An Additive Graphical Model for Discrete Data

Jun Tao, Bing Li, and Lingzhou Xue

Department of Statistics, The Pennsylvania State University, University Park, PA

ABSTRACT

We introduce a nonparametric graphical model for discrete node variables based on additive conditional independence. Additive conditional independence is a three-way statistical relation that shares similar properties with conditional independence by satisfying the semi-graphoid axioms. Based on this relation we build an additive graphical model for discrete variables that does not suffer from the restriction of a parametric model such as the Ising model. We develop an estimator of the new graphical model via the penalized estimation of the discrete version of the additive precision operator and establish the consistency of the estimator under the ultrahigh-dimensional setting. Along with these methodological developments, we also exploit the properties of discrete random variables to uncover a deeper relation between additive conditional independence and conditional independence than previously known. The new graphical model reduces to a conditional independence graphical model under certain sparsity conditions. We conduct simulation experiments and analysis of an HIV antiretroviral therapy dataset to compare the new method with existing ones. Supplementary materials for this article are available online.

1. Introduction

A graphical model uses a graph-based representation as the basis for analyzing multivariate data. It has been gaining popularity in various applied fields. The most frequently studied graphical models are those based on conditional independence (CI) between node variables, which are also called Markov random fields (MRF). Let \( X \) be a \( p \)-dimensional random vector, and \( \mathcal{E} \subseteq \{(i, j) \in V \times V : i \neq j \} \). Then \( X \) follows a Markov random field with respect to the undirected graph \( \mathcal{G} = (V, \mathcal{E}) \) if and only if \( X^i \) and \( X^j \) are independent given the rest of \( X \), that is,

\[
(i, j) \notin \mathcal{E} \iff X^i \perp \perp X^j | X^{\sim(i,j)},
\]

where \( X^{\sim(i,j)} \) represents \( X \) with its \( i \)th and \( j \)th components removed.

Discrete graphical models, where \( X^1, X^2, \ldots, X^p \) are discrete random variables, naturally appear in many fields. The simplest of such models is the binary graphical model, where the node vector \( X \) is a member of \([a_0, a_1]^p\), with \( a_0 \) and \( a_1 \) being the possible labels. One particular example is the Ising model, where \( X \in \{-1, 1\}^p \). The Ising model (Ising 1925) was originally introduced as a mathematical model of ferromagnetism in statistical mechanics. Nowadays it has been applied to such diverse fields as image restoration (Geman and Geman 1993), biophysics (Ahsan, Rudnick, and Bruinsma 1998), genetics (Fierst and Phillips 2015), and network psychometrics (Marsman et al. 2018).

The Ising model has the probability mass function (pmf) of the form:

\[
f_X(x^1, x^2, \ldots, x^p) = \frac{1}{z(\beta)} \exp \left( \sum_i \beta_i x^i + \sum_{i<j} \beta_{ij} x^i x^j \right),
\]

where \( z(\beta) = \sum_{x \in \{-1, 1\}^p} \exp \left( \sum_i \beta_i x^i + \sum_{i<j} \beta_{ij} x^i x^j \right) \) is the partition function. Let \( \beta_{ij} := \beta_{ji} \), for \( i > j \). It is easier to view the parameter as the symmetric matrix \( \beta = (\beta_{ij})_{i,j=1}^p \) in \( \mathbb{R}^{p \times p} \).

For any pair \((i, j)\) with \( i \neq j \) the parameter \( \beta_{ij} \) characterizes the conditional independence between \( X^i \) and \( X^j \) given \( X^{\sim(i,j)} \) by the equivalence

\[
X^i \perp \perp X^j | X^{\sim(i,j)} \iff \beta_{ij} = 0,
\]

as can be seen from the conditional pmf

\[
f(x^1, x^2 | X^{\sim(i,j)}) \propto \exp \left( \beta_{ij} x^i x^j + x^i (\beta_{ij} + \sum_{k \neq i,j} \beta_{ik} x^k) \right) + x^j (\beta_{ij} + \sum_{k \neq i,j} \beta_{jk} x^k).
\]

Wherever \( \beta_{ij} = 0 \), the conditional pmf is separable for \( x^i \) and \( x^j \). Thus, by (1) and (3), for the Ising model, estimating \( \mathcal{G} \) amounts to identifying the zero entries or, equivalently, the sparsity pattern of the symmetric matrix \( \beta \). There have been some related works for learning the edge set of the Ising model. For example, Höfling and Tibshirani (2009), Wang, Chao, and Hsu (2011), and Xue, Zou, and Cai (2012) developed the penalized estimation procedure based on a pseudo-likelihood, while Ravikumar, Wainwright, and Lafferty (2010) suggested fitting separate \( \ell_1 \)-penalized logistic regressions for each node to learn its neighborhood. Cheng et al. (2014) proposed a sparse covari-
ate dependent Ising model. Guo et al. (2015) and Lee et al. (2022) studied the graphical models for discrete and ordered data.

For many applications, the Ising distribution assumption can be violated, making the Ising model inapplicable. An intuitive explanation is that an Ising model has limited degrees of freedom: it can only explain $(p(p + 1))/2$ out of a total of $(2^p - 1)$ degrees of freedom in the pmf. Thus, for a large graph, the Ising models often become inadequate. In this article, we seek a new discrete graphical model without the Ising distribution assumption, which retains the fundamental simplicity of the Ising dependency structure in (3). This is achieved by replacing conditional independence with additive conditional independence (ACI), which is a new statistical relation introduced by Li, Chun, and Zhao (2014). Inspired by this idea, we propose a discrete graphical model derived from ACI, which we refer to as the discrete additive semi-graphoid model (DASG).

Our methods can be summarized as follows. Between the Hilbert spaces of the functions of the node variables, we introduce the cross-covariance operators and use them to define a discrete additive precision operator (DAPO). The DAPO is a matrix of linear operators that inherits the relation (3) at the operator level: the $(i,j)$th entry of the DAPO is the zero operator if and only if $X_i$ and $X_j$ are additively conditionally independent given $X_{-i,j}$. The nonzero entries of the DAPO then determine the edges of the graph.

Li, Chun, and Zhao (2014) and Lee, Li, and Zhao (2016) developed an estimator based on the additive precision operator (APO) to learn the edge set of the ACI graphs with continuous-valued node variables. The APO-based estimator is a hard-thresholding estimator, which asymptotically converges at a relatively slow rate. Furthermore, the tuning of the thresholding parameter is based on generalized cross-validation, which can be difficult to carry out thoroughly when $p$ is very large. There are some important differences between discrete graphs and continuous ones. For example, in the discrete case, the Hilbert spaces of node functions are finite-dimensional and the properties of the related operators—particularly those pertaining to their inverses—are a lot more pleasant. We use a penalized method to obtain a sparse matrix estimation of the DAPO, which has a faster convergence rate. Additionally, our estimator can be easily tuned with common tuning methods like cross-validations. It is worth pointing out that we also investigate the relation between ACI and CI to a deeper degree than previously understood, which fills a gap in the literature. This result is made possible by carefully studying the special structure of the discrete distribution.

The rest of the article is organized as follows. In Section 2, we introduce the DASG and its operator characterization. In Section 3, we investigate the relation between ACI and CI under the Ising model. In Section 4, we derive sample-level coordinate representations of DASG linear operators and develop the new estimator. In Section 5, we establish the asymptotic properties of the proposed estimator. In Section 6, we conduct simulation experiments to evaluate our estimator. In Section 7, we apply our estimator to an HIV antiretroviral therapy dataset. All the proofs are presented in the supplementary materials.

## 2. Discrete Additive Semi-Graphoid Model

We first review the definition of additive conditional independence and propose the discrete semi-graphoid model in Section 2.1. We then introduce the equivalent form of additive conditional independence in terms of linear operators in Section 2.2.

### 2.1. Additive Conditional Independence

Let $\\mathcal{R} \subseteq 2^V \times 2^V \times 2^V$ be a three-way relation on $V$, where $2^V$ consists of all subsets of $V$.

**Definition 1 (Pearl and Verma 1987).** A three-way relation $\\mathcal{R}$ is called a semi-graphoid if it satisfies the following conditions:

1. (symmetry) $(A, C, B) \in \mathcal{R} \Rightarrow (B, C, A) \in \mathcal{R}$;
2. (decomposition) $(A, C, B \cup D) \in \mathcal{R} \Rightarrow (A, C, B) \in \mathcal{R}$;
3. (weak union) $(A, C, B \cup D) \in \mathcal{R} \Rightarrow (A, C \cup B, D) \in \mathcal{R}$;
4. (contraction) $(A, C \cup B, D) \in \mathcal{R}$, $(A, C, B) \in \mathcal{R} \Rightarrow (A, C, B \cup D) \in \mathcal{R}$.

These axioms are extracted from the conditional independence (CI) to convey the general idea of “$B$ is irrelevant for understanding $A$ once $C$ is known,” or “$C$ separates $A$ and $B$.” The CI is a special case of the semi-graphoid three-way relation (Dawid 1979).

Li, Chun, and Zhao (2014) proposed the notion of the additive conditional independence (ACI) as a promising alternative to the CI for constructing graphical models. The ACI satisfies the axioms for a semi-graphoid (Pearl and Verma 1987), which captures the essence of a graph.

Let $X = (X^1, X^2, \ldots, X^p)^T \in \mathbb{R}^p$ be a random vector. For any node $i \in V$, let $L^2(P_X)$ denote the class of functions of $X^i$ such that $\mathbb{E}\phi(X^i) = 0$, $\mathbb{E}\phi^2(X^i) < \infty$. Let $\mathcal{A}_X \subseteq L^2(P_X)$ be a Hilbert subspace. For any nonempty node subset $A \subseteq V$, the subvector of $X$ on $A$ means $X^A := \{X^i\}_{i \in A}$. Let $\mathcal{A}_X^A$ denote the additive family

$$\bigoplus_{i \in A} \mathcal{A}_X^i = \left\{ \sum_{i \in A} \phi_i : \phi_i \in \mathcal{A}_X^i, \ i \in A \right\}$$

For two subspaces $\mathcal{A}$ and $\mathcal{B}$ of $L^2(P_X)$, let $\mathcal{A} \oplus \mathcal{B} = \mathcal{A} \cap \mathcal{B}^\perp$, where the orthogonality is in terms of the $L^2(P_X)$-inner product.

