
- $\mathbf{x}_i \sim \text{elliptical}(\mathbf{0}, \boldsymbol{\Sigma})$, what shall we do?
- Normalized sample $\mathbf{s}_i \triangleq \frac{\mathbf{x}_i}{\|\mathbf{x}_i\|_2}$

pdf

$$f(\mathbf{s}) = C \det(\mathbf{R})^{-\frac{1}{2}} \left(\mathbf{s}^T \mathbf{R}^{-1} \mathbf{s}\right)^{-K/2}$$

Loss function

$$\frac{N}{2} \log\det(\boldsymbol{\Sigma}) + \frac{K}{2} \sum_{i=1}^{N} \log \underbrace{\left(\mathbf{s}_i^T \boldsymbol{\Sigma}^{-1} \mathbf{s}_i\right)}_{\mathbf{x}_i^T \boldsymbol{\Sigma}^{-1} \mathbf{x}_i}$$

- Tyler [Tyl'J87] proposed covariance estimator $\hat{\boldsymbol{\Sigma}}$ as solution to

$$\sum_{i=1}^{N} w_i \mathbf{x}_i \mathbf{x}_i^T = \boldsymbol{\Sigma}, \quad w_i = \frac{K}{N\left(\mathbf{x}_i^T \boldsymbol{\Sigma}^{-1} \mathbf{x}_i\right)}.$$

- Why is Tyler's estimator robust to outliers? ☺

Tyler's M-estimator

- Given $f(\mathbf{x}) \rightarrow$ use MLE.
- $\mathbf{x}_i \sim \text{elliptical}(\mathbf{0}, \mathbf{\Sigma})$, what shall we do?
- Normalized sample $\mathbf{s}_i \triangleq \frac{\mathbf{x}_i}{\|\mathbf{x}_i\|_2}$

pdf

$$f(\mathbf{s}) = C \det(\mathbf{R})^{-\frac{1}{2}} \left(\mathbf{s}^T \mathbf{R}^{-1} \mathbf{s}\right)^{-K/2}$$

Loss function

$$\frac{N}{2} \log\det(\mathbf{\Sigma}) + \frac{K}{2} \sum_{i=1}^{N} \log \underbrace{\left(\mathbf{s}_i^T \mathbf{\Sigma}^{-1} \mathbf{s}_i\right)}_{\mathbf{x}_i^T \mathbf{\Sigma}^{-1} \mathbf{x}_i}$$

- Tyler [Tyl'J87] proposed covariance estimator $\hat{\mathbf{\Sigma}}$ as solution to

$$\sum_{i=1}^{N} w_i \mathbf{x}_i \mathbf{x}_i^T = \mathbf{\Sigma}, \quad w_i = \frac{K}{N\left(\mathbf{x}_i^T \mathbf{\Sigma}^{-1} \mathbf{x}_i\right)}.$$

- Why is Tyler's estimator robust to outliers? ☺

Tyler's M-estimator

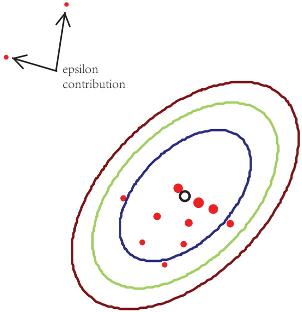
- Consider distance

  $$d_i = \sqrt{\mathbf{x}_i^T \boldsymbol{\Sigma}^{-1} \mathbf{x}_i}.$$

- $w_i \propto 1/d_i^2$

  Outliers are down-weighted.

- Logarithmic loss.



Tyler's M-estimator

- Tyler's M-estimator solves fixed-point equation

$$\mathbf{\Sigma} = \frac{K}{N} \sum_{i=1}^{N} \frac{\mathbf{x}_i \mathbf{x}_i^T}{\mathbf{x}_i^T \mathbf{\Sigma}^{-1} \mathbf{x}_i}.$$

- Existence condition: $N > K$.
- No closed-form solution.
- Iterative algorithm

$$\tilde{\mathbf{\Sigma}}_{t+1} = \frac{K}{N} \sum_{i=1}^{N} \frac{\mathbf{x}_i \mathbf{x}_i^T}{\mathbf{x}_i^T \mathbf{\Sigma}_t^{-1} \mathbf{x}_i}$$

$$\mathbf{\Sigma}_{t+1} = \tilde{\mathbf{\Sigma}}_{t+1} / \mathrm{Tr}\left(\tilde{\mathbf{\Sigma}}_{t+1}\right).$$

Outline

Unsolved Problems

Problem 1

What if the mean value is unknown?

Problem 2

How to deal with small sample scenario?

Problem 3

How to incorporate prior information?

Unsolved Problems

Problem 1

What if the mean value is unknown?

Problem 2

How to deal with small sample scenario?

Problem 3

How to incorporate prior information?

Unsolved Problems

Problem 1

What if the mean value is unknown?

Problem 2

How to deal with small sample scenario?

Problem 3

How to incorporate prior information?

Outline

Robust M-estimators

- Maronna's M-estimators [Mar'J76]:

$$\frac{1}{N}\sum_{i=1}^{N} u_1\left(\left(\mathbf{x}_i-\mu\right)^T\mathbf{R}^{-1}\left(\mathbf{x}_i-\mu\right)\right)\left(\mathbf{x}_i-\mu\right)=\mathbf{0}$$

$$\frac{1}{N}\sum_{i=1}^{N} u_2\left(\left(\mathbf{x}_i-\mu\right)^T\mathbf{R}^{-1}\left(\mathbf{x}_i-\mu\right)\right)\left(\mathbf{x}_i-\mu\right)\left(\mathbf{x}_i-\mu\right)^T=\mathbf{R}.$$

- Special examples:
  - Huber's loss function.
  - MLE for Student's t-distribution.

