Symmetric Constrained Sparse Precision Matrix Estimation Via Lasso Penalized D-trace Loss

Yiyi Yin

School of Statistics
University of Minnesota, Twin Cities

yinxx307@umn.edu

December 7, 2017

Motivation - Gaussian Graphical Model

- Suppose p-dimensional multivariate normal random vector $X = (X_1, X_2, \ldots, X_p)^T \sim N(\mu^*, \Sigma^*)$, $\Theta^* = (\Sigma^*)^{-1}$. When $i \neq j$,

$$\Theta^*_{ij} = 0 \iff X_i \perp\!\!\!\perp X_j | X_{\{1,2,\ldots,p\} \backslash \{i,j\}}$$

- Correspondingly for graph $G = (V, E)$,

$$(i, j) \in E \iff \Theta^*_{ij} \neq 0$$

  which means that the edge (i, j) is in the graph G if and only if $\Theta^*_{ij} = 0$.
- Obviously, the sparse precision matrix estimation method is indispensable when we want to build sparse graph based on the sample covariance matrix.

Sparse Precision Matrix Estimation Application Example

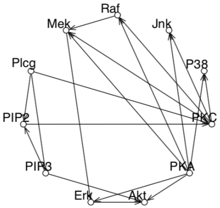
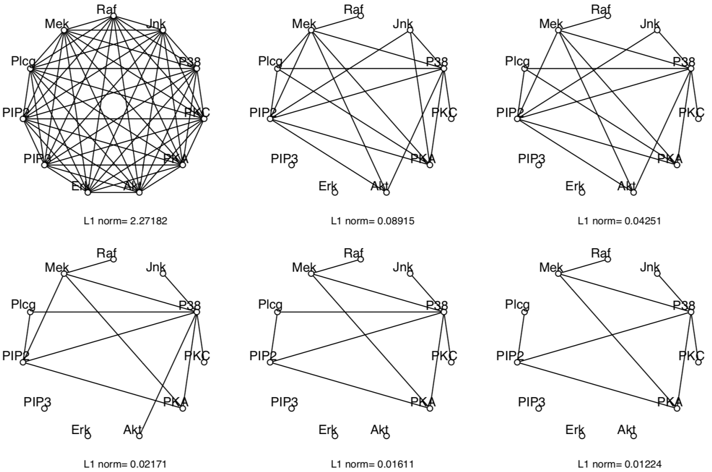


Figure 2: *Directed acylic graph from cell-signaling data, from Sachs et al. (2003).*

- A flow cytometry dataset on $p = 11$ proteins and $n = 7466$ cells, from [Sachs et al.(2005)Sachs, Perez, Pe'er, Lauffenburger, and Nolan].
- They fit a directed acyclic graph (DAG) to the data.

Sparse Precision Matrix Estimation Application Example



Sparse Precision Matrix Estimation Application Example

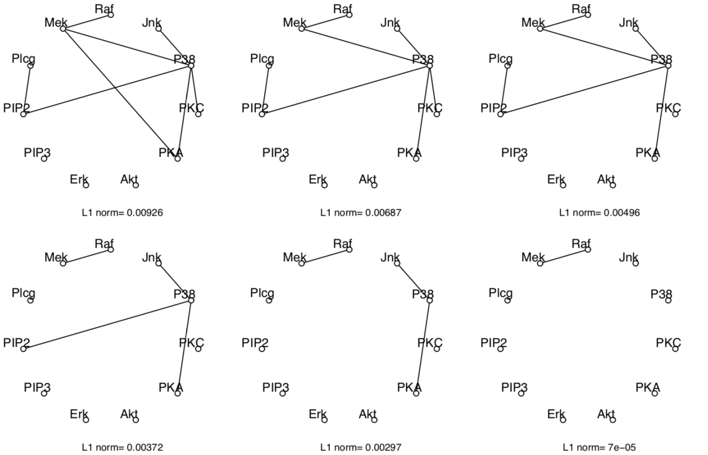


Figure 3: *Cell-signaling data: undirected graphs from graphical lasso with different values of the penalty parameter* ρ.

Graphical Lasso

[Friedman et al.(2008)Friedman, Hastie, and Tibshirani] proppposed the graphical lasso as the precision matrix estimator.

$$\hat{\Theta} = \underset{\Theta \succ 0}{\operatorname{argmin}} \langle \Theta, \hat{\Sigma} \rangle - \operatorname{logdet}(\Theta) + \lambda ||\Theta||_{1,\mathrm{off}} \tag{1}$$

- Let Σ^* and $\hat{\Sigma}$ denote the population and sample covariance matrix. Let $\Theta^* = (\Sigma^*)^{-1}$ denote the population precision matrix.
- Graphical lasso is also the maximum likelihood estimator.

More General Framework

In fact, the graphical lasso can be viewed as a special case of a more general definition of precision matrix estimator.

$$\hat{\Theta} = \underset{\Theta \succ 0}{\operatorname{argmin}} \, L(\Theta, \hat{\Sigma}) + \lambda ||\Theta||_{1,\text{off}} \tag{2}$$

where the loss function $L(\Theta, \hat{\Sigma})$ must satisfy two conditions:
(i) The loss function $L(\Theta, \hat{\Sigma})$ is a smooth convex function of Θ.
(ii) The unique minimizer of $L(\Theta, \hat{\Sigma})$ is $(\hat{\Sigma})^{-1}$.

