Notes on Graphs and Structure Learning

Graphical models come in two basic varieties, corresponding to directed acyclic graphs and undirected graphs. In these notes we focus on undirected graphical models.

1. Introduction

Let X, Y and Z be three random variables. We write $X \perp\!\!\!\perp Y$ to mean that X and Y are independent. We write $X \perp\!\!\!\perp Y \mid Z$ to mean that X and Y are independent given Z. A single random vector drawn from a distribution P is denoted by $X = (X_1, \ldots, X_d)$. Besides directed graphs, another way to explore the structure of the distribution P for the random vector X is to estimate its *undirected graph* G, which consists of a set of vertices V and an edge set E of unordered pairs of vertices. Each vertex here corresponds to one random variable from $X_1, \ldots, X_d$. When we observe n random vectors, $x_1, \ldots, x_n$, drawn from P, we can write this in more detail as:

$$
\begin{array}{ccccc}
x_1, & x_2, & \ldots, & x_n & \text{i.i.d. drawn from} \quad P \\
\| & \| & & \| & \\
\begin{pmatrix} x_{11} \\ \vdots \\ x_{1d} \end{pmatrix} & \begin{pmatrix} x_{21} \\ \vdots \\ x_{2d} \end{pmatrix} & \vdots & \begin{pmatrix} x_{n1} \\ \vdots \\ x_{nd} \end{pmatrix}. &
\end{array}
$$

There are n independent observations with d features each. The nodes of the graph denote the d features but not the individual observations. For example, suppose we measure blood pressure, age and cholesterol on 100 people. Then $n = 100$ and $d = 3$. A possible graph for this example is provided in Figure 1. If we define $X = (X_1, X_2, X_3) = (\mathsf{blood\ pressure}, \mathsf{age}, \mathsf{cholesterol})$ then we

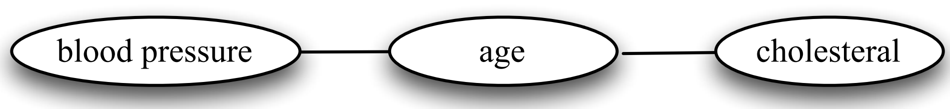


FIG 1. *An undirected graph of three random variables: blood pressure, age, and cholesterol.*

can draw the graph as in 1. Keep in mind that the observations (random vectors) are independent, but the features (coordinates of each random vector) are not. The key idea of the undirected graph representation of probability distributions is the Markov property:

Markov Property

The undirected graph G associated with P has d vertices corresponding to the components $X_1, \ldots, X_d$ of the random vector $X = (X_1, \ldots, X_d)$. We omit an edge between two nodes X_i and X_j if and only if X_i and X_j are conditionally independent given the other variables. This is called the *Markov property* encoded in the graph.

A simple example is shown in Figure 2. In this figure, there is a random vector $X = (X_1, X_2, X_3, X_4)$. Since there is no edge between X_1 and X_4 we conclude that X_1 and X_4 are independent given X_2 and X_3. Similarly, since there is no edge between X_2 and X_3 we conclude that X_2 and X_3 are independent given X_1 and X_4.

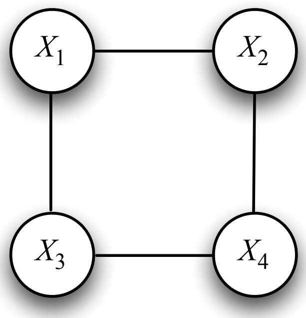


FIG 2. *Graph for a random vector $X = (X_1, X_2, X_3, X_4)$. Since there is no edge between X_1 and X_4 we conclude that X_1 and X_4 are independent given X_2 and X_3. Similarly, since there is no edge between X_2 and X_3 we conclude that X_2 and X_3 are independent given X_1 and X_4.*

Undirected graphs are used as a natural representation of the dependencies between random variables in many applications. One example is image processing, where each pixel in an image may correspond to a random variable. Such lattice based models were first studied in statistical physics, where they were used to represent spin configurations of atoms in a crystal. Other applications where undirected graphical models arise naturally include error correcting codes, protein interaction networks, and social networks; for suggestive illustrations see Figures 3, 4, and 5.

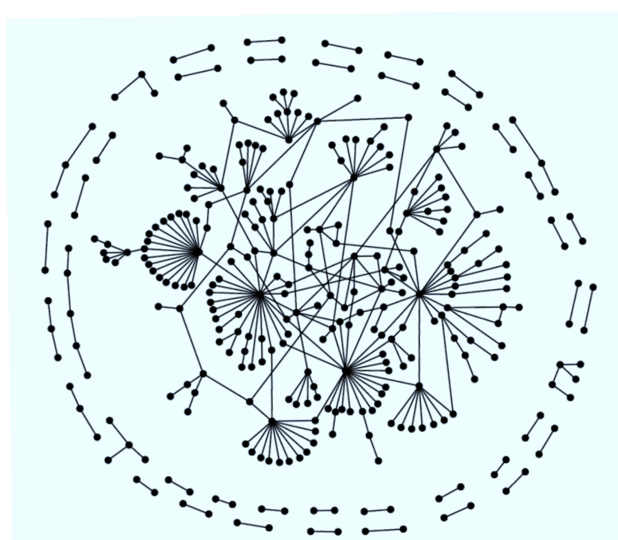


FIG 3. *The graph depicts certain protein networks (Maslov and Sneppen, 2002).*

The estimation and structure learning methods discussed in this chapter are based on parametric models. In later chapters we introduce nonparametric methods for estimating undirected graphs. In the case where each X_i is a continuous random variable, taking values in $\mathbf{R}$, perhaps the most

