Learning discrete graphical models via generalized inverse covariance matrices

Duzhe Wang, Yiming Lv, Yongjoon Kim, Young Lee
Department of Statistics
University of Wisconsin-Madison
{dwang282, lv23, ykim676, ylee546}@wisc.edu

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 corrupted observations and missing data.

1 Introduction

One of the important topics in graphical model is learning the edge structure of an underlying graph. The classical corollary of the Hammersley-Clifford theorem states zeros in the inverse covariance matrix of a multivariate Gaussian distribution indicate absent edges in the corresponding graph. This fact, combined with techniques in high-dimensional statistical inference, has been leveraged by many authors to recover the structure of a Gaussian graphical model when the edge set is sparse.

However, it is unclear whether a relationship exists or not between the conditional independence and the structure of the inverse covariance matrix for non-Gaussian graphical models. Loh et al. [5] focus on discrete graphical models and establish links between covariance matrices and the edge structure of an underlying graph. They show augmenting the usual covariance matrix with higher-order interaction terms will demystify this relationship. Their main population-level results have a striking corollary in the context of tree-structured graphs: for any discrete graphical model, the inverse of a generalized covariance matrix is always graph-structured. In particular, for binary variables, the inverse of the usual covariance matrix corresponds exactly to the edge structure of the tree.

Graph selection methods for Gaussian data include the graphical Lasso which maximizes an $l_1$-regularized likelihood and the neighborhood selection[6]. Other related work on graphical model selection for discrete graphs includes the classic Chow-Liu algorithm for trees[1], nodewise logistic regression for discrete models with pairwise interactions[7]. Loh et al. [5] propose minor variants of the graphical Lasso and neighborhood selection methods remain consistent for trees and the broader class of graphical models with singleton separator sets. In addition, their methods for graph selection may be adapted to handle noisy or missing data in a seamless manner.

The remainder of the report is organized as follows: In Section 2, we provide a brief background and some notations on graphical models, and then describe the augmented covariance matrices with an intuitive example. In Section 3, we state the main results on the relationship between the support of generalized inverse covariance matrices and the edge structure of a discrete graphical model. In Section 4, we list the concrete algorithms that are guaranteed to recover the edge structure of a discrete graph with high probability, we also compare the structure learning methods in [5] with some
previous work. We report the results of simulations in Section 5. For detailed proofs, we refer the
reader to [5].

2 Background

In this section, we provide background on graphical models and define one specific class of discrete
distributions parameterized by an exponential family.

2.1 Undirected graphical models

An undirected graph \( G = (V, E) \) consists of a collection of nodes \( V = \{1, ..., p\} \) and a collection of
edges \( E \subset V \times V \). For each node \( s \in V \), there is an associated random variable \( X_s \) taking values in
some space \( \mathcal{X} \).

**Definition 1** (Markov property). The random vector \( X = (X_1, ..., X_p) \) is a Markov random field
with respect to \( G \) if \( X_A | X_B \neq X_S \) whenever \( S \) is a cutset of \( A \) and \( B \), which means every path from
\( A \) to \( B \) in \( G \) must pass through \( S \).

**Definition 2** (Factorization property). The distribution of \( X \) factorizes according to \( G \) if it may be
written as a product of clique functions:

\[
q(x_1, ..., x_p) \propto \prod_{C \in \mathcal{C}} \Psi_C(x_C)
\]

By the Hammersley-Clifford theorem, the Markov and factorization properties are equivalent for
any strictly positive distribution. Loh and Wainwright [5] focus on the analysis of strictly positive
distributions whose factorization can be represented in terms of an exponential family associated with
the clique structure of \( G \). They are rigorously defined in mathematics as follows.

