Sparse Estimation of a Covariance Matrix (2011)

Jacob Bien
Robert Tibshirani

June 5, 2014
Presented by Sen Zhao

Outline

1. Motivation
2. The proposed ℓ_1 penalized method
3. How to solve the optimization problem
4. Simulation examples
5. Real data example

Scientific Motivation

- 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 estimate the covariance matrix between those p covariates.
  - To determine the gene / gene or species / species interaction.
  - Specifically, we may want to estimate whether two covariates are marginally independent (i.e. covariance $= 0$).

Statistical Motivation

- 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 $\boldsymbol{S}$ is the sample covariance.
- When p is relatively large compared to n, we want estimates that are:
  - Accurate and precise
  - Sparse (sparsistent?)

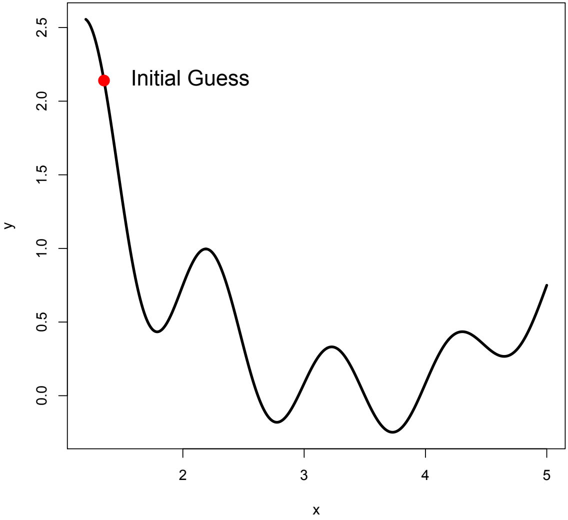
The Method

- Impose an ℓ_1 penalty on the ML problem.

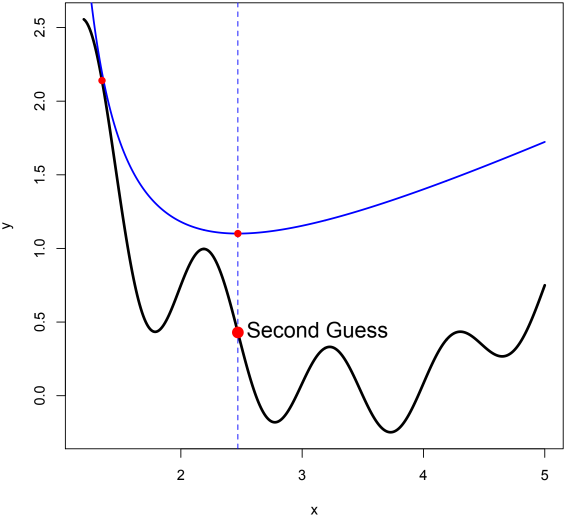
$$\hat{\boldsymbol{\Sigma}} = arg \min_{\boldsymbol{\Sigma} \succ 0} (\log \det(\boldsymbol{\Sigma}) + tr(\boldsymbol{\Sigma}^{-1} \boldsymbol{S}) + \lambda \|\boldsymbol{P} * \boldsymbol{\Sigma}\|_1)$$

  - "$*$" is the component-wise multiplication: $\|\boldsymbol{P} * \boldsymbol{\Sigma}\|_1 = \sum_i \sum_j P_{ij} \Sigma_{ij}$
  - $\boldsymbol{P}$ is the weight matrix of the penalty
  - λ is the "well-chosen" tuning parameter.

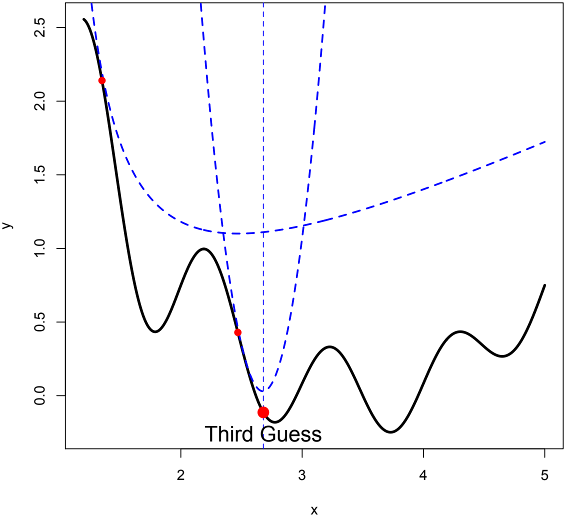
Majorize-Minimization (MM) Algorithm (Lange, 2004, Hunter and Li, 2005)