**Definition 2.** Let $X^A, X^B, X^C$ be subvectors of $X$. $X^A$ and $X^B$ are additively conditionally independent given $X^C$ with respect to $(\mathcal{A}_X^A, \mathcal{A}_X^B, \mathcal{A}_X^C)$ if and only if

$$(\mathcal{A}_X^A + \mathcal{A}_X^C) \cap \mathcal{A}_X^C \perp \perp (\mathcal{A}_X^B + \mathcal{A}_X^C) \cap \mathcal{A}_X^C,$$

where $\perp \perp$ indicates orthogonality with respect to the $L^2(P_X)$-inner product. The above relation is written as $X^A \perp_{\mathcal{A}^B} X^B | X^C$.

Suppose $X = (X^1, X^2, \ldots, X^p)^T \in \{0, 1, \ldots, m\}^p$, $m \geq 1$. The discrete additive semi-graphoid model (DASG) is defined through the following equivalence

$$X^i \perp_{\mathcal{A}^B} X^j | X^{-i,j} \iff (i, j) \notin E.$$
2.2. Discrete Additive Precision Operator

We now introduce a linear operator that characterizes the ACI.

Definition 3 (Baker 1973; Fukumizu, Bach, and Jordan 2009). The cross-covariance operator of \((X_i, X_j)\), denoted by \(\Sigma_{X_iX_j}\), is a mapping from \(\mathcal{A}_{X_i}\) to \(\mathcal{A}_{X_j}\), such that for any \(\phi \in \mathcal{A}_{X_i}\), \(\psi \in \mathcal{A}_{X_j}\)

\[
\langle \phi, \Sigma_{X_iX_j} \psi \rangle_{\mathcal{A}_{X_i}} = \text{cov}[\phi(X_i), \psi(X_j)].
\]  

(4)

By the Cauchy-Schwarz inequality and \(X\) being finitely discrete, the bilinear form \(\mathcal{A}_{X_i} \times \mathcal{A}_{X_j} \rightarrow \mathbb{R}: (\phi, \psi) \mapsto \text{cov}[\phi(X_i), \psi(X_j)]\) is bounded. The existence and uniqueness of \(\Sigma_{X_iX_j}\) defined through (4) is guaranteed by Riesz’s representation theorem.

Let ker\((\Sigma_{X_iX_j}) = \{ h \in \mathcal{A}_{X_j} : \Sigma_{X_iX_j} h = 0 \}\) be the kernel space of \(\Sigma_{X_iX_j}\). Note that \(\phi \in \text{ker}(\Sigma_{X_iX_j})\) if and only if \(\text{var}[\phi(X_i)] = 0\). This implies that ker\((\Sigma_{X_iX_j})\) is a linear subspace of constant functions. Since \(\mathcal{A}_{X_i}\) only contains mean zero functions, we have ker\((\Sigma_{X_iX_j}) = \{0\}\), which implies that the operator \(\Sigma_{X_iX_j}\) is invertible.

The following definitions contain two key matrices of operators in our theory.

Definition 4. The operator \(\Sigma_{XX} := \{ \Sigma_{X_iX_j} \}_{i,j=1}^p\) that satisfies \(\Sigma_{XX} \phi = \sum_{i,j=1}^p \Sigma_{X_iX_j} \phi_i\) for any \(\phi = \phi_1 + \cdots + \phi_p \in \mathcal{A}_{X}\) is called the discrete additive precision operator (DAPO).

By this definition, for any \(\phi = \phi_1 + \cdots + \phi_p\), \(\psi = \psi_1 + \cdots + \psi_p \in \mathcal{A}_{X}\),

\[
\langle \phi, \Sigma_{XX} \psi \rangle_{\mathcal{A}_{X}} = \langle \phi, \Sigma_{X_iX_j} \psi \rangle_{\mathcal{A}_{X_i}} = \sum_{i,j=1}^p \langle \phi_i, \Sigma_{X_iX_j} \psi_j \rangle_{\mathcal{A}_{X_i}}.
\]

By the definition of cross-covariance operator, this is \(\sum_{i,j=1}^p \text{cov}[\phi_i(X_i), \psi_j(X_j)]\), which sums up to \(\text{cov}[\phi(X), \psi(X)]\). The variance matrix, or the covariance matrix, can be regarded as the Gram matrix of the set of random variables \((X_1, \ldots, X_p)\) with respect to the \(L^2(\mathcal{P}_X)\)-inner product. Similarly, the DAVO serves as the Gram matrix of operators with respect to the inner product on the additive function space \(\mathcal{A}_{X}\).

Assumption 1. ker\((\Sigma_{XX}) = \{0\}\).

This condition is satisfied by any nondegenerate node variables, where any nonzero function of the node variables will be linearly independent. Under this assumption, the mapping \(\phi \mapsto \Sigma_{XX} \phi\) is invertible.

Definition 5. The operator \(\Sigma_{XX}^{-1}\) is called the discrete additive precision operator (DAPO), and is written as \(\Theta_{XX}\).

By being the inverse operator of \(\Sigma_{XX}\), \(\Theta_{XX}\) satisfies \(\phi = \Theta_{XX} \Sigma_{XX} \phi = \Sigma_{XX} \Theta_{XX} \phi\) for any \(\phi \in \mathcal{A}_{X}\). We use \(\Theta_{XX}\) to denote the \((i,j)\)th block of \(\Theta_{XX}\), which is a mapping from \(\mathcal{A}_{X_i}\) to \(\mathcal{A}_{X_j}\), such that \(\Theta_{XX} \phi = \sum_{i,j=1}^p \Theta_{XX} \phi_i\) for \(\phi = \phi_1 + \cdots + \phi_p\).

The next theorem re-expresses pairwise independence and ACI in terms of the DAVO and the DAPO.

Theorem 1. The DAVO and the DAPO have the following properties:

1. \(X^i \perp \perp X^j\) if and only if \(\Sigma_{X_iX_j} = 0\).
2. \(X^i \perp_{ACI} X^j (X^{-\{i,j\}})\) if and only if \(\Theta_{X_iX_j} = 0\).

The first statement of Theorem 1 gives a probabilistic interpretation of the DAVO, which can be shown by verifying the definition of independence. The relation between DAVO and the DASG is analogous to that between the covariance matrix and the Gaussian graphical model, where the covariance matrix determines pairwise independence, whereas its inverse determines CI. In our context, DAVO still determines pairwise independence, but its inverse, DAPO, determines the new statistical relationship, ACI.

3. Relation between ACI and CI under the Ising Model

Li, Chun, and Zhao (2014) demonstrated some relations between ACI and CI under the copula Gaussian model assumption. Loh and Wainwright (2012) studied the relationship between the zero pattern of inverse covariance matrices and the edge set for Ising models. In this section, we further investigate the relation between ACI and CI under the Ising model, whose distributional simplicity allows us to uncover a deeper relation than previously understood.

Definition 6. Consider a graph \(G = (V, E)\). For any node pair \((i, j) \in V \times V\) with \(i \neq j\), a subset \(R \subseteq V \setminus \{i, j\}\) is an \((i, j)\)-separator if the removal of the edge \((i, j)\) and the edges between \(R\) and \(V \setminus R\) from the graph \(G\) separates \(i\) and \(j\) into distinct connected components.

For any node pair \((i, j)\), the set \(V \setminus \{i, j\}\) is naturally an \((i, j)\)-separator, so the existence of a node separator is guaranteed. Given a node separator \(R\), let \(G'\) be the graph with the edge \((i, j)\) and the edges between \(R\) and \(V \setminus R\) removed. Consider the maximal connected subgraph of \(G'\) including \(i\), whose node set is denoted by \(C_i\). Similarly, let \(C_j\) denote the node set of the maximal connected subgraph of \(G'\) including \(j\). Then \(C_i\) and \(C_j\) are two mutually exclusive and disconnected components of the graph \(G'\).

Definition 7. Let \(X = (X^1, X^2, \ldots, X^p)^\top \in \mathbb{R}^p\) be a random node vector. For node \(i \in V\) and subset \(D \subseteq V \setminus \{i\}\), \(\mathcal{A}_{X_i}^D\) is said to have linear conditional mean with respect to \(\mathcal{A}_{X}^D\), if either \(D = \emptyset\) or for any \(\phi \in \mathcal{A}_{X_i}^D\),

\[
E(\phi(X^i)|X^D) \in \mathcal{A}_{X_i}^D.
\]  

(5)

When \(D = \emptyset\), \(X^D = \{X^i\}_{i \in D}\) is an empty subvector and \(\mathcal{A}_{X_i}^{\emptyset} = \{0\}\). Then the meaning of linear conditional mean is that \(E(\phi(X^i)) = 0\) by the definition of \(\mathcal{A}_{X_i}^D\). Let \(|D|\) denote its cardinality. When \(X\) is binary, (5) means there exists a constant vector \(\xi \in \mathbb{R}^{|D|}\) such that

\[
E(X^i - EX^i|X^D) = \xi^\top (X^D - EX^D).
\]

In this case, for simplicity, the binary node variable \(X^i\) is said to have a linear conditional mean with respect to \(X^D\), as \(\mathcal{A}_{X_i}\) is one dimensional. The next theorem gives a relation between ACI and CI under a binary graphical model with a linear conditional mean.
Theorem 2. Suppose $X = (X^1, X^2, \ldots, X^P)^\top \in \{a_0, a_1\}^P$, $p \geq 3$, follows a MRF with respect to $(V, E)$. For any node pair $(i, j)$ with $i \neq j$, if there exists an $(i, j)$-separator $R$ on $(V, E)$ such that $X^R$ has linear conditional mean with respect to $X^k$ for any $k \in C_i$, then $X^i \perp \!\!\!\!\perp X^j|X^{-ij}(i, j)$. The equivalence between CI and CI can be obtained.

