# Structure estimation for discrete graphical models: Generalized covariance matrices and their inverses

**Po-Ling Loh**
Department of Statistics
University of California, Berkeley
Berkeley, CA 94720
ploh@berkeley.edu

**Martin J. Wainwright**
Departments of Statistics and EECS
University of California, Berkeley
Berkeley, CA 94720
wainwrig@stat.berkeley.edu

## Abstract

We investigate a curious relationship between the structure of a discrete graphical model and the support of the inverse of a generalized covariance matrix. We show that for certain graph structures, the support of the inverse covariance matrix of indicator variables on the vertices of a graph reflects the conditional independence structure of the graph. Our work extends results that have previously been established only in the context of multivariate Gaussian graphical models, thereby addressing an open question about the significance of the inverse covariance matrix of a non-Gaussian distribution. Based on our population-level results, we show how the graphical Lasso may be used to recover the edge structure of certain classes of discrete graphical models, and present simulations to verify our theoretical results.

## 1 Introduction

Graphical model inference is now prevalent in many fields, running the gamut from computer vision and civil engineering to political science and epidemiology. In many applications, learning the edge structure of an underlying graphical model is of great importance—for instance, a graphical model may be used to represent friendships between people in a social network, or links between organisms with the propensity to spread an infectious disease [1]. It is well known that zeros in the inverse covariance matrix of a multivariate Gaussian distribution indicate the absence of an edge in the corresponding graphical model. 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 (e.g., see the papers [2, 3, 4, 5] and references therein). Recently, Liu et al. [6, 7] introduced the notion of a nonparanormal distribution, which generalizes the Gaussian distribution by allowing for univariate monotonic transformations, and argued that the same structural properties of the inverse covariance matrix carry over to the nonparanormal.

However, the question of whether a relationship exists between conditional independence and the structure of the inverse covariance matrix in a general graph remains unresolved. In this paper, we focus on discrete graphical models and establish a number of interesting links between covariance matrices and the edge structure of an underlying graph. Instead of only analyzing the standard covariance matrix, we show that it is often fruitful to augment the usual covariance matrix with higher-order interaction terms. Our main result has a striking corollary in the context of tree-structured graphs: for *any* discrete graphical model, the inverse of a generalized covariance matrix is always (block) graph-structured. In particular, for binary variables, the inverse of the usual covariance matrix corresponds exactly to the edge structure of the tree. We also establish several corollaries that apply to more general discrete graphs. Our methods are capable of handling noisy or missing data in a seamless manner.

Other related work on graphical model selection for discrete graphs includes the classic Chow-Liu algorithm for trees [8]; nodewise logistic regression for discrete models with pairwise interactions [9, 10]; and techniques based on conditional entropy or mutual information [11, 12]. Our main contribution is to present a clean and surprising result on a simple link between the inverse covariance matrix and edge structure of a discrete model, which may be used to derive inference algorithms applicable even to data with systematic corruptions.

The remainder of the paper is organized as follows: In Section 2, we provide brief background and notation on graphical models, and describe the classes of augmented covariance matrices we will consider. In Section 3, we state our main results on the relationship between the support of generalized inverse covariance matrices and the edge structure of a discrete graphical model. We relate our population-level results to concrete algorithms that are guaranteed to recover the edge structure of a discrete graph with high probability. In Section 4, we report the results of simulations used to verify our theoretical claims. For detailed proofs, we refer the reader to the technical report [13].

## 2 Background and problem setup

In this section, we provide background on graphical models and exponential families. We then work through a simple example that illustrates the phenomena and methodology studied in this paper.

### 2.1 Graphical models

An undirected graph $G = (V, E)$ consists of a collection of vertices $V = \{1, 2, \ldots, p\}$ and a collection of unordered vertex pairs $E \subseteq V \times V$, meaning no distinction is made between edges $(s, t)$ and $(t, s)$. We associate to each vertex $s \in V$ a random variable $X_s$ taking values in some space $\mathcal{X}$. The random vector $X := (X_1, \ldots, X_p)$ is a *Markov random field* with respect to $G$ if $X_A \perp\!\!\!\perp X_B \mid X_S$ whenever $S$ is a *cutset* of $A$ and $B$, meaning every path from $A$ to $B$ in $G$ must pass through $S$. We have used the shorthand $X_A := \{X_s : s \in A\}$. In particular, $X_s \perp\!\!\!\perp X_t \mid X_{\setminus\{s,t\}}$ whenever $(s, t) \notin E$.

