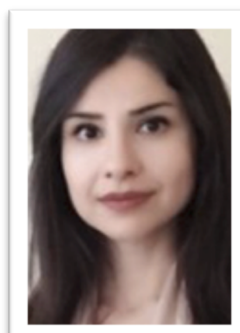
Large-scale Sparse Inverse Covariance Estimation via Thresholding and Max-Det Matrix Completion

Richard Y. Zhang[1], Salar Fattahi[1], Somayeh Sojoudi[2]
[1]Industrial Engineering and Operations Research
[2]Electrical Engineering & Computer Sciences
University of California, Berkeley
{ryz, fattahi, sojoudi}@berkeley.edu





Given n-dimensional random variable

$$X = [X_1, X_2, \ldots, X_n]^T \sim \text{Distribution}$$

Consider estimating the covariance matrix

$$\mu_i = \mathbb{E}[X_i], \qquad \Sigma_{i,j} = \mathbb{E}\left[(X_i - \mu_i)(X_j - \mu_j)\right]$$

from N samples (or realizations)

$$x^{(1)} = [x_1^{(1)}, x_2^{(1)}, \ldots, x_n^{(1)}]^T,$$
$$x^{(2)} = [x_1^{(2)}, x_2^{(2)}, \ldots, x_n^{(2)}]^T,$$
$$x^{(3)} = [x_1^{(3)}, x_2^{(3)}, \ldots, x_n^{(3)}]^T,$$
$$\vdots$$
$$x^{(N)} = [x_1^{(N)}, x_2^{(N)}, \ldots, x_n^{(N)}]^T.$$

Zhang, Fattahi, Sojoudi. Large-scale Sparse Inverse Covariance Estimation via Thresholding and Max-Det Matrix Completion. ICML 2018.

Given n-dimensional random variable

$$X = [X_1, X_2, \ldots, X_n]^T \sim \text{Distribution}$$

Consider estimating the covariance matrix

$$\mu_i = \mathbb{E}[X_i], \qquad \Sigma_{i,j} = \mathbb{E}\left[(X_i - \mu_i)(X_j - \mu_j)\right]$$

from N samples (or realizations).

Maximium likelihood estimator. N = O(n) samples.

Graphical lasso estimator. N = O(log(n)) samples, assuming *sparse inverse covariance* matrix

$$\Theta = \Sigma^{-1} \text{ exists and contains } O(1) \text{ nonzeros per column}$$

Assumption frequently valid in real-life applications.

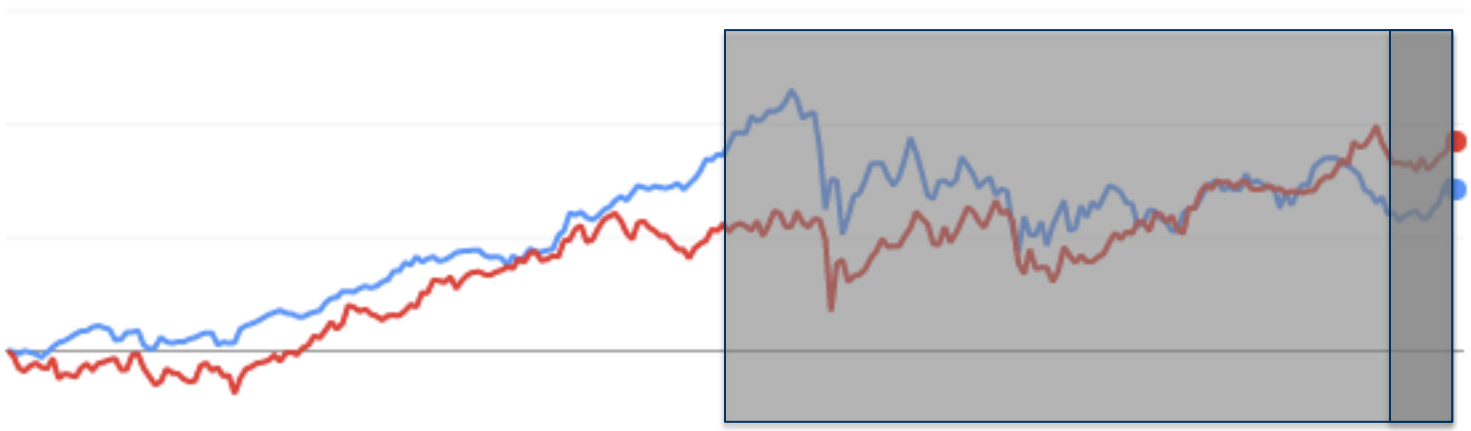
log(n) factor *optimal* due to coupon collector effect.