MLE of the Student's *t*-distribution

- Student's *t*-distribution with degree of freedom ν:

$$f(\mathbf{x}) = C \det(\mathbf{R})^{-\frac{1}{2}} \left(1 + \frac{1}{\nu}(\mathbf{x} - \mu)^T \mathbf{R}^{-1} (\mathbf{x} - \mu)\right)^{-\frac{K+\nu}{2}}.$$

- Negative log-likelihood

$$L^{\nu}(\mu, \mathbf{R}) = \frac{N}{2} \log \det(\mathbf{R}) + \frac{K+\nu}{2} \sum_{i=1}^{N} \log\left(\nu + (\mathbf{x}_i - \mu)^T \mathbf{R}^{-1} (\mathbf{x}_i - \mu)\right).$$

MLE of the Student's t-distribution

- Estimating equations

$$\frac{K+\nu}{N}\sum_{i=1}^{N}\frac{\mathbf{x}_i-\mu}{\nu+(\mathbf{x}_i-\mu)^T\mathbf{R}^{-1}(\mathbf{x}_i-\mu)}=\mathbf{0}$$

$$\frac{K+\nu}{N}\sum_{i=1}^{N}\frac{(\mathbf{x}_i-\mu)(\mathbf{x}_i-\mu)^T}{\nu+(\mathbf{x}_i-\mu)^T\mathbf{R}^{-1}(\mathbf{x}_i-\mu)}=\mathbf{R}.$$

- Weight $w_i(\nu)=\frac{K+\nu}{N}\cdot\frac{1}{\nu+(\mathbf{x}_i-\mu)^T\mathbf{R}^{-1}(\mathbf{x}_i-\mu)}$ decreases in ν.
- Unique solution for $\nu\geq 1$.

Outline

Joint Mean-Covariance Estimation

- Assumption: $\mathbf{x}_i \sim \text{elliptical}(\boldsymbol{\mu}_0, \mathbf{R}_0)$.
- Goal: jointly estimate mean and covariance
  - Robust to outliers.
  - Easy to implement.
  - Provable convergence.
- A natural idea:

  MLE of heavy-tailed distributions.

Joint Mean-Covariance Estimation

- Method: fitting $\{\mathbf{x}_i\}$ to Cauchy (Student's t-distribution with $\nu = 1$) likelihood function.
  - Conservative fitting.
  - Trade-off: robustness $\Leftrightarrow$ efficiency.
  - Tractability.
- $\hat{\mathbf{R}} \to c\mathbf{R}_0$
  c depends on the unknown shape of the underlying distribution $\Longrightarrow$ estimate $\mathbf{R}/\mathrm{Tr}(\mathbf{R})$ instead.
- Existence condition $N > K + 1$ [Ken'J91].

Algorithm

- No closed-form solution.
- Numerical algorithm [Ken-Tyl-Var'J94]:

$$\boldsymbol{\mu}_{t+1} = \frac{\sum_{i=1}^{N} w_i(\boldsymbol{\mu}_t, \mathbf{R}_t)\,\mathbf{x}_i}{\sum_{i=1}^{N} w_i(\boldsymbol{\mu}_t, \mathbf{R}_t)}$$

$$\mathbf{R}_{t+1} = \frac{K+1}{N}\sum_{i=1}^{N} w_i(\boldsymbol{\mu}_t, \mathbf{R}_t)\left(\mathbf{x}_i - \boldsymbol{\mu}_{t+1}\right)\left(\mathbf{x}_i - \boldsymbol{\mu}_{t+1}\right)^T$$

with

$$w_i(\boldsymbol{\mu}, \mathbf{R}) = \frac{1}{1 + (\mathbf{x}_i - \boldsymbol{\mu})^T \mathbf{R}^{-1} (\mathbf{x}_i - \boldsymbol{\mu})}.$$

Outline

Regularization-Known Mean

- Problem:

insufficient observations → estimator does not exist → algorithms fail to converge

- Methods:
  - Diagonal loading.
  - Penalized or regularized loss function.

Diagonal Loading

- Modified Tyler's iteration [Abr-Spe'C07]

$$\tilde{\mathbf{\Sigma}}_{t+1} = \frac{K}{N} \sum_{i=1}^{N} \frac{\mathbf{x}_i \mathbf{x}_i^T}{\mathbf{x}_i^T \mathbf{\Sigma}_t^{-1} \mathbf{x}_i} + \rho \mathbf{I}$$

$$\mathbf{\Sigma}_{t+1} = \tilde{\mathbf{\Sigma}}_{t+1} / \mathrm{Tr}\left(\tilde{\mathbf{\Sigma}}_{t+1}\right).$$

- Provable convergence [Che-Wie-Her'J11].
- Systematic way of choosing parameter ρ [Che-Wie-Her'J11].
- But without a clear motivation.

Penalized Loss Function I

- Wiesel's penalty [Wie'J12]

$$h(\boldsymbol{\Sigma}) = \log\det(\boldsymbol{\Sigma}) + K \log \operatorname{Tr}\left(\boldsymbol{\Sigma}^{-1}\mathbf{T}\right),$$

$\boldsymbol{\Sigma} \propto \mathbf{T}$ minimizes $h(\boldsymbol{\Sigma})$.

- Penalized loss function

$$L^{\text{Wiesel}}(\boldsymbol{\Sigma}) = \frac{N}{2}\log\det(\boldsymbol{\Sigma}) + \frac{K}{2}\sum_{i=1}^{N}\log\left(\mathbf{x}_i^T\boldsymbol{\Sigma}^{-1}\mathbf{x}_i\right)$$
$$+\alpha\left(\log\det(\boldsymbol{\Sigma}) + K\log\operatorname{Tr}\left(\boldsymbol{\Sigma}^{-1}\mathbf{T}\right)\right).$$

- Algorithm

$$\boldsymbol{\Sigma}_{t+1} = \frac{N}{N+2\alpha}\frac{K}{N}\sum_{i=1}^{N}\frac{\mathbf{x}_i\mathbf{x}_i^T}{\mathbf{x}_i^T\boldsymbol{\Sigma}_t^{-1}\mathbf{x}_i} + \frac{2\alpha}{N+2\alpha}\frac{K\mathbf{T}}{\operatorname{Tr}\left(\boldsymbol{\Sigma}_t^{-1}\mathbf{T}\right)}.$$

Penalized Loss Function II

- Alternative penalty: KL-divergence

$$h(\boldsymbol{\Sigma}) = \log\det(\boldsymbol{\Sigma}) + \operatorname{Tr}\left(\boldsymbol{\Sigma}^{-1}\mathbf{T}\right),$$

  $\boldsymbol{\Sigma} = \mathbf{T}$ minimizes $h(\boldsymbol{\Sigma})$.

