Sparse Estimation of a Covariance Matrix (2011)

Jacob Bien
Robert Tibshirani

April 24, 2014
Presented by Sen Zhao

Covariance Estimation

- Suppose we have measured p covariates on n subjects. For example:
  - The expression levels of p genes on n people;
  - The relative abundances of p species at n locations.
- We want to know the covariance matrix between those p covariates.
  - To determine the gene / gene or species / species interaction.
  - Specifically, we may want to know whether two covariates are marginally independent (i.e. covariance $= 0$).

Introduction

- Suppose $\boldsymbol{X}_1, ..., \boldsymbol{X}_n \sim_{iid} N_p(\mathbf{0}, \boldsymbol{\Sigma})$. We want to estimate $\boldsymbol{\Sigma}$.
- Relatively easy when $n >> p$. Use MLE.

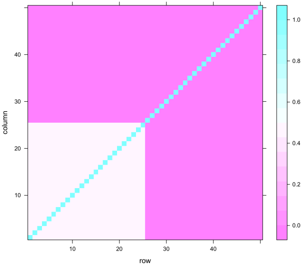
$$l(\boldsymbol{\Sigma}) = -\frac{np}{2}\log(2\pi) - \frac{n}{2}\log\det(\boldsymbol{\Sigma}) - \frac{n}{2}tr(\boldsymbol{\Sigma}^{-1}\boldsymbol{S}),$$

  where $\boldsymbol{S} = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{X}_i \boldsymbol{X}_i^T$.
- When $p > n$, things become much harder.
- Heuristically, we have np values in the dataset, but we have $\frac{p(p+1)}{2}$ parameters to estimate.

A Simple Simulation Study

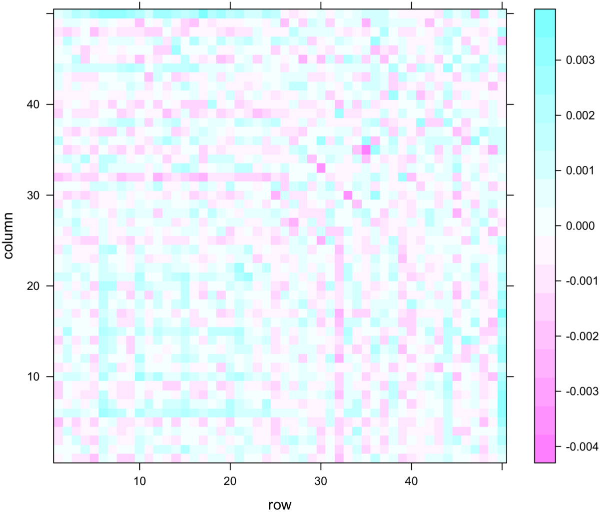
- Let $\boldsymbol{X}_1, ..., \boldsymbol{X}_n \sim_{iid} N_p(\mathbf{0}, \boldsymbol{\Sigma})$. $(n, p) \in \{(1000, 50), (50, 1000)\}$.

$$Cov(\boldsymbol{X}_i, \boldsymbol{X}_j) = \begin{cases} 1 & : i = j \\ 0.5 & : i \neq j, i, j \in \{1, ..., \frac{p}{2}\} \\ 0 & : otherwise \end{cases}$$

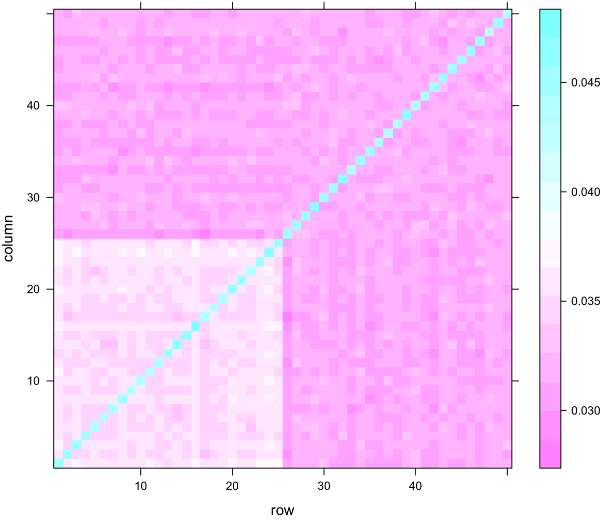


- Estimate $\boldsymbol{\Sigma}$ using MLE. Average over 1000 repetitions.