Corollary 1. Suppose $X = (X^1, X^2, \ldots, X^P)^\top$, $p \geq 3$, follows a symmetric Ising model (6) with respect to $(V, E)$. For any node pair $(i, j)$ with $i \neq j$, if there exists an $(i, j)$-separator $R$ on $(V, E)$ with $|R| \leq 2$, then $X^i \perp \!\!\!\!\perp X^j|X^{-ij}(i, j)$. The equivalence between ACI and CI can be established if $R$ is chosen as the largest separator $V \setminus \{i, j\}$ under the linear conditional mean assumption for this separator.

Theorem 3. Suppose $X = (X^1, X^2, \ldots, X^P)^\top$, $p \geq 3$, follows a symmetric Ising model (6) with respect to $(V, E)$. For any node pair $(i, j)$ with $i \neq j$, if there exists an $(i, j)$-separator $R$ on $(V, E)$ with $|R| \leq 2$, then $X^i \perp \!\!\!\!\perp X^j|X^{-ij}(i, j)$. The linear conditional mean can be easily violated by Ising models. To have a closer look at the reason, let $N_i := \{k \in V \setminus \{i\} : (i, k) \in E\}$ be the neighborhood of node $i \in V$ on the given graph $(V, E)$. Then $N_i \setminus \{j\}$ and $N_j \setminus \{i\}$ are $(i, j)$-separators as well, which can have a much smaller size than $V \setminus \{i, j\}$.

Corollary 2. Suppose $X = (X^1, X^2, \ldots, X^P)^\top$, $p \geq 3$, follows a symmetric Ising model (6) with respect to $(V, E)$. For any node pair $(i, j)$ with $i \neq j$, if $N_i$ is the neighborhood of node $i \in V$ on $|N_i \setminus \{j\}| \leq 2$, then $X^i \perp \!\!\!\!\perp X^j|X^{-ij}(i, j)$. Note that all these results are pairwise statements with respect to a pair of nodes $(i, j)$ rather than graph-level statements. That is, for an interested $(i, j)$, as long as $|N_i \setminus \{j\}| \leq 2$ or $|N_j \setminus \{i\}| \leq 2$, the equivalence between CI and ACI will hold regardless of the other nodes. The pairwise statements are turned into graph-level statements if the node degree $|N_i| \leq 2$ for all $i \in V$, in which case the entire MRF will be the same as DASG. Such graphs may consist of some loops and threads. Figure 1 shows some special cases where such graph-level equivalence holds.

For example, consider the left most graph shown in Figure 1, where the set of nodes is $V = \{1, 2, 3, 4\}$ and the edge set is $E = \{(1, 2), (2, 3), (3, 4), (1, 4), (2, 4)\}$. If the joint distribution is the symmetric Ising distribution specified by the first formula in (7), then the DAPO has blockwise norms given by the second formula in (7). Thus, the graphs determined by ACI and CI are equivalent.

$$\beta = \frac{\log(2)}{2} \times \begin{pmatrix} 0 & 1 & 0 & 1 \\ 0 & 1 & 1 & 0 \\ 1 & 1 & 0 & 0 \end{pmatrix}.$$ 

The condition $|N_i \setminus \{j\}| \leq 2$ or $|R| \leq 2$ requires a rather sparse edge set. The equivalence will not be guaranteed if they are violated. A simple counter-example can be constructed when the graph has five nodes and is fully connected except for one edge, as shown in Figure 2.

In this case, the blockwise norm of $\Theta_{XX}$ is given by

$$\{\|\Theta_{XX}\|_{\text{HS}}\}_{i,j=1}^4 = \begin{pmatrix} 11 & 33 & 33 & 33 \\ 33 & 11 & 33 & 33 \\ 33 & 33 & 11 & 33 \\ 33 & 33 & 33 & 11 \end{pmatrix}.$$ 

The Hilbert-Schmidt norm of $\Theta_{XX}$ is $\|\Theta_{XX}\|_{\text{HS}} = 27\sqrt{15}$, which is strictly greater than 0. Unlike the conditional independence graph, the node variables here are all additive conditionally dependent on each other. For the conditional means,

$$\mathbb{E}[X^i|X^{[3,4,5]}] = \mathbb{E}[X^3|X^{[3,4,5]}]$$

$$= 1 - \frac{2}{1 + \exp\{\log(2)(X^3 + X^4 + X^5)\}},$$

neither of which is an element of $\Theta_{XX}^{[3,4,5]}$.

Next, for the augmented graph with node vector $Y = (X^1, X^2, X^3, X^4, X^5)^\top$, it is easy to see that the MRF is
Figure 2. The fully connected symmetric Ising model except for the edge \((1, 2)\). It has the unique \((1, 2)\)-separator \(R = N_1 \setminus \{2\} = N_2 \setminus \{1\} = \{3, 4, 5\}\). In this case, \(X_1 \perp \perp_A X_2 | X_3, X_4, X_5\) is not true. The DASG yields a fully connected graph.

the complete graph with the edge \((1, 2)\) removed. The blockwise norms of \(\Theta_{YY}\) is given by

\[
\left\|\Theta_{YY}\right\|_{HS}^2 = \begin{pmatrix} 27 & 0 & 15 & 15 & 3 & 28 \\ 6 & 15 & 14 & 12 & 20 & 21 \\ 0 & 7 & 15 & 19 & 21 & 26 \\ 27 & 28 & 15 & 18 & 21 & 28 \\ 28 & 15 & 26 & 25 & 21 & 28 \\ 28 & 28 & 28 & 28 & 28 & 28 \end{pmatrix}.
\]

Thus, the DASG agrees with the MRF again. The key difference between the original graph and the augmented graph lies in whether the linear conditional mean assumption (5) is satisfied. On the other hand, \(Y_{[3,4,5,6]} = (X_1, X_3, X_4, X_5)\) is actually a complete basis of odd function space of \(Y_{[3,4,5]} = (X_1, X_3, X_5)\), which is slightly larger than \(\mathcal{A}_{Y_{[3,4,5]}}\). The conditional expectation can then be expressed by

\[
\mathbb{E}[Y_1 | Y_{[3,4,5,6]}] = \frac{5}{18} (Y_3 + Y_4 + Y_5) - \frac{1}{18} Y_6.
\]

That is to say, \(Y_1\) has a linear conditional mean with respect to \(Y_{[3,4,5,6]}\).

Loh and Wainwright (2012) studied the relationship between CI and the inverse of a generalized covariance matrix by augmenting a graph with interaction terms. We exploit this idea to get a more complete picture of the relation between ACI and CI. Let \(D \subseteq V\) be a subset of nodes with \(|D| \geq 3\). Let

\[
\mathcal{F}_D(X) := \left\{ \prod_{i \in D} u^{n_i}(X_i) : n_i \in \{0, 1\} \text{ and } \sum_{i \in D} n_i > 1, \text{ and } \sum_{i \in D} n_i \equiv 1 \pmod{2} \right\},
\]

where \(u^{(0)}(x) = 1\) and \(u^{(1)}(x) = x\). Note that these terms are still binary. For example, if \(D = \{3, 4, 5, 6\}\), \(\mathcal{F}_D(X) = \{X_1 X_3 X_4 X_5, X_1 X_3 X_4 X_6, X_1 X_3 X_5 X_6, X_1 X_4 X_5, X_1 X_4 X_6, X_1 X_5 X_6\}\); if \(D = \{3, 4, 5, 6, 7\}\), the fifth order interaction \(X_1 X_3 X_4 X_5 X_6 X_7\) is also contained in \(\mathcal{F}_D(X)\).

Now, we may summarize the relation between ACI and CI in the next theorem.

**Theorem 4.** Suppose \(X = (X_1, X_2, \ldots, X_p)^T, p \geq 5\), follows a symmetric Ising model (6) with respect to \([V, E]\). For any pair of nodes \((i, j)\) with \(i \neq j\), let \(R\) be a nonempty \((i, j)\)-separator on \([V, E]\). Then \(X_i \perp \perp_A X_j | X_{(i,j)} \Rightarrow X_i \perp \perp_A X_j | X_{(i,j)} \cup \mathcal{F}_R(X)\). Furthermore, if \(R = N_i \setminus \{j\}\), \(X_i \perp \perp_A X_j | X_{(i,j)} \Rightarrow X_i \perp \perp_A X_j | X_{(i,j)} \cup \mathcal{F}_R(X)\).

This theorem provides some intuition of how much ACI and CI differ. Note that Theorem 4 focuses on a specific pair of nodes \((i, j)\) and the construction of \(\mathcal{F}_R(X)\) depends on \((i, j)\). To obtain the graph-level equivalence, we should augment the graph with nodes of interactions based on each pair \((i, j)\).

**Corollary 3.** Suppose \(X = (X_1, X_2, \ldots, X_p)^T, p \geq 5\), follows a symmetric Ising model (6) with respect to \([V, E]\). Let \(\mathcal{G}(X)\) be a vector function of \(X\) such that for any pair of nodes \((i, j)\) with \(i \neq j\), there exists a nonempty \((i, j)\)-separator \(R_{ij}\) on \([V, E]\) satisfying \(\mathcal{F}_{R_{ij}}(X) \subseteq \mathcal{G}(X)\). Then for any \(1 \leq i, j \leq p\) with \(i \neq j\), \(X_i \perp \perp_A X_j | X_{(i,j)} \Rightarrow X_i \perp \perp_A X_j | X_{(i,j)} \cup \mathcal{G}(X)\). Furthermore, if \(\mathcal{F}_{N_i \setminus \{j\}}(X) \subseteq \mathcal{G}(X), X_i \perp \perp_A X_j | X_{(i,j)} \Rightarrow X_i \perp \perp_A X_j | X_{(i,j)} \cup \mathcal{G}(X)\).

All the discussions so far are about pairwise Markov property. As shown in Li, Chun, and Zhao (2014), the pairwise Markov property in terms of ACI is equivalent to the global Markov property in terms of CI. On the other hand, it is well known that if the pmf of \(X\) is positive for all values of \(X\), then pairwise Markov property is equivalent to global Markov property (Lauritzen 1996). Hence, under the pairwise equivalence between CI and ACI, the graph derived from ACI satisfies the global Markov property in terms of CI. We summarize this fact as the following corollary.

