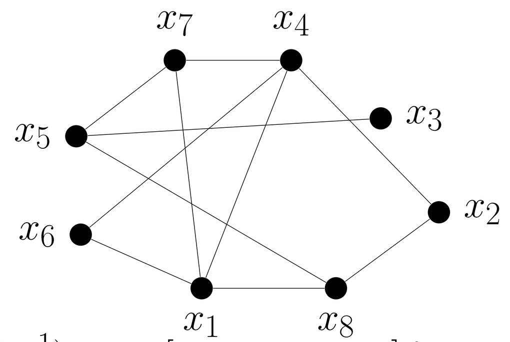
Learning Large-Scale MTP$_2$ Gaussian Graphical Models via Bridge-Block Decomposition

Xiwen Wang, Jiaxi Ying, and Daniel P. Palomar

THE HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY

Codes are available at: https://github.com/Xiwen1997/mtp2-bbd; For more information visit: https://www.danielppalomar.com

MTP$_2$ Graphical Model $\mathcal{G} = (\mathcal{V}, \mathcal{E})$



- Nodes: $\mathcal{V}$ correspond to the **entities**.
- Edges: $\mathcal{E}$ encode the **relationships** between entities.
- When the data follows a Gaussian distribution $\boldsymbol{x} \sim \mathcal{N}(0, \boldsymbol{\Theta}^{-1})$, $\mathbf{x} = [x_1, x_2, \ldots, x_p]$ is a Gaussian Graphical Model.
- The entries of Θ determines a **conditional** graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$:

$$x_i \perp x_j | \boldsymbol{x}/(x_i, x_j) \iff \Theta_{ij} = 0, \qquad \Theta_{ij} \neq 0 \iff \{i,j\} \in \mathcal{E} \; \forall \; i \neq j. \tag{1}$$

- The term **"multivariate total positivity of order 2" (MTP$_2$)** reflects a certain "positive association" between the random variables. All pairs of variables in the system are **positively correlated** when conditioned on the remaining variables.
- For Gaussian distributions, MTP$_2$ is equivalent to the precision matrix being a **M-matrix**, namely, $\Theta_{ij} \leq 0, \forall i \neq j$.
- MTP$_2$ property is particularly useful in many applications, such as finance and genomics, where **positive dependence between variables is a common phenomenon**.

Problem Formulation

Given sample covariance matrix $\boldsymbol{S}$ and regularization matrix $\boldsymbol{\Lambda}$, we are interested in

$$\underset{\boldsymbol{\Theta} \in \mathcal{M}^p}{\text{minimize}} \; -\log\det(\boldsymbol{\Theta}) + \langle \boldsymbol{\Theta}, \boldsymbol{S} \rangle + \textstyle\sum_{i \neq j} \Lambda_{ij} |\Theta_{ij}|, \tag{2}$$

where $\mathcal{M}^p$ refers to a set of M-matrices with dimension p, i.e.,

$$\mathcal{M}^p = \left\{ \boldsymbol{\Theta} \in \mathbb{S}^p \middle| \boldsymbol{\Theta} \succ \mathbf{0} \text{ and } \Theta_{ij} \leq 0, \forall i \neq j \right\}. \tag{3}$$

We aim at leveraging the sparsity properties behind the **support graph** of the **thresholded matrix** to find **optimal precision matrix** more efficiently, in which

- **Support graph**: For a symmetric matrix $\boldsymbol{A} \in \mathbb{S}^p$, its support graph, supp $(\boldsymbol{A})$, consists of vertex set $\mathcal{V} = \{1, \ldots, p\}$ and edge set $\mathcal{E}$, where an edge $(i,j) \in \mathcal{E}$ exists if $A_{ij} \neq 0$ for distinct vertices $i, j \in \mathcal{V}$.
- **Thresholded graph**: The support graph of thresholded matrix $\boldsymbol{T}$, which is defined as

$$T_{ij} = \begin{cases} S_{ij} - \Lambda_{ij} & \text{if } i \neq j \text{ and } S_{ij} > \Lambda_{ij}, \\ 0 & \text{otherwise.} \end{cases} \tag{4}$$

