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

Richard Y. Zhang\textsuperscript{1}, Salar Fattahi\textsuperscript{1}, Somayeh Sojoudi\textsuperscript{2}
\textsuperscript{1}Industrial Engineering and Operations Research
\textsuperscript{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

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

from $N$ samples (or realizations)

$$x^{(1)} = [x^{(1)}_1, x^{(1)}_2, \ldots, x^{(1)}_n]^T,$$
$$x^{(2)} = [x^{(2)}_1, x^{(2)}_2, \ldots, x^{(2)}_n]^T,$$
$$x^{(3)} = [x^{(3)}_1, x^{(3)}_2, \ldots, x^{(3)}_n]^T,$$
$$\vdots$$
$$x^{(N)} = [x^{(N)}_1, x^{(N)}_2, \ldots, x^{(N)}_n]^T.$$
Given $n$-dimensional random variable

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

Consider estimating the covariance matrix

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

from $N$ samples (or realizations).

**Maximum 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

\[ \text{dimension } n \gg \text{num. samples } N. \]

- 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.

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

\[
\Theta_{i,j} = 0 \iff X_i \perp X_j \mid \text{rest}
\]
Graphical lasso most useful in high-dimensional settings

dimension $n \gg$ 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

$O(n + n^2/p)$ time and $O(n)$ memory

on $p$ parallel processors, assuming modestly large $\lambda$ and bounded degree chordal embedding

We solved $n = 200k$ in $<70$ minutes on a Macbook Air.
Estimate the $n \times n$ covariance matrix

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

from $N$ samples.

Approximate expectation with average, obtain MLE

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

Solve graphical lasso optimization problem

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

Bottleneck is the solution of this problem.
Review. Threshold and MDMC (Fattahi & Sojoudi 2017)

$$\overline{x}_i = \frac{1}{N} \sum_{k=1}^{N} x^{(k)}_i,$$

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

1. Estimate sparsity pattern. Soft-threshold in $O(n^2/p)$ time

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

$$\lambda = 2$$

\[\begin{array}{cccc}
4 & 4 & 2 & 1 \\
4 & 3 & 2 & 4 \\
2 & 2 & 2 & 1 \\
1 & 4 & 1 & 1 \\
4 & 1 & 4 & 3 \\
\end{array}\]

\[\begin{array}{cccc}
4 & 2 & 0 & 0 \\
2 & 3 & 0 & 2 \\
0 & 0 & 2 & 0 \\
0 & 2 & 0 & 1 \\
2 & 0 & 2 & 1 \\
\end{array}\]
2. **Estimate parameters.** Solve max-det matrix completion

\[
\begin{align*}
\text{minimize } & \quad \text{trace}(S_\lambda \Theta) - \log \det \Theta \\
\text{subject to } & \quad \Theta_{i,j} = 0 \quad \text{wherever } (S_\lambda)_{i,j} = 0
\end{align*}
\]

Compare with the original graphical lasso problem:

\[
\hat{\Theta} = \underset{\Theta > 0}{\text{minimize}} \quad \text{trace}(S\Theta) - \log \det \Theta + \lambda \sum_{i \neq j} |\Theta_{i,j}|
\]
Our new bottleneck:

\[
\begin{align*}
\text{minimize } & \text{trace}(S_\lambda \Theta) - \log \det \Theta \\
\text{subject to } & \Theta_{i,j} = 0 \text{ wherever } (S_\lambda)_{i,j} = 0
\end{align*}
\]

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

If sparsity graph of $S_\lambda$ is bounded degree chordal, then $O(n)$ time and $O(n)$ space via recursive closed-form solution (Dahl et al. 2008)

\[
f^*_*(C) = \min_{\Theta > 0} \{ \text{trace}(C \Theta) - \log \det \Theta : \Theta_{i,j} = 0 \quad \forall (i, j) \notin G \}
\]

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

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

Use insights to solve MDMC in $O(n)$ time and space.
Main contribution. Newton-CG for MDMC

\[
\text{minimize } \text{trace}(S_\lambda \Theta) - \log \det \Theta \\
\Theta > 0 \\
\text{subject to } \Theta_{i,j} = 0 \quad \text{wherever } (S_\lambda)_{i,j} = 0
\]

1. **Embed** nonchordal sparsity graph \( G \) of \( S_\lambda \) within a chordal graph \( G\text{-tilde} \).

\[
S = LL^T
\]
Main contribution. Newton-CG for MDMC

\[
\begin{align*}
\text{minimize } & \quad \text{trace}(S_\lambda \Theta) - \log \det \Theta \\
\text{subject to } & \quad \Theta_{i,j} = 0 \quad \text{wherever } (S_\lambda)_{i,j} = 0
\end{align*}
\]

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

\[
\begin{align*}
\text{minimize } & \quad \text{tr}(S_\lambda \Theta) - \log \det \Theta \\
\text{subject to } & \quad \Theta_{i,j} = 0 \quad \forall (i, j) \in \tilde{G} \setminus G \\
& \quad \Theta \in S^n_{\tilde{G}}, \quad \Theta \succ 0
\end{align*}
\]

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

\[
\begin{align*}
\min_{\Theta > 0} & \quad \text{trace}(S_{\lambda} \Theta) - \log \det \Theta \\
\text{subject to} & \quad \Theta_{i,j} = 0 \quad \text{wherever} \quad (S_{\lambda})_{i,j} = 0
\end{align*}
\]

3. Solve the **dual** problem

\[
\begin{align*}
\max & \quad -f_*(S_{\lambda} + Y) \\
\text{subject to} & \quad Y \in S^n_{\mathcal{G} \setminus G}
\end{align*}
\]

Self-concordance guarantees $\varepsilon$-accuracy in $O(\log \log (1/\varepsilon))$ Newton iterations.

Self-concordant barrier on the cone of sparse matrices

Edges added to make graph chordal
Main contribution. Newton-CG for MDMC

\[
\min_{\Theta > 0} \text{trace}(S_\lambda \Theta) - \log \det \Theta
\]
subject to \( \Theta_{i,j} = 0 \) wherever \( (S_\lambda)_{i,j} = 0 \)

4. Solve Newton direction using conjugate gradients

\[
\max - f_*(S + Y)
\]
subject to \( Y \in S^n_{\hat{G}\backslash G} \)

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

$O(n)$
$O(n^2)$
$O(n^{1.5})$
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 $\Sigma$ from 5000 i.i.d. samples from $\mathcal{N}(0, \Sigma)$

$O(n^{2.5})$ $O(n^{1.5})$
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.
\[ \hat{\Theta} = \min_{\Theta > 0} \text{trace}(S\Theta) - \log \det \Theta + \lambda \sum_{i \neq j} |\Theta_{i,j}|. \]

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