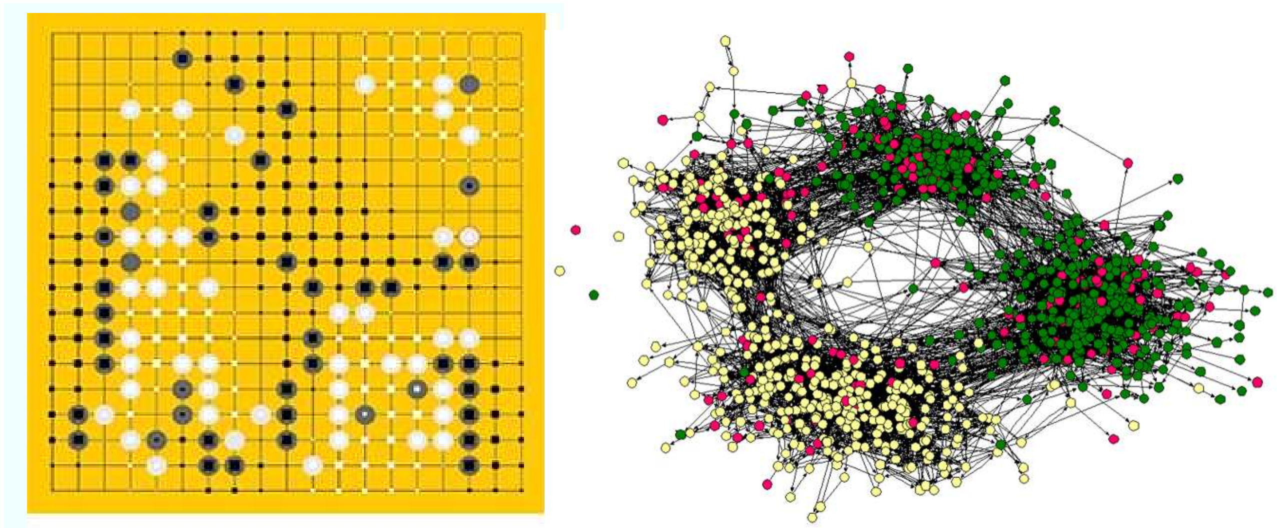


FIG 4. *Left: The game of Go is modeled probabilistically using undirected graphical models in Stern et al. (2004). Here the graph is the grid of the game board, and each node takes on the binary value* `black` *or* `white` *depending on which player owns the position at the end of the game. Right: Graphs have been used for many years to represent social networks (Moody, 2001). Recent work as renewed interest in probabilistic graphical models in social networks.*

popular and useful model is the multivariate Gaussian $X \sim N(\mu, \Sigma)$. As will be explained in the following sections, the graphical structure in this case is encoded in the inverse covariance matrix $\Omega = \Sigma^{-1}$; the edge between X_j and X_k is missing if and only if $\Omega_{jk} = 0$. The case of discrete data, where each variable $X_i \in \{0, 1\}$ is binary, is much more difficult to work with. For this case, the analogue of the Gaussian graphical model is the *Ising model*.

2. Probability and Undirected Graphs

Let $G = (V, E)$ be an undirected graph with vertex set V and edge set E, and let A, B, and C be subsets of vertices. We say that C separates A and B if every path from a node in A to a node in B passes through a node in C. Now consider a random vector $X = (X_1, \ldots, X_d)$ where X_j corresponds to node j in the graph. If $A \subset \{1, \ldots, d\}$ then we write $X_A = (X_j : j \in A)$. In this section, we discuss the relationship between probability distributions and undirected graphs.

2.1. Markov Properties on Undirected Graphs

A probability distribution P for a random vector $X = (X_1, \ldots, X_d)$ may satisfy a range of different Markov properties with respect to a graph $G = (V, E)$:

