Learning Discrete Graphical Models via Generalized Inverse Covariance Matrices

Duzhe Wang, Yiming Lv, Jongkim Kim, Young Lee

STAT679/992: Graphical Model Class Project

Abstract

This project investigates the relationship between the structure of a discrete graphical model and the support of the inverse of a generalized covariance matrix. This work extends results that have previously been established only in the context of multivariate Gaussian graphical models. The population-level results in Loh and Wainwright 2013 have theoretically rigorous consequences for global graph selection methods and local neighborhood selection methods. Furthermore, these methods are easily adapted to permuted observations and missing data.

Motivation

• Hammersley-Clifford theorem: the Markov and factorization properties are equivalent for any strictly positive distribution.

• When \((X_1, \ldots, X_p) \sim N(0, \Sigma), \Sigma_{ij} = \Sigma_{ji} = 0 \iff (x_i, x_j) \notin E\).

• In non-Gaussian setting, what’s the relationship between entries of \(\theta\) and edges of \(G\)?

An example: Binary Ising model

\[ q(x_1, \ldots, x_p) = \sum_{x \in \{0, 1\}^p} e^{-\beta \sum_{i=1}^p h_i x_i + \sum_{i<j} a_{ij} x_i x_j} \]

where \((x_1, \ldots, x_p) \in \{0, 1\}^p\).

Now consider the case \(p = 4, \theta_0 = 0.1, \beta = 2\).

\[\begin{align*}
X_1: & \quad 0.90 & \quad -0.59 & \quad 0.0
\end{align*}\]

\[\begin{align*}
X_2: & \quad -5.39 & \quad 4.77 & \quad -3.59 & \quad 9.00
\end{align*}\]

Figure 1: \(G\) is graph-connected for chain, but not loop.

However, define \(\tau_{\Sigma, 2} = \frac{1}{n} \langle (X_i - \mu_i)(X_j - \mu_j)^\top \rangle \) for loop.

\[\begin{align*}
X_1: & \quad 1.10 & \quad 1.90 & \quad -0.02 & \quad -0.02
\end{align*}\]

\[\begin{align*}
X_2: & \quad 0.09 & \quad -1.14 & \quad 0.52 & \quad -0.51
\end{align*}\]

\[\begin{align*}
X_3: & \quad -0.02 & \quad 0.01 & \quad 0.01 & \quad 0.01
\end{align*}\]

\[\begin{align*}
X_4: & \quad 0.01 & \quad 0.01 & \quad 0.01 & \quad 0.01
\end{align*}\]

Conclusion: The usual inverse covariance matrix is not always graph-structured, but inverse of augmented matrices involving higher-order interaction terms may reveal graph structure.

Some notations and definitions

• \(\text{Support}(X_1, \ldots, X_p) \in \mathbb{X} = \{0, \ldots, 1\}^p\).

• Let \(C\) denote the set of all cliques in \(G = (V, E)\).

• For each clique \(C \subseteq V\), define the subset of configurations \(\mathcal{X}_C = \{X_C \in \mathcal{X} : \forall x \in C, X_C(x) = \text{constant}\}\).

• Let \(\mathcal{X} = \{X \in \mathbb{X}^p : \text{no two variables are independent}\}\).

• For each \(C \subseteq V\), let \(\mathcal{X}_C = \{X_C \in \mathcal{X} : \forall x \in C, X_C(x) = \text{constant}\}\), consisting of all the sufficient statistics indexed by elements of \(C\).

• For any \(S \subseteq V\), define \(\tau_{\Sigma, S} = \frac{1}{n} \langle (X_S - \mu_S)(X_S - \mu_S)^\top \rangle\).

• Given an undirected graph \(G = (V, E)\), a triangulation is a graph \(\hat{G} = (V, \hat{E})\) that contains no chordless cycles of length greater than 3.

Population-level results

• Triangulation and block graph structure

Theorem 1: Let \(G\) be the set of all cliques in any triangulation of \(G\). Then the regularized covariance matrix \(\Sigma_0^{-1}(\hat{G}, \tau_{\Sigma, \hat{G}})\) is invertible and in reverse \(\theta\) is block-graph-structured.

1. for any two subsets \(S, B \subseteq \hat{V}\) where \(S\) is not included in the minimal clique, the block \((\hat{G}, \tau_{\Sigma, \hat{G}})_{S, B}\) is identically zero.

