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
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 $\hat{\Sigma}$ of the pos. def. covariance matrix $\Sigma$
  \[\Sigma = \mathbb{E}[(\mathbf{x} - \mu)(\mathbf{x} - \mu)^\top] \in S_{++}^{p \times p}\]
  where $\mu = \mathbb{E}[\mathbf{x}]$.
- The sample covariance matrix (SCM),
  \[S = \frac{1}{n-1} \sum_{i=1}^{n} (\mathbf{x}_i - \overline{\mathbf{x}})(\mathbf{x}_i - \overline{\mathbf{x}})^\top,\]
  is the most commonly used estimator of $\Sigma$.

Challenges in HD:

1. Insufficient sample support (ISS) case: $p > n$.
   $\implies$ $S$ is singular (non-invertible).
2. Low sample support (LSS) (i.e., $p$ of the same magnitude as $n$)
   $\implies$ estimate $\hat{\Sigma}$ 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 $\hat{\Sigma}$ of the pos. def. covariance matrix

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

where $\mu = \mathbb{E}[\mathbf{x}]$.
- 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 $\Sigma$.
- Challenges in HD:
  1. Insufficient sample support (ISS) case: $p > n$.  
     $\implies$ $\mathbf{S}$ is singular (non-invertible).
  2. Low sample support (LSS) (i.e., $p$ of the same magnitude as $n$)  
     $\implies$ estimate $\hat{\Sigma}$ 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

Gaussian graphical model

$n$-dimensional Gaussian vector $x = (x_1, \ldots, x_n) \sim N(0, \Sigma)$

$xi, xj$ are conditionally independent (given the rest of $x$) if $(\Sigma^{-1})_{ij} = 0$

$\Sigma^{-1} = \begin{bmatrix}
\bullet & \bullet & 0 & \bullet & \bullet \\
\bullet & \bullet & \bullet & \bullet & 0 \\
0 & \bullet & \bullet & \bullet & 0 \\
\bullet & 0 & \bullet & \bullet & 0 \\
\bullet & \bullet & 0 & 0 & \bullet
\end{bmatrix}$
Bias-variance trade-off

- Any estimator $\hat{\Sigma} \in S_{++}^{p \times p}$ of $\Sigma$ verifies

$$
\text{MSE}(\hat{\Sigma}) \triangleq \mathbb{E}[\|\hat{\Sigma} - \Sigma\|_F^2] = \text{var}(\hat{\Sigma}) + \text{bias}^2(\hat{\Sigma})
$$

- Since $S$ is unbiased, $\text{bias}^2(S) = \|\mathbb{E}[S] - \Sigma\|_F^2 = 0$, one has that

$$
\text{MSE}(S) = \text{var}(S)
$$

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

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

  - $\text{MSE}(\hat{\Sigma})$ can be reduced by introducing some bias.
  - Positive definiteness of $\hat{\Sigma}$ can be ensured.
Any estimator $\hat{\Sigma} \in \mathbb{S}^{p \times p}_{++}$ of $\Sigma$ verifies

$$
\text{MSE}(\hat{\Sigma}) \overset{\Delta}{=} \mathbb{E}[\|\hat{\Sigma} - \Sigma\|_F^2] = \text{var}(\hat{\Sigma}) + \text{bias}^2(\hat{\Sigma})
$$

Since $S$ is unbiased, $\text{bias}^2(S) = \|\mathbb{E}[S] - \Sigma\|_F^2 = 0$, one has that

$$
\text{MSE}(S) = \text{var}(S)
$$

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

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

- $\text{MSE}(\hat{\Sigma})$ can be reduced by introducing some bias.
- Positive definiteness of $\hat{\Sigma}$ can be ensured.
Regularized SCM (RSCM) a l’a Ledoit and Wolf:

\[ S_\beta = \beta S + (1 - \beta) \left[ \frac{\text{tr}(S)}{p} \right] I, \]

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

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
1. Portfolio optimization

2. Ell-RSCM estimators

3. Estimates of oracle parameter

4. Compressive Regularized Discriminant Analysis
Modern portfolio theory (MPT)


- 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 $\mu$ and the covariance matrix $\Sigma$ of asset returns.