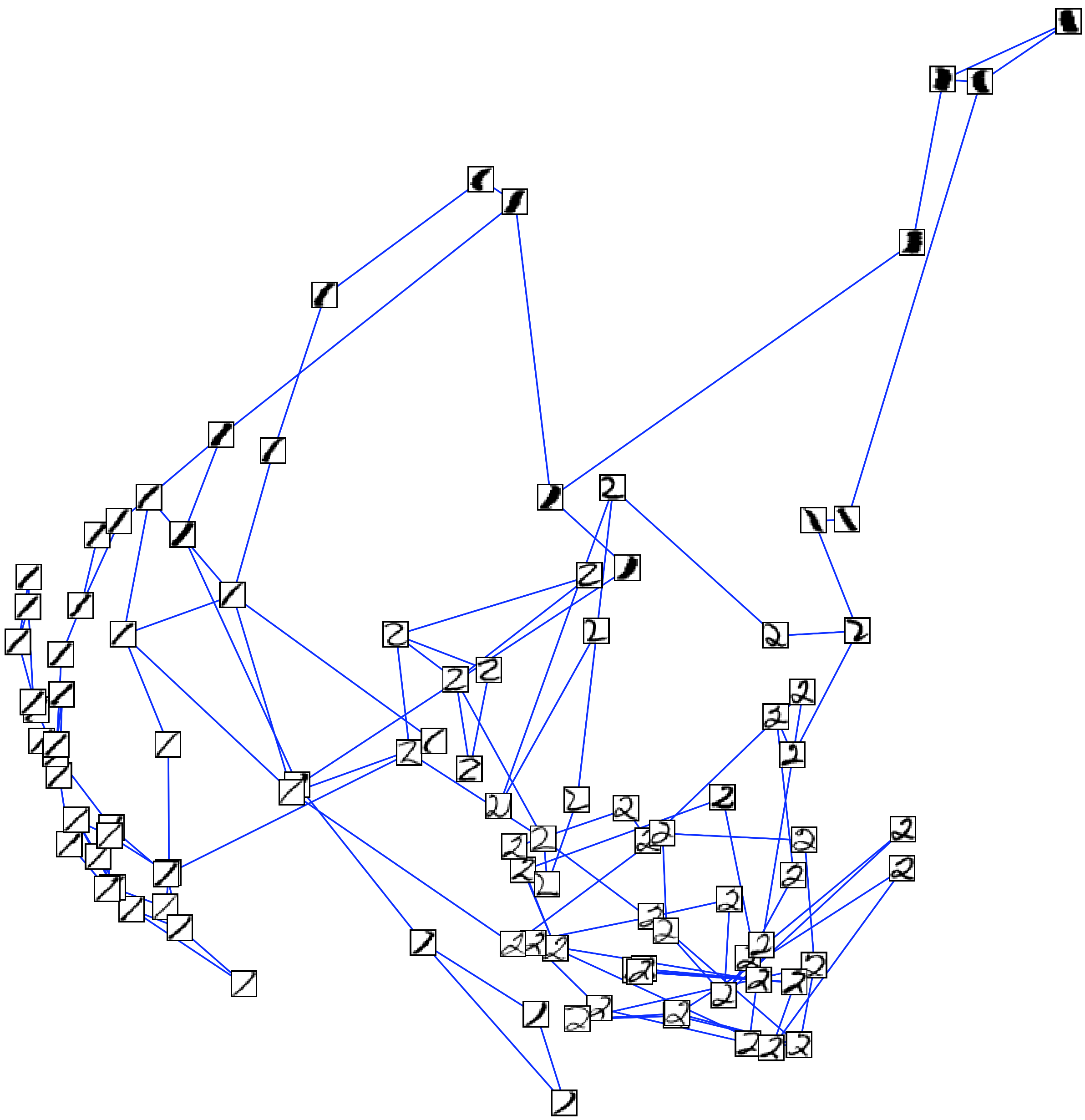


FIG 5. *Undirected graphs have been used in semi-supervised learning; the plot on the right shows a subset of scanned digits of ones and twos, with edges constructed in terms of a nearest neighbor metric, projected into two dimensions using PCA (courtesy of Jerry Zhu).*

Global Markov Property

A probability distribution P for a random vector $X = (X_1, \ldots, X_d)$ satisfies the *global Markov property* with respect to a graph G if for any disjoint vertex subsets A, B, and C such that C separates A and B, the random variables X_A are conditionally independent of X_B given X_C.

Local Markov Property

A probability distribution P for a random vector $X = (X_1, \ldots, X_d)$ satisfies the *local Markov property* with respect to a graph G if the conditional distribution of a variable given its neighbors is independent of the remaining nodes. That is, let $N(s) = \{t \in V \mid (s,t) \in E\}$ denote the set of neighbors of a node $s \in V$. Then the local Markov property is that

$$p(x_s \mid x_t,\ t \neq s) = p\left(x_s \mid x_t, t \in N(s)\right) \tag{1}$$

for each node s.

Pairwise Markov Property

A probability distribution P for a random vector $X = (X_1, \ldots, X_d)$ satisfies the *pairwise Markov property* with respect to a graph G if for any pair of non-adjacent nodes $s, t \in V$, we have

$$X_s \perp\!\!\!\perp X_t \mid X_{V \setminus \{s,t\}}. \tag{2}$$

Consider for example the graph in Figure 6. Here the set C separates A and B. Thus, a distribution that satisfies the global Markov property for this graph must have the property that the random variables in A are conditionally independent of the random variables in B given the random variables C. This is seen to generalize the usual Markov property for simple chains, where $X_A \longrightarrow X_C \longrightarrow X_B$ forms a Markov chain in case X_A and X_B are independent given X_C. A distribution that satisfies the global Markov property is said to be a *Markov random field* or *Markov network* with respect to the graph. The *local Markov property* is depicted in Figure 7.

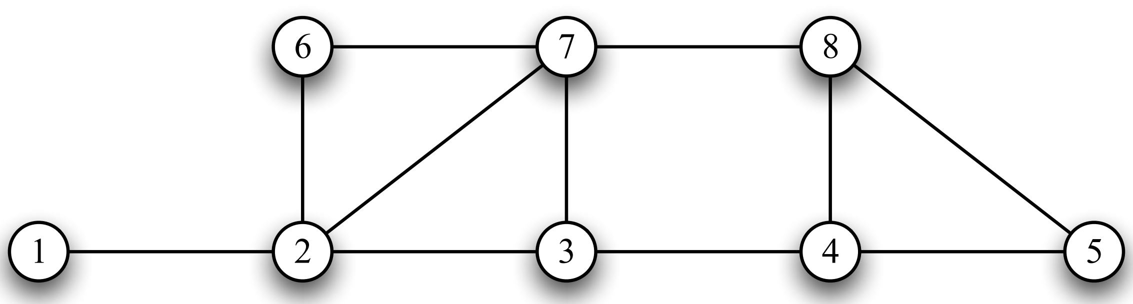


FIG 6. *An undirected graph.* $C = \{3, 7\}$ *separates* $A = \{1, 2\}$ *and* $B = \{4, 8\}$.