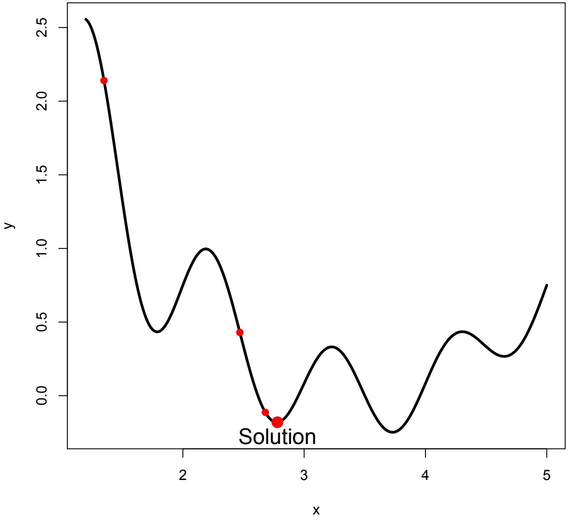
Majorize-Minimization (MM) Algorithm (Lange, 2004, Hunter and Li, 2005)



Majorize-Minimization (MM) Algorithm (Lange, 2004, Hunter and Li, 2005)



Majorize-Minimization (MM) Algorithm (Lange, 2004, Hunter and Li, 2005)



Difference of Convex Functions Programming (An and Tao, 2005)

- Suppose we want to minimize $g(x) = a(x) - b(x)$.
  - $a(x)$ and $b(x)$ are convex functions.
- Suppose $b'_{x_0}(x)$ is the tangent line of $b(x)$ at x_0.
- $f(x) = a(x) - b'_{x_0}(x)$ is the convex surrogate function.
- The convex surrogate function in this case:

$$f(\boldsymbol{\Sigma}) = \log(\det(\boldsymbol{\Sigma}_0)) + tr(\boldsymbol{\Sigma}_0^{-1}\boldsymbol{\Sigma}) - p + tr(\boldsymbol{\Sigma}^{-1}\boldsymbol{S}) + \lambda\|\boldsymbol{P} * \boldsymbol{\Sigma}\|_1$$

What's Wrong with the Newton-Raphson Method?

- The convex surrogate function is not differentiable.
- There is an implicit constraint that $\mathbf{\Sigma}$ is positive semi-definite.

Simulation Setup

- Three methods to consider:
  - Soft-thresholded sample covariance matrix. (Rothman et al., 2009)
    - Off-diagonal entries are shrunken towards 0 by an additive factor c, until they reach 0.
  - Proposed method with $\boldsymbol{P}_{ij} = 1$ for $i \neq j$, $\boldsymbol{P}_{ii} = 0$
    - Equal penalties for all off-diagonal entries.
  - Proposed method with $\boldsymbol{P}_{ij} = \boldsymbol{S}_{ij}^{-1}$ for $i \neq j$, $\boldsymbol{P}_{ii} = 0$
    - Stronger penalties for entries with small sample covariances.
- 2 different structures of $\boldsymbol{\Sigma}$
- $n = 100$, $p = 50$
- 10 repetitions

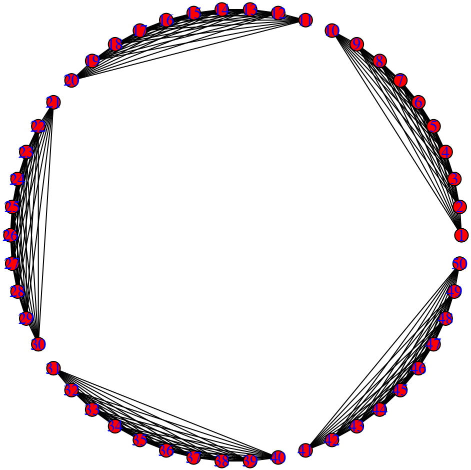
Runtime

- Intel 4^{th} generation Core i7 processor (2013), 2.0GHz.
- 25 candidate shrinkage/tuning parameters.
- Use 5-fold CV to choose the shrinkage/tuning parameters.
- Time for 1 model (125 model fits):

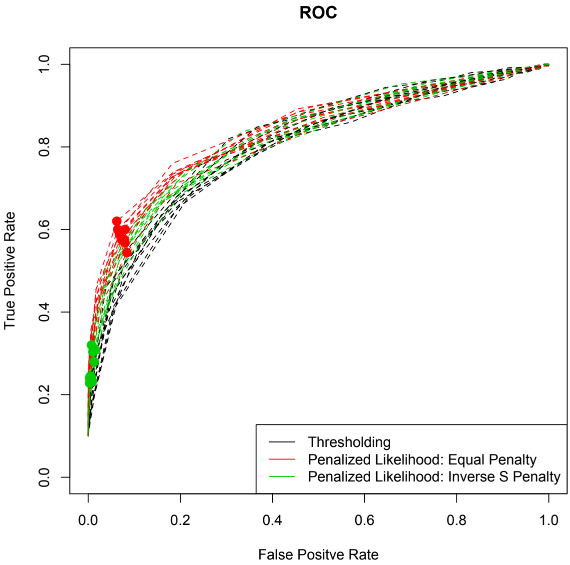
Method	Runtime
Thresholding of Sample Covariance	< 1 sec
Maximum ℓ_1-Penalized Likelihood	10 min

- The runtime is proportional to p^3.