- Penalized loss function

$$L^{\mathrm{KL}}(\boldsymbol{\Sigma}) = \frac{N}{2}\log\det(\boldsymbol{\Sigma}) + \frac{K}{2}\sum_{i=1}^{N}\log\left(\mathbf{x}_i^T\boldsymbol{\Sigma}^{-1}\mathbf{x}_i\right) + \alpha\left(\log\det(\boldsymbol{\Sigma}) + \operatorname{Tr}\left(\boldsymbol{\Sigma}^{-1}\mathbf{T}\right)\right).$$

- Algorithm?

Questions

Existence & Uniqueness?

Which one is better?

Algorithm convergence?

Questions

Existence & Uniqueness?

Which one is better?

Algorithm convergence?

Questions

Existence & Uniqueness?

Which one is better?

Algorithm convergence?

Existence and Uniqueness for Wiesel's Shrinkage Estimator

Theorem [Sun-Bab-Pal'J14a]

Wiesel's shrinkage estimator exists a.s., and is also unique up to a positive scale factor, if and only if the underlying distribution is continuous and $N > K - 2\alpha$.

- Existence condition for Tyler's estimator: $N > K$
  - Regularization relaxes the requirement on the number of samples.
  - Setting $\alpha = 0$ (no regularization) reduces to Tyler's condition.
  - Stronger confidence on the prior information $\Rightarrow$ less number of samples required.

Existence and Uniqueness for Wiesel's Shrinkage Estimator

Theorem [Sun-Bab-Pal'J14a]

Wiesel's shrinkage estimator exists a.s., and is also unique up to a positive scale factor, if and only if the underlying distribution is continuous and $N > K - 2\alpha$.

- Existence condition for Tyler's estimator: $N > K$
  - Regularization relaxes the requirement on the number of samples.
  - Setting $\alpha = 0$ (no regularization) reduces to Tyler's condition.
  - Stronger confidence on the prior information $\Rightarrow$ less number of samples required.

Existence and Uniqueness for KL-Shrinkage Estimator

Theorem [Sun-Bab-Pal'J14a]

KL-shrinkage estimator exists a.s., and is also unique, if and only if the underlying distribution is continuous and $N > K - 2\alpha$

Compared with Wiesel's shrinkage estimator:

- Share the same existence condition.
- Without scaling ambiguity.

Any connection? Which one is better?

Existence and Uniqueness for KL-Shrinkage Estimator

Theorem [Sun-Bab-Pal'J14a]

KL-shrinkage estimator exists a.s., and is also unique, if and only if the underlying distribution is continuous and $N > K - 2\alpha$

Compared with Wiesel's shrinkage estimator:

- Share the same existence condition.
- Without scaling ambiguity.

Any connection? Which one is better?

Existence and Uniqueness for KL-Shrinkage Estimator

Theorem [Sun-Bab-Pal'J14a]

KL-shrinkage estimator exists a.s., and is also unique, if and only if the underlying distribution is continuous and $N > K - 2\alpha$

Compared with Wiesel's shrinkage estimator:

- Share the same existence condition.
- Without scaling ambiguity.

Any connection? Which one is better?

Equivalence

Theorem [Sun-Bab-Pal'J14a]

Wiesel's shrinkage estimator and KL-shrinkage estimator are equivalent.

- Fixed-point equation for KL-shrinkage estimator

$$\boldsymbol{\Sigma} = \frac{N}{N+2\alpha}\frac{K}{N}\sum_{i=1}^{N}\frac{\mathbf{x}_i\mathbf{x}_i^T}{\mathbf{x}_i^T\boldsymbol{\Sigma}^{-1}\mathbf{x}_i} + \frac{2\alpha}{N+2\alpha}\mathbf{T}.$$

- The solution satisfies equality

$$\mathrm{Tr}\left(\boldsymbol{\Sigma}^{-1}\mathbf{T}\right) = K.$$

- Fixed-point equation for Wiesel's shrinkage estimator

$$\boldsymbol{\Sigma} = \frac{N}{N+2\alpha}\frac{K}{N}\sum_{i=1}^{N}\frac{\mathbf{x}_i\mathbf{x}_i^T}{\mathbf{x}_i^T\boldsymbol{\Sigma}^{-1}\mathbf{x}_i} + \frac{2\alpha}{N+2\alpha}\frac{K\mathbf{T}}{\mathrm{Tr}\left(\boldsymbol{\Sigma}^{-1}\mathbf{T}\right)}.$$

Equivalence

Theorem [Sun-Bab-Pal'J14a]

Wiesel's shrinkage estimator and KL-shrinkage estimator are equivalent.

- Fixed-point equation for KL-shrinkage estimator

$$\boldsymbol{\Sigma} = \frac{N}{N+2\alpha}\frac{K}{N}\sum_{i=1}^{N}\frac{\mathbf{x}_i\mathbf{x}_i^T}{\mathbf{x}_i^T\boldsymbol{\Sigma}^{-1}\mathbf{x}_i} + \frac{2\alpha}{N+2\alpha}\mathbf{T}.$$

- The solution satisfies equality

$$\mathrm{Tr}\left(\boldsymbol{\Sigma}^{-1}\mathbf{T}\right) = K.$$

- Fixed-point equation for Wiesel's shrinkage estimator

$$\boldsymbol{\Sigma} = \frac{N}{N+2\alpha}\frac{K}{N}\sum_{i=1}^{N}\frac{\mathbf{x}_i\mathbf{x}_i^T}{\mathbf{x}_i^T\boldsymbol{\Sigma}^{-1}\mathbf{x}_i} + \frac{2\alpha}{N+2\alpha}\frac{K\mathbf{T}}{\mathrm{Tr}\left(\boldsymbol{\Sigma}^{-1}\mathbf{T}\right)}.$$