From the definitions, the relationships of different Markov properties can be characterized as:

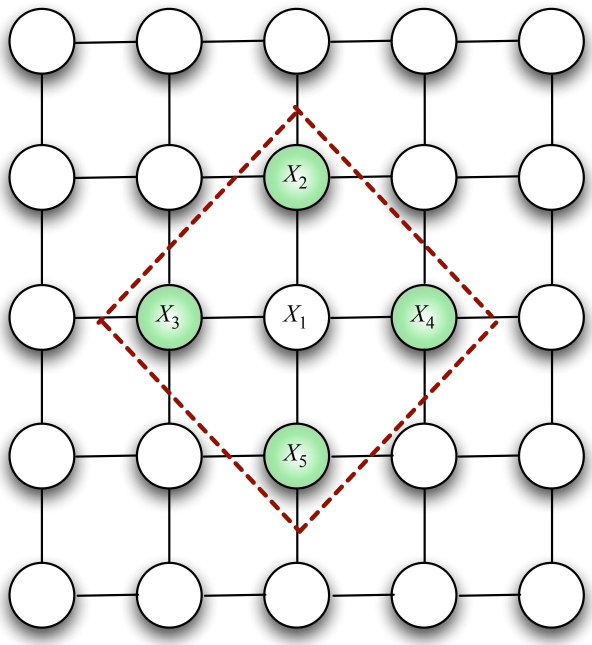


FIG 7. *The local Markov property: Conditioned on its four neighbors X_2, X_3, X_4, and X_5, node X_1 is independent of the remaining nodes in the graph.*

Theorem 2.1. *For any undirected graph G and any distribution P, we have*

$$\text{global Markov property} \implies \text{local Markov property} \implies \text{pairwise Markov property}.$$

Proof. The global Markov property implies the local Markov property because for each node $s \in V$, its neighborhood $N(s)$ separates $\{s\}$ and $V \setminus \{N(s) \cup \{s\}\}$. Assume next that the local Markov property holds. Any t that is not adjacent to s is an element of $t \in V \setminus \{N(s) \cup \{s\}\}$. Therefore

$$N(s) \cup [(V \setminus \{N(s) \cup \{s\}\}) \setminus \{t\}] = V \setminus \{s, t\}, \tag{3}$$

and it follows from the local Markov property that

$$X_s \perp\!\!\!\perp X_{V \setminus \{N(s) \cup \{s\}\}} \mid X_{V \setminus \{s,t\}}. \tag{4}$$

This implies $X_s \perp\!\!\!\perp X_t \mid X_{V \setminus \{s,t\}}$, which is the pairwise Markov property. □

In general, the global, local, and pairwise Markov properties are different, as illustrated by the following examples 2.1 and 2.1. However, if the distributions have positive continuous densities, all three Markov properties are equivalent.

Example 2.2. Define the joint distribution of five binary random variables U, W, X, Y, Z as follows: U and Z are independent with

$$\mathbb{P}(U = 1) = \mathbb{P}(Z = 1) = 1/2, \tag{5}$$

$W = U$, $Y = Z$, and $X = WY$. The joint distribution so defined is easily seen to satisfy the local Markov property but not the global Markov property for the chain graph U—W—X—Y—Z.

Example 2.3. Let $X = Y = Z$ be three binary random variables with $\mathbb{P}(X = 1) = 1/2$. We consider a graph G which only contains one edge that connects Y and Z. It is easy to see that this distribution satisfies the pairwise Markov property with respect to the graph G but it does not satisfy the local Markov property with respect to G, Since X is not independent of Y.

The next theorem, due to Pearl and Paz (1986), provides a sufficient condition for equivalence.

Theorem 2.4. (Pearl and Paz, 1986) *If it holds that for all disjoint subsets $A, B, C, D \subset V$, we have*

$$\text{if } X_A \perp\!\!\!\perp X_B \,|\, X_{C \cup D} \text{ and } X_A \perp\!\!\!\perp X_C \,|\, X_{B \cup D}, \text{ then } X_A \perp\!\!\!\perp X_{B \cup C} \,|\, X_D, \tag{6}$$

then the global, local, and pairwise Markov properties are equivalent.

Proof. It is enough to show that the pairwise Markov property implies the global Markov property under the given condition. Let $S, A, B \subset V$ witn S separating A from B in the graph G. Without loss of generality both A and B are assumed to be non-empty. The proof can be carried out using backward induction on the number of nodes in S, denoted by $m = |S|$. Let $d = |V|$, for the base case, if $m = d - 1$ then both A and B only consist of single vertex and the result follows from pairwise Markov property.

Now assume that $m < d - 1$ and separation implies conditional independence for all separating sets S with more than m nodes. We proceed in two cases: (i) $A \cup B \cup S = V$ and (ii) $A \cup B \cup S \subset V$.

For case (i), we know that at least one of A and B must have more than one element. Without loss of generality, we assume A has more than one element. If $s \in A$, then $S \cup \{s\}$ separates $A \setminus \{s\}$ from B and also $S \cup (A \setminus \{s\})$ separates s from B. Thus by the induction hypothesis

$$X_{A \setminus \{s\}} \perp\!\!\!\perp X_B \,|\, X_{S \cup \{s\}} \text{ and } X_s \perp\!\!\!\perp X_B \,|\, S \cup (A \setminus \{s\}). \tag{7}$$

Now the condition (6) implies $X_A \perp\!\!\!\perp X_B \,|\, X_S$. For case (ii), we could choose $s \in V \setminus (A \cup B \cup S)$. Then $S \cup \{s\}$ separates A and B, implying $A \perp\!\!\!\perp B \,|\, S \cup \{s\}$. We then proceed in two cases, either $A \cup S$ separates B from s or $B \cup S$ separates A from s. For both cases, the condition (6) implies that $A \perp\!\!\!\perp B \,|\, S$. □

The next proposition provides a stronger condition that implies (6).