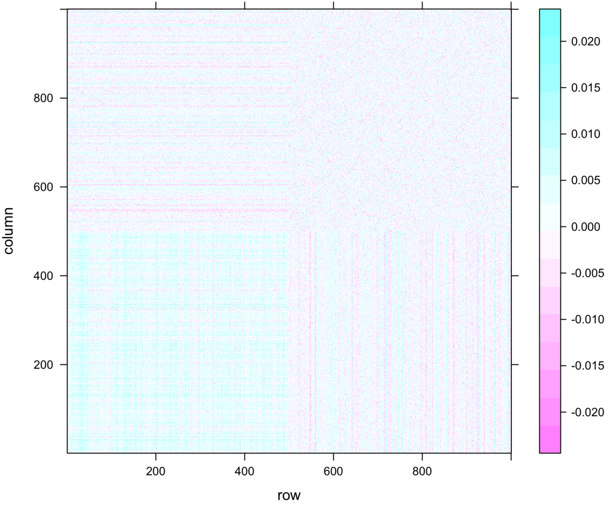
Simulation: n = 1000, p = 50 (Bias)



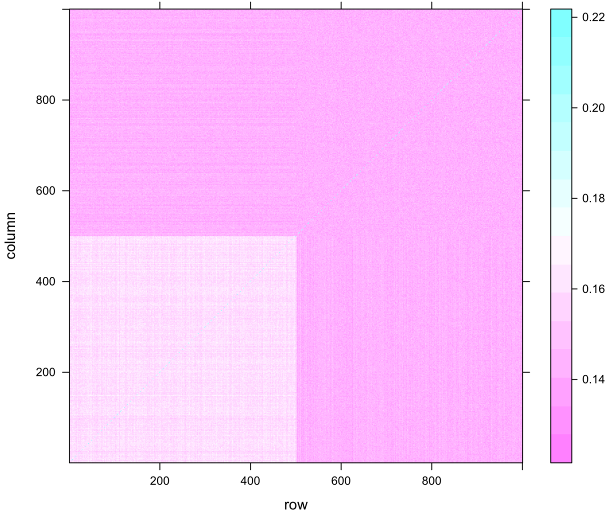
Simulation: n = 1000, p = 50 (Standard Deviation)



Simulation: n = 50, p = 1000 (Bias)



Simulation: $n = 50$, $p = 1000$ (Standard Deviation)



2 Problems

- When $p > n$, the estimates are highly variable.
- Sometimes we care about the independence of the covariates, namely, whether $Cov(\boldsymbol{X}_i, \boldsymbol{X}_j) = 0$.
  - With MLE, $Pr(Cov(\boldsymbol{X}_i, \boldsymbol{X}_j) = 0) = 0$.

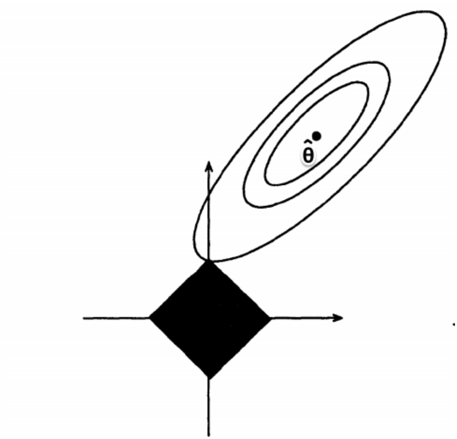
The Lasso (Tibshirani, 1996)

- Suppose $l(\boldsymbol{\theta})$ is the objective function to be maximized.
  - l is the log-likelihood function, negative of the squared error loss, etc.
  - $\boldsymbol{\theta}$ is the parameter of interest, in this case, $\boldsymbol{\Sigma}$.
- Lasso solves the following constrained optimization problem:

$$\hat{\boldsymbol{\theta}} = arg \max_{\boldsymbol{\theta}} l(\boldsymbol{\theta}) \qquad \text{subject to } \|\boldsymbol{\theta}\|_1 = \sum_i |\theta_i| \leq t$$

  for some "well-chosen" t.
- If t is large, then the constraint is loose, $\hat{\boldsymbol{\theta}}$ is close to $\hat{\boldsymbol{\theta}}_{MLE}$.
- If t is small, then the constraint is strict, $\hat{\boldsymbol{\theta}}$ is close to 0.