- For graphical lasso, the loss function $L_{\text{glasso}}(\Theta, \hat{\Sigma}) = \langle \Theta, \hat{\Sigma} \rangle - \text{logdet}(\Theta)$. It can be proved that L_{glasso} satisfies condition (i) and (ii).

Sparse precision matrix estimation via lasso penalized D-trace loss

[Zhang and Zou(2014)] suggested using $L_D(\Theta, \hat{\Sigma}) = \frac{1}{2}\langle \Theta^2, \hat{\Sigma} \rangle - \text{tr}(\Theta)$, where the the precision matrix estimator is the minimizer of lasso penalized D-trace loss.

$$\hat{\Theta} = \underset{\Theta \succ 0}{\text{argmin}} \, \frac{1}{2}\langle \Theta^2, \hat{\Sigma} \rangle - \text{tr}(\Theta) + \lambda ||\Theta||_{1,\text{off}} \tag{3}$$

- It can be proved that L_D also satisfies condition (i) and (ii).
- They solved the problem by alternating direction method of multipliers (ADMM).
- The algorithm is relative slow in terms of computation time.

Our suggested method

We relax the positive definiteness constraint of the precision matrix but only keep the symmetric constraint.

$$\hat{\Theta} = \underset{\Theta^T = \Theta}{\operatorname{argmin}} \frac{1}{2}\langle \Theta^2, \hat{\Sigma} \rangle - \operatorname{tr}(\Theta) + \lambda ||\Theta||_{1,\mathrm{off}} \tag{4}$$

- People don't care that much about the positive definiteness in real world applications.
- We don't have to solve it by ADMM.

We rewrite the lasso penalized D-trace loss with symmetric constrain minimization problem in the following way.

$$\begin{aligned}\hat{\Theta} &= \underset{\Theta^T=\Theta}{\operatorname{argmin}} \frac{1}{2}\langle \Theta^2, \hat{\Sigma}\rangle - \operatorname{tr}(\Theta) + \lambda ||\Theta||_{1,\mathrm{off}} \\ &= \underset{\Theta^T=\Theta}{\operatorname{argmin}} \sum_{j=1}^{p} \frac{1}{2}\theta_j^T \hat{\Sigma}\theta_j - \sum_{i=1}^{p}\theta_{ii} + 2\lambda \sum_{1\le i<j\le p}\theta_{ij}\end{aligned} \tag{5}$$

The univariate optimization problem of θ_{ij} when $i = j$, i.e. θ_{ii} is

$$\begin{aligned}\hat{\theta}_{ii} &= \underset{\theta_{ii}}{\operatorname{argmin}} \frac{1}{2}\hat{\sigma}_{ii}\theta_{ii}^2 + (\sum_{k\neq i}\hat{\sigma}_{ik}\theta_{ik})\theta_{ii} - \theta_{ii} \\ &= \frac{1 - \sum_{k\neq i}\hat{\sigma}_{ik}\theta_{ik}}{\hat{\sigma}_{ii}}\end{aligned} \tag{6}$$

where $(\hat{\Sigma})_{ij} = \hat{\sigma}_{ij}$.

Coordinate Descent

The univariate optimization problem of θ_{ij} when $i \neq j$ is

$$\begin{aligned}\hat{\theta}_{ij} &= \underset{\theta_{ij}}{\operatorname{argmin}} \frac{1}{2}(\hat{\sigma}_{ii} + \hat{\sigma}_{jj})\theta_{ij}^2 + (\sum_{k \neq i} \hat{\sigma}_{ik}\theta_{jk} + \sum_{k \neq j} \hat{\sigma}_{jk}\theta_{ik})\theta_{ij} + 2\lambda|\theta_{ij}| \\ &= s(-\frac{\sum_{k \neq i} \hat{\sigma}_{ik}\theta_{jk} + \sum_{k \neq j} \hat{\sigma}_{jk}\theta_{ik}}{\hat{\sigma}_{ii} + \hat{\sigma}_{jj}}, \frac{2\lambda}{\hat{\sigma}_{ii} + \hat{\sigma}_{jj}}) \\ &= s(-\sum_{k \neq i} \hat{\sigma}_{ik}\theta_{jk} - \sum_{k \neq j} \hat{\sigma}_{jk}\theta_{ik}, 2\lambda)/(\hat{\sigma}_{ii} + \hat{\sigma}_{jj})\end{aligned} \tag{7}$$

where $s(z, \lambda) = sign(z)(|z| - \lambda)_+$ represent the soft thresholding function. For each λ, we cyclically update one parameter at a time until convergence.

Solution Path

We consider providing a solution path which includes a list of N (say $N = 100$) estimated precision matrices corresponding to a list of N different λ values taking $\lambda_1 = \lambda_{\max}, \lambda_2 = \frac{N-1}{N}\lambda_{\max}, \ldots, \lambda_N = \frac{1}{N}\lambda_{\max}$.