Theorem 2.5. *Let $X = (X_1, \ldots, X_d)$ be a random vector with distribution P and joint density $p(x)$. If the joint density $p(x)$ is positive and continuous with respect to a product measure, then condition (6) holds.*

Proof. Without loss of generality, it suffices to assume that $d = 3$. We want to show that

$$\text{if } X_1 \perp\!\!\!\perp X_2 \mid X_3 \text{ and } X_1 \perp\!\!\!\perp X_3 \mid X_2 \text{ then } X_1 \perp\!\!\!\perp \{X_2, X_3\}. \tag{8}$$

Since the density is positive and $X_1 \perp\!\!\!\perp X_2 \mid X_3$ and $X_1 \perp\!\!\!\perp X_3 \mid X_2$, we know that there must exist some positive functions $f_{13}, f_{23}, g_{12}, g_{23}$ such that the joint density takes the following factorization:

$$p(x_1, x_2, x_3) = f_{13}(x_1, x_3)f_{23}(x_2, x_3) = g_{12}(x_1, x_2)g_{23}(x_2, x_3). \tag{9}$$

Since the density is continuous and positive, we have

$$g_{12}(x_1, x_2) = \frac{f_{13}(x_1, x_3)f_{23}(x_2, x_3)}{g_{23}(x_2, x_3)}. \tag{10}$$

For each fixed $X_3 = x_3'$, we see that $g_{12}(x_1, x_2) = h(x_1)\ell(x_2)$ where $h(x_1) = f_{13}(x_1, x_3')$ and $\ell(x_2) = f_{23}(x_2, x_3')/g_{23}(x_2, x_3')$. This implies that

$$p(x_1, x_2, x_3) = h(x_1)\ell(x_2)g_{23}(x_2, x_3) \tag{11}$$

and hence $X_1 \perp\!\!\!\perp \{X_2, X_3\}$ as desired. □

From Proposition 2.1, we see that for distributions with positive continuous densities, the global, local, and pairwise Markov properties are all equivalent. If a distribution P satisfies global Markov property with respect to a graph G, we say that P is *Markov to G*

2.2. Clique Decomposition

Unlike a directed graph which encodes a factorization of the joint probability distribution in terms of conditional probability distributions. An undirected graph encodes a factorization of the joint probability distribution in terms of clique potentials. Recall that a *clique* in a graph is a fully connected subset of vertices. Thus, every pair of nodes in a clique is connected by an edge. A clique is a *maximal clique* if it is not contained in any larger clique. Consider, for example, the graph shown in the right plot of Figure 8. The pairs $\{X_4, X_5\}$ and $\{X_1, X_3\}$ form cliques; $\{X_4, X_5\}$ is a maximal clique, while $\{X_1, X_3\}$ is not maximal since it is contained in a larger clique $\{X_1, X_2, X_3\}$.

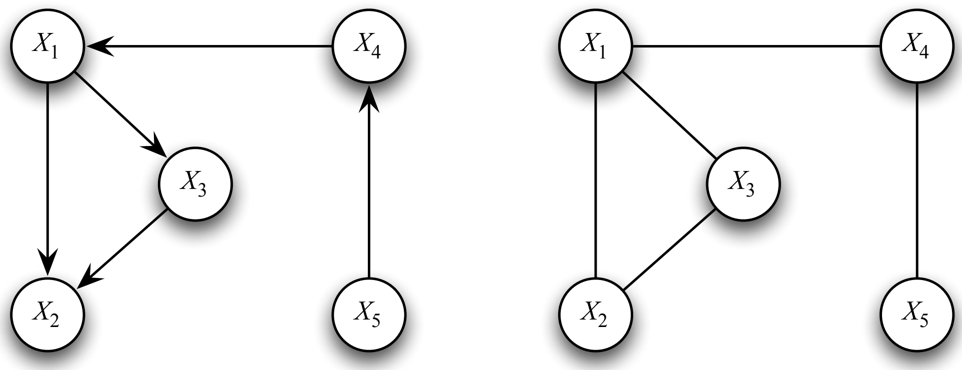


FIG 8. *A directed graph encodes a factorization of the joint probability distribution in terms of conditional probability distributions. An undirected graph encodes a factorization of the joint probability distribution in terms of clique potentials.*

Let $\mathcal{C}$ be the set of all maximal cliques in a graph. A probability distribution *factors with respect to this graph* in case it can be written as a product of factors, one for each of the maximal cliques in the graph:

$$p(x_1, \ldots, x_{|V|}) = \prod_{C \in \mathcal{C}} \psi_C(x_C). \tag{12}$$

Similarly, a set of clique potentials $\{\psi_C(x_C) \geq 0\}_{C \in \mathcal{C}}$ determines a probability distribution that factors with respect to the graph by normalizing:

$$p(x_1, \ldots, x_{|V|}) = \frac{1}{Z} \prod_{C \in \mathcal{C}} \psi_C(x_C). \tag{13}$$

The *normalizing constant* or *partition function* Z sums (or integrates) over all settings of the random variables:

$$Z = \int_{x_1, \ldots, x_{|V|}} \prod_{C \in \mathcal{C}} \psi_C(x_C) dx_1 \ldots dx_{|V|}. \tag{14}$$

Thus, the family of distributions represented by the undirected graph in Figure 8 can be written as

$$p(x_1, x_2, x_3, x_4, x_5) = \psi_{1,2,3}(x_1, x_2, x_3)\, \psi_{1,4}(x_1, x_4)\, \psi_{4,5}(x_4, x_5). \tag{15}$$