By the Hammersley-Clifford theorem for strictly positive distributions [14], the Markov properties imply a factorization of the distribution of $X$:

$$\mathbb{P}(x_1, \ldots, x_p) \propto \prod_{C \in \mathcal{C}} \psi_C(x_C), \tag{1}$$

where $\mathcal{C}$ is the set of all cliques (fully-connected subsets of $V$) and $\psi_C(x_C)$ are the corresponding clique potentials. The factorization (1) may alternatively be represented in terms of an *exponential family* associated with the clique structure of $G$. For each clique $C \in \mathcal{C}$, we define a family of sufficient statistics $\{\phi_{C;\alpha} : \mathcal{X}^{|C|} \to \mathbb{R},\ \alpha \in \mathcal{I}_C\}$ associated with variables in $C$, where $\mathcal{I}_C$ indexes the sufficient statistics corresponding to $C$. We also introduce a canonical parameter $\theta_{C;\alpha} \in \mathbb{R}$ associated with each sufficient statistic $\phi_{C;\alpha}$. For a given assignment of canonical parameters $\theta$, we may express the clique potentials as

$$\psi_C(x_C) = \sum_{\alpha \in \mathcal{I}_C} \theta_{C;\alpha} \phi_{C;\alpha}(x_C) := \langle \theta_C,\, \phi_C \rangle,$$

so equation (1) may be rewritten as

$$\mathbb{P}_\theta(x_1, \ldots, x_p) = \exp\big\{ \sum_{C \in \mathcal{C}} \langle \theta_C,\, \phi_C \rangle - A(\theta) \big\}, \tag{2}$$

where $A(\theta) := \log \sum_{x \in \mathcal{X}^p} \exp\big( \sum_{C \in \mathcal{C}} \langle \theta_C,\, \phi_C \rangle \big)$ is the (log) partition function.

Note that for a graph with only pairwise interactions, we have $\mathcal{C} = V \cup E$. If we associate the function $\phi_s(x_s) = x_s$ with clique $\{s\}$ and the function $\phi_{st}(x_s, x_t) = x_s x_t$ with edge $(s, t)$, the factorization (2) becomes

$$\mathbb{P}_\theta(x_1, \ldots, x_p) = \exp\big\{ \sum_{s \in V} \theta_s x_s + \sum_{(s,t) \in E} \theta_{st} x_s x_t - A(\theta) \big\}. \tag{3}$$

When $\mathcal{X} = \{0, 1\}$, this family of distributions corresponds to the inhomogeneous Ising model. When $\mathcal{X} = \mathbb{R}$ (and with certain additional restrictions on the weights), the family (3) corresponds to a Gauss-Markov random field. Both of these models are minimal exponential families, meaning the sufficient statistics are linearly independent [15].

For a discrete graphical model with $\mathcal{X} = \{0, 1, \ldots, m-1\}$, it is convenient to make use of sufficient statistics involving indicator functions. For clique $C$, define the subset of configurations

$$\mathcal{X}_0^{|C|} = \{J = (j_1, \ldots, j_{|C|}) \mid j_\ell \neq 0 \quad \text{for all } \ell = 1, \ldots, |C|\},$$

for which no variables take the value 0. Then $|\mathcal{X}_0^{|C|}| = (m-1)^{|C|}$. For any configuration $J \in \mathcal{X}_0^{|C|}$, we define the indicator function

$$\phi_{C;J}(x_C) = \begin{cases} 1 & \text{if} \quad x_C = J, \\ 0 & \text{otherwise,} \end{cases}$$

and consider the family of models

$$\mathbb{P}_\theta(x_1, \ldots, x_p) = \exp\Big\{ \sum_{C \in \mathcal{C}} \langle \theta_C, \, \phi_C \rangle - A(\theta) \Big\}, \quad \text{where } x_j \in \mathcal{X} = \{0, 1, \ldots, m-1\}, \tag{4}$$

with $\langle \theta_C, \, \phi_C \rangle = \sum_{J \in \mathcal{X}_0^{|C|}} \theta_{C;J} \phi_{C;J}(x_C)$. Note in particular that when $m = 2$, $X_0^{|C|}$ is a singleton state containing the vector of all ones, and the sufficient statistics are given by

$$\phi_{C;J}(x_C) = \prod_{s \in C} x_s, \qquad \text{for} \quad C \in \mathcal{C} \quad \text{and} \quad J = \{1\}^{|C|};$$

i.e., the indicator functions may simply be expressed as products of variables appearing in the clique. When the graphical model has only pairwise interactions, elements of $\mathcal{C}$ have cardinality at most two, and the model (4) clearly reduces to the Ising model (3). Finally, as with the equation (3), the family (4) is a minimal exponential family.