Graphical lasso most useful in high-dimensional settings

$$\text{dimension } n \quad \gg \quad \text{num. samples } N.$$

- Shrinkage estimator, e.g. Markowitz portfolio

  Goal: minimize number of samples.



- Markov graphical models, e.g. in neuroscience

  Goal: impose sparsity on inverse covariance matrix.

$$X = [X_1, X_2, \ldots, X_n]^T \sim N(\mu, \Theta^{-1}).$$

$$\Theta_{i,j} = 0 \quad \Longleftrightarrow \quad X_i \perp X_j \mid \text{rest}$$

Graphical lasso most useful in high-dimensional settings

$$\text{dimension } n \quad \gg \quad \text{num. samples } N.$$

State-of-the-art solvers usually $O(n^3)$ time and $O(n^2)$ space

- GLASSO (Friedman et al. 2008)
- CVXOPT (Dahl et al. 2008)
- (BIG)-QUIC (Hsieh et al. 2013)

BIG-QUIC solved n = 200k in 5 hours on 4 x 8-core CPUs

Complexity motivates other estimators, e.g. EEGM (Yang et al. 2014).

This work. Solve graphical lasso in

$$O(n + n^2/p) \text{ time and } O(n) \text{ memory}$$

on p parallel processors, assuming modestly large λ and bounded degree chordal embedding

We solved n = 200k in <70 minutes on a Macbook Air.

Review. Graphical lasso

Estimate the n x n covariance matrix

$$\mu_i = \mathbb{E}[X_i], \qquad \Sigma_{i,j} = \mathbb{E}\left[(X_i - \mu_i)(X_j - \mu_j)\right]$$

from N samples.

Approximate expection with average, obtain MLE

$$\overline{x}_i = \frac{1}{N}\sum_{k=1}^{N} x_i^{(k)}, \qquad S = \frac{1}{N}\sum_{k=1}^{N}(x_i^{(k)} - \overline{x}_i)(x_j^{(k)} - \overline{x}_j),$$

Solve graphical lasso optimization problem

$$\hat{\Theta} = \underset{\Theta \succ 0}{\text{minimize}} \ \text{trace}(S\Theta) - \log\det\Theta + \lambda \sum_{i \neq j} |\Theta_{i,j}|.$$

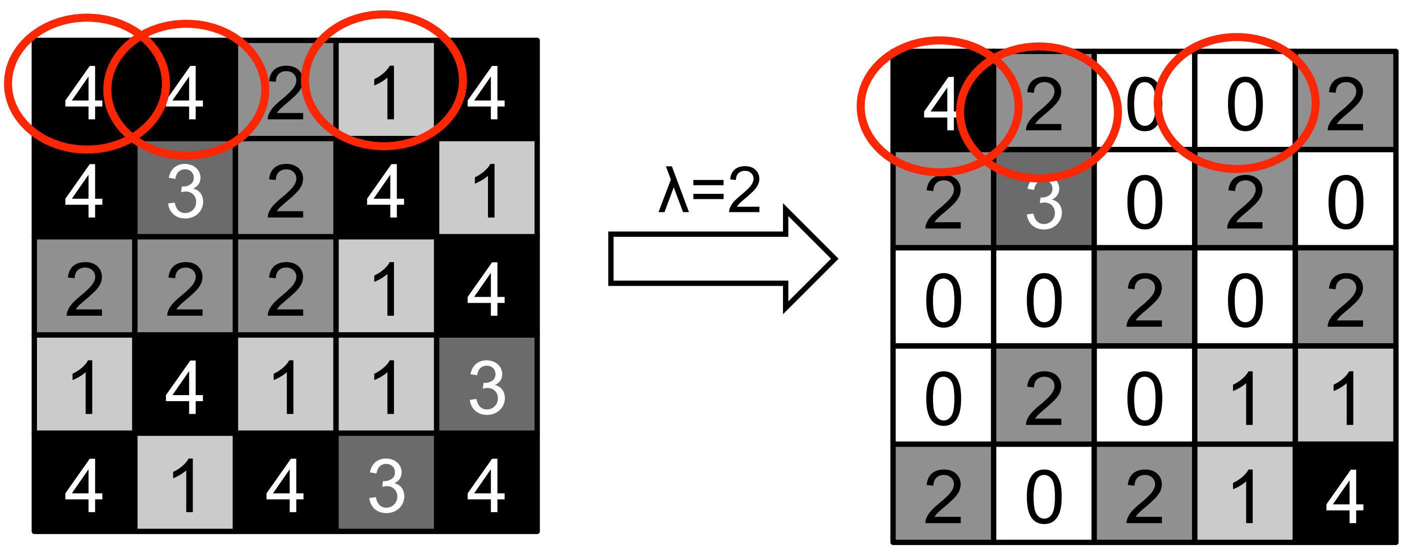
Bottleneck is the solution of this problem.