In contrast, the family of distributions represented by the directed graph in Figure 8 can be factored into conditional distributions according to

$$p(x_1, x_2, x_3, x_4, x_5) = p(x_5)\, p(x_4 \,|\, x_5)\, p(x_1 \,|\, x_4)\, p(x_3 \,|\, x_1)\, p(x_2 \,|\, x_1, x_3). \tag{16}$$

Theorem 2.6. *For any undirected graph $G = (V, E)$, a distribution P that factors with respect to the graph also satisfies the global Markov property on the graph.*

Proof. Let $A, B, S \subset V$ such that S separates A and B. We want to show $X_A \perp\!\!\!\perp X_B \,|\, X_S$. For a subset $D \subset V$, we denote G_D to be the subgraph induced by the vertex set D. We define $\tilde{A}$ to be the connectivity components in $G_{V \setminus S}$ which contain A and $\tilde{B} = V \setminus (\tilde{A} \cup S)$. Since A and B are separated by S, they must belong to different connectivity components of $G_{V \setminus S}$ and any clique of G must be a subset of either $\tilde{A} \cup S$ or $\tilde{B} \cup S$. Let $\mathcal{C}_A$ be the set of cliques contained in $\tilde{A} \cup S$, the joint density $p(x)$ takes the following factorization

$$p(x) = \prod_{C \in \mathcal{C}} \psi_C(x_C) = \prod_{C \in \mathcal{C}_A} \psi_C(x_C) \prod_{C \in \mathcal{C} \setminus \mathcal{C}_A} \psi_C(x_C). \tag{17}$$

This implies that $\tilde{A} \perp\!\!\!\perp \tilde{B} \,|\, S$ and thus $A \perp\!\!\!\perp B \,|\, S$. $\square$

It is worth remembering that while we think of the set of maximal cliques as given in a list, the problem of enumerating the set of maximal cliques in a graph is NP-hard, and the problem of determining the largest maximal clique is NP-complete (Akkoyunlu, 1973; Bomze et al., 1999). However, many graphs of interest in statistical analysis are sparse, with the number of cliques of size $O(|V|)$.

Theorem 2.2 shows that factoring with respect to a graph implies global Markov property. The next question is, under what conditions the Markov properties imiply factoring with respect to a graph. In fact, in the case where P has a positive and continuous density we can show that the pairwise Markov property implies factoring with respect to a graph. Thus all Markov properties are equivalent. The results have been discovered by many authors but is usually referred to as Hammersley and Clifford due to one of their unpublished manuscript in 1971. They proved the result in the discrete case. The following result is usually referred to as the *Hammersley-Clifford theorem*, a proof appears in (Besag, 1974). The extension to the continuous case is left as an exercise (See Exercise 3).

Theorem 2.7. *[Hammersley-Clifford-Besag] Suppose that $G = (V, E)$ is a graph and X_i, $i \in V$ are random variables that take on a finite number of values. If $\mathbb{P}(x) > 0$ is strictly positive and satisfies the local Markov property with respect to G, then it factors with respect to G.*

Proof. Let $d = |V|$. By re-indexing the values of X_i, we may assume without loss of generality that each X_i takes on the value 0 with positive probability, and $\mathbb{P}(0, 0, \ldots, 0) > 0$. Let $X_{0\setminus i}$ denote the vector $X_{0\setminus i} = (X_1, X_2, \ldots, X_{i-1}, 0, X_{i+1}, \ldots, X_d)$ obtained by setting $X_i = 0$, and let $X_{\setminus i} = (X_1, X_2, \ldots, X_{i-1}, X_{i+1}, \ldots, X_d)$ denote the vector of all components except X_i. Then

$$\frac{\mathbb{P}(x)}{\mathbb{P}(x_{i\setminus 0})} = \frac{\mathbb{P}(x_i \,|\, x_{\setminus i})}{\mathbb{P}(0 \,|\, x_{\setminus i})}. \tag{18}$$

Now, let

$$Q(x) = \log\left(\frac{\mathbb{P}(x)}{\mathbb{P}(0)}\right). \tag{19}$$

Then for any $i \in \{1, 2, \ldots, d\}$ we have that

$$
\begin{aligned}
Q(x) &= \log\left(\frac{\mathbb{P}(x)}{\mathbb{P}(0)}\right) && (20)\\
&= \log\left(\frac{\mathbb{P}(0,\ldots,x_i,0,\ldots,0)}{\mathbb{P}(0)}\right) + \log\left(\frac{\mathbb{P}(x)}{\mathbb{P}(0,\ldots,x_i,0,\ldots,0)}\right) && (21)\\
&= \frac{1}{d}\sum_{i=1}^{d}\left\{\log\left(\frac{\mathbb{P}(0,\ldots,x_i,0,\ldots,0)}{\mathbb{P}(0)}\right) + \log\left(\frac{\mathbb{P}(x)}{\mathbb{P}(0,\ldots,x_i,0,\ldots,0)}\right)\right\}. && (22)
\end{aligned}
$$