**Corollary 4.** Suppose

1. the pmf of \(X = (X_1, X_2, \ldots, X_p)^T\) is positive for all values of \(X\);
2. CI and ACI are equivalent for all distinct pairs, that is, \(X_i \perp \perp_A X_j | X_{(i,j)} \Rightarrow X_i \perp \perp_A X_j | X_{(i,j)} \) for all \(i \neq j\);
3. the edge set \(E\) is determined by pairwise ACI, that is, \((i, j) \notin E \Rightarrow X_i \perp \perp_A X_j | X_{(i,j)} \) for all \(i \neq j\).

Then \(X\) satisfies the global Markov property in terms of CI with respect to \([V, E]\); that is, for any subsets \(A, B, C\) of \(V\) such that \(C\) separates \(A\) and \(B\), we have \(X_A \perp \perp X_B | X_C\).

### 4. Penalized Estimation

Given that \(X\) follows a DASG with respect to \(\mathcal{G} = (V, E)\), we need to estimate the edge set \(E\) from a sample of \(X\). Since the relationship defined in Definition 2 is difficult to check directly, we look for a DAPO-based estimator. We derive the matrix form of this operator in Section 4.1 and introduce a group penalized D-trace estimator in Section 4.2.

#### 4.1. Coordinate Representation

The discussion so far is in operator form. In this section, we represent the operators as matrices using coordinate representation. Two versions of such coordinate representations are derived, which are more direct than those given in Li, Chun, and Zhao (2014) thanks to the simplicity offered by the discrete \(X\). We first give the matrix representations at the population level, which lead naturally to a sample estimate.
For each node variable $X^i$, consider its function space $\mathcal{A}_{X^i} = L^2(P_{X^i})$, the centered $L^2$ class. That is, for any $\phi \in \mathcal{A}_{X^i}$, $\mathbb{E}\phi(X^i) = 0$ and $\mathbb{E}\phi(X^i)^2 < \infty$. A natural way to represent such a function is by simply listing its values on the support, that is, $(\phi(0), \ldots, \phi(m))^\top$. Thus, $\mathcal{A}_{X^i}$ has a finite dimension $\dim(\mathcal{A}_{X^i}) \leq m + 1$. Since the node variable $X^i$ takes its values in a finite set, $\phi(X^i)$ is a bounded random variable, whose second moment always exists. Therefore, the only restriction on $\phi$ is that its mean is zero and this renders the dimension of $\mathcal{A}_{X^i}$ as $m$. We select an orthonormal basis $\{u_i^a(\cdot), \ldots, u_i^{(m)}(\cdot)\}$ of $\mathcal{A}_{X^i}$; $(u_i^a(\cdot), u_i^b(\cdot))_{\mathcal{A}_{X^i}} = 0$, for any $a, b = 1, \ldots, m$ and $a \neq b$. Let $U_i(\cdot) = (u_i^1(\cdot), \ldots, u_i^{(m)}(\cdot))^\top$. Then any function $\phi$ in $\mathcal{A}_{X^i}$ can be represented by a vector $c_\phi \in \mathbb{R}^m$ as $\phi(\cdot) = U_i(\cdot)^\top c_\phi$. With such a representation of functions, the $L^2(P_{X^i})$-inner product is just Euclidean inner product $(\phi, \psi)_{\mathcal{A}_{X^i}} = c_\phi^\top c_\psi$. The function basis $U_i(\cdot), i \in V$, will give rise to the orthonormal representation of the DASG operator.

**Definition 8 (Orthonormal Representation).** 1. With $U_i(\cdot), i \in V$ defined above, the matrix $[\Sigma_{XX}]_0 := \text{cov}[U_i(X^i), U_j(X^j)] \in \mathbb{R}^{m \times m}$ is called an orthonormal representation of $\Sigma_{XX}$.

2. The matrix $[\Sigma_{XX}]_0 := \{[\Sigma_{XX}]_0\}_{i,j=1}^{mp \times mp}$ is called an orthonormal representation of the DAVO.

3. The matrix $[\Theta_{XX}]_0 := \{[\Theta_{XX}]_0\}_{i,j=1}^{mp \times mp}$ is called an orthonormal representation of the DAVO. Let $[\Theta_{XX}]_0 \in \mathbb{R}^{m \times m}$ denote the $(i,j)$th block of $[\Theta_{XX}]_0$.

The orthonormal representation is a natural way to represent the operators. Note that "orthonormal" refers to the function basis. Such a representation is not unique since it depends on the choice of $U_i(\cdot)$. If we are only interested in the signal (zero or nonzero) of a block in the operators, there is a more convenient representation, as defined below.

**Definition 9 (Vertex Representation).** 1. Let $V(\cdot) = (I_{i(1)}(\cdot), \ldots, I_{i(m)}(\cdot))^\top$ be a vector-valued function on $\{0,1,\ldots,m\}$, whose $\ell$th coordinate $I_{i(\ell)}(\cdot)$ is the indicator function on $\{0,1,\ldots,m\}$. We call the matrix $[\Sigma_{XX}]_v := \text{cov}[V(X^i), V(X^j)] \in \mathbb{R}^{m \times m}$ as the vertex representation of $\Sigma_{XX}$.

2. The matrix $[\Sigma_{XX}]_v := \{[\Sigma_{XX}]_v\}_{i,j=1}^{mp \times mp}$ is called the vertex representation of the DAVO.

3. The matrix $[\Theta_{XX}]_v := \{[\Theta_{XX}]_v\}_{i,j=1}^{mp \times mp}$ is called the vertex representation of the DAVO. Let $[\Theta_{XX}]_v \in \mathbb{R}^{m \times m}$ denote the $(i,j)$th block of $[\Theta_{XX}]_v$.

The word "vertex" in the name comes from the fact that $V$ maps $\{0,1,\ldots,m\}$ into $\{0,1\}^m$, which is the set of all vertices of the standard $m$-dimensional cube. Replacing $U_i(\cdot)$s with indicator function vector $V(\cdot)$ is computationally efficient. One may notice that $V(\cdot)$ itself is not a mean zero function basis. But it still works since the constant functions vanish in the cross-covariance operator. The actual input function vector for node $i$ is $V(\cdot) - \mathbb{E}V(X^i)$. Even though the constant functions are excluded when defining the operators, they can still be borrowed back for computational purposes.

**Theorem 5.** The operators $\Sigma_{X^iX^j}$ and $\Theta_{X^iX^j}$ and their coordinate representations satisfy:

1. $\Sigma_{X^iX^j} = 0 \leftrightarrow [\Sigma_{X^iX^j}]_0 = 0 \leftrightarrow [\Sigma_{X^iX^j}]_v = 0$.

2. $\Theta_{X^iX^j} = 0 \leftrightarrow [\Theta_{X^iX^j}]_0 = 0 \leftrightarrow [\Theta_{X^iX^j}]_v = 0$.

Within each statement, the first 0 is the zero operator that maps any function in $\mathcal{A}_{X^i}$ to the zero function in $\mathcal{A}_{X^j}$. The second and the third 0s are $m \times m$ zero matrices. We give two examples of coordinate representations and their roles in constructing the DASG.

**Example 1.** Suppose that $X = (X_1^1, X_2^2, X_3^3)^\top \in \{0,1\}^3$ is a random vector with pmf $f(x) = 3x^2/12$. The orthonormal and vertex representations of the DAVO and the DAPO are shown below:

$[\Sigma_{XX}]_0 = \begin{pmatrix} 1 & \frac{1}{4} & 0 \\ \frac{1}{4} & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$

$[\Theta_{XX}]_0 = \begin{pmatrix} 1 & \frac{1}{15} & -\frac{4}{15} & 0 \\ \frac{15}{15} & 1 & 0 & 0 \\ -4/15 & 0 & 0 & 1 \\ 0 & 0 & 1/4 \end{pmatrix}$

$[\Sigma_{XX}]_v = \begin{pmatrix} \frac{24}{5} & -6 & 0 \\ -6 & \frac{24}{5} & 0 \\ 0 & 0 & 4 \end{pmatrix}$

$[\Theta_{XX}]_v = \begin{pmatrix} 8.2 & -16.1 & 23.4 \\ -10.5 & 0 \end{pmatrix}$

The matrix $[\Sigma_{XX}]_0$ has 1’s as its diagonal entries since the orthonormal basis $u_i$ actually normalizes the node variable. As a result, for a binary DASG, the edge set is determined by $[\Theta_{XX}]_0 = (\text{corr}(X))^\top$, the inverse of the correlation matrix. In comparison, it is easy to verify that the vertex representation $[\Sigma_{XX}]_v$ is the covariance matrix $\operatorname{var}(X)$ and $[\Theta_{XX}]_v = (\text{var}(X))^{-1}$. Clearly, both representations lead to the same graph.

**Example 2.** Suppose that $X = (X_1^1, X_2^2, X_3^3)^\top \in \{0,1,2\}^3$ is a random vector with pmf $f(x) = 2^x(x^2+x^3)/499$. Take $V(\cdot) = (I_{1(1)}(\cdot), I_{2(2)}(\cdot))^\top$. For instance, the vertex representation of the cross-covariance operator $\Sigma_{X^1X^2}$ is $[\Sigma_{X^1X^2}]_v = \text{cov}[V(X^1), V(X^2)] = 10^{-3} \times \begin{pmatrix} 8.2 & -16.1 & 23.4 \\ -10.5 & 0 \end{pmatrix}$.

The vertex representation of the DAPO is $[\Theta_{XX}]_v = \begin{pmatrix} \Sigma_{X^1X^1} & [\Sigma_{X^1X^2}]_v & [\Sigma_{X^1X^3}]_v \\ [\Sigma_{X^2X^1}]_v & \Sigma_{X^2X^2} & [\Sigma_{X^2X^3}]_v \\ [\Sigma_{X^3X^1}]_v & [\Sigma_{X^3X^2}]_v & \Sigma_{X^3X^3} \end{pmatrix}^{-1}$
Note that the entire (2, 3)th block of $[\Theta_{XX}]_v$ is zero, which means additive conditional independence holds between $X^2$ and $X^3$ given $X^1$.

