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

Richard Y. Zhang[1], Salar Fattahi[1], Somayeh Sojoudi[2] [1]IEOR, UC Berkeley, [2]EECS, UC Berkeley

- Graphical lasso (GL) estimates covariance matrix assuming that its inverse is sparse.
- Issues with complexity. Estimate n x n matrix in ≈ $O(n^3)$ time and ≈ $O(n^2)$ memory.
- **Contribution.** New algorithm solves GL on p parallel processors in $O(n + n^2 / p)$ time and $O(n)$ memory, assuming that the estimated model has bounded treewidth.
- Solve n = 200k problem in slightly over an hour on a laptop.

Sparse Inverse Covariance Estimation

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

Problem. Estimate the n x n covariance matrix assuming that the inverse covariance matrix is sparse

$$\Sigma \text{ is invertible} \qquad \Theta \equiv \Sigma^{-1} \text{ contains } O(n) \text{ nonzero elements}$$

Graphical Lasso. Lasso-regularized Gaussian MLE

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

Theoretical guarantees. (Ravikumar, Wainwright, Raskutti, Yu 2011)

$$\|\Theta - \hat{\Theta}\|_2 \le C\sqrt{\frac{d^2 \log n}{N}}$$

recover graph with sample size $N = \Omega(d^2 \log n)$

d is max degree, C is constant

Related work

GLASSO (Friedman et al. 2008). $O(n^3)$ time per iteration. Slow convergence due to block coordinate descent.

(BIG)-QUIC (Hsieh et al. 2014). $O(n^2)$ to $O(n^3)$ time per iteration. Bottlenecked by the computation of the proximal Newton direction.

CVXOPT (Dahl et al. 2008; Andersen et al. 2010). $O(n^3)$ time per iteration. Bottlenecked by the computation of the Newton direction.

Threshold+MDMC (Fattahi & Sojoudi 2018; Zhang, Fattahi & Sojoudi 2018). Only outputs GL estimator if thresholded pattern is chordal.

Elementary Estimator (Yang et al. 2014). Similar in schematics. However, EE is not a GL estimator.

Thresholding and MDMC

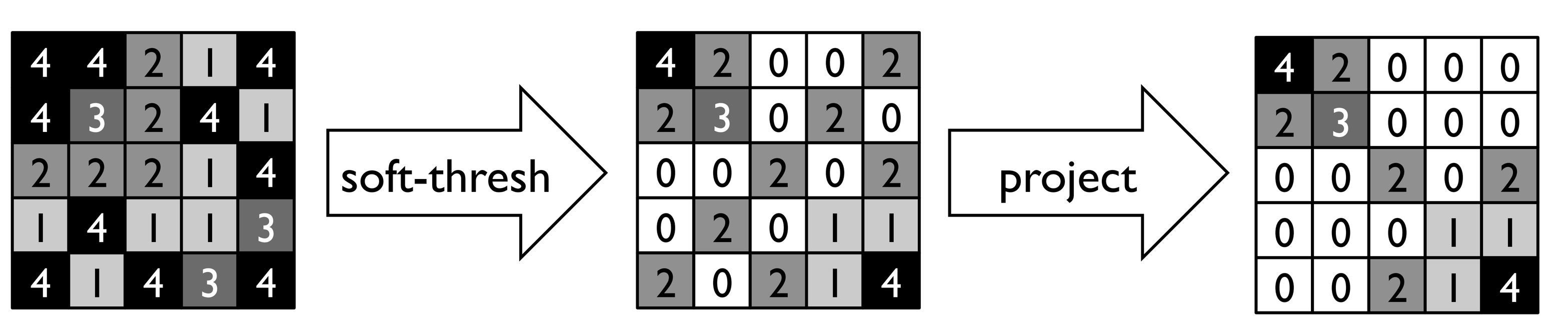
Restricted graphical lasso. GL with sparsity prior H

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

$$\text{subject to } \Theta_{i,j} = 0 \quad \text{for all } (i,j) \in H$$

1. Sparsity pattern. Soft-threshold + Project

$$(S_\lambda)_{i,j} = \begin{cases} S_{i,j} & i = j \\ S_{i,j} - \text{sign}(S_{i,j}) \cdot \lambda_{i,j} & |S_{i,j}| > \lambda_{i,j},\ i \neq j \\ 0 & |S_{i,j}| \le \lambda_{i,j},\ i \neq j \end{cases} \qquad [P_H(S_\lambda)]_{i,j} = \begin{cases} (S_\lambda)_{i,j} & (i,j) \in H \\ 0 & (i,j) \notin H \end{cases}$$