Recursively, we obtain

$$
Q(x) = \sum_i \phi_i(x_i) + \sum_{i<j}\phi_{ij}(x_i,x_j) + \sum_{i<j<k}\phi_{ijk}(x_i,x_j,x_k) + \cdots + \phi_{12\cdots d}(x)
$$

for functions ϕ_A that satisfy $\phi_A(x_A) = 0$ if $i \in A$ and $x_i = 0$. Consider node $i = 1$, we have

$$
\begin{aligned}
Q(x) - Q(x_{0\backslash i}) &= \log\left(\frac{\mathbb{P}(x_i \,|\, x_{\backslash i})}{\mathbb{P}(0 \,|\, x_{\backslash i})}\right) && (23)\\
&= \phi_1(x_1) + \sum_{i>1}\phi_{1i}(x_1,x_i) + \sum_{j>i>1}\phi_{1ij}(x_1,x_i,x_j) + \cdots + \phi_{12\cdots d}(x)
\end{aligned}
$$

depends only on x_1 and the neighbors of node 1 in the graph. Thus, from the local Markov property, if k is not a neighbor of node 1, then the above expression does not depend of x_k. In particular, $\phi_{1k}(x_1, x_k) = 0$, and more generally all $\phi_A(x_A)$ with $1 \in A$ and $k \in A$ are identically zero. Similarly, if i, j are not neighbors in the graph, then $\phi_A(x_A) = 0$ for any A containing i and j. Thus, $\phi_A \neq 0$ only holds for the subsets A that form cliques in the graph. Since it is obvious that $\exp(\phi_A(x)) > 0$, we finish the proof. $\square$

Since factoring with respect to the graph implies the global Markov property, we may summarize this result as follows:

For positive distributions, global Markov ⇔ local Markov ⇔ factored

For strictly positive distributions, the global Markov property, the local Markov property, and factoring with respect to the graph are equivalent.

In the following, if a probability distribution factors with respect to a graph, we call it *Gibbs distribution*.

Gibbs Distribution

A probability distribution P with density or probability mass function $p(x)$ on an undirected graph G is called a Gibbs distribution if it can be factorized into positive functions defined on cliques that cover all the nodes and edges of G. That is,

$$p(x) = \frac{1}{Z} \prod_{C \in \mathcal{C}} \psi_C(x_C) = \frac{1}{Z} \exp\left(\sum_{C \in \mathcal{C}} \log \psi_C(x_C) \right) \tag{24}$$

where $\mathcal{C}$ is the set of all (maximal) cliques in G and Z is the normalization constant.

3. Directed vs. Undirected Graphs

Directed graphical models are naturally viewed as generative; the graph specifies a straightforward (in principle) procedure for sampling from the underlying distribution. For instance, a sample from a distribution represented from the DAG in left plot of Figure 9 can be sampled as follows:

$$\begin{aligned}
X_1 &\sim P(X_1) && (25)\\
X_2 &\sim P(X_2) && (26)\\
X_3 &\sim P(X_3) && (27)\\
X_5 &\sim P(X_5) && (28)\\
X_4 \mid X_1, X_2 &\sim P(X_4 \mid X_1, X_2) && (29)\\
X_6 \mid X_3, X_4, X_5 &\sim P(X_6 \mid X_3, X_4, X_5). && (30)
\end{aligned}$$

As long as each of the conditional probability distributions can be efficiently sampled, the full model can be efficiently sampled as well. In contrast, there is no straightforward way to sample from an distribution from the family specified by an undirected graph. We will return to this when we discuss simulation.

3.1. Converting Directed Graphs to Undirected Graphs

In certain cases, a Bayesian network (or DAGs) can also be written as a Gibbs distribution, note that the distribution in (16) also takes the form in (15) with

$$\begin{aligned}
\psi_{1,2,3}(x_1, x_2, x_3) &\equiv p(x_2 \mid x_1, x_3)\, p(x_3 \mid x_1) && (31)\\
\psi_{1,4}(x_1, x_4) &\equiv p(x_1 \mid x_4) && (32)\\
\psi_{4,5}(x_4, x_5) &\equiv p(x_4 \mid x_5)\, p(x_5). && (33)
\end{aligned}$$

However, more generally edges must be added to the skeleton of a DAG in order for the distribution to satisfy the global Markov property on the graph. Consider the example in Figure 9. Here the directed model has a distribution

$$p(x_1)\, p(x_2)\, p(x_3)\, p(x_5)\, p(x_4 \mid x_1, x_2)\, p(x_6 \mid x_3, x_4, x_5). \tag{34}$$

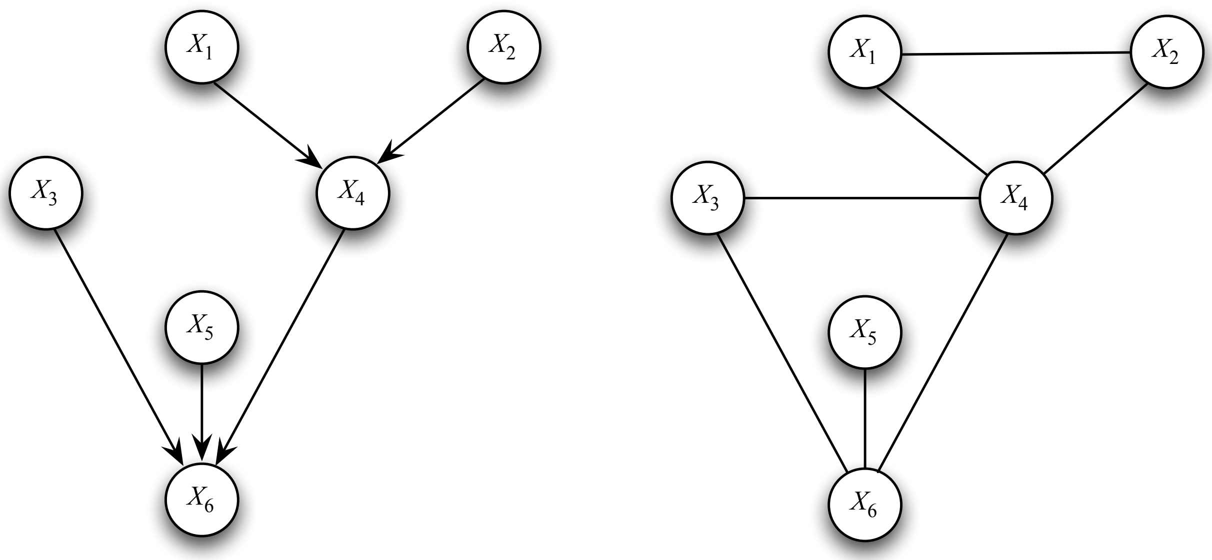


FIG 9. *A DAG and its corresponding moral graph. A probability distribution that factors according to a DAG obeys the global Markov property on the undirected moral graph.*