Bridge-Block Decomposition on Thresholded Graph

An edge is called a **Bridge** if and only if its deletion increases the number of graph components. Therefore, an edge is a bridge only when it is not contained in any cycles. **Bridge-block decomposition** [1] is conducted as follows. By removing all the bridges, we compute the clusters $\mathcal{V}_k$ corresponding to the components of $G = (\mathcal{V}, \mathcal{E} \setminus \mathcal{B})$. This process results in a vertex-partition, known as the bridge-block decomposition

$$\mathcal{P}^{\text{bbd}} = \{\mathcal{V}_1, \mathcal{V}_2, \ldots, \mathcal{V}_K\},$$

where K refers to the number of clusters, also the number of components in the graph $(\mathcal{V}, \mathcal{E} \setminus \mathcal{B})$.

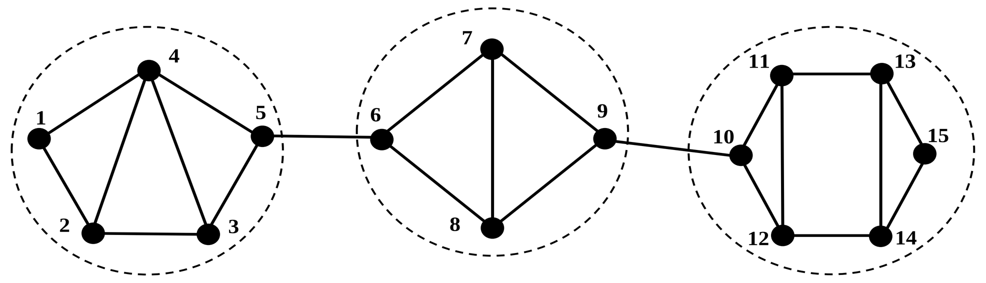


Figure 1: Bridge-block decomposition. $(5,6)$ and $(9,10)$ are bridges. By removing them, we obtain $\mathcal{P}^{\text{bbd}} = \{\{1,2,3,4,5\}, \{7,8,9,10\}, \{11,12,13,14,15,16\}\}$.

Main Results

Given $\boldsymbol{A} \in \mathbb{S}^p$, we define $\boldsymbol{A}_{\mathcal{V}_k} \in \mathbb{S}^{p_k}$ as the principal sub-matrix of $\boldsymbol{A}$ keeping the rows and columns indexed by $\mathcal{V}_k$. For each $i \in \mathcal{V}_k$, we mark $\pi(i)$ as its corresponding index in $\mathcal{V}_k$. Hence, we define $\bar{\boldsymbol{\Theta}}_k$ as the optimal solution of k-th sub-problem, i.e.,

$$\bar{\boldsymbol{\Theta}}_k = \arg\min_{\boldsymbol{\Theta}_k \in \mathcal{M}^{p_k}} -\log\det(\boldsymbol{\Theta}_k) + \langle \boldsymbol{\Theta}_k, \boldsymbol{S}_{\mathcal{V}_k} - \boldsymbol{\Lambda}_{\mathcal{V}_k} \rangle. \tag{5}$$

Theorem 1. *Given the bridge-block decomposition of the thresholded graph supp $(\boldsymbol{T})$ as $\mathcal{P}^{bbd}$, and the optimal solution of each sub-problem (5) as $\bar{\boldsymbol{\Theta}}_k$, the optimal solution of Problem (2), i.e., $\Theta^\star$, can be obtained as*

$$\Theta^\star_{i,j} = \begin{cases} [\bar{\boldsymbol{\Theta}}_k]_{\pi(i),\pi(i)} + \zeta_i & \text{if } i = j \in \mathcal{V}_k, \\ [\bar{\boldsymbol{\Theta}}_k]_{\pi(i),\pi(j)} & \text{if } i \neq j \text{ and } i,j \in \mathcal{V}_k, \\ -T_{ij}/(S_{ii}S_{jj} - T_{ij}^2) & \text{if } (i,j) \in \mathcal{B}, \\ 0 & \text{otherwise.} \end{cases} \tag{6}$$