Majorization-minimization

- Problem:

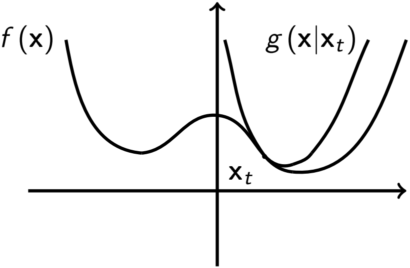
$$\begin{array}{ll} \underset{\mathbf{x}}{\text{minimize}} & f(\mathbf{x}) \\ \text{subject to} & \mathbf{x} \in \mathcal{X} \end{array}$$

- Majorization-minimization:

$$\mathbf{x}_{t+1} = \arg\min_{\mathbf{x}\in\mathcal{X}} g(\mathbf{x}|\mathbf{x}_t)$$

with

$$\begin{aligned} & f(\mathbf{x}_t) = g(\mathbf{x}_t|\mathbf{x}_t) \\ & f(\mathbf{x}) \leq g(\mathbf{x}|\mathbf{x}_t) \;\; \forall \mathbf{x} \in \mathcal{X} \\ & f'(\mathbf{x}_t;\mathbf{d}) = g'(\mathbf{x}_t;\mathbf{d}|\mathbf{x}_t) \;\; \forall \mathbf{x}_t + \mathbf{d} \in \mathcal{X} \end{aligned}$$



Modified Algorithm for Wiesel's Shrinkage Estimator

- Surrogate function

$$g\left(\boldsymbol{\Sigma}|\boldsymbol{\Sigma}_t\right) = \frac{N}{2}\log\det\left(\boldsymbol{\Sigma}\right) + \frac{K}{2}\sum_{i=1}^{N}\frac{\mathbf{x}_i^T\boldsymbol{\Sigma}^{-1}\mathbf{x}_i}{\mathbf{x}_i^T\boldsymbol{\Sigma}_t^{-1}\mathbf{x}_i}$$

$$+\alpha\left(\log\det\left(\boldsymbol{\Sigma}\right) + K\frac{\mathrm{Tr}\left(\boldsymbol{\Sigma}^{-1}\mathbf{T}\right)}{\mathrm{Tr}\left(\boldsymbol{\Sigma}_t^{-1}\mathbf{T}\right)}\right)$$

- Update

$$\tilde{\boldsymbol{\Sigma}}_{t+1} = \frac{N}{N+2\alpha}\frac{K}{N}\sum_{i=1}^{N}\frac{\mathbf{x}_i\mathbf{x}_i^T}{\mathbf{x}_i^T\boldsymbol{\Sigma}_t^{-1}\mathbf{x}_i} + \frac{2\alpha}{N+2\alpha}\frac{K\mathbf{T}}{\mathrm{Tr}\left(\boldsymbol{\Sigma}_t^{-1}\mathbf{T}\right)}$$

- Normalization

$$\boldsymbol{\Sigma}_{t+1} = \tilde{\boldsymbol{\Sigma}}_{t+1}/\mathrm{Tr}\left(\tilde{\boldsymbol{\Sigma}}_{t+1}\right)$$

Algorithm Convergence

Theorem [Sun-Bab-Pal'J14a]

Under the existence conditions, the modified algorithm for Wiesel's shrinkage estimator converges to the unique solution.

Proof idea:

- Majorization-minimization decreases the value of objective function.
- Normalization does not change the value of objective function.
- There is a unique minimizer of the objective function.

Algorithm Convergence

Theorem [Sun-Bab-Pal'J14a]

Under the existence conditions, the modified algorithm for Wiesel's shrinkage estimator converges to the unique solution.

Proof idea:

- Majorization-minimization decreases the value of objective function.
- Normalization does not change the value of objective function.
- There is a unique minimizer of the objective function.

Algorithm for KL-Shrinkage Estimator

- Surrogate function

$$g(\boldsymbol{\Sigma}|\boldsymbol{\Sigma}_t) = \frac{N}{2}\log\det(\boldsymbol{\Sigma}) + \frac{K}{2}\sum_{i=1}^{N}\frac{\mathbf{x}_i^T\boldsymbol{\Sigma}^{-1}\mathbf{x}_i}{\mathbf{x}_i^T\boldsymbol{\Sigma}_t^{-1}\mathbf{x}_i} + \alpha\left(\log\det(\boldsymbol{\Sigma}) + \mathrm{Tr}\left(\boldsymbol{\Sigma}^{-1}\mathbf{T}\right)\right)$$

- Update

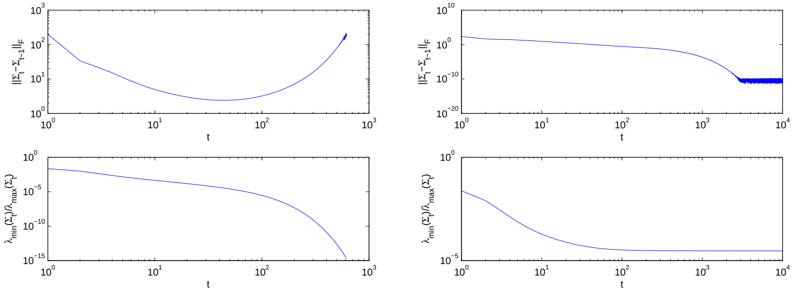
$$\boldsymbol{\Sigma}_{t+1} = \frac{N}{N+2\alpha}\frac{K}{N}\sum_{i=1}^{N}\frac{\mathbf{x}_i\mathbf{x}_i^T}{\mathbf{x}_i^T\boldsymbol{\Sigma}_t^{-1}\mathbf{x}_i} + \frac{2\alpha}{N+2\alpha}\mathbf{T}$$

Theorem [Sun-Bab-Pal'J14a]

Under the existence conditions, the algorithm for KL-shrinkage estimator converges to the unique solution.