- $\lambda_{\max}$: the smallest value such that $\hat{\Theta}$ have as many zeros as possible
- Warm Start: When computing $\hat{\Theta}(\hat{\Sigma}, \lambda_i)$, we firstly initialize it as $\hat{\Theta}(\hat{\Sigma}, \lambda_{i-1})$
- Active Set: Iteratively check the KKT condition, add elements that violate KKT condition into active set and update until the KKT condition is satisfied.

Numerical Results

- $N(0, \Sigma^*)$, where $\Theta^* = (\Sigma^*)^{-1}$. $\Theta^*_{ii} = 1, \Theta^*_{ij} = 0.2$ for $1 \leq |i - j| \leq 2$ and $\Theta^*_{ij} = 0$ otherwise.
- We generated data iid from $N(0, \Sigma^*)$ taking sample size $n = 1000$ and different p values.
- The algorithm was coded in C and called in R.

Computation Time for Solution Path

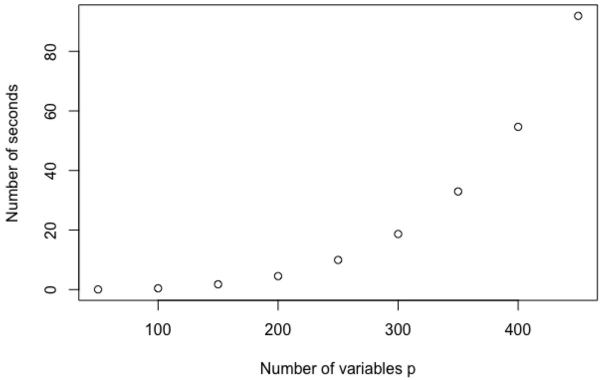


Figure: Number of seconds needed for computing the estimated precision matrices solution path ($N = 100$) when p take $50, 100, \dots, 450$

Computation Time for Single Precision Matrix Estimation

λ	Non-zero Fraction	Time (seconds)
0.001	0.9851	8.068
0.017	0.6669	4.079
0.033	0.3370	1.509
0.274	0.0025	0.070

Table: When $p = 400$, computing time for different penalization parameter λ values, fraction of non-zeros elements in estimated precision matrices are about 1, 2/3, 1/3 and 0

Estimation Accuracy Comparison

We still consider the same setting that generating data i.i.d. from $N(0, \Sigma^*)$, where $\Theta^*_{ii} = 1, \Theta^*_{ij} = 0.2$ for $1 \leq |i-j| \leq 2$ and $\Theta^*_{ij} = 0$ otherwise. We take sample size $n = 1000$ and $p = 100$. We compared graphical lasso with our method in five quantities:

- Frobenius norm $E||\hat{\Theta} - \Theta^*||_F$ where $||X||_F = \sqrt{\sum_{i,j} X_{ij}^2}$
- Spectral norm $E||\hat{\Theta} - \Theta^*||_2$, where $||X||_2 = \sqrt{\lambda_{max}(X^T X)} = \sigma_{max}(X)$, which is the largest singular value
- Infinity norm $E||\hat{\Theta} - \Theta^*||_\infty$, where $||X||_\infty = max_i \sum_j |X_{ij}|$, which is the maximum absolute row sum
- True positive rate (TPR) of found non-zero elements, where TPR $=$ TP/(TP+FN)
- True negative rate (TNR) of found zero elements, where TNR $=$ TN/(TN + FP).

Estimation Accuracy Comparison

	Frobenius	Spectral	Infinity	TPR	TNR
Graphical lasso	1.929	0.470	0.720	1	0.902
	(0.043)	(0.016)	(0.041)	(0)	(0.005)
Our method	1.679	0.416	0.613	1	0.937
	(0.050)	(0.022)	(0.035)	(0)	(0.003)

Table: Comparison between graphical lasso and our method using three norms and two correctness ratios. The smaller the norms the better, the larger the correctness ratios the better. Values are means and values in the parenthesis are standard deviations of 100 independent runs.

Future Work

- Instead of using the same penalization parameter λ on all the off-diagonal elements, we are interested in using different penalization parameter on different off-diagonal elements.
- Everything would be very similar as in the previous case.

$$\hat{\Theta} = \underset{\Theta^T = \Theta}{\operatorname{argmin}} \frac{1}{2}\langle \Theta^2, \hat{\Sigma} \rangle - \operatorname{tr}(\Theta) + \sum_{1 \le i < j \le p} w_{ij}\theta_{ij} \tag{8}$$

References

Jerome Friedman, Trevor Hastie, and Robert Tibshirani.
Sparse inverse covariance estimation with the graphical lasso.
Biostatistics, 9(3):432–441, 2008.

Karen Sachs, Omar Perez, Dana Pe'er, Douglas A Lauffenburger, and Garry P Nolan.
Causal protein-signaling networks derived from multiparameter single-cell data.
Science, 308(5721):523–529, 2005.

Teng Zhang and Hui Zou.
Sparse precision matrix estimation via lasso penalized d-trace loss.
Biometrika, 101(1):103–120, 2014.

Thanks!