We have assumed each node $X^i$ has the same support $\{0, \ldots , m\}$ for simplicity. But this is not necessary for constructing the DASG. Suppose for $X = (X^1, X^2, \ldots , X^p)\top$, the node variable $X^i$ has the support $\{0, 1, \ldots , m\}$ for $i \in V$. Let $V_i(\cdot ) = (\mathbb{1}_{[1]}(\cdot ), \ldots , \mathbb{1}_{[m]}(\cdot ))\top$. The vertex representation of the DAVO is $\{\Sigma_{XX}\}_v := \{[\Sigma_{X_iX_j}]_v\}_{i,j=1}^p$ with its $(i,j)$th block being $[\Sigma_{X_iX_j}]_v := \text{cov}(V_i(X^i), V_j(X^j)) \in \mathbb{R}^{m \times m}$. Hence, the off-diagonal sub-blocks may not be square matrices, as in the case of a common support $\{0, 1, \ldots , m\}$. $\Sigma_{X_iX_j}$, its Frobenius norm $\|\|\Sigma_{X_iX_j}\|_F$ coincides with the Hilbert-Schmidt norm $\|\|\Sigma_{X_iX_j}\|_{HS}$.

### 4.2. Group Penalized D-Trace Estimator

Suppose $X$ follows a DASG with respect to the graph $G = \{V, E\}$. We now develop an estimator of $E$ based on an iid sample of $X$ of size $n$. In Li, Chun, and Zhao (2014) and Lee, Li, and Zhao (2016), $E$ is estimated by thresholding the small entries of the estimated APO. Here, we propose a penalized sparse procedure to estimate it.

In the Gaussian (copula) graphical model setting, a well-known method for estimating a positive definite sparse precision matrix is via the penalized Gaussian likelihood estimator or its rank-based alternative (Yuan and Lin 2007; Friedman, Hastie, and Tibshirani 2008; Fan, Feng, and Wu 2009; Xue and Zou 2012). This is so because, from the estimation point of view, where estimating the edge set $E$ reduces to a sparse estimation of $[\Theta_{XX}]_v$, with its vanishing blocks determining the absence of edges. Instead of the negative Gaussian log-likelihood, we choose the loss function as the difference between two traces of operators (D-trace), $L_D(\Theta, \Sigma) = \frac{1}{2}\text{tr}(\Theta^2, \Sigma) - \text{tr}(\Theta)$, proposed by Zhang and Zou (2014), where $\langle \cdot, \cdot \rangle$ is the Frobenius inner product. This loss function is a smooth and convex function of $\Theta$ with a unique minimizer $\Sigma^{-1}$. The convexity ensures computational efficiency.

We first need an estimate of the coordinate representation of the DAVO. Since $\{\Sigma_{XX}\}_v$ is the covariance matrix of $(V(X^1), \ldots , V(X^p))\top$, its sample version is a natural choice. Let $\{\hat{\Sigma}_{XX}\}_v := \{[\hat{\Sigma}_{X_iX_j}]_v\}_{i,j=1}^p$, whose $(i,j)$th block is $[\hat{\Sigma}_{X_iX_j}]_v := \text{cov}_n(V(X^i), V(X^j))$, the sample covariance between $V(X^i)$ and $V(X^j)$, with $[\hat{\Sigma}_{XX}]_v$ thus constructed, we propose to estimate the DAVO by

$$\hat{\Theta}_{XX} := \text{arg min}_{\Theta = \Theta^\top} \left\{ L_D(\Theta, [\hat{\Sigma}_{XX}]_v) + \lambda_n P(\Theta) \right\},$$

where $\lambda_n > 0$ is the regularization parameter, and $P(\Theta)$ is a penalty function, which is chosen to be $\sum_{1 \leq i,j,p,i \neq j} \|\|\Theta_{ijp}\|_F$ with $[i,j]$ representing the index set corresponding to the $(i,j)$th block of a $p \times p$ block matrix. That is, $[i,j] = \{(a, b) \in \mathbb{N} \times \mathbb{N} : (a-1)m < a \leq im, (j-1)m < b \leq jm\}$, where $\mathbb{N} = \{1, 2, \ldots\}$ stands for the set of natural numbers. With the norm in the penalty being chosen as the Frobenius norm, $P(\Theta)$ is actually a group-Lasso penalty (Yuan and Lin 2006). By the definition of the orthonormal representation

$$\Theta_{XX} = D_A\Sigma_A D_A^\top,$$

where $\Sigma_A$ is the eigenvalue decomposition with ordered eigenvalues $\sigma_1 \geq \cdots \geq \sigma_{mp}$, then $H(A, B)$ can be written down explicitly as

$$H(A, B) := D_A\{D_A^\top B D_A\} \circ C D_A^\top,$$

where $\circ$ denotes the Hadamard product and $C$ is the matrix $\{\frac{1}{\sigma_i + \sigma_j}\}_{i,j=1}^{mp}$. For (10), we have the stepwise solution

$$\Theta_0^{(t+1)} = \text{arg min}_{\Theta \in \Theta^\top} \left\{ \frac{1}{2} \Theta^2, A_F - \langle \Theta, B \rangle_F \right\}.$$
where $S(A, \lambda)$ is the function

$$S(A, \lambda) := \arg \min_{\Theta_0} \left\{ \frac{1}{2} \langle \Theta_0, I_{mp} \rangle_F - \langle \Theta_0, A \rangle_F + \lambda P(\Theta_0) \right\}.$$ 

Given a symmetric matrix $A$, the $(i,j)$th block of $S(A, \lambda)$ can be obtained by soft-thresholding:

$$S(A, \lambda)_{ij} = \begin{cases} A_{ij} & \text{if } i = j, \\ (1 - \frac{\lambda}{\|A_{ij}\|_F}) A_{ij} & \text{if } i \neq j, \|A_{ij}\|_F > \lambda, \\ 0 & \text{if } i \neq j, \|A_{ij}\|_F \leq \lambda. \end{cases}$$

We summarize the procedure developed above as the following algorithm. For any matrix $M$ in $\mathbb{R}^{mp \times mp}$, let $\text{diag}(M)$ denote its diagonal matrix, that is, $\text{diag}(M) = M \circ I_{mp}$.

**Algorithm 1** D-trace group-Lasso

Input: Sample DAVO $\hat{\Sigma}_{XX}^V$, block size $m$ and regularization parameter $\lambda_0$.

Initialize: Set $t = 0$, $\theta^{(0)} = \Theta_0^{(0)} = [\text{diag}(\hat{\Sigma}_{XX}^V)]^{-1}$ and $\Lambda^{(0)} = 0$.

while not converge do

$$\theta^{(t+1)} = H(\hat{\Sigma}_{XX}^V + \eta I_{mp}, I_{mp} + \eta \Theta_0^{(t)} - \Lambda^{(t)}).$$

$$\Theta^{(t+1)} = S(\theta^{(t+1)} + \Lambda^{(t)} / \eta, \lambda_0 / \eta).$$

$$\Lambda^{(t+1)} = \Lambda^{(t)} + \eta(\theta^{(t+1)} - \Theta_0^{(t)}).$$

$t = t + 1$.

end while

Return: $\Theta^{(t)}$.

The above algorithm can be easily extended to the case where the node variables $X^1, \ldots, X^p$ have different (but finite) supports $\{0, 1, \ldots, m_i\}, i \in V$, by allowing the blocks of $[\Theta_{XX}]_0$ to be nonsquared matrices of different sizes. We can allow the penalty in (8) to be adaptive to different block sizes, similar to the group Lasso (Yuan and Lin 2006) in linear regression. That is,

$$P(\Theta) = \sum_{i \neq j} w_{ij} \|\Theta_{ij}\|_F,$$

where $w_{ij}$ accounts for the varying block sizes, such as $w_{ij} = \sqrt{m_im_j}$.

**5. Asymptotic Properties**

In this section we derive the convergence rate of the penalized D-trace estimator (8). In the ultrahigh-dimensional setting, where $\log(p)$ is comparable to the sample size $n$, we will show that under an irrepresentable condition, the proposed estimator $[\hat{\Theta}_{XX}]_0$ is consistent. It then follows by the consistency of $\hat{\Sigma} = \{(i,j) \in V \times V : [\Theta_{XX}]_0 \neq 0\}$, the estimated edge set based on the estimated DAVO.

We first study the behavior of the sample-level coordinate representation of the DAVO. Since we use $[\hat{\Sigma}_{XX}]_0$ to estimate $[\Sigma_{XX}]_0$, the consistency requires a bound on the difference $[\hat{\Sigma}_{XX}]_0 - [\Sigma_{XX}]_0$ in the entrywise $\ell_\infty$-norm. For a matrix $M$, $\|M\|_\infty = \max_{ij} |M_{ij}|$.

**Lemma 1.** For any constant $\tau > 2$, with probability at least $1 - 1/p^{\tau-2}$, we have

$$\|\hat{\Sigma}_{XX}^V - [\Sigma_{XX}]_0\|_\infty \leq \frac{3}{\sqrt{2}} \left( \frac{\log(6n^2) + \tau \log(p)}{n} \right) .$$

As with the classical Lasso, we need a type of irrepresentable condition for it to be consistent. For this purpose, we introduce some additional notations. Let $\mathcal{S}$ be the true blockwise support of $[\Theta_{XX}]_0$, that is $\mathcal{S} := \{(i,j) \in V \times V : [\Theta_{XX}]_0 \neq 0\}$. It is easy to see that $\mathcal{S} = \mathcal{E} \cup \{(i,j) : i \in V\}$, which is the union of the edge set and the diagonal elements. Let $[i,j]$ be the index set defined in the previous section. Furthermore, for any subset $A$ of $V \times V$, let $[A] := \bigcup_{(i,j) \in A} [i,j]$ be the index set of the union of the blocks corresponding to the members of $A$. Denoting the Kronecker product by $\otimes$, an important quantity involved in the irrepresentable condition is

$$\Gamma := \frac{1}{2}([\Sigma_{XX}]_0 \otimes I_{mp} + I_{mp} \otimes [\Sigma_{XX}]_0).$$

For any two subsets $B_1$ and $B_2$ of $\{1, \ldots, mp\} \times \{1, \ldots, mp\}$, let $\Gamma_{B_1, B_2}$ denote the submatrix of $\Gamma$ with rows and columns indexed by $B_1$ and $B_2$, that is $\Gamma_{B_1, B_2} = \frac{1}{2} ([\Sigma_{XX}]_0)_{\delta B_1 \delta B_2} + ([\Sigma_{XX}]_0)_{\delta B_1 \delta \bar{B}_2} + ([\Sigma_{XX}]_0)_{\delta \bar{B}_1 \delta B_2}$, where $\delta$ is the Kronecker delta function. The next assumption adapts the irrepresentable condition (Ravikumar et al. 2011; Zhang and Zou 2014) to our blockwise setting. Note that, for any subset $A \subseteq V \times V$, $|[A]| = |A|m^2$.