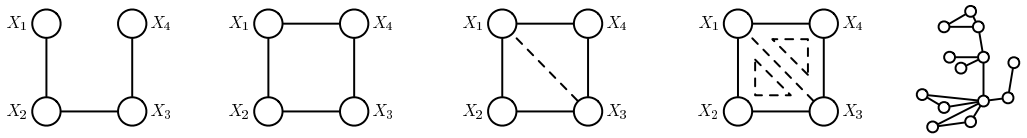
### 2.2 Covariance matrices and beyond

Consider the usual covariance matrix $\Sigma = \operatorname{cov}(X_1, \ldots, X_p)$. When $X$ is Gaussian, it is a well-known consequence of the Hammersley-Clifford theorem that the entries of the precision matrix $\Gamma = \Sigma^{-1}$ correspond to rescaled conditional correlations [14]. The magnitude of $\Gamma_{st}$ 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 $(s, t)$. In particular, the sparsity pattern of $\Gamma_{st}$ reflects the edge structure of the graph: $\Gamma_{st} = 0$ if and only if $X_s \perp\!\!\!\perp X_t \mid X_{\setminus\{s,t\}}$. For more general distributions, however, $\operatorname{Corr}(X_s, X_t \mid X_{\setminus\{s,t\}})$ is a function of $X_{\setminus\{s,t\}}$, and it is not known whether the entries of $\Gamma$ have any relationship with the strengths of edges in the graph.

Nonetheless, it is tempting to conjecture that inverse covariance matrices, and more generally, inverses of higher-order moment matrices, might be related to graph structure. Let us explore this possibility by considering a simple example, namely the binary Ising model (3) with $\mathcal{X} = \{0, 1\}$.

**Example 1.** Consider a simple chain graph on four nodes, as illustrated in Figure 1(a). In terms of the 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 $(s, t) \in E$. For a multivariate Gaussian graphical model defined on $G$, standard theory predicts that the inverse covariance matrix $\Gamma = \Sigma^{-1}$ of the distribution is graph-structured: $\Gamma_{st} = 0$ if and only if $(s, t) \notin E$. Surprisingly, this is also the case for the chain graph with binary variables: a little computation show that $\Gamma$ takes the form shown in panel (f). However, this statement is *not* true for the single-cycle graph shown in panel (b), with added edge $(1, 4)$. Indeed, as shown in panel (g), the inverse covariance matrix has no nonzero entries at all. But for a more complicated graph, say the one in (e), we again observe a graph-structured inverse covariance matrix.

Still focusing on the single-cycle graph in panel (b), suppose that instead of considering the ordinary covariance matrix, we compute the covariance matrix of 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 the dotted edge shown




(a) Chain (b) Single cycle (c) Edge augmented (d) With 3-cliques (e) Dino

$$\Gamma_{\text{chain}} = \begin{bmatrix} 9.80 & -3.59 & 0 & 0 \\ -3.59 & 34.30 & -4.77 & 0 \\ 0 & -4.77 & 34.30 & -3.59 \\ 0 & 0 & -3.59 & 9.80 \end{bmatrix}. \qquad \Gamma_{\text{loop}} = \begin{bmatrix} 51.37 & -5.37 & -0.17 & -5.37 \\ -5.37 & 51.37 & -5.37 & -0.17 \\ -0.17 & -5.37 & 51.37 & -5.37 \\ -5.37 & -0.17 & -5.37 & 51.37 \end{bmatrix},$$

(f) (g)

**Figure 1.** (a)–(e) Different examples of graphical models. (f) Inverse covariance for chain-structured graph in (a). (g) Inverse covariance for single-cycle graph in (b).

in panel (c). The $5 \times 5$ inverse of this generalized covariance matrix takes the form

$$\Gamma_{\text{aug}} = 10^3 \times \begin{bmatrix} 1.15 & -0.02 & 1.09 & -0.02 & -1.14 \\ -0.02 & 0.05 & -0.02 & 0 & 0.01 \\ 1.09 & -0.02 & 1.14 & -0.02 & -1.14 \\ -0.02 & 0 & -0.02 & 0.05 & 0.01 \\ -1.14 & 0.01 & -1.14 & 0.01 & 1.19 \end{bmatrix}.$$

This matrix safely separates nodes 1 and 4, but the entry corresponding to the phantom edge $(1, 3)$ is *not* equal to zero. Indeed, we would observe a similar phenomenon if we chose to augment the graph by including the edge $(2, 4)$ rather than $(1, 3)$. Note that the relationship between entries of $\Gamma_{\text{aug}}$ and the edge strength is not direct; although the factorization (3) has no potential corresponding to the augmented "edge" $(1, 3)$, the $(1, 3)$ entry of $\Gamma_{\text{aug}}$ is noticeably larger in magnitude than the entries corresponding to actual edges with nonzero potentials. This example shows that the usual inverse covariance matrix is not always graph-structured, but computing generalized covariance matrices involving higher-order interaction terms may indicate graph structure.

