High-dimensional covariance matrix estimation with applications in finance and genomic studies

Esa Ollila

Department of Signal Processing and Acoustics
Aalto University, Finland

Nov 12, 2018, ML coffee seminar

Aalto University

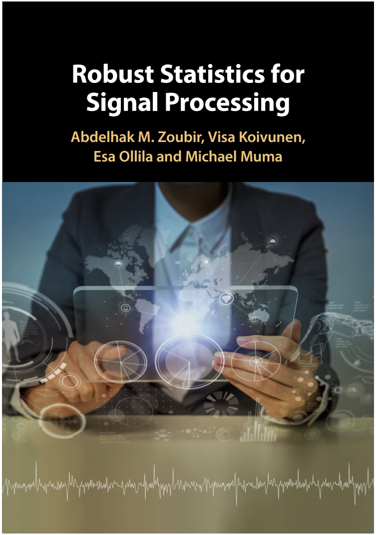
New Book! Published November 2018 by Cambridge University Press

Covers robust methods for

1. sparse regression
2. covariance estimation
3. bootstrap-based statistical inference
4. tensor data analysis
5. filtering
6. spectrum estimation
7. ...

Includes real-life applications and data analysis.

Matlab RobustSP Toolbox:
https://github.com/RobustSP/toolbox



Covariance estimation problem

- $\mathbf{x}$: p-variate (centered) random vector (p large)
- $\mathbf{x}_1, \ldots, \mathbf{x}_n$ i.i.d. realizations of $\mathbf{x}$.
- Problem: Find an estimate $\hat{\boldsymbol{\Sigma}}$ of the pos. def. covariance matrix

$$\boldsymbol{\Sigma} = \mathbb{E}\big[(\mathbf{x} - \boldsymbol{\mu})(\mathbf{x} - \boldsymbol{\mu})^\top\big] \in \mathbb{S}^{p \times p}_{++}$$

  where $\boldsymbol{\mu} = \mathbb{E}[\mathbf{x}]$.
- The sample covariance matrix (SCM),

$$\mathbf{S} = \frac{1}{n-1} \sum_{i=1}^{n} (\mathbf{x}_i - \overline{\mathbf{x}})(\mathbf{x}_i - \overline{\mathbf{x}})^\top,$$

  is the most commonly used estimator of $\boldsymbol{\Sigma}$.
- Challenges in HD:
  1. *Insufficient sample support (ISS)* case: $p > n$.
     $\Longrightarrow$ $\mathbf{S}$ is singular (non-invertible).
  2. *Low sample support (LSS)* (i.e., p of the same magnitude as n)
     $\Longrightarrow$ estimate $\hat{\boldsymbol{\Sigma}}$ has a lot of error.
  3. *Outliers* or heavy-tailed non-Gaussian data

Covariance estimation problem

- $\mathbf{x}$: p-variate (centered) random vector (p large)
- $\mathbf{x}_1, \ldots, \mathbf{x}_n$ i.i.d. realizations of $\mathbf{x}$.
- Problem: Find an estimate $\hat{\boldsymbol{\Sigma}}$ of the pos. def. covariance matrix

$$\boldsymbol{\Sigma} = \mathbb{E}\big[(\mathbf{x} - \boldsymbol{\mu})(\mathbf{x} - \boldsymbol{\mu})^\top\big] \in \mathbb{S}_{++}^{p \times p}$$

  where $\boldsymbol{\mu} = \mathbb{E}[\mathbf{x}]$.
- The sample covariance matrix (SCM),

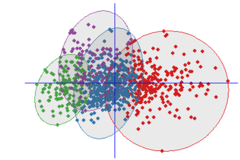
$$\mathbf{S} = \frac{1}{n-1} \sum_{i=1}^{n} (\mathbf{x}_i - \overline{\mathbf{x}})(\mathbf{x}_i - \overline{\mathbf{x}})^\top,$$

  is the most commonly used estimator of $\boldsymbol{\Sigma}$.
- Challenges in HD:
  1. *Insufficient sample support (ISS)* case: $p > n$.
     $\Longrightarrow \mathbf{S}$ is singular (non-invertible).
  2. *Low sample support (LSS)* (i.e., p of the same magnitude as n)
     $\Longrightarrow$ estimate $\hat{\boldsymbol{\Sigma}}$ has a lot of error.
  3. *Outliers* or heavy-tailed non-Gaussian data

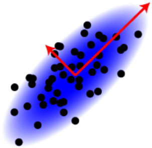
Why covariance estimation?

Portfolio selection

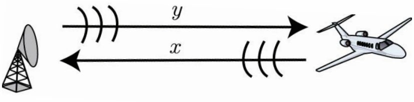
Classification/Clustering



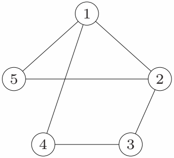
PCA



Radar detection



Graphical models



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

Bias-variance trade-off

- Any estimator $\hat{\mathbf{\Sigma}} \in \mathbb{S}_{++}^{p \times p}$ of $\mathbf{\Sigma}$ verifies

$$\begin{aligned} \mathsf{MSE}(\hat{\mathbf{\Sigma}}) &\triangleq \mathbb{E}\big[\|\hat{\mathbf{\Sigma}} - \mathbf{\Sigma}\|_{\mathrm{F}}^2\big] \qquad\qquad (\|\mathbf{A}\|_{\mathrm{F}}^2 = \mathrm{tr}(\mathbf{A}^2)) \\ &= \mathsf{var}(\hat{\mathbf{\Sigma}}) + \mathsf{bias}^2(\hat{\mathbf{\Sigma}}) \end{aligned}$$

- Since $\mathbf{S}$ is unbiased, $\mathsf{bias}^2(\mathbf{S}) = \|\mathbb{E}\big[\mathbf{S}\big] - \mathbf{\Sigma}\|_{\mathrm{F}}^2 = 0$, one has that

$$\mathsf{MSE}(\mathbf{S}) = \mathsf{var}(\mathbf{S})$$

but $\mathsf{var}(\mathbf{S})$ can be very large when $n \approx p$.

