On robust estimation of high-dimensional covariance matrices

Karina Ashurbekova

GIPSA-lab, INRIA

July 18, 2019

No machine learning today! ☺

Sample Covariance Matrix

Warm-up: covariance matrix ☺

- $\boldsymbol{X} = (X_1, \ldots, X_p)$ - p-variate random vector
- Covariance matrix:

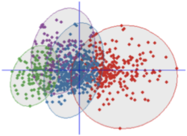
$$\boldsymbol{\Sigma} = \begin{bmatrix} \mathbb{E}\Big[\big(X_1 - \mathbb{E}(X_1)\big)\big(X_1 - \mathbb{E}(X_1)\big)\Big] & \ldots & \mathbb{E}\Big[\big(X_1 - \mathbb{E}(X_1)\big)\big(X_p - \mathbb{E}(X_p)\big)\Big] \\ & \ldots & \\ \mathbb{E}\Big[\big(X_p - \mathbb{E}(X_p)\big)\big(X_1 - \mathbb{E}(X_1)\big)\Big] & \ldots & \mathbb{E}\Big[\big(X_p - \mathbb{E}(X_p)\big)\big(X_p - \mathbb{E}(X_p)\big)\Big] \end{bmatrix} \tag{1}$$

- **Problem: estimate the covariance matrix** from the i.i.d. observations $\boldsymbol{X}_1 = (X_{11}, \ldots, X_{1p}), \ldots, \boldsymbol{X}_n = (X_{n1}, \ldots, X_{np})$

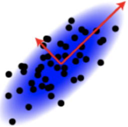
Why covariance estimation?

Portfolio selection

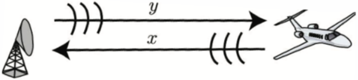
Discriminant Analysis



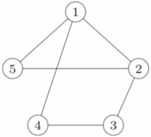
PCA



Radar detection



Graphical models



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

How to estimate the covariance matrix?

- We have a data matrix $\boldsymbol{X} = [\boldsymbol{X}_1, \ldots \boldsymbol{X}_n]^T$ (n rows (samples) and p columns (dimension))
- The sample covariance matrix:

$$\boldsymbol{S} = \frac{1}{n}\sum_{i=1}^{n}(\boldsymbol{X}_i - \widehat{\boldsymbol{X}})(\boldsymbol{X}_i - \widehat{\boldsymbol{X}})^T, \tag{2}$$

$$\widehat{\boldsymbol{X}} = \frac{1}{n}\sum_{i=1}^{n}\boldsymbol{X}_i$$

- **Problem**:
  - The eigenstructure of $\boldsymbol{S}$ tends to be systematically distorted unless $\frac{p}{n}$ is small $\Rightarrow$
  - Larger eigenvalues are overestimated; smaller eigenvalues are underestimated

How to estimate the covariance matrix?

- We have a data matrix $\boldsymbol{X} = [\boldsymbol{X}_1, \ldots \boldsymbol{X}_n]^T$ (n rows (samples) and p columns (dimension))
- The sample covariance matrix:

$$\boldsymbol{S} = \frac{1}{n}\sum_{i=1}^{n}(\boldsymbol{X}_i - \widehat{\boldsymbol{X}})(\boldsymbol{X}_i - \widehat{\boldsymbol{X}})^T, \tag{2}$$

$$\widehat{\boldsymbol{X}} = \frac{1}{n}\sum_{i=1}^{n}\boldsymbol{X}_i$$

- **Problem**:
  - The eigenstructure of $\boldsymbol{S}$ tends to be systematically distorted unless $\frac{p}{n}$ is small $\Rightarrow$
  - Larger eigenvalues are overestimated; smaller eigenvalues are underestimated

How to estimate the covariance matrix?

- We have a data matrix $\boldsymbol{X} = [\boldsymbol{X}_1, \ldots \boldsymbol{X}_n]^T$ (n rows (samples) and p columns (dimension))
- The sample covariance matrix:

$$\boldsymbol{S} = \frac{1}{n}\sum_{i=1}^{n}(\boldsymbol{X}_i - \widehat{\boldsymbol{X}})(\boldsymbol{X}_i - \widehat{\boldsymbol{X}})^T, \tag{2}$$

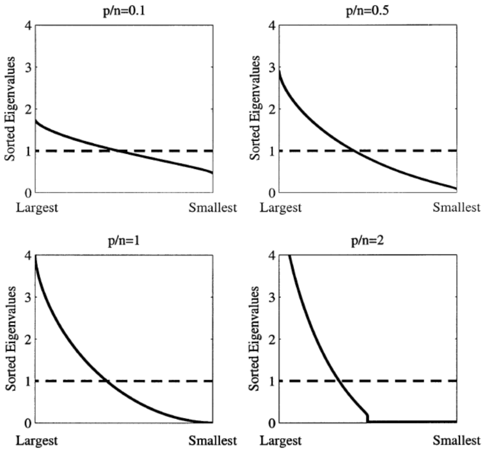
$$\widehat{\boldsymbol{X}} = \frac{1}{n}\sum_{i=1}^{n}\boldsymbol{X}_i$$

- **Problem**:
  - The eigenstructure of $\boldsymbol{S}$ tends to be systematically distorted unless $\frac{p}{n}$ is small $\Rightarrow$
  - Larger eigenvalues are overestimated; smaller eigenvalues are underestimated

Eigenvalues structure for the range $\frac{p}{n}$

Larger eigenvalues are overestimated; smaller eigenvalues are underestimated.

Too bad ☹



$p > n$. What can we do?

Shrink!

- **To ensure non-singularity**, Ledoit and Wolf (2004) proposed a shrinkage estimator of the covariance matrix:

$$\widehat{\boldsymbol{\Sigma}} = \alpha_1 \boldsymbol{S} + \alpha_2 \boldsymbol{I}$$

- Use $\widehat{\Sigma}$ that shrinks $\boldsymbol{S}$ towards to a structure (e.g., a scaled identity matrix) using a tuning (shrinkage) parameter α_2
- Why?
  - Mean Squared Error $\text{MSE}(\widehat{\Sigma}) = \mathbb{E}\left[\left\|\widehat{\Sigma} - \Sigma\right\|_F^2\right]$ can be reduced by introducing some bias!
  - Positive definiteness of $\widehat{\Sigma}$ can be ensured!

Shrink!