Now let us consider a more general graphical model that adds the 3-clique interaction terms shown in panel (d) to the usual Ising terms. We compute the covariance matrix of the augmented vector

$$\Phi(X) = \left\{X_1, X_2, X_3, X_4, X_1X_2,\ X_2X_3,\ X_3X_4,\ X_1X_4,\ X_1X_3,\ X_1X_2X_3,\ X_1X_3X_4\right\} \in \{0, 1\}^{11}.$$

Empirically, we find that the $11 \times 11$ inverse of the matrix $\operatorname{cov}(\Phi(X))$ continues to respect aspects of the graph structure: in particular, there are zeros in position $(\alpha, \beta)$, corresponding to the associated functions $X_\alpha = \prod_{s \in \alpha} X_s$ and $X_\beta = \prod_{s \in \beta} X_\beta$, whenever $\alpha$ and $\beta$ do not lie within the same maximal clique. (For instance, this applies to the pairs $(\alpha, \beta) = (\{2\}, \{4\})$ and $(\alpha, \beta) = (\{2\}, \{1, 4\})$.)

The goal of this paper is to understand when certain inverse covariances do (and *do not*) capture the structure of a graphical model. The underlying principles behind the behavior demonstrated in Example 1 will be made concrete in Theorem 1 and its corollaries in the next section.

## 3 Main results and consequences

We now state our main results on the structure of generalized inverse covariance matrices and graph structure. We present our results in two parts: one concerning statements at the population level, and the other concerning statements at the level of statistical consistency based on random samples.

### 3.1 Population-level results

Our main result concerns a connection between the inverses of generalized inverse covariance matrices associated with the model (4) and the structure of the graph. We begin with some notation.

Recall that a *triangulation* of a graph $G = (V, E)$ is an augmented graph $\widetilde{G} = (V, \widetilde{E})$ with no chordless 4-cycles. (For instance, the single cycle in panel (b) is a chordless 4-cycle, whereas panel

(c) shows a triangulated graph. The dinosaur graph in panel (e) is also triangulated.) The edge set $\widetilde{E}$ corresponds to the original edge set $E$ plus the additional edges added to form the triangulation. In general, $G$ admits many different triangulations; the results we prove below will hold for any fixed triangulation of $G$.

We also require some notation for defining generalized covariance matrices. Let $\mathcal{S}$ be a collection of subsets of vertices, and define the random vector

$$\Phi(X; \mathcal{S}) = \big\{\phi_{S;J}, J \in \mathcal{X}_0^{|C|}, \quad S \in \mathcal{S} \cap \mathcal{C}\big\}, \tag{5}$$

consisting of all sufficient statistics over cliques in $\mathcal{S}$. We will often be interested in situations where $\mathcal{S}$ contains all subsets of a given set. For a subset $A \subseteq V$, we let $\text{pow}(A)$ denote the set of all non-empty subsets of $A$. (For instance, $\text{pow}(\{1,2\}) = \{1, 2, (1,2)\}$.) Furthermore, for a collection of subsets $\mathcal{S}$, we let $\text{pow}(\mathcal{S})$ be the set of all subsets $\{\text{pow}(S), S \in \mathcal{S}\}$, discarding any duplicates that arise. We are now ready to state our main theorem regarding the support of a certain type of generalized inverse covariance matrix.

**Theorem 1.** *[Triangulation and block graph-structure.] Consider an arbitrary discrete graphical model of the form (4), and let $\mathcal{T}$ be the set of maximal cliques in any triangulation of $G$. Then the inverse $\Gamma$ of the augmented covariance matrix $\text{cov}(\Phi(X; pow(\mathcal{T})))$ is* block graph-structured *in the following sense:*

- *(a) For any two subsets $A$ and $B$ which are not subsets of the same maximal clique, the block $\Gamma(pow(A), pow(B))$ is zero.*
- *(b) For almost all parameters $\theta$, the entire block $\Gamma(pow(A), pow(B))$ is nonzero whenever $A$ and $B$ belong to a common maximal clique.*

The proof of this result relies on convex analysis and the geometry of exponential families [15, 16]. In particular, in any minimal exponential family, there is a one-to-one correspondence between exponential parameters ($\theta_\alpha$ in our notation) and mean parameters ($\mu_\alpha = \mathbb{E}[\phi_\alpha(X)]$). This correspondence is induced by the Fenchel-Legendre duality between the log partition function $A$ and its dual $A^*$, and allows us to relate $\Gamma$ to the graph structure.

Note that when the original graph $G$ is a tree, the graph is already triangulated and the set $\mathcal{T}$ in Theorem 1 is equal to the edge set $E$. Hence, Theorem 1 implies that the inverse $\Gamma$ of the augmented covariance matrix with sufficient statistics for all vertices and edges is graph-structured, and blocks of nonzeros in $\Gamma$ correspond to edges in the graph. In particular, the $(m-1)p \times (m-1)p$ submatrix $\Gamma_{V,V}$ corresponding to sufficient statistics of vertices is block graph-structured; in the case when $m = 2$, the submatrix $\Gamma_{V,V}$ is simply the $p \times p$ block corresponding to the vector $(X_1, \ldots, X_p)$. When $G$ is not triangulated, however, we may need to invert a larger augmented covariance matrix and include sufficient statistics over pairs $(s,t) \notin E$, as well.