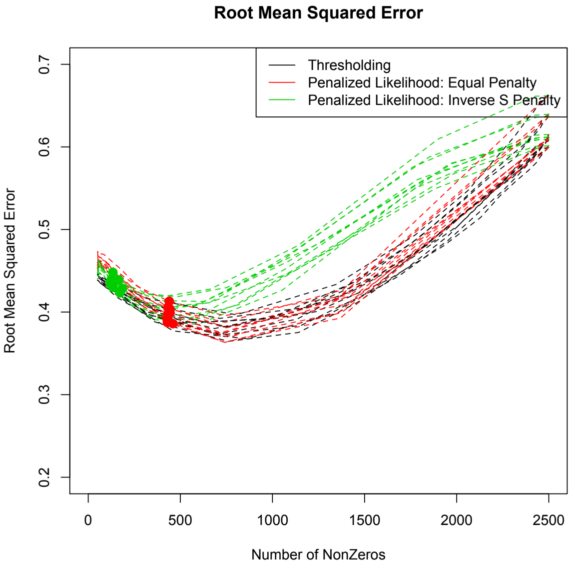
Cliques: Graph Structure



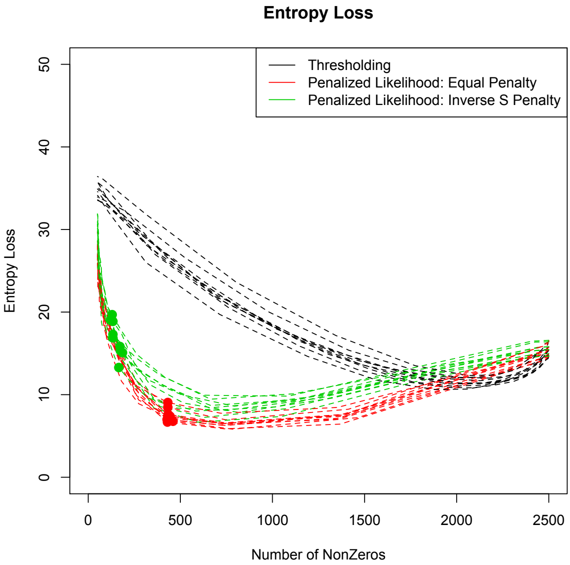
Cliques: ROC


ROC
True Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
False Positve Rate
0.0
0.2
0.4
0.6
0.8
1.0
Thresholding
Penalized Likelihood: Equal Penalty
Penalized Likelihood: Inverse S Penalty