**Assumption 2.** Let $\Upsilon := [\mathcal{S}]_0 \{[\mathcal{S}]_0 \}^{-1} \in \mathbb{R}^{[\mathcal{S}]_0 \times [\mathcal{S}]_0 \times [\mathcal{S}]_0 \times [\mathcal{S}]_0}$. Assume the following irrepresentable condition for the D-trace group-Lasso estimator (8):

$$\max_{e \in [\mathcal{S}]_0} \sum_{f \in [\mathcal{S}]_0} \left( \left\| \Upsilon_{f,e} \right\|_2 \right)^2 < 1,$$

where $\Upsilon_{f,e}$ is the $m^2$-dimensional vector $[\Upsilon_{f,g} : g \in [e]]$ and $\left\| \Upsilon_{f,e} \right\|_2$ is its Euclidean norm. In the binary case with $m = 1$, (12) reduces to $\max_{e \in [\mathcal{S}]_0} \sum_{f \in [\mathcal{S}]_0} \left\| \Upsilon_{f,e} \right\|_2 < 1$.

Let $\gamma := 1 - \max_{e \in [\mathcal{S}]_0} \sqrt{\sum_{f \in [\mathcal{S}]_0} \left( \left\| \Upsilon_{f,e} \right\|_2 \right)^2}$, then an equivalent way to express the irrepresentable condition is that $\gamma > 0$. Assume the edge set $\mathcal{E}$ is sparse in the sense that the maximum degree is bounded from $p$, where the maximum degree is defined by $d := \max_{e \in \mathcal{E}} |N_e| = \max_{e \in \mathcal{E}} |\{k \in V \setminus \{i,j\} : (i,k) \in \mathcal{E}\}|$, which corresponds to its maximum number of nonzero blocks in any row of $[\Theta_{XX}]_0$. For a matrix $M$, the norm $\|M\|_{1,\infty}$ means $\max_{i,j} |M_{ij}|$. Let $\kappa_p := \|([\mathcal{S}]_0)^{-1}\|_{1,\infty}$ and $\kappa_p := \|\Sigma_{XX}\|_{1,\infty}$. The three quantities, $d$, $\kappa_p$, and $\kappa_p$, control the degree of the sparsity of the graph. Let $\delta := \min_{(i,j) \in \mathcal{E}} \|\Theta_{XX}\|_{1,\infty}$ be the minimum Frobenius norm of the nonzero off-diagonal blocks of $[\Theta_{XX}]_0$.

We now establish the convergence rate under the assumption that the $n$ observations of $X$ are independently and identically sampled from a certain DASG.
Theorem 6. Suppose κΓ and κΣ are bounded and the irrepre-
sentable condition (12) holds. There exists C > 0 such that for any con-
tant τ > 2, if
\[
\begin{align*}
  n &> \max \left( m^{-2}, \gamma^{-2}(\kappa_\Sigma \kappa_\Gamma + 1)^2 \right) \cdot \frac{81}{2} \kappa_\Sigma^2 m^4 d^2 \log(6m^2) + \tau \log(p), \\
  \lambda_n &= \frac{2}{\sqrt{\gamma}} \gamma^{-1} (\kappa_\Sigma \kappa_\Gamma + \kappa_\Sigma \kappa_\Gamma) m^2 dn^{-\frac{1}{2}} \log(6m^2) + \tau \log(p),
\end{align*}
\]
then with probability at least 1 − 1/pτ−2, the solution [\(\widehat{\Theta}_{\text{XX}}\)] is unique and satisfies:
\[
\|\widehat{\Theta}_{\text{XX}} - [\Theta_{\text{XX}}]\|_\infty \leq C m^\frac{5}{2} d \left( \frac{\log(6m^2) + \tau \log(p)}{n} \right).
\]
In addition, \(\widehat{\Theta} \subseteq \varepsilon\). Moreover, if
\[
n > \tau^{-2} C^2 m^7 d^2 \log(6m^2) + \tau \log(p),
\]
then \(\widehat{\Theta} = \varepsilon\).

The error bound of the binary DASG can be obtained as a special case. Assume the node vector \(X\) is a member of \([a_0, a_1]^p\), where \(a_0\) and \(a_1\) are the two labels. The vertex representation of its DAPO is a scaled version of the inverse covariance matrix of \(X\). For a matrix \(M\), let \(\|M\|_{\text{op}}\) be its operator norm.

Corollary 5. The binary random vector \(X\) follows the DASG with respect to \(G = (V, E)\). Suppose κΓ and κΣ are bounded and the irrepre-
tentable condition (12) holds. There exists C > 0 such that for any constant τ > 2, if
\[
\begin{align*}
  n &> \max \left( m^{-2}, \gamma^{-2}(\kappa_\Sigma \kappa_\Gamma + 1)^2 \right) \cdot \frac{81}{2} \kappa_\Sigma \kappa_\Gamma m^4 d^2 \log(6) + \tau \log(p), \\
  \lambda_n &= \frac{2}{\sqrt{\gamma}} \gamma^{-1} (\kappa_\Sigma \kappa_\Gamma + \kappa_\Sigma \kappa_\Gamma) m^2 dn^{-\frac{1}{2}} \log(6) + \tau \log(p),
\end{align*}
\]
then with probability at least 1 − 1/pτ−2, the solution [\(\widehat{\Theta}_{\text{XX}}\)] is unique and satisfies:
1. \(\|\widehat{\Theta}_{\text{XX}} - [\Theta_{\text{XX}}]\|_\infty \leq C d \left( \frac{\log(p)}{n} \right)^{\frac{1}{2}}\),
2. \(\|\widehat{\Theta}_{\text{XX}} - [\Theta_{\text{XX}}]\|_{1, \infty} \leq C d \left( \frac{\log(p)}{n} \right)^{\frac{1}{2}}\),
3. \(\|\widehat{\Theta}_{\text{XX}} - [\Theta_{\text{XX}}]\|_{\text{op}} \leq C \min \left( |S|^{-\frac{1}{2}}, d \right) d \left( \frac{\log(p)}{n} \right)^{\frac{1}{2}}\).

This corollary shows that the convergence rate of our proposed estimator is the same as D-trace lasso estimator for Gaussian graphical model (Zhang and Zou 2014). It is also close to the graphical Lasso (Ravikumar et al. 2011) up to the maximum node degree \(d\). This is significant as our method is model-free, whereas the graphical Lasso is a parametric method. Finally, Li, Chun, and Zhao (2014) and Li and Solea (2018) developed consistency and convergence rates for the APO-based thresholding estimators only for fixed \(p\). In comparison, as shown in Theorem 6 and Corollary 5, our penalized method allows us to develop its consistency and convergence rate in the ultra-high dimensional setting. The improved rate also lies in the fact that there is no need to deal with nonparametric smoothing in the discrete setting.

6. Simulation Study
In this section we compare our new D-trace group-DASG estimator (DLasso) with the APO-based estimator (APO; Li, Chun, and Zhao 2014; Lee, Li, and Zhao 2016) and graphical Lasso (GLasso; Friedman, Hastie, and Tibshirani 2008) for binary data. Since binary data does not follow the Gaussian model, the parametric assumption of GLasso is not satisfied. Nonetheless, its algorithm can be adapted to the blockwise setting to produce a DASG estimator. That is, the GLasso estimator in this section is
\[
[\Theta_{\text{XX}}]_{ij}^{(\text{GLasso})} := \arg\min_{\Theta = \Theta^T} \left\{ \langle \Theta, [\Sigma_{\text{XX}}]_{ij} \rangle_F - \log \det(\Theta) + \lambda_n \sum_{i\neq j} \|\Theta_{[i,j]}\|_F \right\}.
\]
We also provide examples where the DASG and the MRF are equivalent, in which case the sparse Ising model (SPising; Xue, Zou, and Cai 2012) is also included for comparison.

For any given edge set \(E\), we construct a random vector \(X\) following a DASG with respect to \(E\) using the following procedure:
1. Choose a positive definite \(A = [A_{ij}]_{p \times p} \) such that \((i,j) \notin E \Rightarrow A_{ij} = 0\) and let \(B = A^{-1}\).
2. Let \(\Sigma = C_B^{-\frac{1}{2}} B C_B^{-\frac{1}{2}}\), where \(C_B = \text{diag}(B)\) is a diagonal matrix.
3. Transform \(\Sigma\) to \(\Sigma' \in \mathbb{R}^{p \times p}\) by \(\Sigma'_{ij} = \sin\left( \frac{\pi}{2} \Sigma_{ij} \right)\). Suppose \(\Sigma'\) is also positive definite.
4. Generate a Gaussian random vector \(W = [W_1, \ldots, W_p]^T \sim N(0, \Sigma')\).
5. Obtain the binary node vector \(X = \text{sign}(W) = (\text{sign}(W_1), \ldots, \text{sign}(W_p))^T\).

By construction, the DAPO for \(X\) has the orthonormal representation as \([\Theta_{XX}]_{0} = C_B^{-\frac{1}{2}} A C_B^{-\frac{1}{2}}\), which agrees with \(E\). The matrix \(A\) in the first step is called a pattern matrix.