Algorithm convergence of Wiesel's shrinkage estimator

- Parameters: $K = 10$, $N = 8$.



(a) (b)

Figure: (a) when the existence conditions are not satisfied with $\alpha_0 = 0.96$, (b) when the existence conditions are satisfied with $\alpha_0 = 1.04$.

Algorithm convergence of KL-shrinkage estimator

- Parameters: $K = 10$, $N = 8$.

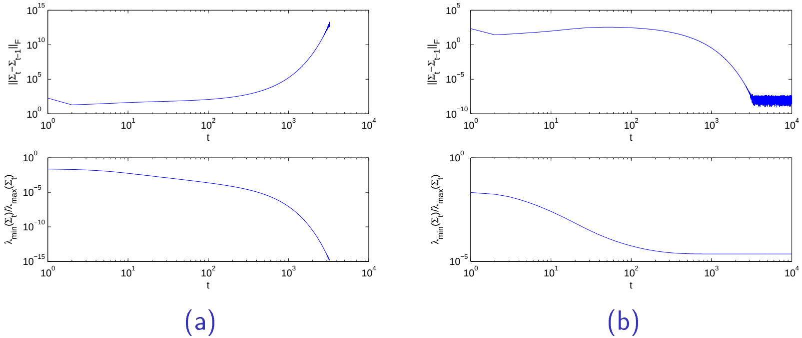


Figure: (a) when the existence conditions are not satisfied with $\alpha_0 = 0.96$, and (b) when the existence conditions are satisfied with $\alpha_0 = 1.04$.

Outline

Regularization-Unknown Mean

- Problem:
  μ_0 is unknown!
- A simple solution: plug-in $\hat{\mu}$
  - Sample mean
  - Sample median
- But...
  - Two-step estimation, not jointly optimal.
  - Estimation error of $\hat{\mu}$ propagates.
- To be done: shrinkage estimator for joint mean-covariance estimation with target $(\mathbf{t}, \mathbf{T})$.

Regularization-Unknown Mean

- Method: adding shrinkage penalty $h(\mu, \mathbf{R})$ to loss function (negative log-likelihood of Cauchy distribution).
- Design criteria:
  - $h(\mu, \mathbf{R})$ attains minimum at prior $(\mathbf{t}, \mathbf{T})$.
  - $h(\mathbf{t}, \mathbf{T}) = h(\mathbf{t}, r\mathbf{T})$, $\forall r > 0$.
- Reason:
  - $\mathbf{R}$ can be estimated up to an unknown scale factor.
  - $\mathbf{T}$ is a prior for the parameter $\mathbf{R}/\mathrm{Tr}(\mathbf{R})$.

Regularization-Unknown Mean

Proposed penalty function

$$h(\mu,\mathbf{R}) = \alpha\left(K\log\left(\mathrm{Tr}\left(\mathbf{R}^{-1}\mathbf{T}\right)\right) + \log\det(\mathbf{R})\right) + \gamma\log\left(1 + (\mu - \mathbf{t})^T\mathbf{R}^{-1}(\mu - \mathbf{t})\right)$$

Proposition [Sun-Bab-Pal'J14b]

$(\mathbf{t}, r\mathbf{T})$, $\forall r > 0$ are the minimizers of $h(\mu,\mathbf{R})$.

Regularization-Unknown Mean

Proposed penalty function

$$h(\mu, \mathbf{R}) = \alpha \left(K \log \left(\mathrm{Tr}\left(\mathbf{R}^{-1} \mathbf{T} \right) \right) + \log\det\left(\mathbf{R} \right) \right) + \gamma \log \left(1 + (\mu - \mathbf{t})^T \mathbf{R}^{-1} (\mu - \mathbf{t}) \right)$$

Proposition [Sun-Bab-Pal'J14b]

$(\mathbf{t}, r\mathbf{T})$, $\forall r > 0$ are the minimizers of $h(\mu, \mathbf{R})$.

Regularization-Unknown Mean

- Resulting optimization problem:

$$\begin{aligned}\underset{\mu,\mathbf{R}\succ\mathbf{0}}{\text{minimize}} \quad & \frac{(K+1)}{2}\sum_{i=1}^{N}\log\left(1+(\mathbf{x}_i-\mu)^T\mathbf{R}^{-1}(\mathbf{x}_i-\mu)\right)\\ & +\alpha\left(K\log\left(\mathrm{Tr}\left(\mathbf{R}^{-1}\mathbf{T}\right)\right)+\log\det\left(\mathbf{R}\right)\right)\\ & +\gamma\log\left(1+(\mu-\mathbf{t})^T\mathbf{R}^{-1}(\mu-\mathbf{t})\right)+\frac{N}{2}\log\det\left(\mathbf{R}\right).\end{aligned}$$

- A minimum satisfies the stationary condition $\frac{\partial L^{\text{shrink}}(\mu,\mathbf{R})}{\partial\mu}=\mathbf{0}$ and $\frac{\partial L^{\text{shrink}}(\mu,\mathbf{R})}{\partial\mathbf{R}}=\mathbf{0}$.

Regularization-Unknown Mean

- $d_i(\mu, \mathbf{R}) = \sqrt{(\mathbf{x}_i - \mu)^T \mathbf{R}^{-1} (\mathbf{x}_i - \mu)}$,
  $d_{\mathbf{t}}(\mu, \mathbf{R}) = \sqrt{(\mathbf{t} - \mu)^T \mathbf{R}^{-1} (\mathbf{t} - \mu)}$.
- $w_i(\mu, \mathbf{R}) = \frac{1}{1 + d_i^2(\mu, \mathbf{R})}$, $w_{\mathbf{t}}(\mu, \mathbf{R}) = \frac{1}{1 + d_{\mathbf{t}}^2(\mu, \mathbf{R})}$.
- Stationary condition:

$$\mathbf{R} = \frac{K+1}{N+2\alpha} \sum_{i=1}^{N} w_i(\mu, \mathbf{R}) (\mathbf{x}_i - \mu)(\mathbf{x}_i - \mu)^T + \frac{2\gamma}{N+2\alpha} w_{\mathbf{t}}(\mu, \mathbf{R}) (\mu - \mathbf{t})(\mu - \mathbf{t})^T + \frac{2\alpha K}{N+2\alpha} \frac{\mathbf{T}}{\mathrm{Tr}(\mathbf{R}^{-1}\mathbf{T})}$$

$$\mu = \frac{(K+1)\sum_{i=1}^{N} w_i(\mu, \mathbf{R}) \mathbf{x}_i + 2\gamma w_{\mathbf{t}}(\mu, \mathbf{R}) \mathbf{t}}{(K+1)\sum_{i=1}^{N} w_i(\mu, \mathbf{R}) + 2\gamma w_{\mathbf{t}}(\mu, \mathbf{R})}$$