✓ Use an estimator $\hat{\mathbf{\Sigma}} = \mathbf{S}_\beta$ that shrinks $\mathbf{S}$ towards a structure (e.g., a scaled identity matrix) using a tuning (shrinkage) parameter β

- $\mathsf{MSE}(\hat{\mathbf{\Sigma}})$ can be reduced by introducing some bias.
- Positive definiteness of $\hat{\mathbf{\Sigma}}$ can be ensured.

Bias-variance trade-off

- Any estimator $\hat{\boldsymbol{\Sigma}} \in \mathbb{S}_{++}^{p\times p}$ of $\boldsymbol{\Sigma}$ verifies

$$\begin{aligned}\mathsf{MSE}(\hat{\boldsymbol{\Sigma}}) &\triangleq \mathbb{E}\big[\|\hat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}\|_{\mathrm{F}}^2\big] \qquad\qquad (\|\mathbf{A}\|_{\mathrm{F}}^2 = \mathrm{tr}(\mathbf{A}^2)) \\ &= \mathsf{var}(\hat{\boldsymbol{\Sigma}}) + \mathsf{bias}^2(\hat{\boldsymbol{\Sigma}})\end{aligned}$$

- Since $\mathbf{S}$ is unbiased, $\mathsf{bias}^2(\mathbf{S}) = \|\mathbb{E}\big[\mathbf{S}\big] - \boldsymbol{\Sigma}\|_{\mathrm{F}}^2 = 0$, one has that

$$\mathsf{MSE}(\mathbf{S}) = \mathsf{var}(\mathbf{S})$$

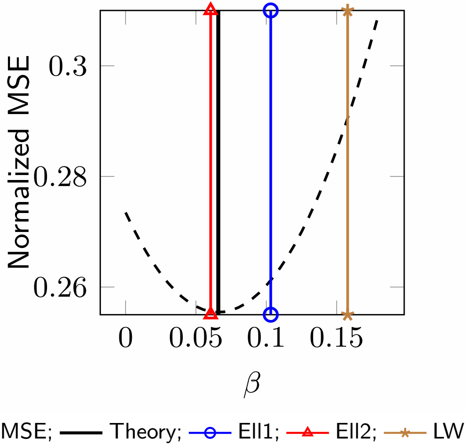
  but $\mathsf{var}(\mathbf{S})$ can be very large when $n \approx p$.

✓ Use an estimator $\hat{\boldsymbol{\Sigma}} = \mathbf{S}_\beta$ that shrinks $\mathbf{S}$ towards a structure (e.g., a scaled identity matrix) using a tuning (shrinkage) parameter β
  - $\mathsf{MSE}(\hat{\boldsymbol{\Sigma}})$ can be reduced by introducing some bias.
  - Positive definiteness of $\hat{\boldsymbol{\Sigma}}$ can be ensured.

Regularized SCM (RSCM) a lá Ledoit and Wolf:

$$\mathbf{S}_\beta = \beta\mathbf{S} + (1-\beta)[\mathrm{tr}(\mathbf{S})/p]\mathbf{I},$$

where $\beta \in [0,1)$ denotes the shrinkage (regularization) parameter.



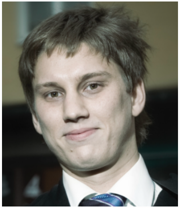
References

Optimal shrinkage covariance matrix estimation under random sampling from elliptical distributions arXiv:1808.10188 [stat.ME], August 2018.

MATLAB® toolbox: http://users.spa.aalto.fi/esollila/regscm/

Compressive regularized discriminant analysis of high-dimensional data with applications to microarray studies, Proc. *ICASSP'18*, Calgary, Canada, 2017, pp. 4204 –4208.

R-package: compressiveRDA @ https://github.com/mntabassm/compressiveRDA



joint work with Elias Raninen

joint work with M.N. Tabassum

Menu

1. Portfolio optimization
2. Ell-RSCM estimators
3. Estimates of oracle parameter
4. Compressive Regularized Discriminant Analysis

Modern portfolio theory (MPT)

- Mathematical framework by Markowitz [1952, 1959] for portfolio allocations that balances the return-risk tradeoff. MPT further developed by Tobin [1958], Sharpe [1964], Malkiel and Fama [1970]*
- A portfolio consist of p assets, e.g.:
  - equity securities (stocks), market indexes
  - fixed-income securities (e.g., government or corporate bonds)
  - currencies (exchange rates),
  - ...
- To use MPT one needs to estimate the mean vector $\boldsymbol{\mu}$ and the covariance matrix $\boldsymbol{\Sigma}$ of asset returns.

✗ often p, the number of assets is larger (or of similar magnitude) to n, the number of historical returns.