Review. Threshold and MDMC (Fattahi & Sojoudi 2017)

$$\bar{x}_i = \frac{1}{N}\sum_{k=1}^{N} x_i^{(k)}, \qquad S = \frac{1}{N}\sum_{k=1}^{N}(x_i^{(k)} - \bar{x}_i)(x_j^{(k)} - \bar{x}_j),$$

1. Estimate sparsity pattern. Soft-threshold in **O(n²/p) time**

$$(S_\lambda)_{i,j} = \begin{cases} S_{i,j} & i = j \\ S_{i,j} - \operatorname{sign}(S_{i,j}) \cdot \lambda_{i,j} & |S_{i,j}| > \lambda_{i,j}, \quad i \neq j \\ 0 & |S_{i,j}| \leq \lambda_{i,j}, \quad i \neq j \end{cases}$$



Review. Threshold and MDMC (Fattahi & Sojoudi 2017)

2. Estimate parameters. Solve max-det matrix completion

Soft-thresholded MLE

$$\begin{aligned}&\underset{\Theta\succ 0}{\text{minimize}}\ \operatorname{trace}(\underline{S_\lambda}\Theta) - \log\det\Theta \\ &\text{subject to } \Theta_{i,j} = 0 \quad \text{wherever } (S_\lambda)_{i,j} = 0\end{aligned}$$

Compare with the original graphical lasso problem:

Original MLE

Nonsmooth term

$$\hat{\Theta} = \underset{\Theta\succ 0}{\text{minimize}}\ \operatorname{trace}(\underline{S}\Theta) - \log\det\Theta + \lambda\sum_{i\neq j}|\Theta_{i,j}|.$$

Our new bottleneck:

$$\underset{\Theta \succ 0}{\text{minimize}} \; \text{trace}(S_\lambda \Theta) - \log\det\Theta$$

$$\text{subject to } \Theta_{i,j} = 0 \quad \text{wherever } (S_\lambda)_{i,j} = 0$$

State-of-the-art solvers usually $O(n^3)$ time and $O(n^2)$ space

If sparsity graph of S_λ is bounded degree chordal, then

O(n) time and O(n) space

via recursive closed-form solution (Dahl et al. 2008)

$$f_*(C) = \min_{\Theta \succ 0}\{\text{trace}(C\Theta) - \log\det\Theta : \Theta_{i,j} = 0 \quad \forall (i,j) \notin G\}$$

This is a self-concordant barrier function on the space of sparse matrices (Andersen et al. 2010)

$$\mathbb{S}^n_G = \{\Theta \in \mathbb{S}^n : \Theta_{i,j} = 0 \quad \forall (i,j) \notin G\}.$$

Use insights to solve MDMC in O(n) time and space.