In fact, it is not necessary to take the set of sufficient statistics over all maximal cliques, and we may consider a slightly smaller augmented covariance matrix. Recall that any triangulation $\mathcal{T}$ gives rise to a *junction tree* representation of $G$, where nodes of the junction tree are subsets of $V$ corresponding to maximal cliques in $\mathcal{T}$, and the edges are intersections of adjacent cliques known as *separator sets* [15]. The following corollary involves the generalized covariance matrix containing only sufficient statistics for nodes and separator sets of $\mathcal{T}$:

**Corollary 1.** *Let $\mathcal{S}$ be the set of separator sets in any triangulation of $G$, and let $\Gamma$ be the inverse of $\text{cov}(\Phi(X; V \cup pow(\mathcal{S})))$. Then $\Gamma_{V,V}$ is block graph-structured: $\Gamma_{s,t} = 0$ whenever $(s,t) \notin \widetilde{E}$.*

The proof of this corollary is based on applying the block matrix inversion formula [17] to express $\Gamma_{V,V}$ in terms of the matrix $\Gamma$ from Theorem 1. Panel (c) of Example 1 and the associated matrix $\Gamma_{\text{aug}}$ provides a concrete instance of this corollary in action. In panel (c), the single separator set in the triangulation is $\{1,3\}$, so augmenting the usual covariance matrix with the additional sufficient statistic $X_1X_3$ and taking the inverse should yield a graph-structured matrix. Indeed, edge $(2,4)$ does not belong to $\widetilde{E}$, and as predicted by Corollary 1, we observe that $\Gamma_{\text{aug}}(2,4) = 0$.

Note that $V \cup \text{pow}(\mathcal{S}) \subseteq \text{pow}(\mathcal{T})$, and the set of sufficient statistics considered in Corollary 1 is generally much smaller than the set of sufficient statistics considered in Theorem 1. Hence, the generalized covariance matrix of Corollary 1 has a smaller dimension than the generalized covariance matrix of Theorem 1, and is much more tractable for estimation.

Although Theorem 1 and Corollary 1 are clean results at the population level, however, forming the proper augmented covariance matrix requires some prior knowledge of the graph—namely, which edges are involved in a suitable triangulation. In the case of a graph with only singleton separator sets, Corollary 1 specializes to the following useful corollary, which only involves the covariance matrix over indicators of vertices of $G$:

**Corollary 2.** *For any graph with singleton separator sets, the inverse matrix $\Gamma$ of the ordinary covariance matrix $\mathrm{cov}(\Phi(X;V))$ is graph-structured. (This class includes trees as a special case.)*

Again, we may relate this corollary to Example 1—the inverse covariance matrices for the tree graph in panel (a) and the dinosaur graph in panel (e) are exactly graph-structured. Indeed, although the dinosaur graph is not a tree, it possesses the nice property that the only separator sets in its junction tree are singletons.

Corollary 1 also guarantees that inverse covariances may be partially graph-structured, in the sense that $(\Gamma_{V,V})_{st} = 0$ for any pair of vertices $(s,t)$ separable by a singleton separator set. This is because for any such pair $(s,t)$, we form a junction tree with two nodes, one containing $s$ and one containing $t$, and apply Corollary 1 to conclude that $(\Gamma_{V,V})_{st} = 0$. Indeed, the matrix $\Gamma_{V,V}$ over singleton vertices is agnostic to which triangulation we choose for the graph.

In settings where there exists a junction tree representation of the graph with only singleton separator sets, Corollary 2 has a number of useful implications for the consistency of methods that have traditionally only been applied for edge recovery in Gaussian graphical models. In such settings, Corollary 2 implies that it suffices to estimate the support of $\Gamma_{V,V}$ from the data.

### 3.2 Consequences for graphical Lasso for trees

Moving beyond the population level, we now establish results concerning the statistical consistency of methods for graph selection in discrete graphical models, based on i.i.d. draws from a discrete graph. We describe how a combination of our population-level results and some concentration inequalities may be leveraged to analyze the statistical behavior of log-determinant methods for discrete tree-structured graphical models, and suggest extensions of these methods when observations are systematically corrupted by noise or missing data.