Consider a discrete graphical model with \( (X_1, ..., X_p) \in \mathcal{X} = \{0, ..., m - 1\}^p \). For clique \( C \), define
the subset of configurations

\[
\mathcal{X}_0^{[C]} = \{(j_s, s \in C) \in \mathcal{X}^{[C]} : j_s \neq 0, \forall s \in C\},
\]

where \( \mathcal{X}_0 = \mathcal{X} \setminus \{0\} = \{1, 2, ..., m - 1\} \). For any configuration \( J = \{j_s, s \in C\} \in \mathcal{X}_0^{[C]} \), define the
sufficient statistics

\[
I_{C; J}(x_C) = \begin{cases} 1, & \text{if } x_C = J, \\ 0, & \text{otherwise}. \end{cases}
\]

Then the specific class of discrete distributions we will study in this report is parametrized by a
minimal exponential family:

\[
q_\theta(x_1, ..., x_p) = \exp \left\{ \sum_{C \in \mathcal{C}} \langle \theta_C, I_C \rangle - \Phi(\theta) \right\}
\]

where \( \langle \theta_C, I_C \rangle = \sum_{J \in \mathcal{X}_0^{[C]}} \theta_{C; J} I_{C; J}(x_C) \). Note when \( m = 2 \) and the graphical model has only
pairwise interactions, then model (2) reduces to the binary Ising model.

2.2 Covariance matrices

When \( X \) is Gaussian, it is well-known that the entries of \( \Theta = \Sigma^{-1} \) correspond to rescaled conditional
correlations. The magnitude of \( \Theta_{st} \) is a scalar multiple of the correlation of \( X_s \) and \( X_t \) conditioned
on \( X_{\setminus\{s,t\}} \), and encodes the strength of the edge \((s, t)\), but for general non-Gaussian distributions, it
is not known whether the entries of \( \Theta \) have any relationship with the strengths of edges in the graph.

**Example 1.** Consider the binary Ising model:

\[
q_\theta(x_1, ..., x_p) = \exp \left\{ \sum_{s \in V} \theta_s x_s + \sum_{(s,t) \in E} \theta_{st} x_s x_t - \Phi(\theta) \right\}
\]
where \((x_1, ..., x_p) \in \{0, 1\}^p\). For the Ising model on a simple chain graph of four nodes, in terms of factorization (3), let the node potentials be \(\theta_s = 0.1\) for all \(s \in V\) and the edge potentials be \(\theta_{st} = 2\) for all \((s, t) \in E\). A little computation shows that \(\Theta = \Sigma^{-1}\) takes the form shown in Figure 1. We find the entries of \((1, 3), (1, 4)\) and \((2, 4)\) in \(\Theta\) are 0, which surprisingly implies \(\Theta\) is graph-structured. However, this statement is not true for the loop graph, which is also shown in Figure 1. We find its inverse covariance matrix has nonzero value in each entry.

![Figure 1: \(\Theta\) is graph-structured for chain, but not for single-cycle](image)

On the other side, if we still focus on the loop graph, instead of considering the ordinary covariance matrix, we compute the covariance matrix with the augmented random vector \((X_1, X_2, X_3, X_4, X_1X_3)\), where the extra term \(X_1X_3\) is represented by adding the edge between \(X_1\) and \(X_3\). Then the inverse of this generalized covariance matrix takes the form

\[
\Theta_{aug} \propto \begin{bmatrix}
1.15 & -0.02 & 1.09 & -0.02 & -1.14 \\
-0.02 & 0.05 & -0.02 & 0 & 0.01 \\
1.09 & -0.02 & -1.14 & -0.02 & -1.14 \\
-0.02 & 0 & -0.02 & 0.05 & 0.01 \\
-1.14 & 0.01 & -1.14 & 0.01 & 1.19 \\
\end{bmatrix}
\]

This matrix separates nodes 2 and 4, but doesn’t separate nodes 1 and 3 \((1, 3)\). We could observe a similar phenomenon if we choose to augment the graph by including the edge \((2, 4)\) rather than \((1, 3)\). This example shows that the usual inverse covariance matrix is not always graph-structured, but generalized inverse covariance matrices involving higher-order interactions may reveal graph structure.

### 3 Main results

We state the main population-level results on the structure of generalized inverse covariance matrices and the graph structure in this section.