Existence and Uniqueness

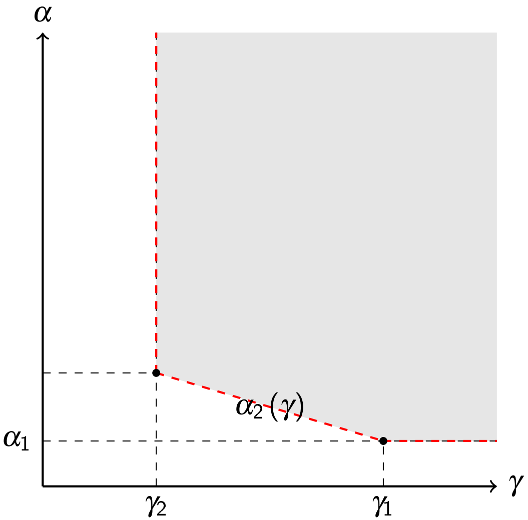
Theorem [Sun-Bab-Pal'J14b]

Assuming continuous underlying distribution, the estimator exists under either of the following conditions:
(i) if $\gamma > \gamma_1$, then $\alpha > \alpha_1$,
(ii) if $\gamma_2 < \gamma \leq \gamma_1$, then $\alpha > \alpha_2(\gamma)$,
where

$$\alpha_1 = \frac{1}{2}(K-N),$$

$$\alpha_2(\gamma) = \frac{1}{2}\left(K+1-N-\frac{2\gamma+N-K-1}{N-1}\right),$$

and $\gamma_1 = \frac{1}{2}(K+1)$, $\gamma_2 = \frac{1}{2}(K+1-N)$.

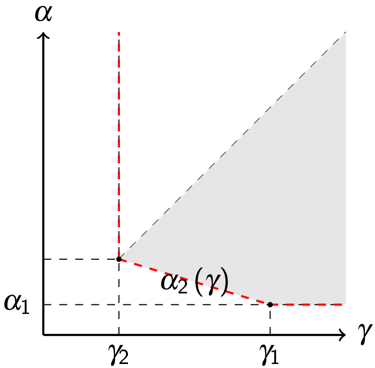


α
α₂(γ)
α₁
γ
γ₂
γ₁

Existence and Uniqueness

Theorem [Sun-Bab-Pal'J14b]

The shrinkage estimator is unique if $\gamma \geq \alpha$.



Algorithm in μ and $\mathbf{R}$

- Surrogate function

$$
\begin{aligned}
L(\mu,\mathbf{R}|\mu_t,\mathbf{R}_t) \quad &= \tfrac{K+1}{2}\sum w_i(\mu_t,\mathbf{R}_t)(\mathbf{x}_i-\mu)^T\mathbf{R}^{-1}(\mathbf{x}_i-\mu) \\
&+\gamma w_{\mathbf{t}}(\mu_t,\mathbf{R}_t)(\mathbf{t}-\mu)^T\mathbf{R}^{-1}(\mathbf{t}-\mu) \\
&+\left(\tfrac{N}{2}+\alpha\right)\log\det(\mathbf{R})+\alpha K\frac{\mathrm{Tr}\left(\mathbf{R}^{-1}\mathbf{T}\right)}{\mathrm{Tr}\left(\mathbf{R}_t^{-1}\mathbf{T}\right)}
\end{aligned}
$$

- Update

$$
\mu_{t+1}=\frac{(K+1)\sum_{i=1}^{N}w_i(\mu_t,\mathbf{R}_t)\mathbf{x}_i+2\gamma w_{\mathbf{t}}(\mu_t,\mathbf{R}_t)\mathbf{t}}{(K+1)\sum_{i=1}^{N}w_i(\mu_t,\mathbf{R}_t)+2\gamma w_{\mathbf{t}}(\mu_t,\mathbf{R}_t)}
$$

$$
\begin{aligned}
\mathbf{R}_{t+1}=&\frac{K+1}{N+2\alpha}\sum_{i=1}^{N}w_i(\mu_t,\mathbf{R}_t)(\mathbf{x}_i-\mu_{t+1})(\mathbf{x}_i-\mu_{t+1})^T \\
&+\frac{2\gamma}{N+2\alpha}w_{\mathbf{t}}(\mu_t,\mathbf{R}_t)(\mathbf{t}-\mu_{t+1})(\mathbf{t}-\mu_{t+1})^T+\frac{2\alpha K}{N+2\alpha}\frac{\mathbf{T}}{\mathrm{Tr}\left(\mathbf{R}_t^{-1}\mathbf{T}\right)}
\end{aligned}
$$

Algorithm in μ and $\mathbf{R}$

Theorem [Sun-Bab-Pal'J14b]

Under the existence conditions, the algorithm in μ and $\mathbf{R}$ for the proposed shrinkage estimator converges to the unique solution.