  $\times$ 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 $t$:

$$w_t = (w_{t,1}, \ldots, w_{t,p})^\top \quad \text{s.t.} \quad 1^\top w_t = 1$$

- 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

$$r_t = (r_{1,t}, \ldots, r_{p,t})^\top$$

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

$$R_{t+1} = w_t^\top r_{t+1} = \sum_{i=1}^{p} w_{i,t} r_{i,t+1}$$
Assume historical returns \( \{r_t\}_{t=1}^n \) are i.i.d., so that

\[
\mu = \mathbb{E}[r_t] \quad \text{and} \quad \Sigma = \mathbb{E}[(r_t - \mu_t)(r_t - \mu_t)^\top]
\]

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

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

- mean return \( \mathbb{E}[R] = w^\top \mu \)
- variance (risk) \( \text{var}(R) = w^\top \Sigma w \).

Global minimum variance portfolio (GMVP) allocation strategy:

\[
\min_{w \in \mathbb{R}^p} w^\top \Sigma w \quad \text{subject to} \quad 1^\top w = 1.
\]

\[
\Rightarrow w_o = \frac{\Sigma^{-1}1}{1^\top \Sigma^{-1}1}.
\]
S&P 500 and Nasdaq-100 indexes for year 2017

S&P 500 index prices

NASDAQ-100 prices

S&P 500 daily net returns

NASDAQ-100 daily net returns
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

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

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**

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

---

**Dow Jones Industrial Average:**

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
Stock data analysis

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

**Data sets**


**Sliding window method**

- At day $t$, we use the previous $n$ days to estimate $\Sigma$ and $w$.
- 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.
1 Portfolio optimization

2 Ell-RSCM estimators

3 Estimates of oracle parameter

4 Compressive Regularized Discriminant Analysis
Problem: We consider an estimator $S_{\beta, \alpha} = \beta S + \alpha I$, where the weight (shrinkage) parameters are determined by solving

$$(\alpha_o, \beta_o) = \arg \min_{\alpha, \beta > 0} \{ \mathbb{E} \left[ \| \beta S + \alpha I - \Sigma \|^2_F \right] \},$$

$(\alpha_o, \beta_o)$ will depend on true unknown $\Sigma$ ⇒ need to estimate $(\alpha_o, \beta_o)$

How to estimate $(\alpha_o, \beta_o)$?

- Ledoit and Wolf [2004] (no assumptions on $x \sim F$)
- Chen et al. [2010] (assumes Gaussianity)

⇒ we avoid strict assumptions, and simply assume that data is sampled from an unspecified elliptically symmetric distribution.
Important statistics

- **Scale measure:**

\[
\eta = \frac{\text{tr}(\Sigma)}{p} = \text{mean of eigenvalues}
\]

- **Sphericity measure:**

\[
\gamma = \frac{p \text{tr}(\Sigma^2)}{\text{tr}(\Sigma)^2} = \frac{\text{mean of (eigenvalue)}^2}{(\text{mean of eigenvalues})^2}
\]

- \( \gamma \in [1, p] \), and
  - \( \gamma = 1 \) iff \( \Sigma \propto I \)
  - \( \gamma = p \) iff \( \text{rank}(\Sigma) = 1 \).
**Optimal shrinkage parameters**

Define normalized MSE of SCM $S$ as

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

<table>
<thead>
<tr>
<th>Result 1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Assume finite 4th-order moments.</td>
</tr>
<tr>
<td>Optimal shrinkage parameters:</td>
</tr>
</tbody>
</table>

$$\beta_o = \frac{(\gamma - 1)}{(\gamma - 1) + \gamma \cdot \text{NMSE}(S)}$$

$$\alpha_o = (1 - \beta_o)\eta.$$ 

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

⇒ one may use $\hat{\alpha}_0 = (1 - \hat{\beta}_0)\frac{\text{tr}(S)}{p}$ and simply find an estimate $\hat{\beta}_0$ of $\beta_0$. 
Elliptically symmetric distributions

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

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

where \( g : [0, \infty) \rightarrow [0, \infty) \) is the density generator:

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

Throughout, we assume finite 4th-order moments.

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

\[
 \kappa = \frac{\mathbb{E}[\| \Sigma^{-1/2}(x - \mu) \|^4]}{p(p + 2)} - 1
 = \frac{1}{3} \cdot \{\text{kurtosis of } x_i\}
\]
Elliptically symmetric distributions

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

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

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

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

Throughout, we assume finite 4th-order moments.