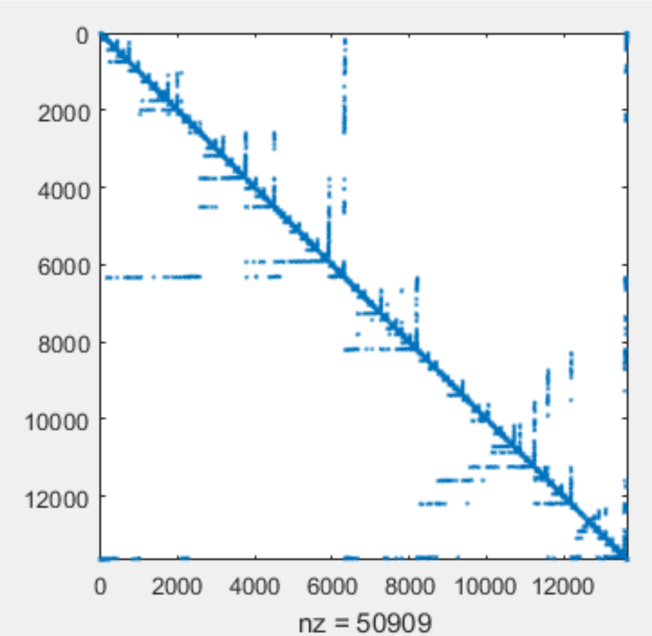
Main contribution. Newton-CG for MDMC

$$\underset{\Theta \succ 0}{\text{minimize}} \ \text{trace}(S_\lambda \Theta) - \log\det\Theta$$

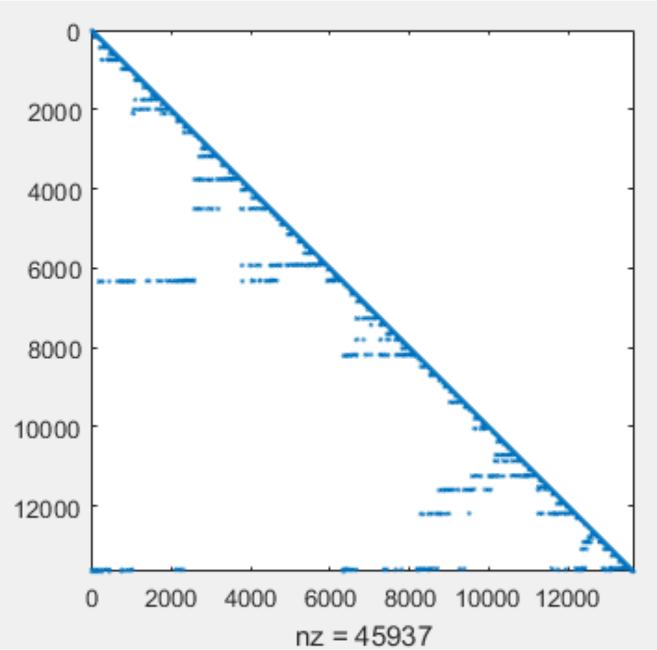
$$\text{subject to } \Theta_{i,j} = 0 \quad \text{wherever } (S_\lambda)_{i,j} = 0$$

1. **Embed** nonchordal sparsity graph G of S_λ within a chordal graph G-tilde.

$$S = LL^T$$



$$L$$



Main contribution. Newton-CG for MDMC

$$\underset{\Theta \succ 0}{\text{minimize}} \ \text{trace}(S_\lambda \Theta) - \log\det\Theta$$

$$\text{subject to } \Theta_{i,j} = 0 \quad \text{wherever } (S_\lambda)_{i,j} = 0$$

2. Pose as optimization problem over the **fill-in**

$$\begin{aligned} \text{minimize } & \text{tr}(S_\lambda \Theta) - \log\det\Theta \\ \text{subject to } & \Theta_{i,j} = 0 \quad \forall (i,j) \in \tilde{G} \backslash G \\ & \Theta \in \mathbb{S}^n_{\tilde{G}}, \quad \Theta \succ 0 \end{aligned}$$

Most sparsity constraints show up here

Extra edges added to to make graph chordal

Optimization problem over the **cone of sparse semidefinite matrices**.

Main contribution. Newton-CG for MDMC

$$\underset{\Theta \succ 0}{\text{minimize}} \; \text{trace}(S_\lambda \Theta) - \log\det\Theta$$
$$\text{subject to } \Theta_{i,j} = 0 \quad \text{wherever } (S_\lambda)_{i,j} = 0$$

3. Solve the **dual** problem

Self-concordant barrier on the cone of sparse matrices

$$\text{maximize} \; -f_*(S_\lambda + Y)$$
$$\text{subject to } Y \in \mathbb{S}^n_{\tilde{G}\setminus G}$$

Edges added to to make graph chordal

Self-concordance guarantees ε-accuracy in O(log log (1/ε)) Newton iterations.