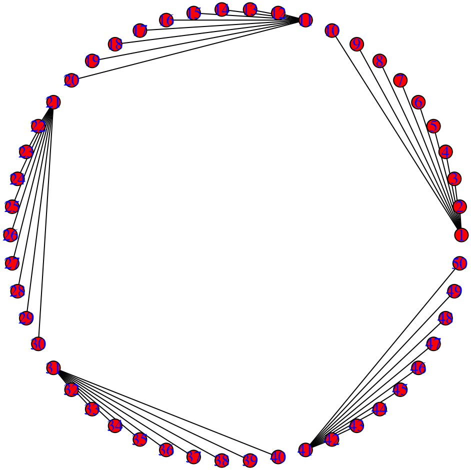
Cliques: RMSE



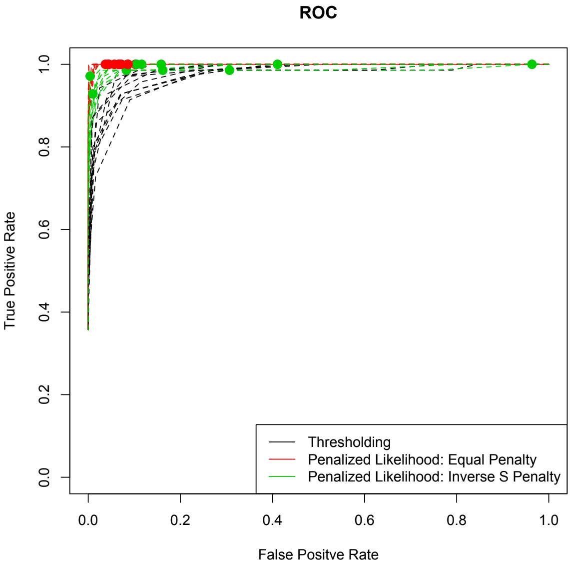
Cliques:Entropy



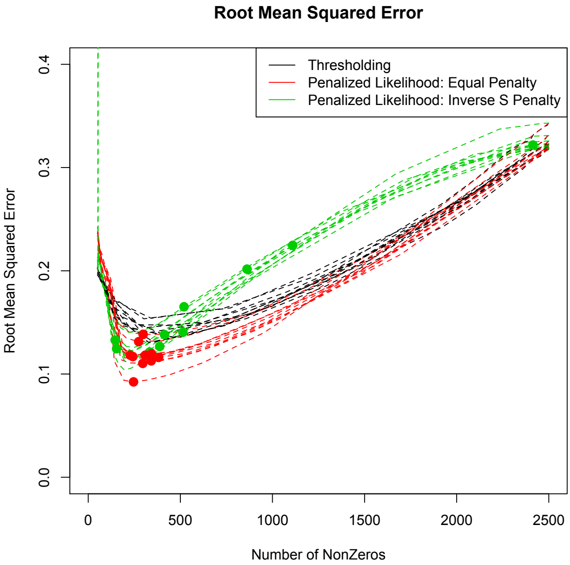
Hubs: Graph Structure



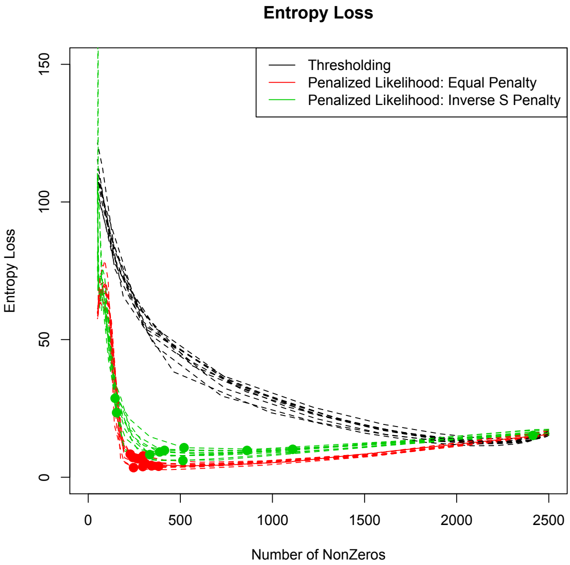
Hubs: ROC


ROC
True Positive Rate
False Positve Rate
0.0
0.2
0.4
0.6
0.8
1.0
Thresholding
Penalized Likelihood: Equal Penalty
Penalized Likelihood: Inverse S Penalty

Hubs: RMSE

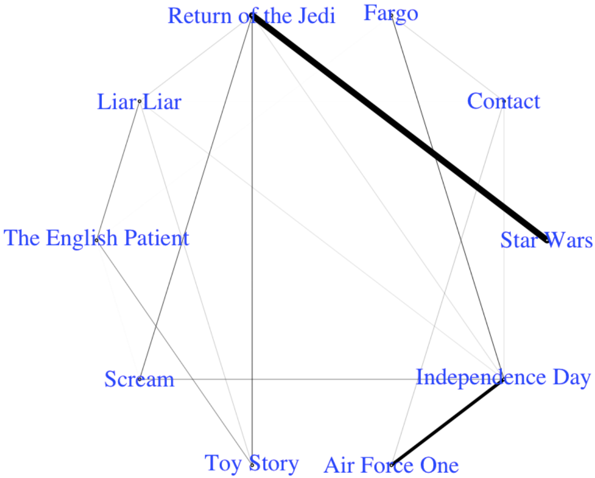

Root Mean Squared Error
Thresholding
Penalized Likelihood: Equal Penalty
Penalized Likelihood: Inverse S Penalty
Root Mean Squared Error
0.0
0.1
0.2
0.3
0.4
0
500
1000
1500
2000
2500
Number of NonZeros

Hubs: Entropy



Application to Movie Ratings

- IMDb rating of 80 top users on 10 movies.
- Proposed method with equal penalty.



Summary

- Bien and Tibshirani (2011) proposed an ℓ_1 penalized maximum likelihood method to find precise and sparse estimates of covariance matrices of normal data.
- They proposed methods to solve the non-convex optimization problem.
- Strengths:
  - Some improvements over an older method through simulations.
  - Estimates are guaranteed to be positive definite.
- Weaknesses:
  - Are those the estimates we want?
  - Do the algorithms solve the optimization problem?
  - The speed of the algorithm is unsatisfactory