*Nobel price recipients: James Tobin (1981), Harry Markovitz (1990) and Willian F. Sharpe (1990), and Eugene F. Fama (2013)

Basic definitions

- Portfolio weight at (discrete) time index t:

$$\mathbf{w}_t = (w_{t,1}, \ldots, w_{t,p})^\top \qquad \text{s.t.} \qquad \mathbf{1}^\top \mathbf{w}_t = 1$$

- Let $C_{i,t} > 0$ be the price of the i^{th} asset
- The net return of the i^{th} asset over one interval is

$$r_{i,t} = \frac{C_{i,t} - C_{i,t-1}}{C_{i,t-1}} = \frac{C_{i,t}}{C_{i,t-1}} - 1 \in [-1, \infty)$$

- Single period net returns of p assets form a p-variate vector

$$\mathbf{r}_t = (r_{1,t}, \ldots, r_{p,t})^\top$$

- The portfolio net return at time $t+1$ is

$$R_{t+1} = \mathbf{w}_t^\top \mathbf{r}_{t+1} = \sum_{i=1}^{p} w_{i,t} r_{i,t+1}$$

- Assume historical returns $\{\mathbf{r}_t\}_{t=1}^n$ are i.i.d., so that

$$\boldsymbol{\mu} = \mathbb{E}[\mathbf{r}_t] \quad \text{and} \quad \boldsymbol{\Sigma} = \mathbb{E}\big[(\mathbf{r}_t - \boldsymbol{\mu}_t)(\mathbf{r}_t - \boldsymbol{\mu}_t)^\top\big]$$

holds for all t (so drop the index t from subscript).

- Let $\mathbf{r}$ denote the (random) vector of returns. Two key statistics of portfolio return $R = \mathbf{w}^\top \mathbf{r}$ are

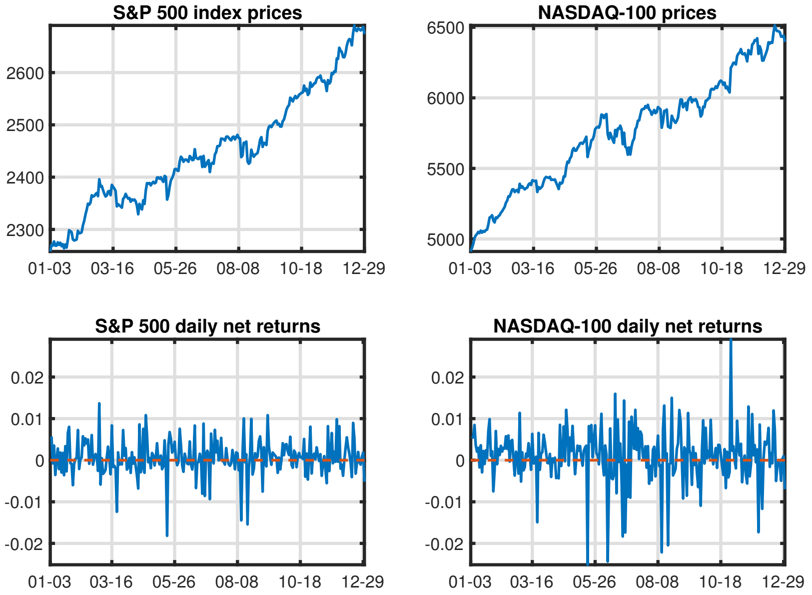
$$\begin{array}{ll} \text{mean return} & \mathbb{E}[R] = \mathbf{w}^\top \boldsymbol{\mu} \\ \text{variance (risk)} & \text{var}(R) = \mathbf{w}^\top \boldsymbol{\Sigma} \mathbf{w}. \end{array}$$

- Global minimum variance portfolio (GMVP) allocation strategy:

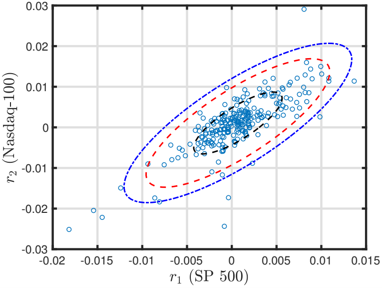
$$\underset{\mathbf{w} \in \mathbb{R}^p}{\text{minimize}}\ \mathbf{w}^\top \boldsymbol{\Sigma} \mathbf{w} \quad \text{subject to} \quad \mathbf{1}^\top \mathbf{w} = 1.$$

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

S&P 500 and Nasdaq-100 indexes for year 2017

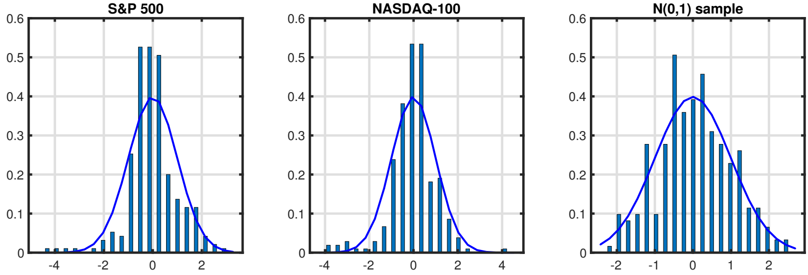


Are historical returns Gaussian?



Scatter plots and estimated 99%, 95% and 50% tolerance ellipses:

inside the 50% ellipse: 65.6% of returns
inside the 95% ellipse: 95.6% of returns



And stocks are unpredictable...

TECH GETS SLAMMED: Here's what you need to know

Follow @BiNordic | Follow @BINordic | 2,169 followers | Follow 28K

Elena Holodny | 09 Jun 2017 10:00 PM | 143

TECH stocks (Facebook, Apple, Amazon, Microsoft, Google) dropped drastically (in seconds) due to "fat finger" or **automated trade**.

...and there is that guy in the white house

And stocks are unpredictable...

TECH GETS SLAMMED: Here's what you need to know

Follow @BiNordic | Follow @BINordic | 2,169 followers | Follow 28K

Elena Holodny | 09 Jun 2017 10:00 PM | 143

TECH stocks (Facebook, Apple, Amazon, Microsoft, Google) dropped drastically (in seconds) due to "fat finger" or **automated trade**.