Main contribution. Newton-CG for MDMC

$$\begin{aligned} &\underset{\Theta \succ 0}{\text{minimize}} \ \text{trace}(S_\lambda \Theta) - \log\det\Theta \\ &\text{subject to } \Theta_{i,j} = 0 \quad \text{wherever } (S_\lambda)_{i,j} = 0 \end{aligned}$$

4. Solve Newton direction using **conjugate gradients**

$$\begin{aligned} &\text{maximize } \ -f_*(S+Y) \\ &\text{subject to } Y \in \mathbb{S}^n_{\tilde{G}\backslash G} \end{aligned}$$

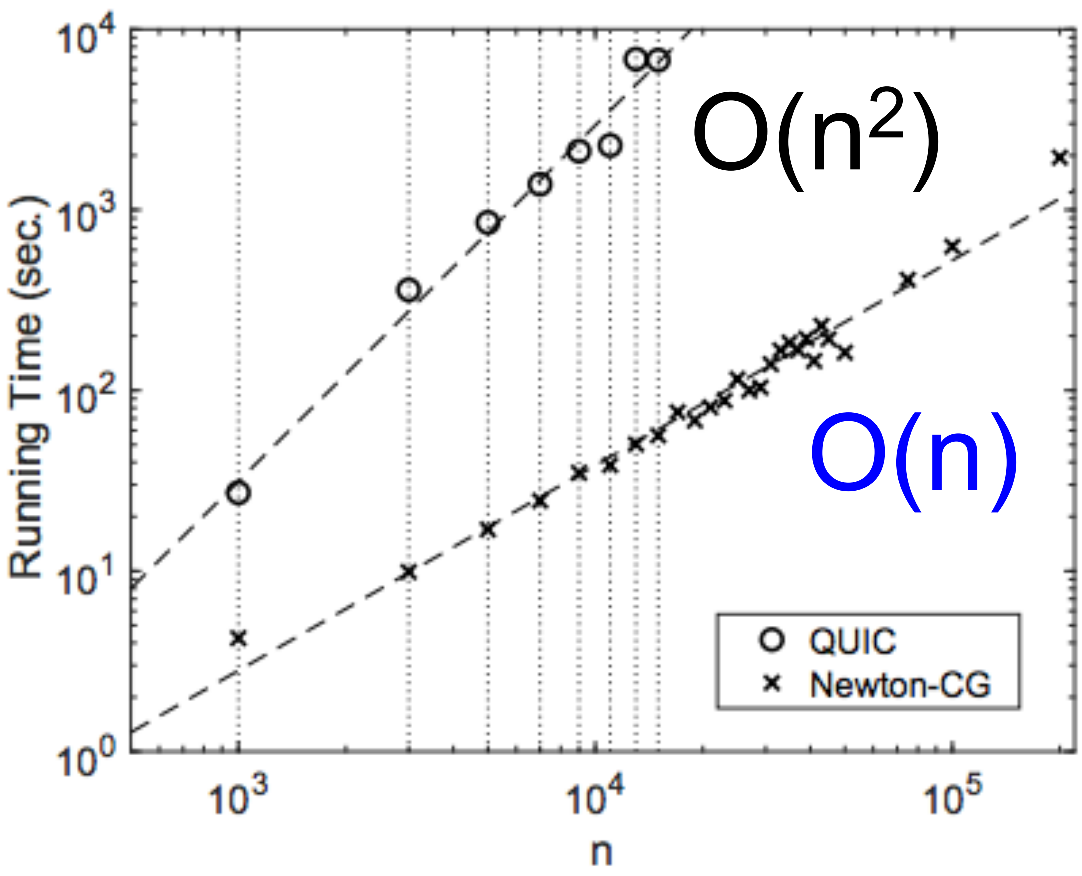
Main Theorem (Informal). CG converges to ε-accuracy in O(log(1/ε)) iterations

Each CG iteration costs O(n) time and O(n) memory.

soft-O(1) CG iters. over soft-O(1) Newton iters. QED.

