Fast Projected Newton-like Method for Precision Matrix Estimation under Total Positivity

Jiaxi Ying

Joint work with Jian-Feng Cai, José Vinícius de M. Cardoso, and Daniel P. Palomar

Department of Mathematics
Hong Kong University of Science and Technology

Thirty-seventh Conference on Neural Information Processing Systems

Graphical models under total positivity

- Total positivity is also called multivariate totally positive of order two (MTP_2). A multivariate Gaussian distribution is MTP_2 if and only if precision matrix satisfies

$$\Theta_{ij} \leq 0, \quad \text{for all } i \neq j. \tag{1}$$

  Equivalently, $\boldsymbol{\Theta}$ is an *M-matrix*, or all partial correlations are nonnegative.

- Total positivity is a special form of positive dependence:

$$\mathrm{MTP}_2 \Longrightarrow \Sigma_{ij} \geq 0, \quad \text{for any } i \neq j. \tag{2}$$

- This model has broad applications including financial time series (Agrawal et al., 2020), taxonomic reasoning (Slawski and Hein, 2015), and factor analysis in psychometrics (Lauritzen et al., 2019).

Problem formulation

- Learning Gaussian graphical models under total positivity can be formulated as

$$\begin{array}{ll} \underset{\boldsymbol{\Theta}\in\mathcal{M}^p}{\text{minimize}} & -\log\det(\boldsymbol{\Theta}) + \text{tr}\left(\boldsymbol{\Theta}\boldsymbol{S}\right) + \sum_{i\neq j}\lambda_{ij}\left|\Theta_{ij}\right|, \\ \text{subject to} & \Theta_{ij} = 0,\ \forall\,(i,j)\in\mathcal{E}, \end{array} \tag{3}$$

  where λ_{ij} is regularization parameter, $\mathcal{E}$ is disconnectivity set that forces the pairs of nodes to be disconnected, and $\mathcal{M}^p := \{\boldsymbol{\Theta}\in\mathbb{S}^p_{++}|\,\Theta_{ij}\leq 0,\ \forall i\neq j\}$.

- Problem (3) is convex and constrained ($\boldsymbol{\Theta}\in\mathcal{M}^p$, $\Theta_{ij}=0\ \forall\,(i,j)\in\mathcal{E}$).

Existing algorithms

- **Block coordinate descent** (Egilmez et al., 2017): Updates one column/row at a time by solving a non-negative quadratic program, and cyclically updates all columns/rows. Each cycle requires $O(p^4)$ operations.
- **Proximal point algorithm** (Deng and So, 2020): Compute an inexact Newton direction from a $p^2 \times p^2$ linear system during each inner loop iteration, which is time-consuming in high-dimensional scenarios.
- **Projected gradient method** (Ying et al., 2023) : Computational efficiency with only $O(p^3)$ operations per iteration, but usually suffers from low convergence rate.

Proposed algorithm

Step 1: Partition variables into *restricted* set $\mathcal{I}_k$ and *free* set $\mathcal{I}_k^c$:

- The *restricted* set $\mathcal{I}_k$ in k-th iteration is set as

$$\mathcal{I}_k := \mathcal{T}(\mathbf{\Theta}_k, \epsilon_k) \cup \mathcal{E}. \tag{4}$$

  where $\mathcal{T}(\mathbf{\Theta}, \epsilon) := \left\{(i,j) \in [p]^2 \,\middle|\, -\epsilon \le \Theta_{ij} \le 0,\ \left[\nabla f(\mathbf{\Theta})\right]_{ij} < 0\right\}$.

- Variables in $\mathcal{T}(\mathbf{\Theta}_k, \epsilon_k)$ are close to boundary and tending to move outside feasible region.
- We impose variables in *restricted* set to be zero, and only update variables in *free* set.

Proposed algorithm

Step 2: Compute the search direction

- We use the special structure of the Hessian of the problem:

$$\left[\boldsymbol{P}_k\right]_{\mathcal{I}_k^c} = \left[\boldsymbol{\Theta}_k \mathcal{P}_{\mathcal{I}_k^c}\left(\nabla f\left(\boldsymbol{\Theta}_k\right)\right) \boldsymbol{\Theta}_k\right]_{\mathcal{I}_k^c}, \tag{5}$$

  where $\left[\mathcal{P}_{\mathcal{I}_k^c}(\boldsymbol{A})\right]_{ij} = A_{ij}$ if $(i,j) \in \mathcal{I}_k^c$, and zero otherwise.