Given $p$-dimensional random variables $(X_1, \ldots, X_p)$ with covariance $\Sigma^*$, consider the estimator

$$\widehat{\Theta} \in \arg\min_{\Theta \succeq 0} \{\mathrm{trace}(\widehat{\Sigma}\Theta) - \log\det(\Theta) + \lambda_n \sum_{s \neq t} |\Theta_{st}|\}, \tag{6}$$

where $\widehat{\Sigma}$ is an estimator for $\Sigma^*$. For multivariate Gaussian data, this program is an $\ell_1$-regularized maximum likelihood estimate known as the *graphical Lasso*, and is a well-studied method for recovering the edge structure in a Gaussian graphical model [18, 19, 20]. Although the program (6) has no relation to the MLE in the case of a discrete graphical model, it is still useful for estimating $\Theta^* := (\Sigma^*)^{-1}$, and our analysis shows the surprising result that the program is consistent for recovering the structure of any tree-structured Ising model. We consider a general estimate $\widehat{\Sigma}$ of the covariance matrix $\Sigma$ such that

$$\mathbb{P}\big[\|\widehat{\Sigma} - \Sigma^*\|_{\max} \geq \varphi(\Sigma^*)\sqrt{\frac{\log p}{n}}\big] \leq c\exp(-\psi(n,p)) \tag{7}$$

for functions $\varphi$ and $\psi$, where $\|\cdot\|_{\max}$ denotes the elementwise $\ell_\infty$-norm. In the case of fully-observed i.i.d. data with sub-Gaussian parameter $\sigma^2$, where $\widehat{\Sigma} = \frac{1}{n}\sum_{i=1}^n x_i x_i^T - \bar{x}\bar{x}^T$ is the usual sample covariance, this bound holds with $\varphi(\Sigma^*) = \sigma^2$ and $\psi(n,p) = c'\log p$.

In addition, we require a certain *mutual incoherence* condition on the true covariance matrix $\Sigma^*$ to control the correlation of non-edge variables with edge variables in the graph. Let $\Gamma^* = \Sigma^* \otimes \Sigma^*$, where $\otimes$ denotes the Kronecker product. Then $\Gamma^*$ is a $p^2 \times p^2$ matrix indexed by vertex pairs. The incoherence condition is given by

$$\max_{e \in S^c} \|\Gamma^*_{eS}(\Gamma^*_{SS})^{-1}\|_1 \leq 1 - \alpha, \qquad \alpha \in (0,1], \tag{8}$$

where $S := \{(s,t) : \Theta^*_{st} \neq 0\}$ is the set of vertex pairs corresponding to nonzero elements of the precision matrix $\Theta^*$—equivalently, the edge set of the graph, by our theory on tree-structured discrete graphs. For more intuition on the mutual incoherence condition, see Ravikumar et al. [4].

Our global edge recovery algorithm proceeds as follows:

**Algorithm 1** (Graphical Lasso)**.**

1. *Form a suitable estimate* $\widehat{\Sigma}$ *of the true covariance matrix* $\Sigma$.
2. *Optimize the graphical Lasso program* (6) *with parameter* $\lambda_n$*, denoting the solution by* $\widehat{\Theta}$.
3. *Threshold the entries of* $\widehat{\Theta}$ *at level* $\tau_n$ *to obtain an estimate of* $\Theta^*$.

We then have the following consistency result, a straightforward consequence of the graph structure of $\Theta^*$ and concentration properties of $\widehat{\Sigma}$:

**Corollary 3.** *Suppose we have a tree-structured Ising model with degree at most* $d$*, satisfying the mutual incoherence condition* (8)*. If* $n \gtrsim d^2 \log p$*, then Algorithm 1 with* $\widehat{\Sigma}$ *the sample covariance matrix and parameters* $\lambda_n \geq \frac{c_1}{\alpha}\sqrt{\frac{\log p}{n}}$ *and* $\tau_n = c_2\{\frac{c_1}{\alpha}\sqrt{\frac{\log p}{n}} + \lambda_n\}$ *recovers all edges* $(s,t)$ *with* $|\Theta^*_{st}| > \tau_n/2$*, with probability at least* $1 - c\exp(-c' \log p)$.

Hence, if $|\Theta^*_{st}| > \tau_n/2$ for all edges $(s,t) \in E$, Corollary 3 guarantees that the log-determinant method plus thresholding recovers the full graph exactly. In the case of the standard sample covariance matrix, this method has been implemented by Banerjee et al. [18]; our analysis establishes consistency of their method for discrete trees. The scaling $n \gtrsim d^2 \log p$ is unavoidable, as shown by information-theoretic analysis [21], and also appears in other past work on Ising models [10, 9, 11]. Our analysis also has a *cautionary message*: the proof of Corollary 3 relies heavily on the population-level result in Corollary 2, which ensures that $\Theta^*$ is tree-structured. For a general graph, we have no guarantees that $\Theta^*$ will be graph-structured (e.g., see panel (b) in Figure 1), so the graphical Lasso (6) *is inconsistent in general*.