Algorithm in $\boldsymbol{\Sigma}$

- Consider case $\alpha = \gamma$, apply transform

$$\boldsymbol{\Sigma} = \begin{bmatrix} \mathbf{R} + \boldsymbol{\mu}\boldsymbol{\mu}^T & \boldsymbol{\mu} \\ \boldsymbol{\mu}^T & 1 \end{bmatrix}$$

$$\bar{\mathbf{x}}_i = [\mathbf{x}_i; 1], \quad \bar{\mathbf{t}} = [\mathbf{t}; 1]$$

- Equivalent loss function

$$L^{\text{shrink}}(\boldsymbol{\Sigma}) = \left(\frac{N}{2} + \alpha\right) \log\det(\boldsymbol{\Sigma}) + \frac{K+1}{2} \sum_{i=1}^{N} \log\left(\bar{\mathbf{x}}_i^T \boldsymbol{\Sigma}^{-1} \bar{\mathbf{x}}_i\right) + \alpha K \log\left(\mathrm{Tr}\left(\mathbf{S}^T \boldsymbol{\Sigma}^{-1} \mathbf{S}\mathbf{T}\right)\right) + \alpha \log\left(\bar{\mathbf{t}}^T \boldsymbol{\Sigma}^{-1} \bar{\mathbf{t}}\right)$$

with $\mathbf{S} = \begin{bmatrix} \mathbf{I}_K \\ \mathbf{0}_{1\times K} \end{bmatrix}$.

- $L^{\text{shrink}}(\boldsymbol{\Sigma})$ is scale-invariant.

Algorithm in $\boldsymbol{\Sigma}$

- Surrogate function

$$L(\boldsymbol{\Sigma}|\boldsymbol{\Sigma}_t) = \left(\frac{N}{2}+\alpha\right)\log\det(\boldsymbol{\Sigma}) + \frac{K+1}{2}\sum_{i=1}^{N}\frac{\bar{\mathbf{x}}_i^T\boldsymbol{\Sigma}^{-1}\bar{\mathbf{x}}_i}{\bar{\mathbf{x}}_i^T\boldsymbol{\Sigma}_t^{-1}\bar{\mathbf{x}}_i}$$
$$+\alpha\left(K\frac{\mathrm{Tr}\left(\mathbf{S}^T\boldsymbol{\Sigma}^{-1}\mathbf{S}\mathbf{T}\right)}{\mathrm{Tr}\left(\mathbf{S}^T\boldsymbol{\Sigma}_t^{-1}\mathbf{S}\mathbf{T}\right)}+\frac{\bar{\mathbf{t}}^T\boldsymbol{\Sigma}^{-1}\bar{\mathbf{t}}}{\bar{\mathbf{t}}^T\boldsymbol{\Sigma}_t^{-1}\bar{\mathbf{t}}}\right)$$

- Update

$$\tilde{\boldsymbol{\Sigma}}_{t+1} = \frac{K+1}{N+2\alpha}\sum_{i=1}^{N}\frac{\bar{\mathbf{x}}_i\bar{\mathbf{x}}_i^T}{\bar{\mathbf{x}}_i^T\boldsymbol{\Sigma}_t^{-1}\bar{\mathbf{x}}_i}$$
$$+\frac{2\alpha}{N+2\alpha}\left(\frac{K\mathbf{S}\mathbf{T}\mathbf{S}^T}{\mathrm{Tr}\left(\mathbf{S}^T\boldsymbol{\Sigma}_t^{-1}\mathbf{S}\mathbf{T}\right)}+\frac{\bar{\mathbf{t}}\bar{\mathbf{t}}^T}{\bar{\mathbf{t}}^T\boldsymbol{\Sigma}_t^{-1}\bar{\mathbf{t}}}\right)$$

$$\boldsymbol{\Sigma}_{t+1} = \tilde{\boldsymbol{\Sigma}}_{t+1}/\left(\tilde{\boldsymbol{\Sigma}}_{t+1}\right)_{K+1,K+1}$$

Algorithm in $\boldsymbol{\Sigma}$

Theorem [Sun-Bab-Pal'J14b]

Under the existence conditions, which simplifies to $N > K + 1 - 2\alpha$ for $\alpha = \gamma$, the algorithm in $\boldsymbol{\Sigma}$ for the proposed shrinkage estimator converges to the unique solution.

Simulation

- Parameters: $K = 10$

$$\boldsymbol{\mu}_0 = 0.1 \times \mathbf{1}_{K \times 1}$$

$$(\mathbf{R}_0)_{ij} = 0.8^{|i-j|}$$

- Error measurement: KL-distance

$$\begin{aligned} \mathrm{err}\left(\hat{\mu}, \hat{\mathbf{R}}\right) = E\Big\{ & D_{KL}\left(\mathcal{N}\left(\hat{\mu}, \hat{\mathbf{R}}\right) \| \mathcal{N}\left(\boldsymbol{\mu}_0, \mathbf{R}_0\right)\right) \\ & + D_{KL}\left(\mathcal{N}\left(\boldsymbol{\mu}_0, \mathbf{R}_0\right) \Big\| \mathcal{N}\left(\hat{\mu}, \hat{\mathbf{R}}\right)\right)\Big\} \end{aligned}$$

Performance Comparison for Gaussian

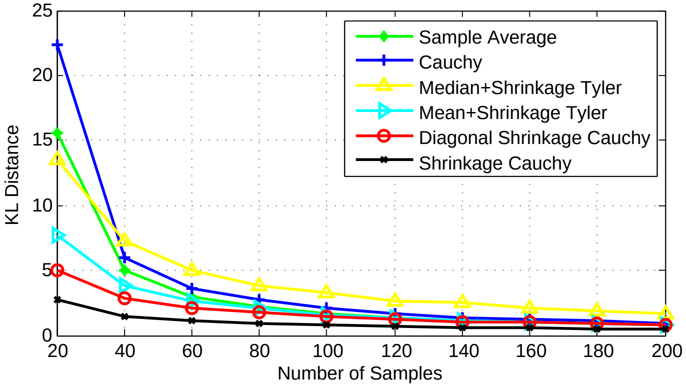


Figure: $\mathcal{N}(\mu_0, \mathbf{R}_0)$

Performance Comparison for t-distribution

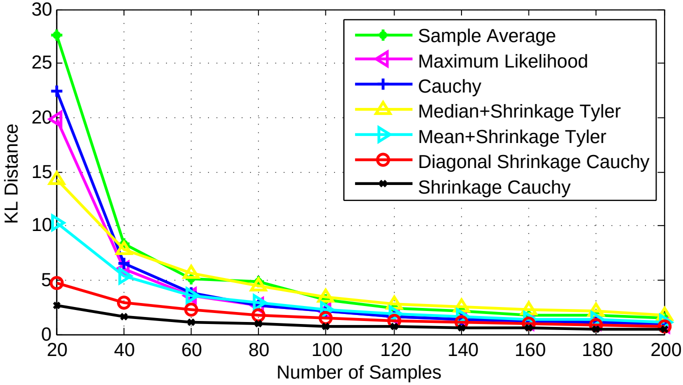


Figure: $t_\nu(\mu_0, \mathbf{R}_0)$, $\nu = 5$.