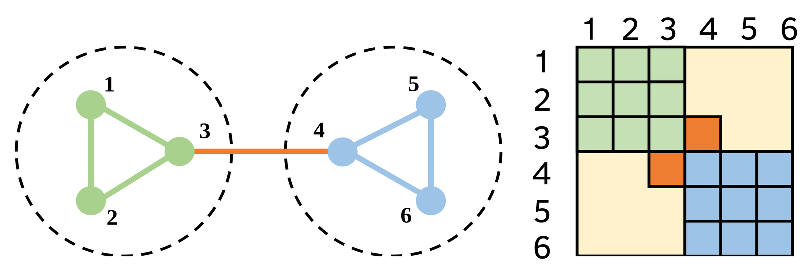
in which $\zeta_i = \frac{1}{S_{ii}} \sum_{(i,m) \in \mathcal{B}} \frac{T_{im}^2}{S_{ii}S_{mm} - T_{im}^2}$ *and* $\zeta_i = 0$ *if* $\forall m : (i,m) \notin \mathcal{B}$.

Especially, when the threhsolded graph is acyclic, we have $\mathcal{V}_k = \{k\}$ for all k. In this case, we obtain an closed-form expression for Θ as

$$\Theta^\star_{ij} = \begin{cases} \frac{1}{S_{ii}} \left(1 + \sum_{m \in \mathcal{N}(i)} \frac{T_{im}^2}{S_{ii}S_{mm} - T_{im}^2} \right) & \text{if } i = j, \\ -\frac{T_{ij}}{S_{ii}S_{jj} - T_{ij}^2} & \text{if } (i,j) \in \mathcal{B}, \\ 0 & \text{otherwise.} \end{cases} \tag{7}$$

Similar results also hold for graphical lasso [2] when more restrict conditions are given. The involvement of MTP$_2$ remove the non-smoothness in the problem and therefore simplify the KKT conditions.

Examples of Applying Bridge-Block Decomposition



- $\Theta_{\{1,2,3\}} = \bar{\Theta}_1 + \text{diag}(\zeta_1, \zeta_2, \zeta_3)$.
- $\Theta_{\{4,5,6\}} = \bar{\Theta}_2 + \text{diag}(\zeta_4, \zeta_5, \zeta_6)$.
- $(i,j) \in \mathcal{B}$: $\Theta_{ij} = -T_{ij}/(S_{ii}S_{jj} - T_{ij}^2)$.
- $(i,j) \notin \mathcal{B}$ and $\psi(i) \neq \psi(j)$: $\Theta_{ij} = 0$.

Figure 2: An example of obtaining optimal solution via Theorem 1. The thresholded graph of 6 nodes can be partitioned into 2 clusters $\mathcal{V}_1 = \{1,2,3\}$ and $\mathcal{V}_2 = \{4,5,6\}$. The optimal solution $\Theta \in \mathbb{S}^6$ can be exactly computed via a set of explicit solution $\{\Theta_{3,4}, \Theta_{4,3}\}$, solutions of smaller-sized sub-problems, i.e., $\bar{\Theta}_1 \in \mathbb{S}^3$ and $\bar{\Theta}_2 \in \mathbb{S}^3$, and zeros in the rest positions.

It is often more efficient to employ the bridge-block decomposition technique instead of directly manipulating full-sized matrices. This approach involves breaking down the problem into smaller isolated sub-problems.

- **Significant reduction in computational cost**. Suppose that we use a BCD solver of complexity $\mathcal{O}(p^4)$, then the total cost is reduced to $\sum_{k=1}^K \mathcal{O}(|\mathcal{V}_k|^4) \ll \mathcal{O}(p^4)$, where $\sum_k |\mathcal{V}_k| = p$. This can prompt an enormous difference.
- **Considerable reduction in memory cost**. Theorem 1 can avoid generating a number of full-dimensional intermediate variables during computation.
- **Potential speed-up via parallel computing**. The sub-problems can be optimized independently, which allows parallel computing for significant speed-up.