On the positive side, if we restrict ourselves to tree-structured graphs, the estimator (6) is attractive, since it relies only on an estimate $\widehat{\Sigma}$ of the population covariance $\Sigma^*$ that satisfies the deviation condition (7). In particular, when the samples $\{x_i\}_{i=1}^n$ are contaminated by noise or missing data, all we require is a sufficiently good estimate $\widehat{\Sigma}$ of $\Sigma^*$. Furthermore, the program (6) is always convex even when the estimator $\widehat{\Sigma}$ is not positive semidefinite (as will often be the case for missing/corrupted data).

As a concrete example of how we may correct the program (6) to handle corrupted data, consider the case when each entry of $x_i$ is missing independently with probability $\rho$, and the corresponding observations $z_i$ are zero-filled for missing entries. A natural estimator is

$$\widehat{\Sigma} = \left(\frac{1}{n}\sum_{i=1}^{n} z_i z_i^T\right) \div M - \frac{1}{(1-\rho)^2}\bar{z}\bar{z}^T, \tag{9}$$

where $\div$ denotes elementwise division by the matrix $M$ with diagonal entries $(1-\rho)$ and off-diagonal entries $(1-\rho)^2$, correcting for the bias in both the mean and second moment terms. The deviation condition (7) may be shown to hold w.h.p., where $\varphi(\Sigma^*)$ scales with $(1-\rho)$ (cf. Loh and Wainwright [22]). Similarly, we may derive an appropriate estimator $\widehat{\Sigma}$ and a subsequent version of Algorithm 1 in situations when the data are systematically contaminated by other forms of additive or multiplicative corruption.

Generalizing to the case of $m$-ary discrete graphical models with $m > 2$, we may easily modify the program (6) by replacing the elementwise $\ell_1$-penalty by the corresponding group $\ell_1$-penalty, where the groups are the indicator variables for a given vertex. Precise theoretical guarantees may be derived from results on the group graphical Lasso [23].

## 4 Simulations

Figure 2 depicts the results of simulations we performed to test our theoretical predictions. In all cases, we generated binary Ising models with node weights 0.1 and edge weights 0.3 (using spin $\{-1, 1\}$ variables). The five curves show the results of our graphical Lasso method applied to the dinosaur graph in Figure 1. Each curve plots the probability of success in recovering the 15

edges of the graph, as a function of the rescaled sample size $\frac{n}{\log p}$, where $p = 13$. The leftmost (red) curve corresponds to the case of fully-observed covariates ($\rho = 0$), whereas the remaining four curves correspond to increasing missing data fractions $\rho \in \{0.05, 0.1, 0.15, 0.2\}$, using the corrected estimator (9). We observe that all five runs display a transition from success probability 0 to success probability 1 in roughly the same range of the rescaled sample size, as predicted by our theory. Indeed, since the dinosaur graph has only singleton separators, Corollary 2 ensures that the inverse covariance matrix is exactly graph-structured. Note that the curves shift right as the fraction $\rho$ of missing data increases, since the problem becomes harder.

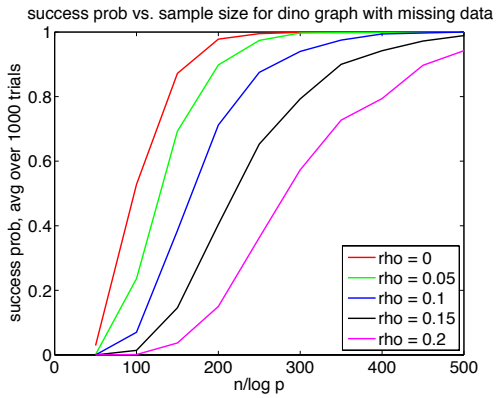



**Figure 2.** Simulation results for our graphical Lasso method on binary Ising models, allowing for missing data in the observations. The figure shows simulation results for the dinosaur graph. Each point represents an average over 1000 trials. The horizontal axis gives the rescaled sample size $\frac{n}{\log p}$.

## 5 Discussion

The correspondence between the inverse covariance matrix and graph structure of a Gauss-Markov random field is a classical fact, with many useful consequences for efficient estimation of Gaussian graphical models. It has long been an open question as to whether or not similar properties extend to a broader class of graphical models. In this paper, we have provided a partial affirmative answer to this question and developed theoretical results extending such relationships to discrete undirected graphical models.

As shown by our results, the inverse of the ordinary covariance matrix is graph-structured for special subclasses of graphs with singleton separator sets. More generally, we have shown that it is worthwhile to consider the inverses of *generalized covariance matrices*, formed by introducing indicator functions for larger subsets of variables. When these subsets are chosen to reflect the structure of an underlying junction tree, the edge structure is reflected in the inverse covariance matrix. Our population-level results have a number of statistical consequences for graphical model selection. We have shown how our results may be used to establish consistency (or inconsistency) of the standard graphical Lasso applied to discrete graphs, even when observations are systematically corrupted by mechanisms such as additive noise and missing data. As noted by an anonymous reviewer, the Chow-Liu algorithm might also potentially be modified to allow for missing or corrupted observations. However, our proposed method and further offshoots of our population-level result may be applied even in cases of non-tree graphs, which is beyond the scope of the Chow-Liu algorithm.