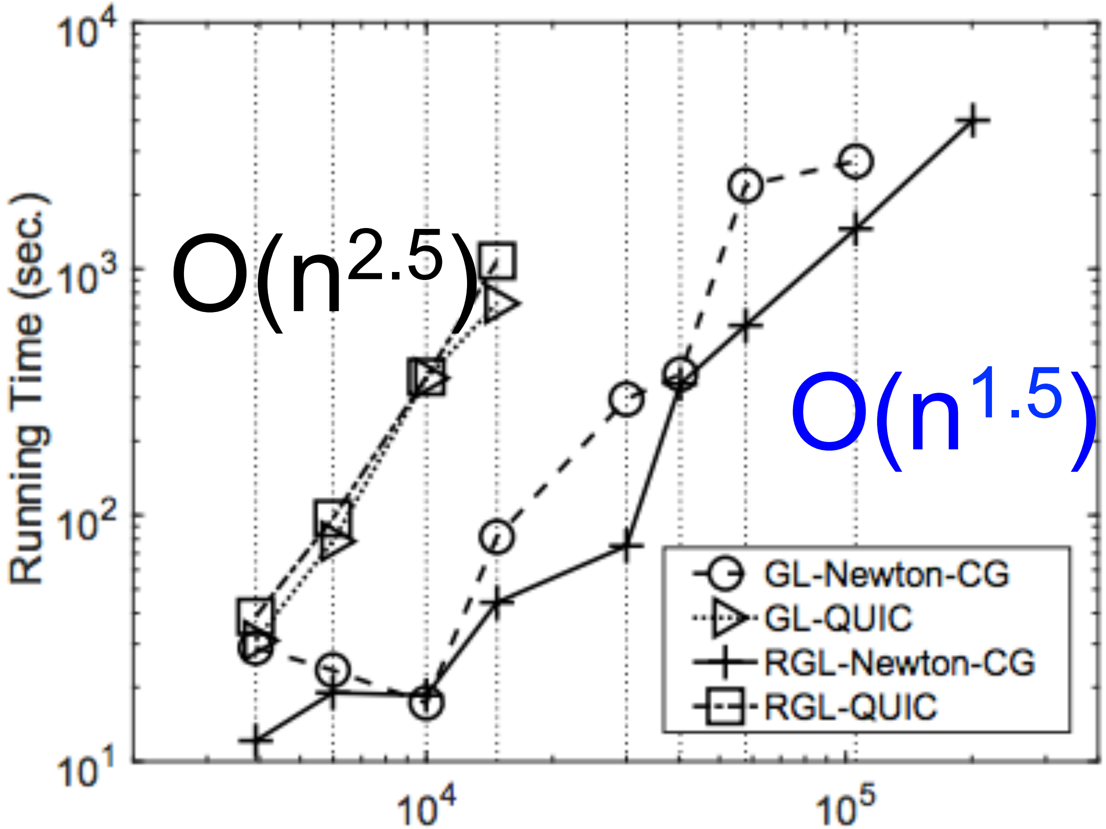
Numerical results on banded graphs

- Synthetic $\Theta=\Sigma^{-1}$ with banded sparsity pattern
- Off-diagonals [-1,+1], corrupted to zero with p=0.3
- Diagonals set to sum of off-diagonals plus one
- **Solve MDMC on this sparsity pattern**



Numerical results on real-life graphs

- Synthetic $\Theta=\Sigma^{-1}$ from real-life graphs.
- Off-diagonals [-1,+1], corrupted to zero with p=0.3
- Diagonals set to sum of off-diagonals plus one
- **Estimate Σ from 5000 i.i.d. samples from N(0, Σ)**



Conclusions

- Graphical lasso estimates covariance matrix assuming that its inverse is sparse. Applications in finance and neuroscience.
- Nice theory, most useful in high-dimensional setting.
- **This paper.** Fast algorithm for graphical lasso
  - O(n) time and space.
- **Numerical results.** Solve n = 200k problem in 70 minutes on a laptop.
- **Next steps.** Benchmark statistical performance for recovering ground-truth.

$$\hat{\Theta} = \underset{\Theta \succ 0}{\text{minimize}} \ \text{trace}(S\Theta) - \log\det\Theta + \lambda \sum_{i \neq j} |\Theta_{i,j}|.$$

Thank you! – Poster #1

Richard Y. Zhang[1], Salar Fattahi[1], Somayeh Sojoudi[2]

[1]Industrial Engineering and Operations Research

[2]Electrical Engineering & Computer Sciences
University of California, Berkeley

{ryz, fattahi, sojoudi}@berkeley.edu