Experimental Results

Synthetic Data

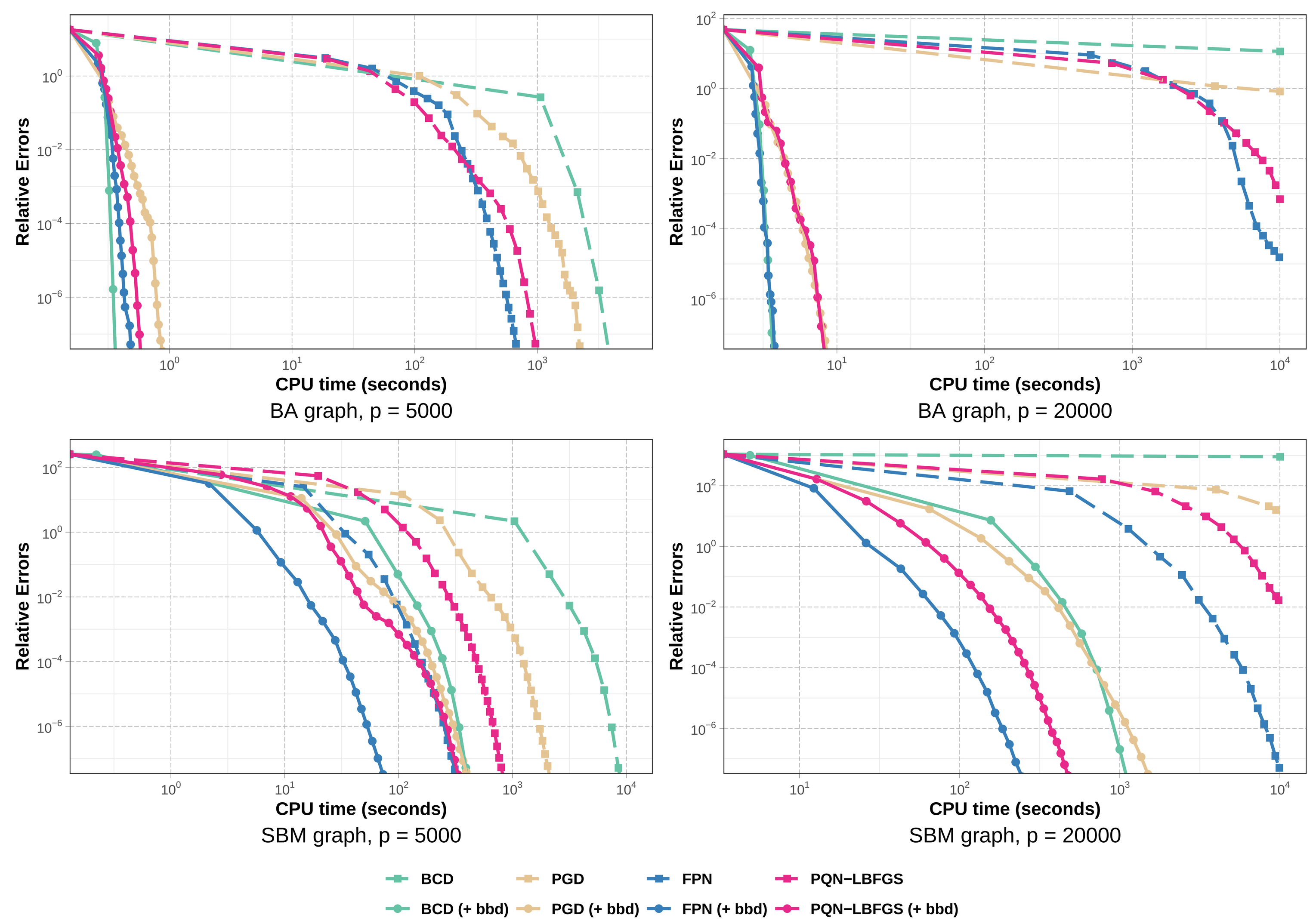


Figure 3: Relative errors of the objective values versus the computational time.