- The proposed algorithm requires $O(p^3)$ operations and $O(p^2)$ memory, which are the same as those of projected gradient method.

- Our algorithm exploits second-order information when computing search direction, leading to a faster convergence than projected gradient method.

Proposed algorithm

Step 3: Compute the step size

- We try the step size $\gamma_k \in \{\beta^0, \beta^1, \beta^2, \ldots\}$, until we find the smallest $m \in \mathbb{N}$ such that the next iterate $\boldsymbol{\Theta}_{k+1}$ with $\gamma_k = \beta^m$ satisfies $\boldsymbol{\Theta}_{k+1} \succ \mathbf{0}$ and

$$f\left(\boldsymbol{\Theta}_{k+1}\right) \leq f\left(\boldsymbol{\Theta}_k\right) - \alpha\beta^m \left\langle \left[\nabla f(\boldsymbol{\Theta}_k)\right]_{\mathcal{I}_k^c}, \left[\boldsymbol{P}_k\right]_{\mathcal{I}_k^c} \right\rangle - \alpha \left\langle \left[\nabla f(\boldsymbol{\Theta}_k)\right]_{\mathcal{I}_k}, \left[\boldsymbol{\Theta}_k\right]_{\mathcal{I}_k} - \left[\boldsymbol{\Theta}_{k+1}\right]_{\mathcal{I}_k} \right\rangle. \quad (6)$$

- Such backtracking line search condition guarantees that the iterate will not terminate until it reaches the minimizer.

Proposed algorithm

Algorithm 1: Fast Projected Newton-like method (FPN)

Input: Sample covariance $\boldsymbol{S}$, λ, $\hat{\boldsymbol{w}}^{(0)}$;

for $k = 1, 2, \ldots$ **do**

Identify the *restricted* set $\mathcal{I}_k$ and *free* set $\mathcal{I}_k^c$;

Compute the approximation Newton direction over *free* set:

$$\left[\boldsymbol{P}_k\right]_{\mathcal{I}_k^c} = \left[\boldsymbol{\Theta}_k \mathcal{P}_{\mathcal{I}_k^c}\left(\nabla f\left(\boldsymbol{\Theta}_k\right)\right)\boldsymbol{\Theta}_k\right]_{\mathcal{I}_k^c};$$

$m \leftarrow 0$;

repeat

Update $\left[\boldsymbol{\Theta}_{k+1}\right]_{\mathcal{I}_k} = \boldsymbol{0}, \quad \left[\boldsymbol{\Theta}_{k+1}\right]_{\mathcal{I}_k^c} = \mathcal{P}_{\Omega}\left(\left[\boldsymbol{\Theta}_k\right]_{\mathcal{I}_k^c} - \beta^m \left[\boldsymbol{P}_k\right]_{\mathcal{I}_k^c}\right)$;

until $\boldsymbol{\Theta}_{k+1} \succ \boldsymbol{0}$, *and*

$$f\left(\boldsymbol{\Theta}_{k+1}\right) \leq f\left(\boldsymbol{\Theta}_k\right) - \alpha\beta^m \left\langle \left[\nabla f(\boldsymbol{\Theta}_k)\right]_{\mathcal{I}_k^c}, \left[\boldsymbol{P}_k\right]_{\mathcal{I}_k^c} \right\rangle - \alpha \left\langle \left[\nabla f(\boldsymbol{\Theta}_k)\right]_{\mathcal{I}_k}, \left[\boldsymbol{\Theta}_k\right]_{\mathcal{I}_k} - \left[\boldsymbol{\Theta}_{k+1}\right]_{\mathcal{I}_k} \right\rangle;$$

end

Output: $\boldsymbol{\Theta}_k$.