Lasso Encourages Sparsity



TIBSHIRANI, R. (1996). Regression shrinkage and selection via the lasso. J. R. Statist. Soc. B 58, 267-88.

Sparse Estimation of a Covariance Matrix

- Impose a Lasso constraint on the ML problem.

$$\hat{\mathbf{\Sigma}} = arg \min_{\mathbf{\Sigma}} (\log \det(\mathbf{\Sigma}) + tr(\mathbf{\Sigma}^{-1} \boldsymbol{S}))$$

$$\text{subject to } \|\boldsymbol{P} * \mathbf{\Sigma}\|_1 = \sum_i \sum_j |P_{ij} \Sigma_{ij}| \leq t$$

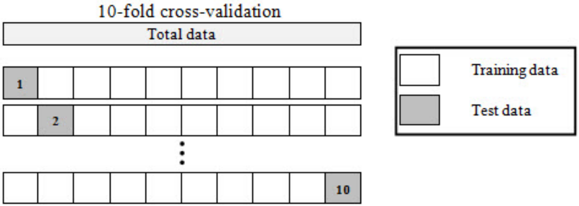
  where "$*$" is the component-wise multiplication.
- We can also rewrite the problem in Lagrangian form

$$\hat{\mathbf{\Sigma}} = arg \min_{\mathbf{\Sigma}} (\log \det(\mathbf{\Sigma}) + tr(\mathbf{\Sigma}^{-1} \boldsymbol{S}) + \lambda \|\boldsymbol{P} * \mathbf{\Sigma}\|_1)$$

  where λ is the "well-chosen" tuning parameter. Larger λ corresponds to smaller t, i.e. stronger penalty.
- This problem is non-convex.

Choose the Tuning Parameter λ

- (10-fold) Cross-Validation:
  1. For each candidate λ, Randomly and evenly separate those n subjects into 10 groups: $A_1, ..., A_5$.
  2. For $i \in \{1, ..., 10\}$:
     1. Calculate $\hat{\boldsymbol{\Sigma}}_i^{\lambda}$ using the data without A_i. Calculate $\boldsymbol{S}_i = \frac{1}{n}\sum_i \boldsymbol{X}_i^T \boldsymbol{X}_i$ using A_i.
     2. Get the estimated likelihood $l_i^{\lambda} = -\log\det(\hat{\boldsymbol{\Sigma}}_i^{\lambda}) - tr((\hat{\boldsymbol{\Sigma}}_i^{\lambda})^{-1}\boldsymbol{S}_i)$
  3. Get the average estimated likelihood $l^{\lambda} = \frac{1}{n}\sum_i l_i^{\lambda}$.
  4. Choose λ that generates the smallest average estimated likelihood l^{λ}.



Earlier Work

- Chaudhuri, Drton and Richardson (2007) consider estimating a covariance matrix given that some pre-specifed entries in $\mathbf{\Sigma}$ are 0.
- Rothman, Levina and Zhu (2009) consider thresholding the sample covariance matrix to get a sparse estimation.

Marginal vs. Conditional

- In multivariate Gaussian, $\boldsymbol{X}_i$ and $\boldsymbol{X}_j$ are **marginally** independent if $\Sigma_{ij} = 0$.
- In multivariate Gaussian, $\boldsymbol{X}_i$ and $\boldsymbol{X}_j$ are **conditionally** independent if $\Sigma_{ij}^{-1} = 0$.
- The proposed method can only be used to infer the marginal independence.
- For conditional associations, we use "graphical Lasso". (Yuan and Lin (2007), Friedman, Hastie, and Tibshirani (2007)).
  - Replace $\|\boldsymbol{P} * \boldsymbol{\Sigma}\|_1$ in the penalty with $\|\boldsymbol{\Sigma}^{-1}\|_1$.
  - Convex problem – easier to solve.

Next Time

- How to solve the optimization problem. (non-convex problem – hard)
- Some simulation results (if I am lucky)
- Some problems (if I am not)