- **To ensure non-singularity**, Ledoit and Wolf (2004) proposed a shrinkage estimator of the covariance matrix:

$$\widehat{\boldsymbol{\Sigma}} = \alpha_1 \boldsymbol{S} + \alpha_2 \boldsymbol{I}$$

- Use $\widehat{\boldsymbol{\Sigma}}$ that shrinks $\boldsymbol{S}$ towards to a structure (e.g., a scaled identity matrix) using a tuning (shrinkage) parameter α_2
- Why?
  - Mean Squared Error $\text{MSE}(\widehat{\Sigma}) = \mathbb{E}\left[\left\|\widehat{\Sigma} - \Sigma\right\|_F^2\right]$ can be reduced by introducing some bias!
  - Positive definiteness of $\widehat{\Sigma}$ can be ensured!

Shrink!

- **To ensure non-singularity**, Ledoit and Wolf (2004) proposed a shrinkage estimator of the covariance matrix:

$$\widehat{\boldsymbol{\Sigma}} = \alpha_1 \boldsymbol{S} + \alpha_2 \boldsymbol{I}$$

- Use $\widehat{\boldsymbol{\Sigma}}$ that shrinks $\boldsymbol{S}$ towards to a structure (e.g., a scaled identity matrix) using a tuning (shrinkage) parameter α_2
- Why?
  - Mean Squared Error $\mathrm{MSE}(\widehat{\boldsymbol{\Sigma}}) = \mathbb{E}\left[\left\|\widehat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}\right\|_F^2\right]$ can be reduced by introducing some bias!
  - Positive definiteness of $\widehat{\boldsymbol{\Sigma}}$ can be ensured!

How to estimate α_1 and α_2?

Find α_1^* and α_2^* that minimize Mean Squared Error (MSE):

$$MSE(\widehat{\mathbf{\Sigma}}) = \mathbb{E}\left[\left\|\widehat{\mathbf{\Sigma}} - \mathbf{\Sigma}\right\|_F^2\right] = \mathbb{E}\left[\|\alpha_1 \boldsymbol{S} + \alpha_2 \boldsymbol{I} - \mathbf{\Sigma}\|_F^2\right]$$

Theorem

The optimal parameters α_1^* and α_2^* are:

$$\alpha_2^* = (1 - \alpha_1^*)\frac{tr(\Sigma)}{p},$$

$$\alpha_1^* = \frac{p\left(\frac{ptr(\Sigma^2)}{tr(\Sigma)^2} - 1\right)\left(\frac{tr(\Sigma)}{p}\right)^2}{\mathbb{E}\left[tr(\boldsymbol{S}^2)\right] - p\left(\frac{tr(\Sigma)}{p}\right)^2}$$

Thus $\widehat{\Sigma}$ is a convex combnation of $\boldsymbol{S}$ and $\frac{tr(\Sigma)}{p}\boldsymbol{I}$

How to estimate α_1 and α_2?

Find α_1^* and α_2^* that minimize Mean Squared Error (MSE):

$$MSE(\widehat{\boldsymbol{\Sigma}}) = \mathbb{E}\left[\left\|\widehat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}\right\|_F^2\right] = \mathbb{E}\left[\|\alpha_1 \boldsymbol{S} + \alpha_2 \boldsymbol{I} - \boldsymbol{\Sigma}\|_F^2\right]$$

Theorem

The optimal parameters α_1^* and α_2^* are:

$$\alpha_2^* = (1 - \alpha_1^*)\frac{tr(\boldsymbol{\Sigma})}{p},$$

$$\alpha_1^* = \frac{p\left(\frac{p\,tr(\boldsymbol{\Sigma}^2)}{tr(\boldsymbol{\Sigma})^2} - 1\right)\left(\frac{tr(\boldsymbol{\Sigma})}{p}\right)^2}{\mathbb{E}\left[tr(\boldsymbol{S}^2)\right] - p\left(\frac{tr(\boldsymbol{\Sigma})}{p}\right)^2}$$

Thus $\widehat{\boldsymbol{\Sigma}}$ is a convex combnation of $\boldsymbol{S}$ and $\frac{tr(\boldsymbol{\Sigma})}{p}\boldsymbol{I}$

Bias-variance trade-off

LW estimator:

$$\widehat{\boldsymbol{\Sigma}} = \alpha \boldsymbol{S} + (1-\alpha)\frac{tr(\boldsymbol{\Sigma})}{p}\boldsymbol{I}$$

- $\frac{tr(\boldsymbol{\Sigma})}{p}\boldsymbol{I}$: all bias no variance.
- $\boldsymbol{S}$: all variance no bias ($\boldsymbol{S}$ is unbiased estimation of $\boldsymbol{\Sigma}$ i. e. $\mathbb{E}(\boldsymbol{S}) = \boldsymbol{\Sigma}$)

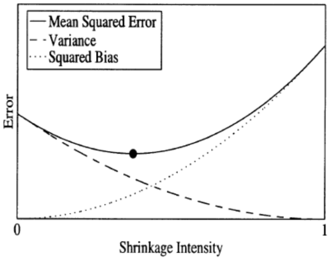


Figure: Shrinkage intensity: $1-\alpha$

A Bayesian interpretation of shrinkage

Bayes' rule:

$$p(\theta|\boldsymbol{X}) = \frac{p(\boldsymbol{X}|\theta) * p(\theta)}{p(\boldsymbol{X})}$$

$$posterior = \frac{likelihood * prior}{evidence}$$

Maximum A Posteriori Estimation:

$$\theta_{MAP} = arg \max_{\theta} p(\boldsymbol{X}|\theta)p(\theta) =$$
$$= arg \max_{\theta} \log p(\boldsymbol{X}|\theta) + \log p(\theta)$$

Thus:

- Laplacian prior $\Rightarrow$ l_1-regularization (lasso)
- Wishart-inverse prior $\Rightarrow$ shrinkage

Difference between lasso and shrinkage:

- lasso impose the sparsity on the elements of covariance matrix!
- shrinkage - no! shrinkage affects only eigenvalues.