We consider the following models:
- Model 1: Ising model with parameter \(\beta_{ij} = 0.3 \cdot 1_{|i-j|=1} + 0.3 \cdot 1_{|i-j|=p-1}\).
- Model 2: Ising model with parameter \(\beta_{i,i+1} = 0.3\) for any \(i \notin D; \beta_{ij} = 0.2\) for \(j \in D \setminus \{1\}\) or \((j-1) \in D\); and \(\beta_{ij} = 0\) otherwise, where \(D = \{1, p/50, 2p/50, \ldots, 49p/50, p\}\), with \(p\) chosen to be 50 for some integer \(q \geq 2\).
- Model 3: binary DASG generated by the above procedure with the pattern matrix \(A\), where \(A_{ij} = 1_{|i-j|=1} + 0.25 \cdot 1_{|i-j|=p-1}\).  \(\beta_{i,i+1} = 0.3\) for any \(i \notin D; \beta_{ij} = 0.2\) for \(j \in D \setminus \{1\}\) or \((j-1) \in D\); and \(\beta_{ij} = 0\) otherwise, where \(D = \{1, p/50, 2p/50, \ldots, 49p/50, p\}\), with \(p\) chosen to be 50 for some integer \(q \geq 2\).
- Model 4: binary DASG generated by the above procedure with the pattern matrix \(A\), where \(A_{ij} = 1_{|i-j|=1} + 0.24 \cdot 0.75^{(i-j)-1} \cdot 1_{1 \leq |i-j| \leq 3}\).

We choose \(n = 300\) and \(p = 200, 400\), and repeat the experiment 100 times for each \((n, p)\). The DLasso and the GLasso are tuned by the 5-fold cross-validation and the APO is tuned by generalized cross-validation.

We use three criteria to compare the prediction performance: the true positive rate (TPR), the true negative rate (TNR), and
the $F_1$ score, as defined below:

$$TPR = \frac{\hat{E} \cap E}{|E|}, \quad TNR = \frac{\hat{E} \cap E^c}{|E^c|}, \quad F_1 = \frac{2|\hat{E} \cap E|}{|\hat{E}| + |E|}.$$

TPR is the proportion of correctly estimated nonzeros; TNR is the proportion of correctly estimated zeros; $F_1$ score is the harmonic mean of $\frac{|\hat{E} \cap E|}{|E|}$ and $\frac{|\hat{E} \cap E^c|}{|E^c|}$.

The simulation results are shown in Table 1 ($p = 200$) and Table 2 ($p = 400$). We see that DLasso has good accuracy. For Model 1, DLasso, as a model-free estimator, yields comparable results to the parametric Splsing with close TPR and TNR. For Model 2, the DASG estimators give higher TPR than Splsing, and DLasso maintains high TNR as well. For Models 3 and 4, DLasso achieves the highest TNR and $F_1$ score, and an acceptably high TPR. APO, the norm thresholding estimator, tends to give the highest TPR at the cost of the lowest TNR. GLasso behaves similarly to DLasso, but its TNR is uniformly lower than DLasso. By taking advantage of the sparsity of the graph, DLasso uncovers significantly more true negatives (true absent edges) than GLasso. On the other hand, DLasso gives much smaller standard errors in TPR and TNR than those of GLasso, reflecting its robustness due to the D-trace loss function.

Table 3 compares the estimation errors in matrix norms by the DLasso and the GLasso. The DLasso has smaller errors with respect to the $\ell_{1,\infty}$-norm and the operator norm. The two estimators behave similarly with respect to the $\ell_{\infty}$-norm.

We also computed the receiver operating characteristic (ROC) curves of the estimators, which are presented in Figure 3. In each panel the horizontal axis shows the false positive rate (1-specificity) and the vertical axis the true positive rate (sensitivity). For Models 3 and 4, we studied four different $(n, p)$ settings, where $n = 300$ is fixed and $p = 200, 400, 600, 800$. Each curve in Figure 3 is the average of ROC curves across 10 simulation runs. We see that, in all cases, the areas under the curve (AUC) for the DLasso are substantially larger than both GLasso and APO.

7. Application to HIV Antiretroviral Therapy Data

High throughput technology in biomedicine produces enormous data, offering an integrated view of life for scientists and clinicians. Processing and interpreting these data are on the frontiers of statistical methodological research. By presenting pairwise connections between variables intuitively, graphical models have been widely applied in disease diagnosis, drug discovery, and prediction of regulatory networks. In this section, the DASG is used to detect the association between mutations in a dataset of HIV-1 protease sequences. It could provide evidence for drug discovery in the future.

We apply our method to an HIV antiretroviral therapy (ART) susceptibility dataset, which is a binary dataset, as described in Rhee et al. (2006). The dataset includes virus mutation information at 99 protease residues for $n = 702$ isolates from the plasma of HIV-1-infected patients. A mutation for a certain protease residue is denoted by $+1$, and no mutation is denoted by $-1$. Our analysis only includes $p = 62$ of the 99 residues that contain at least five mutations.

The CI graphical structure of this dataset was studied by Xue, Zou, and Cai (2012) using the Ising model-based method. However, it is questionable whether the data follows an Ising model. Recent work by Yang et al. (2018) provided a goodness-of-fit test for the Ising model assumption via the Kernelized Discrete Stein Discrepancy (KDS). We perform the KDS test for two null hypotheses, the Ising model (2) and the symmetric Ising model (6). Since the KDS-test is a bootstrap-based test, we repeat it 100 times. For the Ising model (2) hypothesis, we obtain an average $p$-value of 0.0281 and the maximum $p$-value of 0.044; for the symmetric Ising model (6) hypothesis, this average $p$-value is 0.0034 and the maximum $p$-value is 0.008.
Figure 3. The comparison of ROC curves with respect to Model 3 and Model 4.
for the two sets of p-values for the two hypotheses are presented in Figure 4. Both the Ising model hypothesis and the symmetric Ising model hypothesis are rejected with a significance level 0.05. Especially, for the symmetric Ising model, the p-value is less than 0.01. This can be partly explained by the mean values of the data, as zero expectation is a basic property satisfied by the symmetric Ising model (6). In fact, except for residues 10, 63, and 71, the mutation frequency at all the other residues are lower than 0.38, leading to mostly negative column mean values. For that reason, a symmetric Ising model is not a good choice for CI structure learning.

Thus, by the test results, the Ising model does not seem to be a valid assumption. It is then reasonable to turn to a nonparametric method such as the DASG. In Figure 5 we present the bootstrapped stable edges of the Ising model and the DASG: we draw 100 bootstrap samples of size 702 and estimate the Ising model and the DASG repeatedly, and plot the edges selected more than 95 times.

In the following discussion, a string such as “x", “Ax", and “AxB", where “A" and “B" are capital letters and “x" is an integer, corresponds to a node represented by the number “x" in the graphs in Figure 5. With “x" standing for residue x, “Ax" means that its original type of amino acid is A and “AxB" means the mutation into type B amino acid.

The stable edge set of the DASG includes 48 edges. Many of them are meaningful in the context of HIV study, which can be verified by some previous literature. Some of the important pairs of residues are labeled in the ribbon diagram shown in Figure 6. Hoffman, Schiffer, and Swanstrom (2003) discovered and explained many pairs of interactions that are consistent with our findings. For example, in the K20R:M36I double mutant, where K20 and M36 are physically close to each other, the double mutation would compensate for the original interaction between the two residues (Hoffman, Schiffer, and Swanstrom 2003); Residues 35 and 37 are near one another in the protease structure in the hinge region of the flap. E35 forms an ionic bond with R57 and N37D would be stabilized by interacting with R57 (Hoffman, Schiffer, and Swanstrom 2003); An M46I mutant exhibits enhanced catalytic activity over the wild-type enzyme and improves the activity of a protease mutant containing both V82T and I84V (Schock, Garsky, and Kuo 1996; Hoffman, Schiffer, and Swanstrom 2003); The double mutation D30N and N88D can reduce nelfinavir susceptibility by 50-fold (Rhee et al. 2003; Liu, Eyal, and Bahar 2008).

The DASG identifies residue 10 as an important node hub. L10I, although not causing resistance alone (it is a minor resistance residue), plays a critical role in eliciting the cooperative response along with L90M at the dimer interface (Ohtaka, Schön, and Freire 2003; Liu, Eyal, and Bahar 2008). Mutations at both residues 10 and 71 frequently appear in clinical samples from subjects who have failed protease inhibitor therapy and these mutations can be selected by the passage of HIV-1 in the presence of a protease inhibitor in vitro (Hoffman, Schiffer, and Swanstrom 2003).

Compared with the DASG, the Ising model recovers only the most significant pairs with 25 stable edges. Based on 2244 subtype B HIV-1 isolates from 1919 persons with different protease inhibitor experiences, Wu et al. (2003) reported (54, 82) and (32, 47) as two of the most highly correlated pairs. Hoffman, Schiffer, and Swanstrom (2003) pointed out that residues 48, 54, and 82 represent a hierarchy of interactions, where G48V is likely a late mutation added after changes at both 54 and 82 have occurred. The Ising model finds (54, 82) but misses their interactions with 48; the HIV-1 protease with drug-resistant mutations V32I, I47V, and V82I has been evaluated as a model for inhibition of HIV-2 protease to overcome the problem of autoproteolysis of HIV-2 protease. There are hydrophobic interactions of residues 32 and 47 according to molecular biology analysis (Pawar et al. 2019). The Ising model fails to keep (32, 47) as an important edge while the DASG captures it successfully. Another pair uncovered by the DASG only is (12, 19), and the interaction between residues 12 and 19 seems to be necessary to maintain Van der Waals force (Hoffman, Schiffer, and Swanstrom 2003).

8. Discussion

In this work, we developed the additive semi-graphoid model for finite discrete data, which is a nonparametric graphical model based on ACI. We discussed the relationship between ACI and CI under the Ising model and showed some sufficient conditions under which the DASG agrees with the MRF. We proposed the DLasso estimator and derived corresponding asymptotic properties under the high-dimensional setting. We demonstrated the power of DLasso in structure learning on both synthetic and real datasets.