...and there is that guy in the white house

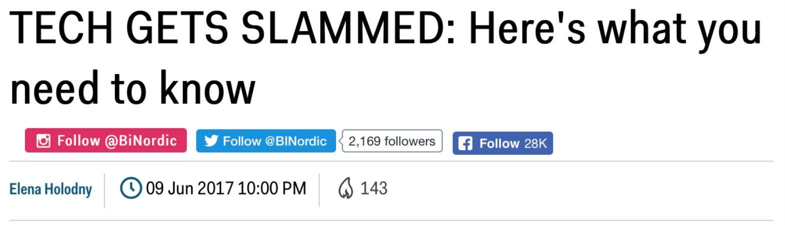
Donald J. Trump
@realDonaldTrump

Just had a long and very good conversation with President Xi Jinping of China. We talked about many subjects, with a heavy emphasis on Trade. Those discussions are moving along nicely with meetings being scheduled at the G-20 in Argentina. Also had good discussion on North Korea!

4:09 PM - Nov 1, 2018

93.8K 32K people are talking about this

And stocks are unpredictable...


TECH GETS SLAMMED: Here's what you need to know

Follow @BiNordic | Follow @BINordic | 2,169 followers | Follow 28K

Elena Holodny | 09 Jun 2017 10:00 PM | 143

TECH stocks (Facebook, Apple, Amazon, Microsoft, Google) dropped drastically (in seconds) due to "fat finger" or **automated trade**.

...and there is that guy in the white house

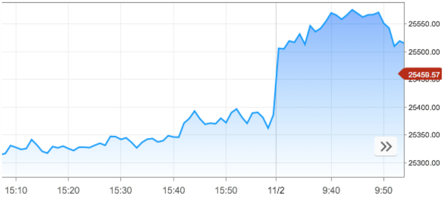

Donald J. Trump
@realDonaldTrump

Just had a long and very good conversation with President Xi Jinping of China. We talked about many subjects, with a heavy emphasis on Trade. Those discussions are moving along nicely with meetings being scheduled at the G-20 in Argentina. Also had good discussion on North Korea!

4:09 PM - Nov 1, 2018

93.8K 32K people are talking about this

Dow Jones Industrial Average:



Stock data analysis

We apply GMVP to stock data set consitsing of daily net returns computed from divident adjusted daily closing prices.

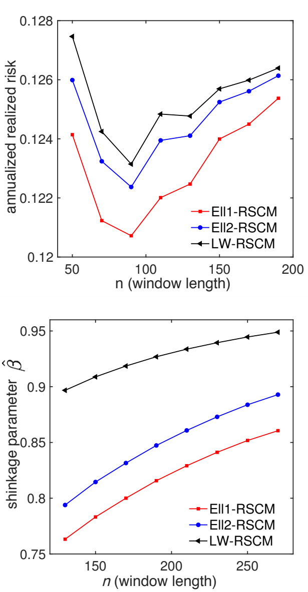
Data sets

- $p = 45$ stocks in Hang Seng Index (HSI), 1/2010 - 12/2011.
- $p = 396$ stocks in S&P500, 1/2016 - 4/2018.

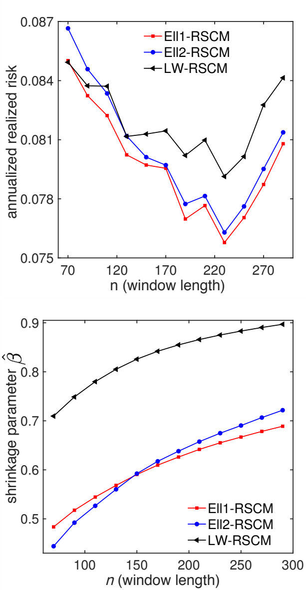
Sliding window method

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

HSI (Jan/2010 - Dec/2011)



SP500 (Jan/2016 - Apr/2018)



Menu

Regularized SCM and MMSE estimator

- Problem: We consider an estimator $\mathbf{S}_{\beta,\alpha} = \beta\mathbf{S} + \alpha\mathbf{I}$, where the weight (shrinkage) parameters are determined by solving

$$(\alpha_o, \beta_o) = \underset{\alpha,\beta>0}{\arg\min}\Big\{\mathbb{E}\Big[\big\|\beta\mathbf{S} + \alpha\mathbf{I} - \mathbf{\Sigma}\big\|_{\mathrm{F}}^2\Big]\Big\},$$

✗ (α_o, β_o) will depend on true *unknown* $\mathbf{\Sigma} \Rightarrow$ need to estimate (α_o, β_o)

- How to estimate (α_o, β_o)?
  - Ledoit and Wolf [2004] (no assumptions on $\mathbf{x} \sim F$)
  - Chen et al. [2010] (assumes Gaussianity)

$\Rightarrow$ we avoid strict assumptions, and simply assume that data is sampled from an unspecified elliptically symmetric distribution.

Important statistics

- Scale measure:

$$\eta = \frac{\mathrm{tr}(\mathbf{\Sigma})}{p} = \text{mean of eigenvalues}$$

- Sphericity measure:

$$\begin{aligned}\gamma &= \frac{p\,\mathrm{tr}(\mathbf{\Sigma}^2)}{\mathrm{tr}(\mathbf{\Sigma})^2} \\ &= \frac{\text{mean of (eigenvalue)}^2}{\text{(mean of eigenvalues)}^2}\end{aligned}$$

- $\gamma \in [1, p]$, and
  - $\gamma = 1$ iff $\mathbf{\Sigma} \propto \mathbf{I}$
  - $\gamma = p$ iff $\mathrm{rank}(\mathbf{\Sigma}) = 1$.