#### 3.1 Triangulation and block graph structure

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

**Remark.** A tree contains no cycles, so it’s trivially triangulated. The chordless 4-cycle in Figure 1 is the simplest example of a nontriangulated graph. By adding the single edge \((1, 3)\) to form the augmented edge set \(\tilde{E} = U \cup \{(1, 3)\}\), we obtain the triangulated graph \(\tilde{G} = (V, \tilde{E})\).

We require some extra notations for defining generalized covariance matrices. For \(S \subseteq C\), define the random vector \(\Psi(X; S) = \{I_{C,J}, J \in X_0^{[C]}, C \in S\}\), consisting of all the sufficient statistics indexed by elements of \(S\). For any \(S\), define \(\text{pow}(S) = \bigcup_{C \in S}\text{pow}(C)\).

Finally, we explore the relationship between the graph structure of some triangulation \(\tilde{G}\) of \(G\) and the inverse \(\Theta\) of the matrix \(\text{cov}(\Psi(X; \hat{C}))\), where \(\hat{C}\) is the set of all cliques arising from \(\tilde{G}\). For any two subsets \(A, B \in \hat{C}\), we write \(\Theta(A, B)\) to denote the sub-block of \(\Theta\) indexed by all indicator statistics on \(A\) and \(B\), respectively.
We now state some results concerning different methods for graph selection in discrete graphical models, based on i.i.d. draws from a discrete graph. For sparse Gaussian models, existing methods that exploit sparsity of the inverse covariance matrix fall into two main categories: global graph selection methods and local neighborhood selection methods. We discuss these two categories of methods in the following.

4 Structure learning

We now state some results concerning different methods for graph selection in discrete graphical models, based on i.i.d. draws from a discrete graph. For sparse Gaussian models, existing methods that exploit sparsity of the inverse covariance matrix fall into two main categories: global graph selection methods and local neighborhood selection methods. We discuss these two categories of methods in the following.

4.1 Graphical Lasso for singleton separator graphs

Given a p-dimensional random vector \((X_1, ..., X_p)\) with covariance \(\Sigma^*\), consider the estimator

\[
\hat{\Theta} \in \arg\min_{\Theta} \left\{ \text{trace}(\hat{\Sigma}\Theta) - \log \det(\Theta) + \lambda_n \sum_{s \neq t} |\Theta_{st}| \right\}
\]

(4)
where $\hat{\Sigma}$ is an estimator for $\Sigma^*$. For multivariate Gaussian random variables, program (4) is a $l_1$-regularized maximum likelihood estimate known as the graphical Lasso. Although the program (4) has no relation to the MLE in the case of a discrete graphical model, it may still be useful for estimating $\Theta^*$. Ravikumar et al. [8] shows the estimator (4) requires only tail condition such as sub-Gaussianity in order to guarantee that the sample minimizer is close to the population minimizer.

The global edge recovery algorithm in Loh et al. [5] proceeds as follows:

**Algorithm 1:** Graphical Lasso

**Data:** i.i.d. draws of sample size $n$ and parameters $(\lambda_n, \tau_n)$

1. Form a suitable pair of estimators $\hat{\Theta}$ of the true covariance matrix $\Sigma$;
2. Optimize the graphical Lasso program with parameter $\lambda_n$ and denote the solution by $\hat{\Theta}$;
3. Threshold the entries of $\hat{\Theta}$ at level $\tau_n$ to obtain an estimate of $\Theta^*$;

**Corollary 4.** Consider an Ising model defined by an undirected graph with singleton separator sets and with degree at most $d$, and suppose that some certain mutual incoherence condition holds. With $n > d^2 \log p$ samples, there are universal constants $(c, c')$ such that with probability at least $1 - c \exp(- c' \log p)$, Algorithm 1 recovers all edges $(s, t)$ with $|\Theta^*_{st}| > \tau/2$.

**Remark.** 1. A suitable estimator $\hat{\Sigma}$ in Algorithm 1 should satisfy

$$P(\|\hat{\Sigma} - \Sigma^*\|_{\max} \geq \phi(\Sigma^*) \sqrt{\frac{\log p}{n}} \leq c \exp(- \psi(n, p))$$

for functions $\phi$ and $\psi$, where $\| \cdot \|_{\max}$ denotes the elementwise $l_\infty$ norm.