2. Parameter values. Max-Det Matrix Completion

$$\hat{\Theta} = \underset{\Theta \succ 0}{\text{minimize}}\ \text{tr}([P_H(S_\lambda)]\Theta) - \log\det\Theta$$

$$\text{subject to } \Theta_{i,j} = 0 \quad \text{for all } [P_H(S_\lambda)]_{i,j} = 0$$

Newton-CG algorithm for MDMC

$$\text{minimize } \text{tr}(S\Theta) - \log\det\Theta$$

$$\text{subject to } \Theta \in \mathbb{S}^n_G, \quad \Theta \succ 0$$

ove space of sparse matrices with sparsity pattern G:

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

Key insight. MDMC has *closed-form solution* over *chordal* pattern

$$f_*(S) = -\min_{\Theta \succ 0}\{\text{tr}(S\Theta) - \log\det\Theta : \Theta \in \mathbb{S}^n_{\tilde{G}}\}$$

This is a self-concordant barrier. Evaluation of function, gradient, and Hessian are O(n) time and space if G-tilde is bounded degree.

1. Embed general pattern G into chordal pattern G-tilde
2. Pose general MDMC as optimization over fill-in

$$\text{minimize } \text{tr}(S\Theta) - \log\det\Theta \qquad \text{maximize } -f_*(S+Y)$$

$$\text{subject to } \Theta_{i,j} = 0 \quad \forall (i,j) \in \tilde{G}\backslash G \qquad \text{subject to } Y \in \mathbb{S}^n_{\tilde{G}\backslash G}$$

$$\Theta \in \mathbb{S}^n_{\tilde{G}}, \quad \Theta \succ 0$$

3. Solve dual using Newton's, noting *self-concordance*

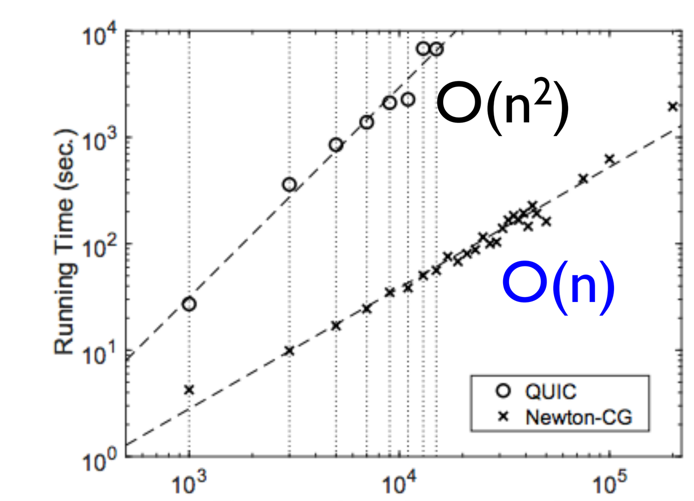
$$\hat{y} = \text{minimize } f_*(S - A(y)) \equiv g(y)$$

4. Compute each Newton direction using CG

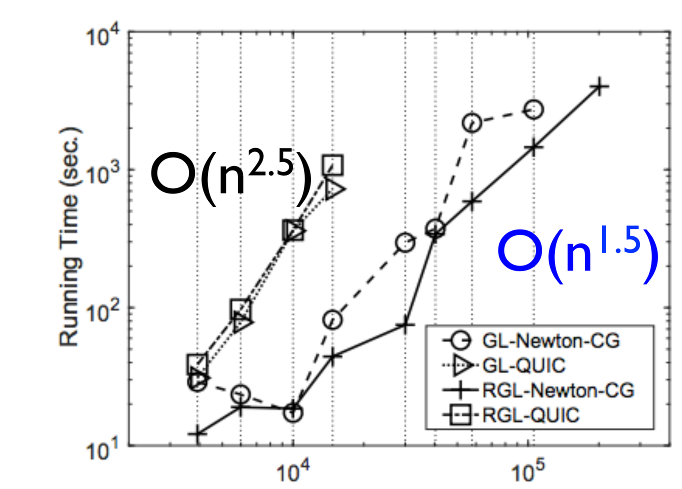
$$\nabla^2 g(y)\, \Delta y = -\nabla g(y).$$