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

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

\[
= \frac{1}{3} \cdot \{ \text{kurtosis of } x_i \}
\]
Result 2

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

$$\beta^\text{Ell}_o = \frac{\gamma - 1}{\gamma - 1 + \kappa (2\gamma + p)/n + (\gamma + p)/(n - 1)}$$

$$\gamma := \text{sphericity} = \frac{p \text{tr}(\Sigma^2)}{\text{tr}(\Sigma)^2} \quad \kappa := \text{elliptical kurtosis}$$

Note: $\beta^\text{Ell}_o = \beta^\text{Ell}_o(\gamma, \kappa)$ depends on unknown $\gamma$ and $\kappa$.

Proof: Use Result 1 and the results:

$$\text{MSE}(S) = \mathbb{E}[\|S - \Sigma\|^2_F] = \text{tr}\{\text{cov}(\text{vec}(S))\},$$

$$\text{cov}(\text{vec}(S)) = \left(\frac{1}{n-1} + \frac{\kappa}{n}\right)(I + K_p)(\Sigma \otimes \Sigma) + \frac{\kappa}{n} \text{vec}(\Sigma)\text{vec}(\Sigma)^\top,$$

where $K_p$ is a commutation matrix ($K_p \text{vec}(A) = \text{vec}(A^\top)$).
Optimal shrinkage parameter when $\mathbf{x} \sim \mathcal{E}_p(\mathbf{\mu}, \Sigma, g)$ is

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

$$
\gamma := \text{sphericity} = \frac{p \text{tr}(\Sigma^2)}{\text{tr}(\Sigma)^2}
$$

$\kappa := \text{elliptical kurtosis}$

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

$$
\text{MSE}(S) = \mathbb{E}[\|S - \Sigma\|_F^2] = \text{tr}\{\text{cov}(\text{vec}(S))\},
$$

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

where $\mathbf{K}_p$ is a commutation matrix ($\mathbf{K}_p \text{vec}(\mathbf{A}) = \text{vec}(\mathbf{A}^\top)$).
Result 2

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

\[
\beta^\text{Ell}_o = \frac{\gamma - 1}{\gamma - 1 + \kappa(2\gamma + p)/n + (\gamma + p)/(n - 1)}
\]

\( \gamma := \text{sphericity} = \frac{p\text{tr}(\Sigma^2)}{\text{tr}(\Sigma)^2} \)

\( \kappa := \text{elliptical kurtosis} \)

Note: \( \beta^\text{Ell}_o = \beta^\text{Ell}_o(\gamma, \kappa) \) depends on unknown \( \gamma \) and \( \kappa \).

Proof: Use Result 1 and the results:

\[
\text{MSE}(S) = \mathbb{E}\left[\|S - \Sigma\|_F^2\right] = \text{tr}\{\text{cov}(\text{vec}(S))\},
\]

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

where \( \mathbf{K}_p \) is a commutation matrix (\( \mathbf{K}_p\text{vec}(\mathbf{A}) = \text{vec}(\mathbf{A}^\top) \)).
1 Portfolio optimization
2 Ell-RSCM estimators
3 Estimates of oracle parameter
4 Compressive Regularized Discriminant Analysis
Estimation of oracle shrinkage parameter

- Ell-RSCM estimator is defined as

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

where

\[ \hat{\beta} = \beta_o^{\text{Ell}}(\hat{\gamma}, \hat{\kappa}) \]

\[ = \frac{\hat{\gamma} - 1}{\hat{\gamma} - 1 + \hat{\kappa}(2\hat{\gamma} + p)/n + (\hat{\gamma} + p)/(n - 1)} \]

- A consistent estimator of \( \kappa = \frac{1}{3} \times \{ \text{kurtosis of } x_i \} \) 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 \( \gamma \).
Ell1-estimator of sphericity $\gamma$

- Sample sign covariance matrix [Visuri et al., 2000] is defined as
  \[
  S_{sgn} = \frac{1}{n} \sum_{i=1}^{n} \frac{(x_i - \hat{\mu})(x_i - \hat{\mu})^\top}{\|x_i - \hat{\mu}\|^2},
  \]
  where $\hat{\mu} = \arg\min_{\mu} \sum_{i=1}^{n} \|x_i - \mu\|$