A Bayesian interpretation of shrinkage

Bayes' rule:

$$p(\boldsymbol{\theta}|\boldsymbol{X}) = \frac{p(\boldsymbol{X}|\boldsymbol{\theta}) * p(\boldsymbol{\theta})}{p(\boldsymbol{X})}$$

$$\textit{posterior} = \frac{\textit{likelihood} * \textit{prior}}{\textit{evidence}}$$

Maximum A Posteriori Estimation:

$$\boldsymbol{\theta}_{MAP} = arg \max_{\boldsymbol{\theta}} p(\boldsymbol{X}|\boldsymbol{\theta}) p(\boldsymbol{\theta}) = $$
$$= arg \max_{\boldsymbol{\theta}} \log p(\boldsymbol{X}|\boldsymbol{\theta}) + \log p(\boldsymbol{\theta})$$

Thus:

- Laplacian prior $\Rightarrow$ l_1-regularization (lasso)
- Wishart-inverse prior $\Rightarrow$ shrinkage

Difference between lasso and shrinkage:

- lasso impose the sparsity on the elements of covariance matrix!
- shrinkage - no! shrinkage affects only eigenvalues.

A Bayesian interpretation of shrinkage

Bayes' rule:

$$p(\boldsymbol{\theta}|\boldsymbol{X}) = \frac{p(\boldsymbol{X}|\boldsymbol{\theta}) * p(\boldsymbol{\theta})}{p(\boldsymbol{X})}$$

$$\textit{posterior} = \frac{\textit{likelihood} * \textit{prior}}{\textit{evidence}}$$

Maximum A Posteriori Estimation:

$$\begin{aligned} \boldsymbol{\theta}_{MAP} &= \textit{arg} \max_{\boldsymbol{\theta}} p(\boldsymbol{X}|\boldsymbol{\theta}) p(\boldsymbol{\theta}) = \\ &= \textit{arg} \max_{\boldsymbol{\theta}} \log p(\boldsymbol{X}|\boldsymbol{\theta}) + \log p(\boldsymbol{\theta}) \end{aligned}$$

Thus:

- Laplacian prior $\Rightarrow$ l_1-regularization (lasso)
- Wishart-inverse prior $\Rightarrow$ shrinkage

Difference between lasso and shrinkage:

- lasso impose the sparsity on the elements of covariance matrix!
- shrinkage - no! shrinkage affects only eigenvalues.

But...

LW shrinkage approach is based on the sample covariance matrix $\boldsymbol{S}$!

- we know how to deal with $p > n$ case only for the **Gaussian data**
- but if the data is corrupted? (outliers, noise)
  - Sample Covariance Matrix is sensitive to outliers.☹ Why?

$$\boldsymbol{S} = \frac{1}{n}\sum_{i=1}^{n}(\boldsymbol{X}_i - \widehat{\boldsymbol{X}})(\boldsymbol{X}_i - \widehat{\boldsymbol{X}})^T = $$
$$= \sum_{i=1}^{n} w_i(\boldsymbol{X}_i - \widehat{\boldsymbol{X}})(\boldsymbol{X}_i - \widehat{\boldsymbol{X}})^T,$$
$$w_i = \frac{1}{n}$$

  - consider the distance $d_i = \sqrt{(\boldsymbol{X}_i - \widehat{\boldsymbol{X}})^T \Sigma^{-1} (\boldsymbol{X}_i - \widehat{\boldsymbol{X}})}$
  - $w_i = \frac{1}{n}$, normal samples and outliers contribute to $\boldsymbol{S}$ equally

LW shrinkage approach is based on the sample covariance matrix $\boldsymbol{S}$!

- we know how to deal with $p > n$ case only for the **Gaussian data**
- but if the data is corrupted? (outliers, noise)
  - Sample Covariance Matrix is sensitive to outliers.☹ Why?

$$\boldsymbol{S} = \frac{1}{n}\sum_{i=1}^{n}(\boldsymbol{X}_i - \widehat{\boldsymbol{X}})(\boldsymbol{X}_i - \widehat{\boldsymbol{X}})^T =$$
$$= \sum_{i=1}^{n} w_i(\boldsymbol{X}_i - \widehat{\boldsymbol{X}})(\boldsymbol{X}_i - \widehat{\boldsymbol{X}})^T,$$
$$w_i = \frac{1}{n}$$

  - consider the distance $d_i = \sqrt{(\boldsymbol{X}_i - \widehat{\boldsymbol{X}})^T \Sigma^{-1} (\boldsymbol{X}_i - \widehat{\boldsymbol{X}})}$
  - $w_i = \frac{1}{n}$, normal samples and outliers contribute to $\boldsymbol{S}$ equally

LW shrinkage approach is based on the sample covariance matrix $\boldsymbol{S}$!

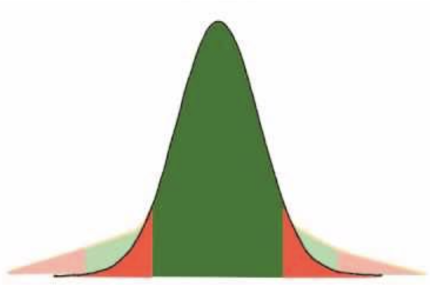
- we know how to deal with $p > n$ case only for the **Gaussian data**
- but if the data is corrupted? (outliers, noise)
  - Sample Covariance Matrix is sensitive to outliers. ☹ Why?

$$\boldsymbol{S} = \frac{1}{n}\sum_{i=1}^{n}(\boldsymbol{X}_i - \widehat{\boldsymbol{X}})(\boldsymbol{X}_i - \widehat{\boldsymbol{X}})^T = $$
$$= \sum_{i=1}^{n} w_i(\boldsymbol{X}_i - \widehat{\boldsymbol{X}})(\boldsymbol{X}_i - \widehat{\boldsymbol{X}})^T,$$
$$w_i = \frac{1}{n}$$

  - consider the distance $d_i = \sqrt{(\boldsymbol{X}_i - \widehat{\boldsymbol{X}})^T\Sigma^{-1}(\boldsymbol{X}_i - \widehat{\boldsymbol{X}})}$
  - $w_i = \frac{1}{n}$, normal samples and outliers contribute to $\boldsymbol{S}$ equally

What to do in the case of noisy data/data with outliers

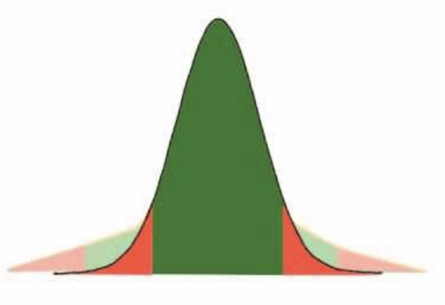
- consider the assumption that data comes form heavy-tailed distribution