### Acknowledgments

PL acknowledges support from a Hertz Foundation Fellowship and an NDSEG Fellowship. MJW and PL were also partially supported by grants NSF-DMS-0907632 and AFOSR-09NL184. The authors thank the anonymous reviewers for helpful feedback.

## References


[1] M.E.J. Newman and D.J. Watts. Scaling and percolation in the small-world network model. *Phys. Rev. E*, 60(6):7332–7342, December 1999.

[2] T. Cai, W. Liu, and X. Luo. A constrained $\ell_1$ minimization approach to sparse precision matrix estimation. *Journal of the American Statistical Association*, 106:594–607, 2011.

[3] N. Meinshausen and P. Bühlmann. High-dimensional graphs and variable selection with the Lasso. *Annals of Statistics*, 34:1436–1462, 2006.

[4] P. Ravikumar, M. J. Wainwright, G. Raskutti, and B. Yu. High-dimensional covariance estimation by minimizing $\ell_1$-penalized log-determinant divergence. *Electronic Journal of Statistics*, 4:935–980, 2011.

[5] M. Yuan. High-dimensional inverse covariance matrix estimation via linear programming. *Journal of Machine Learning Research*, 99:2261–2286, August 2010.

[6] H. Liu, F. Han, M. Yuan, J.D. Lafferty, and L.A. Wasserman. High dimensional semiparametric Gaussian copula graphical models. *arXiv e-prints*, March 2012. Available at http://arxiv.org/abs/1202.2169.

[7] H. Liu, J.D. Lafferty, and L.A. Wasserman. The nonparanormal: Semiparametric estimation of high dimensional undirected graphs. *Journal of Machine Learning Research*, 10:2295–2328, 2009.

[8] C.I. Chow and C.N. Liu. Approximating discrete probability distributions with dependence trees. *IEEE Transactions on Information Theory*, 14:462–467, 1968.

[9] A. Jalali, P.D. Ravikumar, V. Vasuki, and S. Sanghavi. On learning discrete graphical models using group-sparse regularization. *Journal of Machine Learning Research - Proceedings Track*, 15:378–387, 2011.

[10] P. Ravikumar, M.J. Wainwright, and J.D. Lafferty. High-dimensional Ising model selection using $\ell_1$-regularized logistic regression. *Annals of Statistics*, 38:1287, 2010.

[11] A. Anandkumar, V.Y.F. Tan, and A.S. Willsky. High-dimensional structure learning of Ising models: Local separation criterion. *Annals of Statistics*, 40(3):1346–1375, 2012.

[12] G. Bresler, E. Mossel, and A. Sly. Reconstruction of markov random fields from samples: Some observations and algorithms. In *APPROX-RANDOM*, pages 343–356, 2008.

[13] P. Loh and M.J. Wainwright. Structure estimation for discrete graphical models: Generalized covariance matrices and their inverses. *arXiv e-prints*, November 2012.

[14] S.L. Lauritzen. *Graphical Models*. Oxford University Press, 1996.

[15] M. J. Wainwright and M. I. Jordan. Graphical models, exponential families, and variational inference. *Found. Trends Mach. Learn.*, 1(1-2):1–305, January 2008.

[16] R. T. Rockafellar. *Convex Analysis*. Princeton University Press, Princeton, 1970.

[17] R. A. Horn and C. R. Johnson. *Matrix Analysis*. Cambridge University Press, 1990.

[18] O. Banerjee, L. El Ghaoui, and A. d'Aspremont. Model selection through sparse maximum likelihood estimation for multivariate Gaussian or binary data. *Journal of Machine Learning Research*, 9:485–516, 2008.

[19] J. Friedman, T. Hastie, and R. Tibshirani. Sparse inverse covariance estimation with the graphical Lasso. *Biostatistics*, 9(3):432–441, July 2008.

[20] M. Yuan and Y. Lin. Model selection and estimation in the Gaussian graphical model. *Biometrika*, 94(1):19–35, 2007.

[21] Narayana P. Santhanam and Martin J. Wainwright. Information-theoretic limits of selecting binary graphical models in high dimensions. *IEEE Transactions on Information Theory*, 58(7):4117–4134, 2012.

[22] P. Loh and M.J. Wainwright. High-dimensional regression with noisy and missing data: Provable guarantees with non-convexity. *Annals of Statistics*, 40(3):1637–1664, 2012.

[23] L. Jacob, G. Obozinski, and J. P. Vert. Group Lasso with Overlap and Graph Lasso. In *International Conference on Machine Learning (ICML)*, pages 433–440, 2009.