Performance Comparison for Corrupted Gaussian

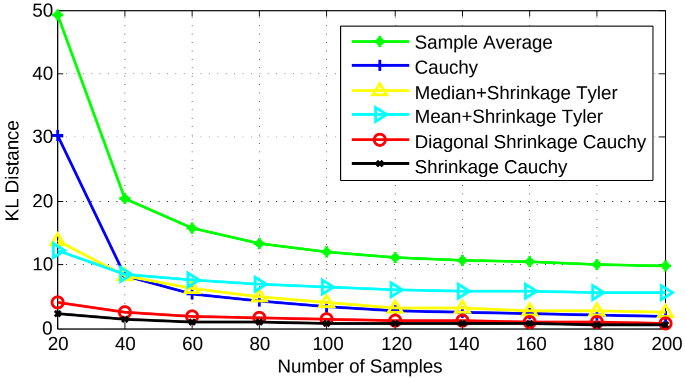
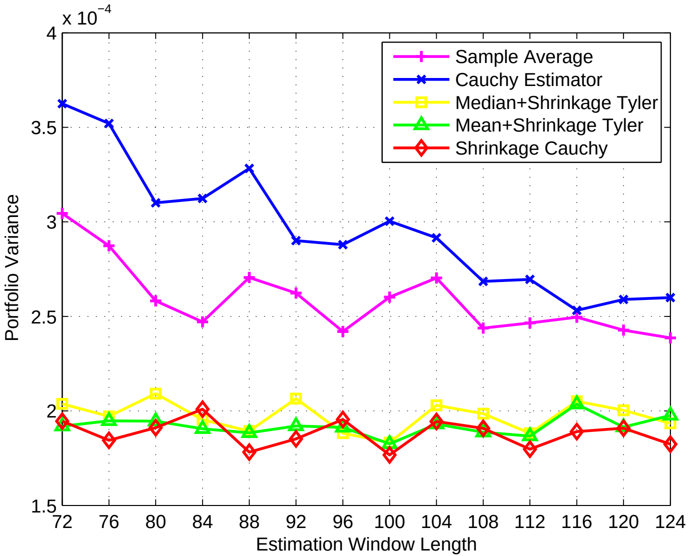


Figure: $0.9 \times \mathcal{N}(\mu_0, \mathbf{R}_0) + 0.1 \times \mathcal{N}(5 \times \mathbf{1}_{K \times 1}, \mathbf{I})$

Real Data Simulation

- Minimum variance portfolio.
- Training : S&P 500 index components weekly log-returns, $K = 40$.
  - Estimate $\mathbf{R}$
  - Construct portfolio weights $\mathbf{w}$
- Parameter selection: choose α yields minimum variance on validation set.
- Collect half a year portfolio returns.

$\cdots$	train	validate	test	$\cdots$



x 10^−4
Sample Average
Cauchy Estimator
Median+Shrinkage Tyler
Mean+Shrinkage Tyler
Shrinkage Cauchy
Portfolio Variance
Estimation Window Length
4
3.5
3
2.5
2
1.5
72
76
80
84
88
92
96
100
104
108
112
116
120
124

Summary

- In this talk, we have discussed
  - Robust mean-covariance estimation for heavy-tailed distributions.
  - Shrinkage estimation in small sample scenario.

- Future work
  - Parameter tuning.
  - Structured covariance estimation.

R. Maronna, D. Martin, and V. Yohai.
Robust Statistics: Theory and Methods.
Wiley Series in Probability and Statistics. Wiley, 2006.

D. E. Tyler.
A distribution-free *M*-estimator of multivariate scatter.
Ann. Statist., volume 15, no. 1, pp. 234–251, 1987.

R. A. Maronna.
Robust M-estimators of multivariate location and scatter.
Ann. Statist., volume 4, no. 1, pp. 51–67, 1976.

J. T. Kent and D. E. Tyler.
Redescending *M*-estimates of multivariate location and scatter.
Ann. Statist., volume 19, no. 4, pp. 2102–2119, 1991.

J. T. Kent, D. E. Tyler, and Y. Vard.
A curious likelihood identity for the multivariate t-distribution.
Communications in Statistics - Simulation and Computation, volume 23, no. 2, pp. 441–453, 1994.

Y. Abramovich and N. Spencer.
Diagonally loaded normalised sample matrix inversion (LNSMI) for outlier-resistant adaptive filtering.
In *Proc. IEEE Int Conf. Acoust., Speech, Signal Process. (ICCASP), 2007.*, volume 3, pp. III–1105–III–1108. Honolulu, HI, 2007.

Y. Chen, A. Wiesel, and A. Hero.
Robust shrinkage estimation of high-dimensional covariance matrices.
IEEE Trans. Signal Process., volume 59, no. 9, pp. 4097–4107, 2011.

References III

A. Wiesel.
Unified framework to regularized covariance estimation in scaled Gaussian models.
IEEE Trans. Signal Process., volume 60, no. 1, pp. 29–38, 2012.

Y. Sun, P. Babu, and D. Palomar.
Regularized Tyler's scatter estimator: Existence, uniqueness, and algorithms.
arXiv preprint arXiv:1407.3079, 2014.

—.
Regularized robust estimation of mean and covariance matrix under heavy tails and outliers.
submitted to IEEE Trans. Signal Process., 2014.

Thanks

For more information visit:

http://www.ece.ust.hk/~palomar