- **Problem: estimate the covariance matrix** in **high-dimensional case** for **heavy-tailed distributions.**

What to do in the case of noisy data/data with outliers

- consider the assumption that data comes form heavy-tailed distribution



- **Problem: estimate the covariance matrix** in **high-dimensional case** for **heavy-tailed distributions.**

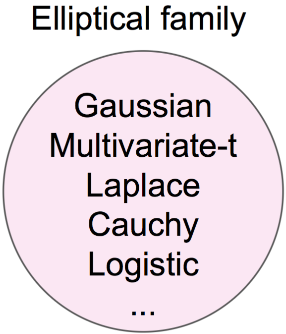
Heavy-tailed distributions

Elliptical distribution

- The probability density function of the elliptical distribution is:

$$f(\boldsymbol{x}) = C_{p,g}|\boldsymbol{\Sigma}|^{-1/2} g\left((\boldsymbol{x} - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1} (\boldsymbol{x} - \boldsymbol{\mu})\right),$$

where $g(t)$ is non-negative generator function, $C_{p,g}$ is the normalisation constant



Tyler's estimator

- $\boldsymbol{X}_i \sim$ elliptical $(0, \boldsymbol{\Sigma})$ (we don't specify which distribution in elliptical family)
- Normalized sample $\boldsymbol{Z}_i \triangleq \frac{\boldsymbol{X}_i - \boldsymbol{\mu}}{\|\boldsymbol{X}_i - \boldsymbol{\mu}\|_2}$
  - pdf, Angular Central Gaussian Distribution:

$$f(\boldsymbol{z}) = C|\Sigma|^{-1/2}(\boldsymbol{z}^T\Sigma^{-1}\boldsymbol{z})^{-p/2}$$

  - negative log-likelihood function:

$$\frac{n}{2}\log|\Sigma| + \frac{p}{2}\sum_{i=1}^{n}\log(\boldsymbol{Z}_i\Sigma^{-1}\boldsymbol{Z}_i)$$

- Tyler [1987] proposed covariance estimator $\widehat{\Sigma}$ as solution to

$$\Sigma = \frac{p}{n}\sum_{i=1}^{n}\frac{\boldsymbol{Z}_i\boldsymbol{Z}_i^T}{\boldsymbol{Z}_i^T\Sigma^{-1}\boldsymbol{Z}_i}$$

Tyler's estimator

- $\boldsymbol{X}_i \sim$ elliptical $(0, \boldsymbol{\Sigma})$ (we don't specify which distribution in elliptical family)
- Normalized sample $\boldsymbol{Z}_i \triangleq \frac{\boldsymbol{X}_i - \boldsymbol{\mu}}{\|\boldsymbol{X}_i - \boldsymbol{\mu}\|_2}$
  - pdf, Angular Central Gaussian Distribution:

$$f(z) = C|\Sigma|^{-1/2}(z^T \Sigma^{-1} z)^{-p/2}$$

  - negative log-likelihood function:

$$\frac{n}{2} \log|\Sigma| + \frac{p}{2} \sum_{i=1}^{n} \log(\boldsymbol{Z}_i \Sigma^{-1} \boldsymbol{Z}_i)$$

- Tyler [1987] proposed covariance estimator $\widehat{\Sigma}$ as solution to

$$\Sigma = \frac{p}{n} \sum_{i=1}^{n} \frac{\boldsymbol{Z}_i \boldsymbol{Z}_i^T}{\boldsymbol{Z}_i^T \Sigma^{-1} \boldsymbol{Z}_i}$$

Tyler's estimator

- $\boldsymbol{X}_i \sim$ elliptical $(0, \boldsymbol{\Sigma})$ (we don't specify which distribution in elliptical family)
- Normalized sample $\boldsymbol{Z}_i \triangleq \frac{\boldsymbol{X}_i - \boldsymbol{\mu}}{\|\boldsymbol{X}_i - \boldsymbol{\mu}\|_2}$
  - pdf, Angular Central Gaussian Distribution:

$$f(\boldsymbol{z}) = C|\boldsymbol{\Sigma}|^{-1/2}(\boldsymbol{z}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{z})^{-p/2}$$

  - negative log-likelihood function:

$$\frac{n}{2}\log|\Sigma| + \frac{p}{2}\sum_{i=1}^{n}\log(\boldsymbol{Z}_i\Sigma^{-1}\boldsymbol{Z}_i)$$

- Tyler [1987] proposed covariance estimator $\widehat{\Sigma}$ as solution to

$$\Sigma = \frac{p}{n}\sum_{i=1}^{n}\frac{\boldsymbol{Z}_i\boldsymbol{Z}_i^T}{\boldsymbol{Z}_i^T\Sigma^{-1}\boldsymbol{Z}_i}$$

Tyler's estimator

- $\boldsymbol{X}_i \sim$ elliptical $(0, \boldsymbol{\Sigma})$ (we don't specify which distribution in elliptical family)
- Normalized sample $\boldsymbol{Z}_i \triangleq \frac{\boldsymbol{X}_i - \boldsymbol{\mu}}{\|\boldsymbol{X}_i - \boldsymbol{\mu}\|_2}$
  - pdf, Angular Central Gaussian Distribution:

$$f(\boldsymbol{z}) = C|\boldsymbol{\Sigma}|^{-1/2}(\boldsymbol{z}^T \boldsymbol{\Sigma}^{-1} \boldsymbol{z})^{-p/2}$$

  - negative log-likelihood function:

$$\frac{n}{2} \log |\boldsymbol{\Sigma}| + \frac{p}{2} \sum_{i=1}^{n} \log(\boldsymbol{Z}_i \boldsymbol{\Sigma}^{-1} \boldsymbol{Z}_i)$$

- Tyler [1987] proposed covariance estimator $\widehat{\boldsymbol{\Sigma}}$ as solution to

$$\boldsymbol{\Sigma} = \frac{p}{n} \sum_{i=1}^{n} \frac{\boldsymbol{Z}_i \boldsymbol{Z}_i^T}{\boldsymbol{Z}_i^T \boldsymbol{\Sigma}^{-1} \boldsymbol{Z}_i}$$