The idea of DASG can be further extended to situations where the discrete random variables on the nodes have finite but countable supports. Roy and Dunson (2020) studied a nonparametric exponential family for count data with countably infinite supports. It is worthwhile to extend the current method to this type of data. Also, Fan, Feng, and Xia (2020) proposed the framework of learning a nonparametric CI graphical model via conditional independence tests. This test-based method is another important direction of future development for the ACI graphical models.
Figure 5. Comparison of the Ising model (MRF) and the DASG on the HIV data.

Figure 6. A ribbon diagram of the HIV-1 protease dimer about the locations of some double mutations found by the DASG, with each pair of residues highlighted in the same color. The diagram was generated from the molecular visualization system PyMOL (DeLano 2002).

Supplementary Materials
The supplementary file includes the technical proofs and notes.

Acknowledgments
We would like to thank the Coeditor, Associate Editor and referees for their helpful comments and suggestions.

Funding
Jun Tao and Lingzhou Xue were partially supported by the National Science Foundation grants DMS-1811552, DMS-1953189, and DMS-2210775. Bing Li was partially supported by the National Science Foundation grant DMS-2210775.

ORCID
Lingzhou Xue © http://orcid.org/0000-0002-8252-0637

References
Ahsan, A., Rudnick, J., and Bruinsma, R. (1998), “Elasticity Theory of the B-DNA to S-DNA Transition,” Biophysical Journal, 74, 132–137. [368]
Baker, C. R. (1973), “Joint Measures and Cross-Covariance Operators,” Transactions of the American Mathematical Society, 186, 273–289. [370]
Cheng, J., Levina, E., Wang, P., and Zhu, J. (2014), “A Sparse Ising Model with Covariates,” Biometrics, 70, 943–953. [368]
Dawid, A. P. (1979), “Conditional Independence in Statistical Theory,” Journal of the Royal Statistical Society, Series B, 41, 1–15. [369]
DeLano, W. L. (2002), The Pymol Molecular Graphics System, San Carlos, CA: De-Lano Scientific. Available at http://www.pymol.org. [380]
Fan, J., Feng, Y., and Wu, Y. (2009), “Network Exploration via the Adaptive Lasso and Scad Penalties,” The Annals of Applied Statistics, 3, 521–541. [374]
Fan, J., Feng, Y., and Xia, L. (2020), “A Projection-based Conditional Dependence Measure with Applications to High-Dimensional Undirected Graphical Models,” Journal of Econometrics, 218, 119–139. [379]
Fierst, J. L., and Phillips, P. C. (2015), “Modeling the Evolution of Complex Genetic Systems: The Gene Network Family Tree,” Journal of Experimental Zoology Part B: Molecular and Developmental Evolution, 324, 1–12. [368]
Friedman, J., Hastie, T., and Tibshirani, R. (2008), “Sparse Inverse Covariance Estimation with the Graphical Lasso,” Biostatistics, 9, 432–441. [374,376]
Fukumizu, K., Bach, F. R., and Jordan, M. I. (2009), “Kernel Dimension Reduction in Regression,” The Annals of Statistics, 37, 1871–1905. [370]
Geman, S., and Geman, D. (1993), “Stochastic Relaxation, Gibbs Distributions and the Bayesian Restoration of Images,” Journal of Applied Statistics, 20, 25–62. [368]
Guo, J., Levina, E., Michailidis, G., and Zhu, J. (2015), “Graphical Models for Ordinal Data,” Journal of Computational and Graphical Statistics, 24, 183–204. [369]
Hoffman, N. G., Schiffer, C. A., and Swanstrom, R. (2003), “Covariation of Amino Acid Positions in HIV-1 Protease,” Virology, 314, 536–548. [379]
Höfling, H., and Tibshirani, R. (2009), “Estimation of Sparse Binary Pairwise Markov Networks using Pseudo-Likelihoods,” Journal of Machine Learning Research, 10, 883–906. [368]
Ising, E. (1925), “Beitrag zur Theorie des Ferromagnetismus,” Zeitschrift für Physik, 31, 253–258. [368]
Lauritzen, S. L. (1996), Graphical Models, Vol. 17, Oxford: Clarendon Press. [372]
Lee, K. H., Chen, Q., DeSarbo, W. S., and Xue, L. (2022), “Estimating Finite Mixtures of Ordinal Graphical Models,” Psychometrika, 87, 83–106. [369]
Lee, K.-Y., Li, B., and Zhao, H. (2016), “On an Additive Partial Correlation Operator and Nonparametric Estimation of Graphical Models,” Biometrika, 103, 513–530. [369,374,376]
Li, B., Chun, H., and Zhao, H. (2014), “On an Additive Semigraphoid Model for Statistical Networks with Application to Pathway Analysis,” Journal of the American Statistical Association, 109, 1188–1204. [369,370,372,374,376]
Li, B., and Solea, E. (2018), “A Nonparametric Graphical Model for Functional Data with Application to Brain Networks based on fMRI,” *Journal of the American Statistical Association*, 113, 1637–1655. [376]

Liu, Y., Eyal, E., andBahar, I. (2008), “Analysis of Correlated Mutations in HIV-1 Protease using Spectral Clustering,” *Bioinformatics*, 24, 1243–1250. [379]

Loh, P.-L., and Wainwright, M. J. (2012), “Structure Estimation for Discrete Graphical Models: Generalized Covariance Matrices and their Inverses,” in *Advances in Neural Information Processing Systems*, pp. 2087–2095. [370,372]

Marsman, M., Borsboom, D., Kruijs, J., Epksamp, S., Van Bork, R., Waldorp, L., Maas, H. v. d., and Maris, G. (2018), “An Introduction to Network Psychometrics: Relating Ising Network Models to Item Response Theory Models,” *Multivariate Behavioral Research*, 53, 15–35. [368]

Ohtaka, H., Schön, A., and Freire, E. (2003), “Multidrug Resistance to HIV-1 Protease Inhibition Requires Cooperative Coupling between Distal Mutations,” *Biochemistry*, 42, 13659–13666. [379]

Pawar, S., Wang, Y.-F., Wong-Sam, A., Agniswamy, J., Ghosh, A. K., Harrison, R. W., and Weber, I. T. (2019), “Structural Studies of Antiviral Inhibitor with HIV-1 Protease Bearing Drug Resistant Substitutions of v32i, i47v and v82i,” *Biochemical and Biophysical Research Communications*, 514, 974–978. [379]

Pearl, J., and Verma, T. (1987), “The Logic of Representing Dependencies by Directed Graphs,” in *Proceedings of the Sixth National conference on Artificial intelligence* (Vol. 1), pp. 374–379. [369]

Ravikumar, P., Wainwright, M. J., and Lafferty, J. D. (2010), “High-Dimensional Ising Model Selection using $\ell_1$-regularized Logistic Regression,” *The Annals of Statistics*, 38, 1287–1319. [368]

Ravikumar, P., Wainwright, M. J., Raskutti, G., and Yu, B. (2011), “High-Dimensional Covariance Estimation by Minimizing $\ell_1$-penalized Log-Determinant Divergence.” *Electronic Journal of Statistics*, 5, 935–980. [375,376]

Rhee, S.-Y., Gonzales, M. J., Kantor, R., Betts, B. J., Ravela, J., and Shafer, R. W. (2003), “Human Immunodeficiency Virus Reverse Transcriptase and Protease Sequence Database,” *Nucleic Acids Research*, 31, 298–303. [379]

Rhee, S.-Y., Taylor, J., Wadhera, G., Ben-Hur, A., Brutlag, D. L., and Shafer, R. W. (2006), “Genotypic Predictors of Human Immunodeficiency Virus Type 1 Drug Resistance,” *Proceedings of the National Academy of Sciences*, 103, 17355–17360. [377]

Roy, A., and Dunson, D. B. (2020), “Nonparametric Graphical Model for Counts,” *Journal of Machine Learning Research*, 21, 1–21. [379]

Schöck, H. B., Garsky, V. M., and Kuo, L. C. (1996), “Mutational Anatomy of an HIV-1 Protease Variant Conferring Cross-Resistance to Protease Inhibitors in Clinical Trials Compensatory Modulations of Binding and Activity,” *Journal of Biological Chemistry*, 271, 31957–31963. [379]

Wang, P., Chao, D. L., and Hsu, L. (2011), “Learning Oncogenic Pathways from Binary Genomic Instability Data,” *Biometrics*, 67, 164–173. [368]

Wu, T. D., Schiffer, C. A., Gonzales, M. J., Taylor, J., Kantor, R., Chou, S., Israelski, D., Zolopa, A. R., Fessel, W. J., and Shafer, R. W. (2003), “Mutation Patterns and Structural Correlates in Human Immunodeficiency Virus Type 1 Protease Following Different Protease Inhibitor Treatments,” *Journal of Virology*, 77, 4836–4847. [379]

Xue, L., and Zou, H. (2012), “Regularized Rank-based Estimation of High-Dimensional Nonparanormal Graphical Models,” *The Annals of Statistics*, 40, 2541–2571. [374]

Xue, L., Zou, H., and Cai, T. (2012), “Nonconcave Penalized Composite Conditional Likelihood Estimation of Sparse Ising Models,” *The Annals of Statistics*, 40, 1403–1429. [368,376,377]

Yang, J., Liu, Q., Rao, V., and Neville, J. (2018), “Goodness-of-Fit Testing for Discrete Distributions via Stein Discrepancy,” in *International Conference on Machine Learning*, pp. 5561–5570. [377]

Yuan, M., and Lin, Y. (2006), “Model Selection and Estimation in Regression with Grouped Variables,” *Journal of the Royal Statistical Society, Series B*, 68, 49–67. [374,375]

— (2007), “Model Selection and Estimation in the Gaussian Graphical Model,” *Biometrika*, 94, 19–35. [374]

Zhang, T., and Zou, H., (2014), “Sparse Precision Matrix Estimation via Lasso Penalized d-Trace Loss,” *Biometrika*, 101, 103–120. [374,375,376]