Optimal shrinkage parameters

Define normalized MSE of SCM $\mathbf{S}$ as

$$\mathrm{NMSE}(\mathbf{S}) = \frac{\mathbb{E}\big[\|\mathbf{S} - \boldsymbol{\Sigma}\|_{\mathrm{F}}^2\big]}{\|\boldsymbol{\Sigma}\|_{\mathrm{F}}^2}$$

Result 1

- Assume finite 4th-order moments.
- Optimal shrinkage parameters:

$$\beta_o = \frac{(\gamma - 1)}{(\gamma - 1) + \gamma \cdot \mathrm{NMSE}(\mathbf{S})}$$
$$\alpha_o = (1 - \beta_o)\eta.$$

and note that $\beta_o \in [0, 1)$.

$\Rightarrow$ one may use $\hat{\alpha}_0 = (1 - \hat{\beta}_0)\dfrac{\mathrm{tr}(\mathbf{S})}{p}$ and simply find an estimate $\hat{\beta}_0$ of β_0

Elliptically symmetric distributions

$\mathbf{x} \sim \mathcal{E}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}, g)$, when its pdf is of the form:

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

where $g : [0, \infty) \to [0, \infty)$ is the density generator:

- Gaussian distribution : $\mathbf{x} \sim \mathcal{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$: $g(t) = \exp(-t/2)$.
- t-distribution with $\nu > 4$ dof: $\mathbf{x} \sim t_\nu(\mathbf{0}, \boldsymbol{\Sigma})$, $g(t) = \dots$

Throughout, we assume finite 4th-order moments.

We also need to introduce the elliptical kurtosis parameter [Muirhead, 1982]:

$$\kappa = \frac{\mathbb{E}[\|\boldsymbol{\Sigma}^{-1/2}(\mathbf{x} - \boldsymbol{\mu})\|^4]}{p(p+2)} - 1$$

$$= \frac{1}{3} \cdot \{\text{kurtosis of } x_i\}$$

Elliptically symmetric distributions

$\mathbf{x} \sim \mathcal{E}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}, g)$, when its pdf is of the form:

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

where $g : [0, \infty) \to [0, \infty)$ is the density generator:

- Gaussian distribution : $\mathbf{x} \sim \mathcal{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$: $g(t) = \exp(-t/2)$.
- t-distribution with $\nu > 4$ dof: $\mathbf{x} \sim t_\nu(\mathbf{0}, \boldsymbol{\Sigma})$, $g(t) = \ldots$

Throughout, we assume finite 4th-order moments.

We also need to introduce the elliptical kurtosis parameter [Muirhead, 1982]:

$$\begin{aligned} \kappa &= \frac{\mathbb{E}[\|\boldsymbol{\Sigma}^{-1/2}(\mathbf{x} - \boldsymbol{\mu})\|^4]}{p(p+2)} - 1 \\ &= \frac{1}{3} \cdot \{\text{kurtosis of } x_i\} \end{aligned}$$

Result 2

Optimal shrinkage parameter when $\mathbf{x} \sim \mathcal{E}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}, g)$ is

$$\beta_o^{\mathsf{Ell}} = \frac{\gamma - 1}{\gamma - 1 + \kappa(2\gamma + p)/n + (\gamma + p)/(n-1)}$$

$\gamma :=$ sphericity $= \dfrac{p \,\mathrm{tr}(\boldsymbol{\Sigma}^2)}{\mathrm{tr}(\boldsymbol{\Sigma})^2}$ $\qquad$ $\kappa :=$ elliptical kurtosis

- Note: $\beta_o^{\mathsf{Ell}} = \beta_o^{\mathsf{Ell}}(\gamma, \kappa)$ depends on unknown γ and κ.
- Proof: Use Result 1 and the results:

$$\mathrm{MSE}(\mathbf{S}) = \mathbb{E}\left[\|\mathbf{S} - \boldsymbol{\Sigma}\|_{\mathrm{F}}^2\right] = \mathrm{tr}\{\mathrm{cov}(\mathrm{vec}(\mathbf{S}))\},$$

$$\mathrm{cov}(\mathrm{vec}(\mathbf{S})) = \left(\frac{1}{n-1} + \frac{\kappa}{n}\right)(\mathbf{I} + \mathbf{K}_p)(\boldsymbol{\Sigma} \otimes \boldsymbol{\Sigma}) + \frac{\kappa}{n}\mathrm{vec}(\boldsymbol{\Sigma})\mathrm{vec}(\boldsymbol{\Sigma})^\top,$$

where $\mathbf{K}_p$ is a commutation matrix ($\mathbf{K}_p \mathrm{vec}(\mathbf{A}) = \mathrm{vec}(\mathbf{A}^\top)$).

Result 2

Optimal shrinkage parameter when $\mathbf{x} \sim \mathcal{E}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}, g)$ is

$$\beta_o^{\mathsf{Ell}} = \frac{\gamma - 1}{\gamma - 1 + \kappa(2\gamma + p)/n + (\gamma + p)/(n-1)}$$

$\gamma :=$ sphericity $= \dfrac{p\,\mathrm{tr}(\boldsymbol{\Sigma}^2)}{\mathrm{tr}(\boldsymbol{\Sigma})^2}$ $\qquad$ $\kappa :=$ elliptical kurtosis

- Note: $\beta_o^{\mathsf{Ell}} = \beta_o^{\mathsf{Ell}}(\gamma, \kappa)$ depends on unknown γ and κ.
- Proof: Use Result 1 and the results:

$$\mathrm{MSE}(\mathbf{S}) = \mathbb{E}\left[\|\mathbf{S} - \boldsymbol{\Sigma}\|_{\mathrm{F}}^2\right] = \mathrm{tr}\{\mathrm{cov}(\mathrm{vec}(\mathbf{S}))\},$$

$$\mathrm{cov}(\mathrm{vec}(\mathbf{S})) = \left(\frac{1}{n-1} + \frac{\kappa}{n}\right)(\mathbf{I} + \mathbf{K}_p)(\boldsymbol{\Sigma} \otimes \boldsymbol{\Sigma}) + \frac{\kappa}{n}\mathrm{vec}(\boldsymbol{\Sigma})\mathrm{vec}(\boldsymbol{\Sigma})^\top,$$

where $\mathbf{K}_p$ is a commutation matrix ($\mathbf{K}_p\mathrm{vec}(\mathbf{A}) = \mathrm{vec}(\mathbf{A}^\top)$).