Tyler's estimator

- $\boldsymbol{X}_i \sim$ elliptical $(0, \boldsymbol{\Sigma})$ (we don't specify which distribution in elliptical family)
- Normalized sample $\boldsymbol{Z}_i \triangleq \frac{\boldsymbol{X}_i - \boldsymbol{\mu}}{\|\boldsymbol{X}_i - \boldsymbol{\mu}\|_2}$
  - pdf, Angular Central Gaussian Distribution:

$$f(\boldsymbol{z}) = C|\boldsymbol{\Sigma}|^{-1/2}(\boldsymbol{z}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{z})^{-p/2}$$

  - negative log-likelihood function:

$$\frac{n}{2}\log|\boldsymbol{\Sigma}| + \frac{p}{2}\sum_{i=1}^{n}\log(\boldsymbol{Z}_i\boldsymbol{\Sigma}^{-1}\boldsymbol{Z}_i)$$

- Tyler [1987] proposed covariance estimator $\widehat{\boldsymbol{\Sigma}}$ as solution to

$$\boldsymbol{\Sigma} = \frac{p}{n}\sum_{i=1}^{n}\frac{\boldsymbol{Z}_i\boldsymbol{Z}_i^T}{\boldsymbol{Z}_i^T\boldsymbol{\Sigma}^{-1}\boldsymbol{Z}_i}$$

Tyler's M-estimator

- Tyler's estimator:

$$\boldsymbol{\Sigma} = \sum_{i=1}^{n} w_i \boldsymbol{Z}_i \boldsymbol{Z}_i^T$$

$$w_i = \frac{p}{n} \frac{1}{\boldsymbol{Z}_i^T \boldsymbol{\Sigma}^{-1} \boldsymbol{Z}_i}$$

- Why is Tyler's estimator robust to outliers? ☺
  - consider distance $d_i = \sqrt{\boldsymbol{Z}_i^T \Sigma^{-1} \boldsymbol{Z}_i}$
  - $w_i \propto \frac{1}{d_i^2}$, outliers are down-weighted
- Fixed-point equation, iterative algortihm:

$$\widetilde{\Sigma}_{t+1} = \frac{p}{n} \sum_{i=1}^{n} \frac{\boldsymbol{Z}_i \boldsymbol{Z}_i^T}{\boldsymbol{Z}_i \widehat{\Sigma}_t^{-1} \boldsymbol{Z}_i}$$

$$\widehat{\Sigma}_{t+1} = \frac{\widetilde{\Sigma}_{t+1}}{tr(\widetilde{\Sigma}_{t+1})/p}$$

- existence condition: $n > p$

Tyler's M-estimator

- Tyler's estimator:

$$\boldsymbol{\Sigma} = \sum_{i=1}^{n} w_i \boldsymbol{Z}_i \boldsymbol{Z}_i^T$$

$$w_i = \frac{p}{n} \frac{1}{\boldsymbol{Z}_i^T \boldsymbol{\Sigma}^{-1} \boldsymbol{Z}_i}$$

- Why is Tyler's estimator robust to outliers? ☺
  - consider distance $d_i = \sqrt{\boldsymbol{Z}_i^T \boldsymbol{\Sigma}^{-1} \boldsymbol{Z}_i}$
  - $w_i \propto \frac{1}{d_i^2}$, outliers are down-weighted
- Fixed-point equation, iterative algortihm:

$$\widetilde{\boldsymbol{\Sigma}}_{t+1} = \frac{p}{n} \sum_{i=1}^{n} \frac{\boldsymbol{Z}_i \boldsymbol{Z}_i^T}{\boldsymbol{Z}_i \widehat{\boldsymbol{\Sigma}}_t^{-1} \boldsymbol{Z}_i}$$

$$\widehat{\boldsymbol{\Sigma}}_{t+1} = \frac{\widetilde{\boldsymbol{\Sigma}}_{t+1}}{tr(\widetilde{\boldsymbol{\Sigma}}_{t+1})/p}$$

- existence condition: $n > p$

Tyler's M-estimator

- Tyler's estimator:

$$\Sigma = \sum_{i=1}^{n} w_i \boldsymbol{Z}_i \boldsymbol{Z}_i^T$$

$$w_i = \frac{p}{n} \frac{1}{\boldsymbol{Z}_i^T \Sigma^{-1} \boldsymbol{Z}_i}$$

- Why is Tyler's estimator robust to outliers? ☺
  - consider distance $d_i = \sqrt{\boldsymbol{Z}_i^T \Sigma^{-1} \boldsymbol{Z}_i}$
  - $w_i \propto \frac{1}{d_i^2}$, outliers are down-weighted
- Fixed-point equation, iterative algortihm:

$$\widetilde{\Sigma}_{t+1} = \frac{p}{n} \sum_{i=1}^{n} \frac{\boldsymbol{Z}_i \boldsymbol{Z}_i^T}{\boldsymbol{Z}_i \widehat{\Sigma}_t^{-1} \boldsymbol{Z}_i}$$

$$\widehat{\Sigma}_{t+1} = \frac{\widetilde{\Sigma}_{t+1}}{tr(\widetilde{\Sigma}_{t+1})/p}$$

- existence condition: $n > p$

Tyler's M-estimator

- Tyler's estimator:

$$\Sigma = \sum_{i=1}^{n} w_i \boldsymbol{Z}_i \boldsymbol{Z}_i^T$$

$$w_i = \frac{p}{n} \frac{1}{\boldsymbol{Z}_i^T \Sigma^{-1} \boldsymbol{Z}_i}$$

- Why is Tyler's estimator robust to outliers? ☺
  - consider distance $d_i = \sqrt{\boldsymbol{Z}_i^T \Sigma^{-1} \boldsymbol{Z}_i}$
  - $w_i \propto \frac{1}{d_i^2}$, outliers are down-weighted
- Fixed-point equation, iterative algortihm:

$$\widetilde{\Sigma}_{t+1} = \frac{p}{n} \sum_{i=1}^{n} \frac{\boldsymbol{Z}_i \boldsymbol{Z}_i^T}{\boldsymbol{Z}_i \widehat{\Sigma}_t^{-1} \boldsymbol{Z}_i}$$

$$\widehat{\Sigma}_{t+1} = \frac{\widetilde{\Sigma}_{t+1}}{tr(\widetilde{\Sigma}_{t+1})/p}$$

- existence condition: $n > p$

How to deal with small sample scenario?
Shrink! As for the sample covariance matrix