2. The authors establish statistical consistency of $\hat{\Theta}$ under the following setting:

$$\lambda_n \geq \frac{c_1}{\alpha} \sqrt{\frac{\log p}{n}}, \tau_n = c_2 \{ \frac{c_1}{\alpha} \sqrt{\frac{\log p}{n}} + \lambda_n \}$$

3. The proof of Corollary 3 relies heavily on the population-level result in Corollary 2, which ensures that $\Theta^*$ is graph-structured when $G$ has only singleton separators. Therefore, the graphical Lasso (4) is inconsistent for a general graph.

4. When the samples $\{x_i\}_{i=1}^n$ are contaminated by noise or missing data, we may still form a good estimate $\hat{\Sigma}$ of $\Sigma^*$, so the program (4) is easily adapted to handle corrupted data.

**4.2 Nodewise regression in trees**

Motivated by the population-level results on the graph structure of the inverse covariance matrix, the authors also propose a method for neighborhood selection in the tree-structured graph.

The method is based on the following steps. For each node $s \in V$, first perform $l_1$-regularized linear regression of $X_s$ against $X_{\backslash s}$ by solving the modified Lasso program

$$\hat{\beta} \in \text{argmin} \{ \frac{1}{2} \beta^T \hat{\Gamma} \beta - \hat{\gamma}^T \beta + \lambda_n \|\beta\|_1, s.t. \|\beta\|_1 \leq b_0 \sqrt{k} \}$$

where $b_0 \geq \|\hat{\beta}\|_1$ and $\hat{\beta} = \Sigma_{\backslash s, s}^{-1} \Sigma_{s, s} (\hat{\Gamma}, \hat{\gamma})$ are suitable estimators for $(\Sigma_{\backslash s, s}, \Sigma_{s, s})$ and $\lambda_n$ is an appropriate parameter.

The following algorithm summarizes the procedure for recovering the neighborhood set $N(s)$ of a given node $s$:

**Algorithm 2:** Nodewise method for trees

**Data:** i.i.d. draws of sample size $n$ and parameters $(\lambda_n, \tau_n)$

1. Form a suitable pair of estimators $(\hat{\Gamma}, \hat{\gamma})$ for covariance submatrices of the true covariance matrix $(\Sigma_{\backslash s, s}, \Sigma_{s, s})$ and $\lambda_n$;
2. Optimize the above modified Lasso program with parameter $\lambda_n$ and denote the solution by $\hat{\beta}$;
3. Threshold the entries of $\hat{\beta}$ at level $\tau_n$ and define the estimated neighborhood set $N(s)$ as the support of the thresholded vector.
Corollary 5. Suppose we have i.i.d. observations \( \{x_i\}_{i=1}^n \) from an Ising model and that \( n > \phi^2 \max\left\{ \frac{1}{\log p}, \frac{1}{\log d} \right\} d^2 \log p \). Then there are universal constants \( (c, c', c'') \) such that with probability greater than \( 1 - c \exp(-c' \log p) \), for any nodes \( s \in V \), Algorithm 2 recovers all neighbors \( t \in N(s) \) for which \( |\hat{\beta}_t| \geq c' \phi |\hat{\beta}|_2 \sqrt{\frac{\log p}{n}} \).

Remark. 1. For fully observed i.i.d. observations, \((\hat{\Gamma}, \hat{\gamma})\) are chosen to be the recentered estimators
\[
\hat{\Gamma} = \frac{X^T X_s}{n} - \bar{x}_s \bar{x}_s^T, \quad \hat{\gamma} = \frac{X^T X_s}{n} - \bar{x}_s \bar{x}_s
\]
and assign the parameters \((\lambda_n, \tau_n)\) according to the scaling
\[
\lambda_n \gs \phi |\hat{\beta}|_2 \sqrt{\frac{\log p}{n}}, \quad \tau_n \gs \phi |\hat{\beta}|_2 \sqrt{\frac{\log p}{n}}.
\]

4.3 Nodewise regression in general graphs

Corollary 3 in Section 3.2 suggests a graph recovery method based on nodewise linear regression for general discrete graphs.