Result 2

Optimal shrinkage parameter when $\mathbf{x} \sim \mathcal{E}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}, g)$ is

$$\beta_o^{\mathsf{Ell}} = \frac{\gamma - 1}{\gamma - 1 + \kappa(2\gamma + p)/n + (\gamma + p)/(n-1)}$$

$\gamma :=$ sphericity $= \dfrac{p\,\mathrm{tr}(\boldsymbol{\Sigma}^2)}{\mathrm{tr}(\boldsymbol{\Sigma})^2}$ $\kappa :=$ elliptical kurtosis

- Note: $\beta_o^{\mathsf{Ell}} = \beta_o^{\mathsf{Ell}}(\gamma, \kappa)$ depends on unknown γ and κ.
- Proof: Use Result 1 and the results:

$$\mathrm{MSE}(\mathbf{S}) = \mathbb{E}\big[\|\mathbf{S} - \boldsymbol{\Sigma}\|_{\mathrm{F}}^2\big] = \mathrm{tr}\{\mathrm{cov}(\mathrm{vec}(\mathbf{S}))\},$$

$$\mathrm{cov}(\mathrm{vec}(\mathbf{S})) = \Big(\frac{1}{n-1} + \frac{\kappa}{n}\Big)(\mathbf{I} + \mathbf{K}_p)(\boldsymbol{\Sigma} \otimes \boldsymbol{\Sigma}) + \frac{\kappa}{n}\mathrm{vec}(\boldsymbol{\Sigma})\mathrm{vec}(\boldsymbol{\Sigma})^\top,$$

where $\mathbf{K}_p$ is a commutation matrix ($\mathbf{K}_p\mathrm{vec}(\mathbf{A}) = \mathrm{vec}(\mathbf{A}^\top)$).

Menu

Estimation of oracle shrinkage parameter

- Ell-RSCM estimator is defined as

$$\mathbf{S}_{\hat{\beta}} = \hat{\beta}\mathbf{S} + (1-\hat{\beta})[\mathrm{tr}(\mathbf{S})/p]\mathbf{I}$$

where

$$\begin{aligned}\hat{\beta} &= \beta_o^{\mathsf{Ell}}(\hat{\gamma}, \hat{\kappa}) \\ &= \frac{\hat{\gamma}-1}{\hat{\gamma}-1+\hat{\kappa}(2\hat{\gamma}+p)/n+(\hat{\gamma}+p)/(n-1)}\end{aligned}$$

- A consistent estimator of $\kappa = \frac{1}{3} \times \{$ kurtosis of $x_i\}$ is easy to find:

$$\hat{\kappa} = \frac{1}{3} \times \text{ average of sample kurtosis of } x_1, \ldots, x_p$$

- Next we consider two different estimates for sphericity γ.

Ell1-estimator of sphericity γ

- Sample sign covariance matrix [Visuri et al., 2000] is defined as

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

$$\text{where} \quad \hat{\boldsymbol{\mu}} = \arg\min_{\boldsymbol{\mu}} \sum_{i=1}^{n} \|\mathbf{x}_i - \boldsymbol{\mu}\|$$

- [Zhang and Wiesel, 2016] proposed a sphericity statistic

$$\hat{\gamma}^{\mathsf{Ell1}} = p\,\mathrm{tr}\left(\mathbf{S}^2_{\mathrm{sgn}}\right) - \frac{p}{n}$$

  and showed that $\hat{\gamma}^{\mathsf{Ell1}} \to \gamma$ under the random matrix theory regime:

$$n, p \to \infty \text{ and } \frac{p}{n} \to c_0,\ 0 < c_0 < \infty.$$

- Ell1-RSCM estimator uses $\hat{\beta} = \beta_o(\hat{\kappa}, \hat{\gamma}^{\mathsf{Ell1}})$.

Ell2-estimator of sphericity γ

Consider the statistic:

$$\hat{\vartheta} = b_n \left(\frac{\operatorname{tr}(\mathbf{S}^2)}{p} - a_n \, \frac{p}{n} \, \left[\frac{\operatorname{tr}(\mathbf{S})}{p} \right]^2 \right),$$

where

$$b_n = \frac{(\kappa + n)(n-1)^2}{(n-2)(3\kappa(n-1) + n(n+1))} \quad \& \quad a_n = \frac{n}{n+\kappa} \left(\frac{n}{n-1} + \kappa \right)$$

Note: For large n: $\hat{\vartheta} \approx \dfrac{\operatorname{tr}(\mathbf{S}^2)}{p} - (1+\kappa) \, \dfrac{p}{n} \, \left[\dfrac{\operatorname{tr}(\mathbf{S})}{p} \right]^2$.

$$\Rightarrow \frac{\operatorname{tr}(\mathbf{S}^2)}{p} \not\to \frac{\operatorname{tr}(\mathbf{\Sigma}^2)}{p} \quad \text{unless } \frac{p}{n} \to 0 \text{ as } p, n \to \infty$$