☺

Shrinkage covariance matrix: modified Tyler's estimator

- Modified Tyler's estimator [Chen *et al.* [2011]]

$$\widetilde{\Sigma}_{t+1} = (1-\rho)\frac{p}{n}\sum_{i=1}^{n}\frac{\boldsymbol{Z}_i\boldsymbol{Z}_i^T}{\boldsymbol{Z}_i\widehat{\Sigma}_t^{-1}\boldsymbol{Z}_i} + \rho\boldsymbol{I}, \;\; \rho \in [0,1]$$

$$\widehat{\Sigma}_{t+1} = \frac{\widetilde{\Sigma}_{t+1}}{tr(\widetilde{\Sigma}_{t+1})/p}$$

- the second step solves the identifiability issue
- Provable convergence
- Systematic way of choosing parameter ρ

Shrinkage covariance matrix: modified Tyler's estimator

- Modified Tyler's estimator [Chen *et al.* [2011]]

$$\widetilde{\Sigma}_{t+1} = (1-\rho)\frac{p}{n}\sum_{i=1}^{n}\frac{\boldsymbol{Z}_i\boldsymbol{Z}_i^T}{\boldsymbol{Z}_i\widehat{\Sigma}_t^{-1}\boldsymbol{Z}_i} + \rho\boldsymbol{I}, \;\; \rho\in[0,1]$$

$$\widehat{\Sigma}_{t+1} = \frac{\widetilde{\Sigma}_{t+1}}{tr(\widetilde{\Sigma}_{t+1})/p}$$

- the second step solves the identifiability issue
- Provable convergence
- Systematic way of choosing parameter ρ

Ok! But what if the mean value is unknown (cannot use Tyler's estimator anymore).
We still want to deal with heavy-tailed data!

My work ☺

- **Problem: estimate the covariance matrix** in **high-dimensional case** for **heavy-tailed distributions** with **the unknown mean vector**
    - Consider the specific distirbution in the class of elliptical

My work ☺

- **Problem: estimate the covariance matrix** in **high-dimensional case** for **heavy-tailed distributions** with **the unknown mean vector**
    - Consider the specific distirbution in the class of elliptical

Small Sample Regime & Robust Mean-Covariance Estimators

Elliptical distribution

- fix the distribution in the elliptical family:

$$f(\boldsymbol{x}) = C_{p,g}|\boldsymbol{\Sigma}|^{-1/2} g\left((\boldsymbol{x}-\boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right),$$

- find the MAP estimator:

$$\widehat{\Sigma}_\rho = (1-\rho)\frac{1}{n}\sum_{i=1}^{n} u(t)(\boldsymbol{X}_i-\boldsymbol{\mu})(\boldsymbol{X}_i-\boldsymbol{\mu})^\top + \rho\boldsymbol{I},$$

$$u(t) = -2\frac{g'(t)}{g(t)},$$

$$t = (\boldsymbol{X}_i-\boldsymbol{\mu})^\top \widehat{\Sigma}_\rho^{-1}(\boldsymbol{X}_i-\boldsymbol{\mu})$$

- we find the ρ which minimize $\mathrm{MSE}(\widehat{\Sigma}) = \mathbb{E}\left[\|\widehat{\Sigma}-\Sigma\|_F^2\right]$
  - no formula here, too long ...

Elliptical distribution

- fix the distribution in the elliptical family:

$$f(\boldsymbol{x}) = C_{p,g}|\boldsymbol{\Sigma}|^{-1/2} g\left((\boldsymbol{x}-\boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right),$$

- find the MAP estimator:

$$\widehat{\boldsymbol{\Sigma}}_\rho = (1-\rho)\frac{1}{n}\sum_{i=1}^{n} u(t)\,(\boldsymbol{X}_i-\boldsymbol{\mu})(\boldsymbol{X}_i-\boldsymbol{\mu})^\top + \rho \boldsymbol{I},$$

$$u(t) = -2\frac{g'(t)}{g(t)},$$

$$t = (\boldsymbol{X}_i-\boldsymbol{\mu})^\top \widehat{\boldsymbol{\Sigma}}_\rho^{-1}(\boldsymbol{X}_i-\boldsymbol{\mu})$$

- we find the ρ which minimize $\mathrm{MSE}(\widehat{\Sigma}) = \mathbb{E}\left[\left\|\widehat{\Sigma}-\Sigma\right\|_F^2\right]$
  - no formula here, too long ...

Elliptical distribution

- fix the distribution in the elliptical family:

$$f(\boldsymbol{x}) = C_{p,g}|\boldsymbol{\Sigma}|^{-1/2} g\left((\boldsymbol{x}-\boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right),$$

- find the MAP estimator:

$$\begin{aligned}
\widehat{\boldsymbol{\Sigma}}_\rho &= (1-\rho)\frac{1}{n}\sum_{i=1}^{n} u(t)\,(\boldsymbol{X}_i-\boldsymbol{\mu})(\boldsymbol{X}_i-\boldsymbol{\mu})^\top + \rho \boldsymbol{I}, \\
u(t) &= -2\frac{g'(t)}{g(t)}, \\
t &= (\boldsymbol{X}_i-\boldsymbol{\mu})^\top \widehat{\boldsymbol{\Sigma}}_\rho^{-1}(\boldsymbol{X}_i-\boldsymbol{\mu})
\end{aligned}$$

- we find the ρ which minimize $\mathrm{MSE}(\widehat{\boldsymbol{\Sigma}}) = \mathbb{E}\left[\|\widehat{\boldsymbol{\Sigma}}-\boldsymbol{\Sigma}\|_F^2\right]$
  - no formula here, too long ...