**Algorithm 3:** Nodewise method for general graphs

**Data:** i.i.d. draws of sample size \( n \) and parameters \((\lambda_n, \tau_n)\)

1. Use the modified Lasso program \((5)\) with a suitable choice of \((\hat{\Gamma}, \hat{\gamma})\) and regularization parameter \(\lambda_n\) to perform a linear regression of \(X_s\) upon all products of subsets of variables of \(X_{\setminus s}\) of size at most \(d\). Denote the solution by \( \hat{\beta} \);
2. Threshold the entries of \( \hat{\beta} \) at level \( \tau_n \) and define the estimated neighborhood set \( N(\hat{s}) \) as the support of the thresholded vector.

Remark. 1. The knowledge of an upper bound \( d \) is necessary for applying Algorithm 3.
2. Since the regression vector has \( O(p^d) \) components, \( 2^d - 1 \) of which are nonzero, so the sample complexity of Lasso regression in step 1 of Algorithm 3 is \( O(2^d \log p^d) = O(2^d \log p)\).

4.4 Comparison with previous work

Ravikumar et al.\(^7\) propose using \( l_1 \)-regularized logistic regression in the high-dimensional Ising model selection. The method performs logistic regression of each variable \( X_s \) on \( X_{\setminus s} \), and it’s known to succeed with high probability with sample size \( n \gs d^3 \log p \), where \( d \) is the maximum degree.

However, the method in Loh et al.\(^5\) performs linear regression of each variable \( X_s \) on the \( d \)-extended version of \( X_{\setminus s} \), it involves \( O(p^d) \) covariates, so it’s only suitable for bounded degree graphs. But one positive side of this method is it can extend naturally to deal with missing or corrupted covariates.

5 Simulations

Figure 2 in the following depicts the results of simulations in order to (i) test the \( n \gs \log p \) scaling of the required sample size and (ii) compare \( l_1 \)-regularized nodewise linear regression to \( l_1 \)-regularized nodewise logistic regression. In all cases, data was generated from a binary Ising model, with node weights \( \theta_s = 0.1 \) and edge weights \( \theta_{st} = 0.3 \). It plots the probability of correct graph recovery vs. the rescaled sample size \( n/\log p \). Solid curves correspond to linear regression and dotted curves correspond to logistic regression. Curves are based on average performance over 500 trials. We observe all six runs display a transition from success probability 0 to success probability 1 in roughly the same range of the rescaled sample size, which is expected by the theory.

6 Discussion and future directions

Loh et al.\(^5\) extends the connection between the inverse covariance matrix and graph structure to the discrete graphical models which can be represented by a minimal exponential family. Although recent
work\cite{3,2} has established such connection for nonparanormal distribution (Gaussian distribution under monotonic univariate transformations), this is the first work to make the connection for discrete distributions, to our knowledge. Motivated by the results of this work, Loh et al. \cite{4} established a new framework of learning causal networks when data are generated from a linear, possibly non-Gaussian structural equations model via inverse covariance estimation.

Loh et al. \cite{5} also highlights the limitations of inverse covariance matrix estimation for discrete models. First, without prior knowledge of the graph structure, the global selection methods using graphical Lasso are restricted to graphs with singleton separator sets. While nodewise selection method for general graphs are proposed, it is only applicable for bounded degree graphs and can be computationally costly. Secondly, the result is limited to the minimal exponential families such as multinomial distribution. In practice, it would be useful to have an extension to more general discrete distributions (i.e. Poisson), continuous distributions and mixed distributions.

References


\cite{2} Han Liu, Fang Han, Ming Yuan, John Lafferty, and Larry Wasserman. High-dimensional semiparametric gaussian copula graphical models. \textit{Ann. Statist.}, 40(4):2293–2326, 08 2012.


\cite{6} Nicolai Meinshausen and Peter Buhlmann. High-dimensional graphs and variable selection with the lasso. \textit{Ann. Statist.}, 34(3):1436–1462, 06 2006.