Theoretical results: Convergence analysis

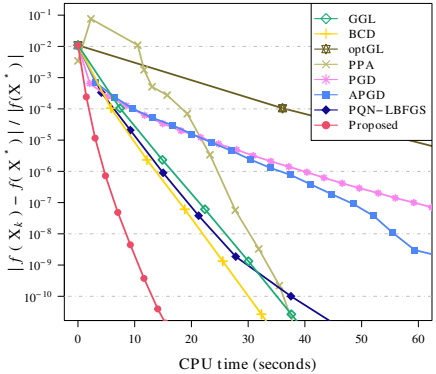
Theorem

The sequence $\{\boldsymbol{\Theta}_k\}$ generated by Algorithm 1 converges to the minimizer $\boldsymbol{\Theta}^\star$, with $\{f(\boldsymbol{\Theta}_k)\}$ monotonically decreasing. Moreover, under a mild assumption, there exists some $k_o \in \mathbb{N}_+$ such that

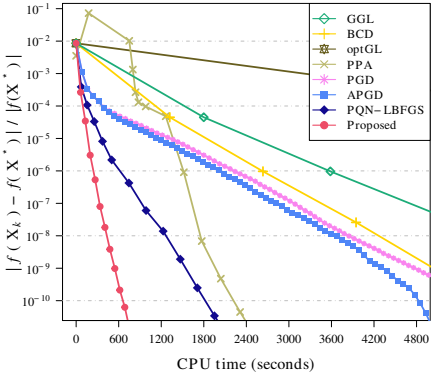
$$\mathcal{I}_k^c = \mathrm{supp}(\boldsymbol{\Theta}^\star), \quad \forall\, k \geq k_o.$$

In other words, for any $k \geq k_o$, the set of free variables is consistent with the support of $\boldsymbol{\Theta}^\star$.

Numerical results: Synthetic data



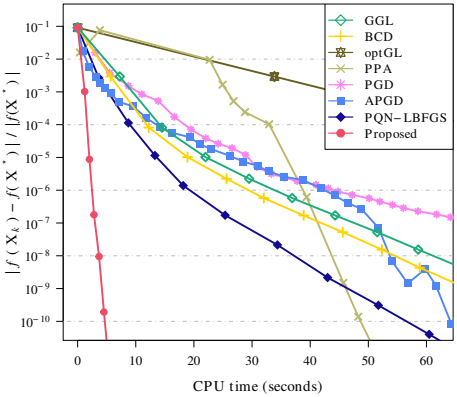
(a) 1000 nodes



(b) 5000 nodes

Convergence comparison of algorithms on BA graphs of degree one.

Numerical results: *concepts* data



Convergence comparison of algorithms on the *concepts* data (Lake and Tenenbaum, 2010), collected by Intel Labs. The data set includes 1000 nodes and 218 semantic features, where each node denotes a concept such as "baby", "chicken", and "house", and each feature is a question such as "Can it fly?", "Is it alive?". The answers are on a five-point scale from "definitely no" to "definitely yes".

References I

Agrawal, R., Roy, U., and Uhler, C. (2020). Covariance Matrix Estimation under Total Positivity for Portfolio Selection. *Journal of Financial Econometrics*.

Deng, Z. and So, A. M.-C. (2020). A fast proximal point algorithm for generalized graph laplacian learning. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5425–5429.

Egilmez, H. E., Pavez, E., and Ortega, A. (2017). Graph learning from data under Laplacian and structural constrints. *IEEE Journal of Selected Topics in Signal Processing*, 11(6):825–841.

Lake, B. and Tenenbaum, J. (2010). Discovering structure by learning sparse graphs. In *Proceedings of the 33rd Annual Cognitive Science Conference*, pages 778–783.

Lauritzen, S., Uhler, C., and Zwiernik, P. (2019). Maximum likelihood estimation in Gaussian models under total positivity. *The Annals of Statistics*, 47(4):1835–1863.

Slawski, M. and Hein, M. (2015). Estimation of positive definite M-matrices and structure learning for attractive Gaussian Markov random fields. *Linear Algebra and its Applications*, 473:145–179.

Ying, J., Cardoso, J. V. d. M., and Palomar, D. P. (2023). Adaptive estimation of graphical models under total positivity. In *International Conference on Machine Learning*, pages 40054–40074.

Thank you!