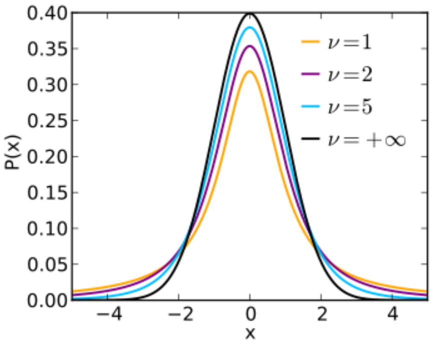
Example: Multivariate t-distribution

- Multivariate t-distribution with degree of freedom ν

$$t_p(\boldsymbol{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma}, \nu) = \frac{\Gamma(\frac{\nu+p}{2})}{(\pi\nu)^{p/2}\Gamma(\frac{\nu}{2})|\boldsymbol{\Sigma}|^{1/2}}\left[1 + \frac{\delta(\boldsymbol{x}, \boldsymbol{\mu}, \boldsymbol{\Sigma})}{\nu}\right]^{-\frac{\nu+p}{2}},$$

$\delta(\boldsymbol{x}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) = (\boldsymbol{x} - \boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x} - \boldsymbol{\mu})$ is the Mahalanobis distance between $\boldsymbol{x}$ and $\boldsymbol{\mu}$

- t-distribution has heavier tail than normal



MAP estimation

- $\boldsymbol{X}_i \sim t_p(\boldsymbol{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma}, \nu)$
- Algorithm:

$$\widehat{\boldsymbol{\mu}}_{t+1} = \frac{\sum_{i=1}^n \tau_i^{t+1} \boldsymbol{X}_i}{\sum_{i=1}^n \tau_i^{t+1}}, \quad \tau_i^{t+1} = \frac{\nu + p}{\nu + (\boldsymbol{X}_i - \widehat{\boldsymbol{\mu}}_t)\widehat{\boldsymbol{\Sigma}}_t^{-1}(\boldsymbol{X}_i - \widehat{\boldsymbol{\mu}}_t)}$$

$$\widehat{\boldsymbol{\Sigma}}_{t+1} = (1-\rho)\frac{p+\nu}{n}\sum_{i=1}^n \frac{(\boldsymbol{X}_i - \widehat{\boldsymbol{\mu}}_t)(\boldsymbol{X}_i - \widehat{\boldsymbol{\mu}}_t)^T}{(\boldsymbol{X}_i - \widehat{\boldsymbol{\mu}}_t)\widehat{\boldsymbol{\Sigma}}_t^{-1}(\boldsymbol{X}_i - \widehat{\boldsymbol{\mu}}_t) + \nu} + \rho \boldsymbol{I},$$

- ρ can be find by minimizing $\mathrm{MSE}(\widehat{\boldsymbol{\Sigma}}_{t+1})$

MAP estimation

- $\boldsymbol{X}_i \sim t_p(\boldsymbol{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma}, \nu)$
- Algorithm:

$$\widehat{\boldsymbol{\mu}}_{t+1} = \frac{\sum_{i=1}^n \tau_i^{t+1} \boldsymbol{X}_i}{\sum_{i=1}^n \tau_i^{t+1}}, \ \ \tau_i^{t+1} = \frac{\nu + p}{\nu + (\boldsymbol{X}_i - \widehat{\boldsymbol{\mu}}_t)\widehat{\boldsymbol{\Sigma}}_t^{-1}(\boldsymbol{X}_i - \widehat{\boldsymbol{\mu}}_t)}$$

$$\widehat{\boldsymbol{\Sigma}}_{t+1} = (1-\rho)\frac{p+\nu}{n}\sum_{i=1}^n \frac{(\boldsymbol{X}_i - \widehat{\boldsymbol{\mu}}_t)(\boldsymbol{X}_i - \widehat{\boldsymbol{\mu}}_t)^T}{(\boldsymbol{X}_i - \widehat{\boldsymbol{\mu}}_t)\widehat{\boldsymbol{\Sigma}}_t^{-1}(\boldsymbol{X}_i - \widehat{\boldsymbol{\mu}}_t) + \nu} + \rho \boldsymbol{I},$$

- ρ can be find by minimizing $\mathrm{MSE}(\widehat{\boldsymbol{\Sigma}}_{t+1})$

Optimal parameter

Theorem

The optimal shrinkage parameter ρ for the t-distribution with $\nu > 0$ degrees of freedom:

$$\rho^* = \frac{tr(\mathbf{\Sigma}^2)\left(1 + \frac{\nu}{p} - \frac{2}{p}\right) + p(\nu + p)}{tr(\mathbf{\Sigma}^2)\left((n+1)\left(\frac{\nu}{p} + 1\right) + \frac{2}{p}(n-1)\right) + (p+\nu)(p-n) - 2n}$$

- ρ^* depends on $tr(\Sigma^2), \nu$ which are unknown
  - to estimate $tr(\Sigma^2)$ we use the normalized sample covariance matrix:

$$\widehat{\boldsymbol{R}} = \frac{p}{n}\sum_{i=1}^{n}\frac{(\boldsymbol{X}_i - \widehat{\boldsymbol{X}})(\boldsymbol{X}_i - \widehat{\boldsymbol{X}})^T}{\|\boldsymbol{X}_i - \widehat{\boldsymbol{X}}\|^2}$$

  - To estimate degrees of freedom parameter ν we use extreme values theory, tail-index

Optimal parameter

Theorem

The optimal shrinkage parameter ρ for the t-distribution with $\nu > 0$ degrees of freedom:

$$\rho^* = \frac{tr(\mathbf{\Sigma}^2)\left(1 + \frac{\nu}{p} - \frac{2}{p}\right) + p(\nu + p)}{tr(\mathbf{\Sigma}^2)\left((n+1)\left(\frac{\nu}{p} + 1\right) + \frac{2}{p}(n-1)\right) + (p+\nu)(p-n) - 2n}$$

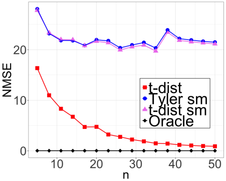
- ρ^* depends on $tr(\mathbf{\Sigma}^2), \nu$ which are unknown
  - to estimate $tr(\mathbf{\Sigma}^2)$ we use the normalized sample covariance matrix:

$$\widehat{\boldsymbol{R}} = \frac{p}{n} \sum_{i=1}^{n} \frac{(\boldsymbol{X}_i - \widehat{\boldsymbol{X}})(\boldsymbol{X}_i - \widehat{\boldsymbol{X}})^T}{\|\boldsymbol{X}_i - \widehat{\boldsymbol{X}}\|^2}$$

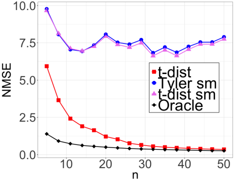
  - To estimate degrees of freedom parameter ν we use extreme values theory, tail-index

Experiments

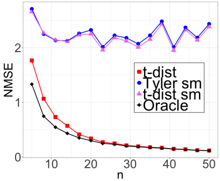
Multivariate t-distribution, $\nu = \{1, 3\}$