2. for almost all parameters \(\theta\), the entire block \((\hat{G}, \tau_{\Sigma, \hat{G}})_{S, B}\) is nonzero whenever \(A\) and \(B\) belong to a common maximal clique.

• Separate sets are effective

Corollary 1 (Estimation): Let \(\theta_0 = \text{for the set of separate sets in any triangulation of } G\), and let the inverse of \(\Sigma_0^{-1}(\hat{G}, \tau_{\Sigma, \hat{G}})\) be \(\Sigma_0^{-1}(\hat{G}, \tau_{\Sigma, \hat{G}}) = 0 \iff (x_i, x_j) \notin E\).

Corollary 2 (Structure learning): For any graph with at least one separate set, the inverse \(\tau_{\Sigma, \hat{G}}^{-1}(\hat{G}, \tau_{\Sigma, \hat{G}})\) consists of all the sufficient statistics indexed by elements of \(G\) that are not subsets of the same maximal clique.

• Using model reestimated

Figure 2: A concrete example of Theorem 1 and Corollary 1

Structure learning

1. Fully observed data

• Global graph selection method

- Graphical Lasso (log-determinant): \(\theta_0 = \arg\min_{\theta} \left\{ \tau_{\Sigma, \hat{G}}^{-1}(\hat{G}, \tau_{\Sigma, \hat{G}}) \right\} = (\hat{G}, \tau_{\Sigma, \hat{G}})^{-1}\).

- Optimize the Graphical Lasso program with \(\lambda = 0.1\) and denote the solution by \(\hat{\theta}\).

- Threshold the entries of \(\hat{\theta}_0\) at level \(\tau_0\) to obtain an estimate of \(\theta_0\).

Algorithm 1: Graphical Lasso

- Statistical consistency:

Corollary 3: Assume the linear model defined by a triangulated graph with separate sets and with degree at most \(d\), and suppose that some certain minimal independence condition holds. With \(n > \log\log d\) samples, there are universal constants \(\mathcal{C}\) such that with probability at least \(1 - \mathcal{C} \log \log d\), Algorithm 1 recovers all edges \(\hat{E}\) with \(\tau_{\Sigma, \hat{G}}\) greater than 2.

- Local neighborhood selection methods

Case 1: nodewise regression in trees

- \(\lambda\)-regularized linear regression of \(X_i\) against \(X_j\):

- \(\beta \in \text{argmin}_{\beta} \left\{ (\beta^\top \Sigma_0^{-1}(\hat{G}, \tau_{\Sigma, \hat{G}}) \beta)^\top \Sigma_0^{-1}(\hat{G}, \tau_{\Sigma, \hat{G}}) \beta \right\} \\text{subject to } \|\beta\|_1 \leq \lambda\).

Case 2: nodewise regression in general graphs

- Neighborhood selection

Corollary 4: Assume the tree is \((\geq l_0)\) and \(\theta_0 = \text{for the set of separate sets in any triangulation of } G\). Then there are universal constants \(\mathcal{C}\) such that with probability greater than \(1 - \mathcal{C} \log\log d\), for all nodes \(X_i \in V\), Algorithm 2 recovers all neighbors \(\hat{E}\) of \(X_i\) in \((\geq l_0)\) such that \(\|\hat{\Sigma}_{i, \hat{E}}\|_{1,\infty} \leq \lambda\).

Corollary 2: Suppose \(E\) is a forest and \(\theta_0 = \text{for the set of separate sets in any triangulation of } G\). Then there are universal constants \(\mathcal{C}\) such that with probability greater than \(1 - \mathcal{C} \log\log d\), for at most \(l_0\) edges \(E\) in \((\geq l_0)\), Algorithm 2 recovers all neighbors \(\hat{E}\) of \(X_i\) in \((\geq l_0)\) such that \(\|\hat{\Sigma}_{i, \hat{E}}\|_{1,\infty} \leq \lambda\).

• Unbiased estimates of \(\Sigma\) and \(\Sigma^{-1}\) (Xin and Lafferty 2010)

- Solution: apply the solution from the linear model selection using \(\lambda\)-regularized linear regression.

Simulations

• Goals: (i) test the \(n = \log\log d\) scaling of the required sample size, (ii) compare \(\lambda\)-regularized nodewise linear regression to \(\lambda\)-regularized nodewise logistic regression.

• Setting: data was generated from a binary linear model, with node weights \(\tau_0 = 0.1\) and edge weights \(\tau_0 = 0.1\).

References