Ell2-estimator of sphericity γ

Consider the statistic:

$$\hat{\vartheta} = b_n \left(\frac{\operatorname{tr}(\mathbf{S}^2)}{p} - a_n \, \frac{p}{n} \, \left[\frac{\operatorname{tr}(\mathbf{S})}{p} \right]^2 \right),$$

where

$$b_n = \frac{(\kappa + n)(n-1)^2}{(n-2)(3\kappa(n-1) + n(n+1))} \quad \& \quad a_n = \frac{n}{n+\kappa} \left(\frac{n}{n-1} + \kappa \right)$$

Note: For large n: $\hat{\vartheta} \approx \dfrac{\operatorname{tr}(\mathbf{S}^2)}{p} - (1+\kappa)\, \dfrac{p}{n} \left[\dfrac{\operatorname{tr}(\mathbf{S})}{p} \right]^2$.

Result 4 (holds for any n and p)

$$\mathbb{E}[\hat{\vartheta}] = \frac{\operatorname{tr}(\mathbf{\Sigma}^2)}{p} = \text{mean of (eigenvalues)}^2$$

$$\Rightarrow \frac{\operatorname{tr}(\mathbf{S}^2)}{p} \not\to \frac{\operatorname{tr}(\mathbf{\Sigma}^2)}{p} \quad \text{unless } \frac{p}{n} \to 0 \text{ as } p, n \to \infty$$

Ell2-estimator of sphericity γ

Consider the statistic:

$$\hat{\vartheta} = b_n \left(\frac{\mathrm{tr}(\mathbf{S}^2)}{p} - a_n \frac{p}{n} \left[\frac{\mathrm{tr}(\mathbf{S})}{p} \right]^2 \right),$$

where

$$b_n = \frac{(\kappa + n)(n-1)^2}{(n-2)(3\kappa(n-1) + n(n+1))} \quad \& \quad a_n = \frac{n}{n+\kappa} \left(\frac{n}{n-1} + \kappa \right)$$

Note: For large n: $\hat{\vartheta} \approx \dfrac{\mathrm{tr}(\mathbf{S}^2)}{p} - (1+\kappa) \dfrac{p}{n} \left[\dfrac{\mathrm{tr}(\mathbf{S})}{p} \right]^2$.

Result 4 (holds for any n and p)

$$\mathbb{E}[\hat{\vartheta}] = \frac{\mathrm{tr}(\boldsymbol{\Sigma}^2)}{p} = \text{mean of (eigenvalues)}^2$$

$$\Rightarrow \frac{\mathrm{tr}(\mathbf{S}^2)}{p} \not\to \frac{\mathrm{tr}(\boldsymbol{\Sigma}^2)}{p} \quad \text{unless } \frac{p}{n} \to 0 \text{ as } p, n \to \infty$$

The sphericity measure

$$\gamma = \frac{\text{mean of (eigenvalues)}^2}{(\text{mean of eigenvalues})^2}$$

can be estimated by

$$\begin{aligned}\hat{\gamma}^{\mathsf{Ell2}} &= \frac{\hat{\vartheta}}{[\mathrm{tr}(\mathbf{S})/p]^2} \\ &= \hat{b}_n \left(\frac{p\,\mathrm{tr}(\mathbf{S}^2)}{\mathrm{tr}(\mathbf{S})^2} - \hat{a}_n \frac{p}{n} \right)\end{aligned}$$

where $\hat{a}_n = a_n(\hat{\kappa})$ and $\hat{b}_n = b_n(\hat{\kappa})$.

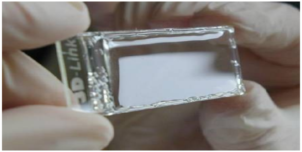
Ell2-RSCM estimator uses $\hat{\beta} = \beta_o(\hat{\kappa}, \hat{\gamma}^{\mathsf{Ell2}})$.

Menu

Microarray data analysis (MDA)

- Inferring large-scale covariance matrices from sparse genomic data is an ubiquitous problem in bioinformatics.

- microarrays measure the expression of genes (which genes are expressed and to what extent) in a given organism.
- A challenging framework:
  - $p = \#$ genes
  - $n = \sum_{g=1}^{G}(\#$ of obs. in class $g)$
  - $G = \#$ of classes



Dataset	n	p	G	Disease/organism
Su *et al.*	102	5,565	4	Multiple mammalian tissues
Yeoh *et al.*	248	12,625	6	Acute lymphoblastic leukemia
Ramaswamy *et al.*	190	16,063	14	Cancer

Table 1. Example of real data sets used in our analysis

Goals:

- Assign $\mathbf{x} \in \mathbb{R}^p$ to a correct class (out of G distinct classes).
- Reduce # of features without sacrificing the classification accuracy.

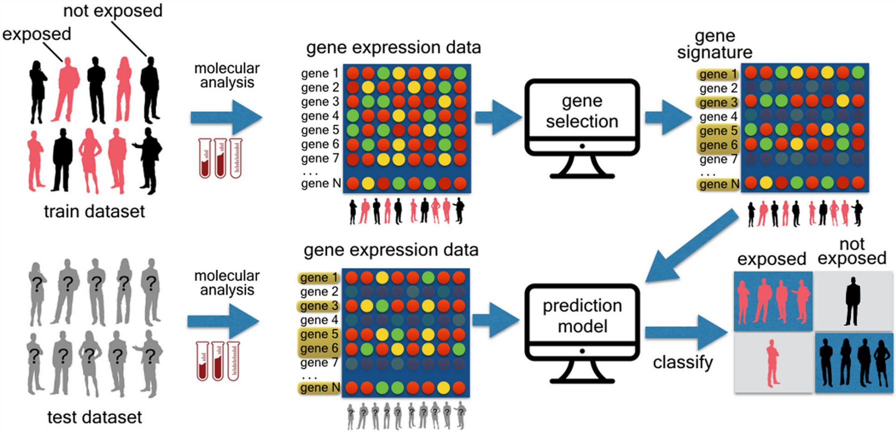


Figure from Giordano et al. [2018]

Benchmark methods:

- nearest shrunken centroid [Tibshirani et al., 2002]
- shrunken centroids regularized discriminant analysis [Guo et al., 2007].