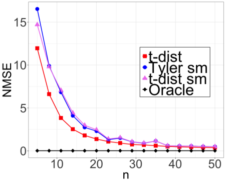
(a) $r = 0.1, \ \nu = 1$



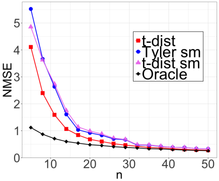
(b) $r = 0.7, \ \nu = 1$



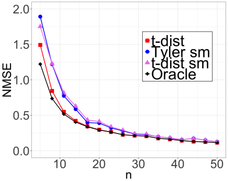
(c) $r = 0.9, \ \nu = 1$



(d) $r = 0.1, \ \nu = 3$



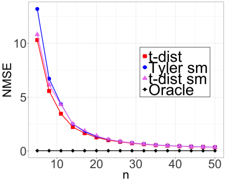
(e) $r = 0.7, \ \nu = 3$



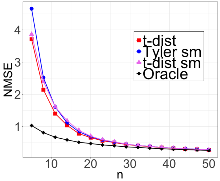
(f) $r = 0.9, \ \nu = 3$

Figure: AR(r) process: comparison of covariance estimators when $p = 50$ and $r \in \{0.1, 0.5, 0.9\}$ and the samples are from multivariate t-distribution distribution with $\nu \in \{1, 2, 3, 6, 10\}$ degrees of freedom; $\boldsymbol{\mu}$ is fixed to be 5 in all simulations

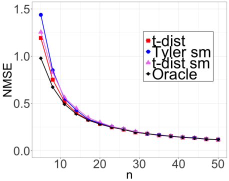
Multivariate t-distribution, $\nu = \{6, 10\}$



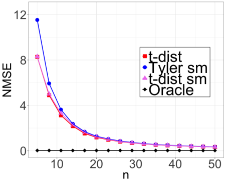
(a) $r = 0.1, \ \nu = 6$



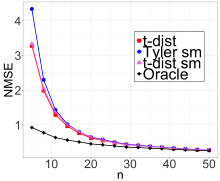
(b) $r = 0.7, \ \nu = 6$



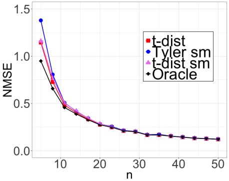
(c) $r = 0.9, \ \nu = 6$



(d) $r = 0.1, \ \nu = 10$



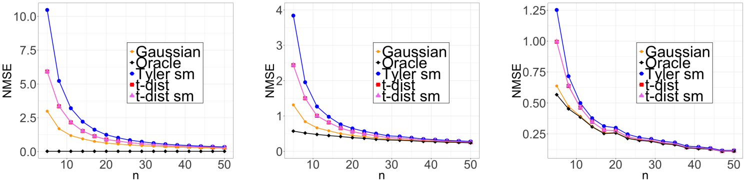
(e) $r = 0.7, \ \nu = 10$



(f) $r = 0.9, \ \nu = 10$

Figure: AR(r) process: comparison of covariance estimators when $p = 50$ and $r \in \{0.1, 0.5, 0.9\}$ and the samples are from multivariate t-distribution distribution with $\nu \in \{1, 2, 3, 6, 10\}$ degrees of freedom; $\boldsymbol{\mu}$ is fixed to be 5 in all simulations

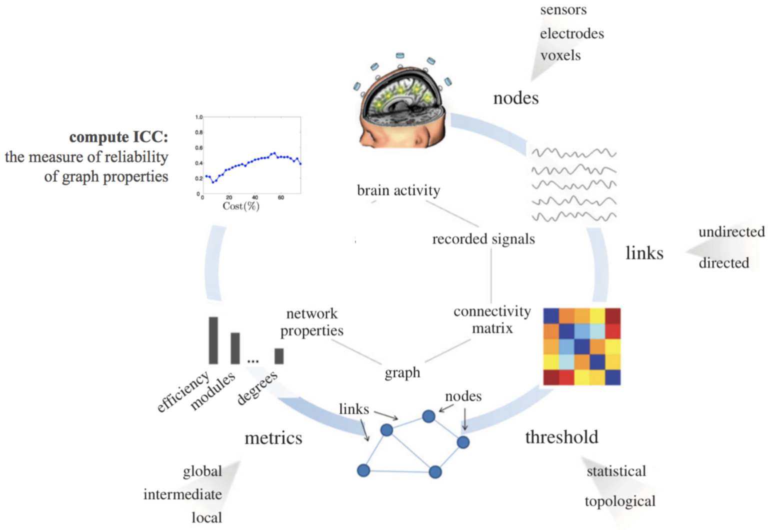
Gaussian distribution



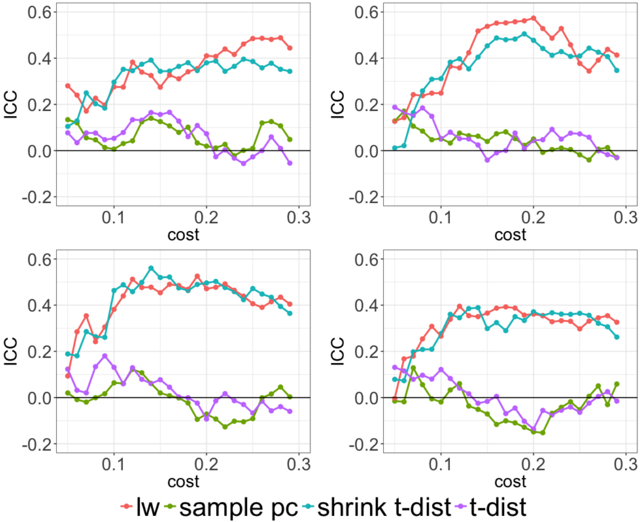
(a) $r = 0.1$ (b) $r = 0.7$ (c) $r = 0.9$

Figure: AR(r) process: $[\Sigma]_{ij} = \varrho^{|i-j|}$, $\varrho \in (0,1)$. Comparison of covariance estimators when $p = 50$ and $r \in \{0.1, 0.5, 0.9\}$ and the samples are from multivariate normal distribution; $\boldsymbol{\mu}$ is fixed to be 5 in all simulations

Application



ICC values



Chen, Yilun, Wiesel, Ami, & Hero, Alfred O. 2011. Robust shrinkage estimation of high-dimensional covariance matrices. *IEEE Transactions on Signal Processing*, **59**(9), 4097–4107.

Tyler, David E. 1987. A distribution-free M-estimator of multivariate scatter. *The Annals of Statistics*, 234–251.