Numerical Results

- Synthetic $\Theta = \Sigma^{-1}$ from given sparsity pattern (banded or real-life)
- Off-diagonals selected [-1,+1] and corrupted to zero with p=0.3
- Diagonals set to sum of off-diagonals plus one
- Attempt to estimate Θ from 5000 samples selected i.i.d. from N(0, Σ)



Banded graphs



Real-life graphs

						Newton-CG			QUIC		
#	file name	type	n	m	m/n	sec	gap	feas	sec	diff. gap	speed-up
1	freeFlyingRobot-7	GL	3918	20196	5.15	28.9	5.7e-17	2.3e-7	31.0	3.9e-4	1.07
1	freeFlyingRobot-7	RGL	3918	20196	5.15	12.1	6.5e-17	2.9e-8	38.7	3.8e-5	3.20
2	freeFlyingRobot-14	GL	5985	27185	4.56	23.5	5.4e-17	1.1e-7	78.3	3.8e-4	3.33
2	freeFlyingRobot-14	RGL	5985	27185	4.56	19.0	6.0e-17	1.7e-8	97.0	3.8e-5	5.11
3	cryg10000	GL	10000	170113	17.0	17.3	5.9e-17	5.2e-9	360.3	1.5e-3	20.83
3	cryg10000	RGL	10000	170113	17.0	18.5	6.3e-17	1.0e-7	364.1	1.9e-5	19.68
4	epb1	GL	14734	264832	18.0	81.6	5.6e-17	4.3e-8	723.5	5.1e-4	8.86
4	epb1	RGL	14734	264832	18.0	44.2	6.2e-17	3.3e-8	1076.4	4.2e-4	24.35
5	bloweya	GL	30004	10001	0.33	295.8	5.6e-17	9.4e-9	*	*	*
5	bloweya	RGL	30004	10001	0.33	75.0	5.5e-17	3.6e-9	*	*	*
6	juba40k	GL	40337	18123	0.44	373.3	5.6e-17	2.6e-9	*	*	*
6	juba40k	RGL	40337	18123	0.44	341.1	5.9e-17	2.7e-7	*	*	*
7	bayer01	GL	57735	671293	11.6	2181.3	5.7e-17	5.2e-9	*	*	*
7	bayer01	RGL	57735	671293	11.6	589.1	6.4e-17	1.0e-7	*	*	*
8	hcircuit	GL	105676	58906	0.55	2732.6	5.8e-17	9.0e-9	*	*	*
8	hcircuit	RGL	105676	58906	0.55	1454.9	6.3e-17	7.3e-8	*	*	*
9	co2010	RGL	201062	1022633	5.08	4012.5	6.3e-17	4.6e-8	*	*	*

- Thresh+MDMC rapidly recovers the GL estimator with suboptimality cerifiably on the order of 10^{-4}
- Largest problem with n = 200k solved in <70 minutes on a laptop.
- True positive rate, false positive rate, false negative rates all identical to GLASSO / QUIC when applicable.

Theorem 1 (Exact recovery)

order the absolute values $|S_{i,j}|$ for $(i,j) \in H$ as $s_1 \ge s_2 \ge \cdots$

Suppose:

- Let $\lambda \in (s_{k+1}, s_k)$ for some k
- $P_H(S_\lambda)$ is positive definite
- $(s_1 - \lambda)^2 \le \lambda - s_{k+1}$

} satisfied if λ is not small.

Then, thresholding and mat-det matrix completion suceeds.

Theorem 2 (CG converges in O(1) iterations)

Suppose:

- $g(y_0) - g(\hat{y}) = O(1)$ where y_0 is the initial point
- $\nabla g(y)^T (y - y_0) \le g(y_0) - g(\hat{y})$ where y is the current iterate

Then, CG converges to ε-accuracy in O(log(1/ε)) iterations.

Conclusions

- New algorithm computes GL estimators on large-scale datasets, based on self-concordant barrier function on sparse matrix cones.
- Statistical performance bottlenecked by the GL estimator.
- **Next steps.** Benchmark performance on real-life data.