Our method, compressive regularized discriminant analysis (CRDA):

- ✓ can be used as fast and accurate gene selection method and classification tool in MDA
- ✓ provides fewer misclassification errors than its competitors while at the same time achieving accurate feature elimination.

Compressive Regularized Discriminant Analysis (CRDA)

Classify $\mathbf{x} \in \mathbb{R}^p$ to class $\hat{g} = \arg\max_g \; d_g(\mathbf{x})$, where

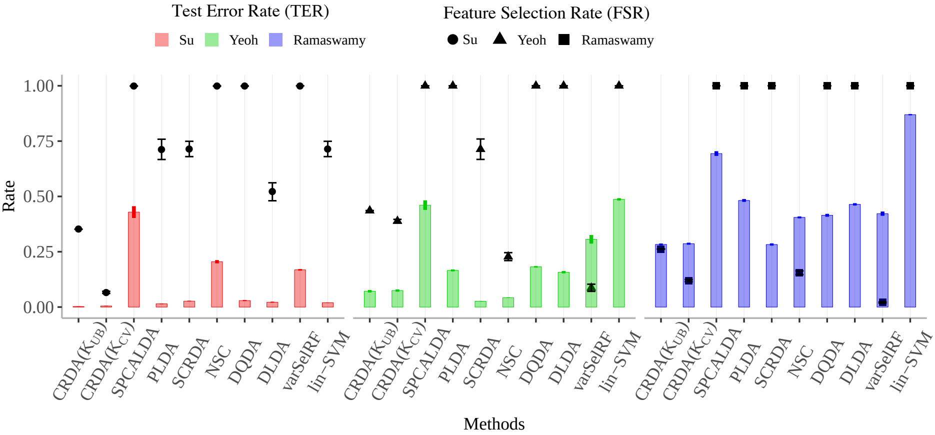
$$\mathbf{d}(\mathbf{x}) = \big(d_1(\mathbf{x}), \ldots, d_g(\mathbf{x}), \ldots, d_G(\mathbf{x})\big)$$
$$= \mathbf{x}^\top \hat{\mathcal{B}} - \frac{1}{2}\operatorname{diag}\big(\hat{\mathbf{M}}^\top \hat{\mathcal{B}}\big),$$

where $\hat{\mathbf{M}} = \begin{pmatrix} \overline{\mathbf{x}}_1 & \ldots & \overline{\mathbf{x}}_G \end{pmatrix}$, where $\overline{\mathbf{x}}_g$ is the sample mean of class g, and

$$\hat{\mathcal{B}} = H_K(\, \mathbf{S}_{\hat{\beta}}^{-1} \hat{\mathbf{M}}, q)$$

hard-thresholding operator $H_K(\cdot, q)$	Ell2-RSCM estimator $\mathbf{S}_{\hat{\beta}}$

- $H_K(\mathcal{B}, q)$ retains the elements of the K rows of $\mathcal{B}$ that possess largest ℓ_q norm and set elements of the other rows to zero.
- Regularization parameter is K (for a fixed ℓ_q-norm $q \in \{1, 2, \infty\}$). Our default choice for q is $q = \infty$.



Classification results for data sets of Table 1. Results are averaged over 10 training-to-test set splits (using 60%-to-40% ratio).

Benefits of CRDA:

a) performs effective gene selection

b) accurate classification

c) very fast to compute

Thank you!

References

Yilun Chen, Ami Wiesel, Yonina C Eldar, and Alfred O Hero. Shrinkage algorithms for mmse covariance estimation. *IEEE Trans. Signal Process.*, 58(10):5016–5029, 2010.

Maurizio Giordano, Kumar Parijat Tripathi, and Mario Rosario Guarracino. Ensemble of rankers for efficient gene signature extraction in smoke exposure classification. *BMC bioinformatics*, 19(2):48, 2018.

Yaqian Guo, Trevor Hastie, and Robert Tibshirani. Regularized linear discriminant analysis and its application in microarrays. *Biostatistics*, 8(1):86–100, 2007.

Olivier Ledoit and Michael Wolf. A well-conditioned estimator for large-dimensional covariance matrices. *Journal of multivariate analysis*, 88(2):365–411, 2004.

Burton G Malkiel and Eugene F Fama. Efficient capital markets: A review of theory and empirical work. *The journal of Finance*, 25(2):383–417, 1970.

Harry Markowitz. Portfolio selection. *The journal of finance*, 7(1):77–91, 1952.

Harry Markowitz. *Portfolio Selection, Efficent Diversification of Investments*. J. Wiley, 1959.

R. J. Muirhead. *Aspects of Multivariate Statistical Theory*. Wiley, New York, 1982. 704 pages.

William F Sharpe. Capital asset prices: A theory of market equilibrium under conditions of risk. *The journal of finance*, 19(3):425–442, 1964.

Robert Tibshirani, Trevor Hastie, Balasubramanian Narasimhan, and Gilbert Chu. Diagnosis of multiple cancer types by shrunken centroids of gene expression. *Proceedings of the National Academy of Sciences*, 99(10):6567–6572, 2002.

James Tobin. Liquidity preference as behavior towards risk. *The review of economic*