- [Zhang and Wiesel, 2016] proposed a sphericity statistic
  \[
  \hat{\gamma}^{Ell1} = p \text{tr} \left( S_{sgn}^2 \right) - \frac{p}{n}
  \]
  and showed that $\hat{\gamma}^{Ell1} \to \gamma$ 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 $\hat{\beta} = \beta_o(\hat{\kappa}, \hat{\gamma}^{Ell1})$. 
Ell2-estimator of sphericity $\gamma$

Consider the statistic:

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

where

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

**Note:** For large $n$: 

$$\hat{\vartheta} \approx \frac{\text{tr}(S^2)}{p} - (1 + \kappa) \frac{p}{n} \left[ \frac{\text{tr}(S)}{p} \right]^2.$$
Ell2-estimator of sphericity $\gamma$

Consider the statistic:

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

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 $n$: $\hat{\vartheta} \approx \frac{\text{tr}(S^2)}{p} - (1 + \kappa) \frac{p}{n} \left[ \frac{\text{tr}(S)}{p} \right]^2$.

**Result 4** (holds for any $n$ and $p$)

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

$$\Rightarrow \frac{\text{tr}(S^2)}{p} \neq \frac{\text{tr}(\Sigma^2)}{p} \quad \text{unless} \quad \frac{p}{n} \to 0 \quad \text{as} \quad p, n \to \infty$$
Ell2-estimator of sphericity $\gamma$

Consider the statistic:

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

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 $n$: $\hat{\vartheta} \approx \frac{\text{tr}(S^2)}{p} - (1 + \kappa) \frac{p}{n} \left[ \frac{\text{tr}(S)}{p} \right]^2$.

Result 4 (holds for any $n$ and $p$)

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

$$\Rightarrow \frac{\text{tr}(S^2)}{p} \not\rightarrow \frac{\text{tr}(\Sigma^2)}{p} \quad \text{unless} \quad \frac{p}{n} \rightarrow 0 \text{ as } p, n \rightarrow \infty.$$
The sphericity measure

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

can be estimated by

\[ \hat{\gamma}_{\text{Ell2}} = \frac{\hat{\vartheta}}{[\text{tr}(S)/p]^2} \]

\[ = \hat{b}_n \left( \frac{p \text{tr}(S^2)}{\text{tr}(S)^2} - \hat{a}_n \frac{p}{n} \right) \]

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

Ell2-RSCM estimator uses \( \hat{\beta} = \beta_o(\hat{\kappa}, \hat{\gamma}_{\text{Ell2}}) \).
1. Portfolio optimization
2. Ell-RSCM estimators
3. Estimates of oracle parameter
4. Compressive Regularized Discriminant Analysis
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:
  - $p = \# \text{ genes}$
  - $n = \sum_{g=1}^{G} (\# \text{ of obs. in class } g)$
  - $G = \# \text{ of classes}$

<table>
<thead>
<tr>
<th>Dataset</th>
<th>$n$</th>
<th>$p$</th>
<th>$G$</th>
<th>Disease/organism</th>
</tr>
</thead>
<tbody>
<tr>
<td>Su et al.</td>
<td>102</td>
<td>5,565</td>
<td>4</td>
<td>Multiple mammalian tissues</td>
</tr>
<tr>
<td>Yeoh et al.</td>
<td>248</td>
<td>12,625</td>
<td>6</td>
<td>Acute lymphoblastic leukemia</td>
</tr>
<tr>
<td>Ramaswamy et al.</td>
<td>190</td>
<td>16,063</td>
<td>14</td>
<td>Cancer</td>
</tr>
</tbody>
</table>

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

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

\[
d(x) = (d_1(x), \ldots, d_g(x), \ldots, d_G(x))
\]

\[
= x^\top \hat{B} - \frac{1}{2} \text{diag}(\hat{M}^\top \hat{B}),
\]

where \( \hat{M} = (\bar{x}_1 \ldots \bar{x}_G) \), where \( \bar{x}_g \) is the sample mean of class \( g \), and

\[
\hat{B} = H_K(\hat{S}_{\hat{\beta}}^{-1} \hat{M}, q)
\]

- \( H_K(B, q) \) retains the elements of the \( K \) rows of \( B \) that possess largest \( \ell_q \) norm and set elements of the other rows to zero.
- Regularization parameter is \( K \) (for a fixed \( \ell_q \)-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