Real-world Data

We perform graph-based clustering to the time series data containing $p = 24000$ instances. Our proposed method enables the learning of such a large-scale problem. To the learned model, the majority of edges exist within the same crop type, while connections between nodes associated with different crops are relatively sparse.

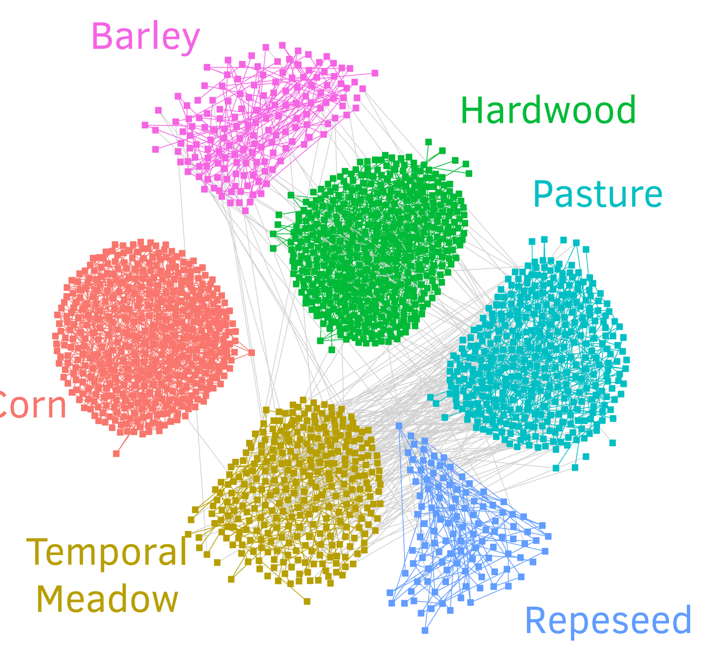


Figure 4: A local structure of optimal graph with 2008 nodes.

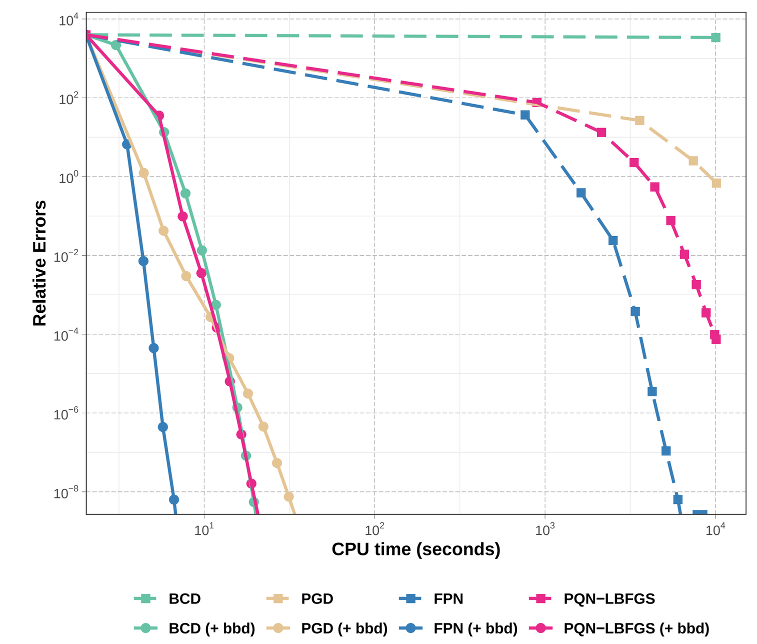


Figure 5: Results of convergence to learn sparse MTP$_2$ GGMs for Crop data set.

References

1. J. Westbrook and R. E. Tarjan, "Maintaining bridge-connected and biconnected components on-line," *Algorithmica*, vol. 7, no. 1, pp. 433–464, 1992.
2. S. Sojoudi, "Equivalence of graphical lasso and thresholding for sparse graphs," *The Journal of Machine Learning Research*, vol. 17, no. 1, pp. 3943–3963, 2016.