The corresponding undirected graphical model has two maximal cliques, and factors as

$$\psi_{1,2,4}(x_1, x_2, x_4)\, \psi_{3,4,5,6}(x_3, x_4, x_5, x_6). \tag{35}$$

More generally, let P be a probability distribution that is Markov to a DAG G. We define the *moralized graph* of G as the following: **Definition 3.8.** (Moral graph) *The moral graph M of a DAG G is an undirected graph that contains an undirected edge between two nodes X_i and X_j if (i) there is a directed edge between X_i and X_j in G, or (ii) X_i and X_j are both parents of the same node.*

Theorem 3.9. *If a probability distribution factors with respected to a DAG G, then it obeys the global Markov property with respect to the undirected moral graph of G.*

Proof. Directly follows from the definition of Bayesian networks and Theorem 2.2. □

Example 3.10. (Basic Directed and Undirected Graphs) To illustrate some basic cases, consider the graphs in Figure 10. Each of the top three graphs encodes the same family of probability distributions. In the two directed graphs, by d-separation the variables X_A and X_B are independent conditioned on the variable X_C. In the corresponding undirected graph, which simply removes the arrows, node C separates A and B.

The two graphs in Figure 11 provide an example of a directed graph which encodes a set of conditional independence relationships that can not be perfectly represented by the corresponding moral graph. In this case, for the directed graph the node C is a collider, and deriving an equivalent undirected graph requires joining the parents by an edge. In the corresponding undirected graph, A and

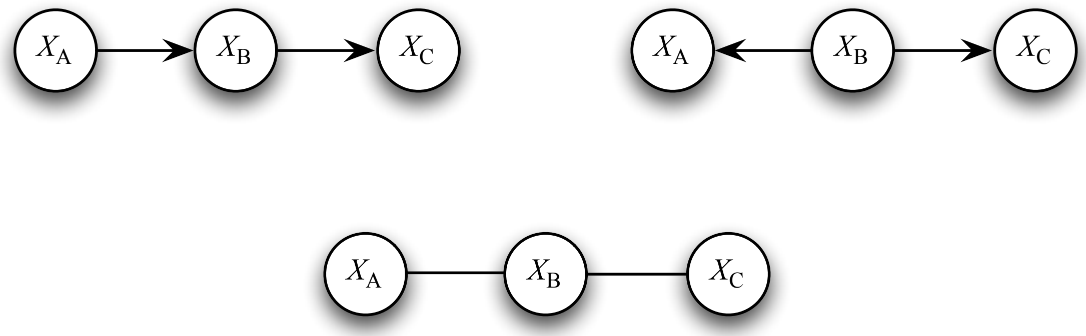


FIG 10. *These three graphs encode distributions with identical independence relations. Conditioned on variable X_C, the variables X_A and X_B are independent; thus C separates A and B in the undirected graph.*

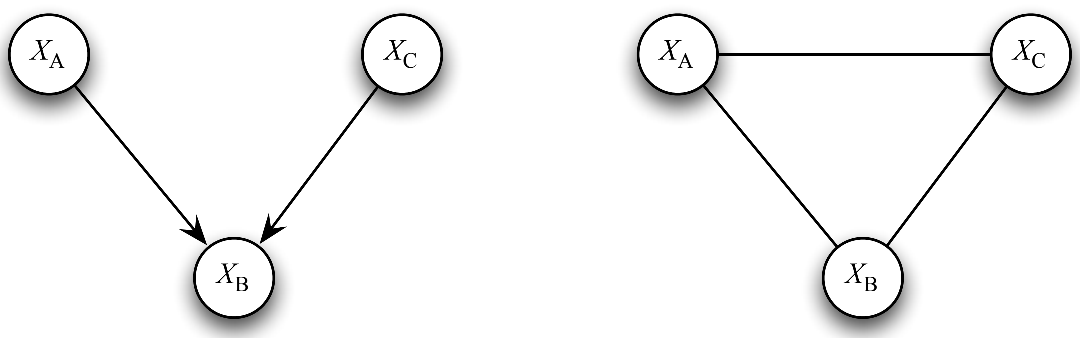


FIG 11. *A directed graph whose conditional independence properties can not be perfectly expressed by its undirected moral graph. In the directed graph, the node C is a collider; therefore, X_A and X_B are not independent conditioned on X_C. In the corresponding moral graph, A and B are not separated by C. However, in the directed graph, we have the independence relationship $X_A \perp\!\!\!\perp X_B$, which is missing in the moral graph.*

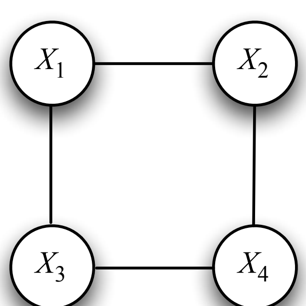


FIG 12. *This undirected graph encodes a family of distributions that cannot be represented by a directed graph on the same set of nodes.*

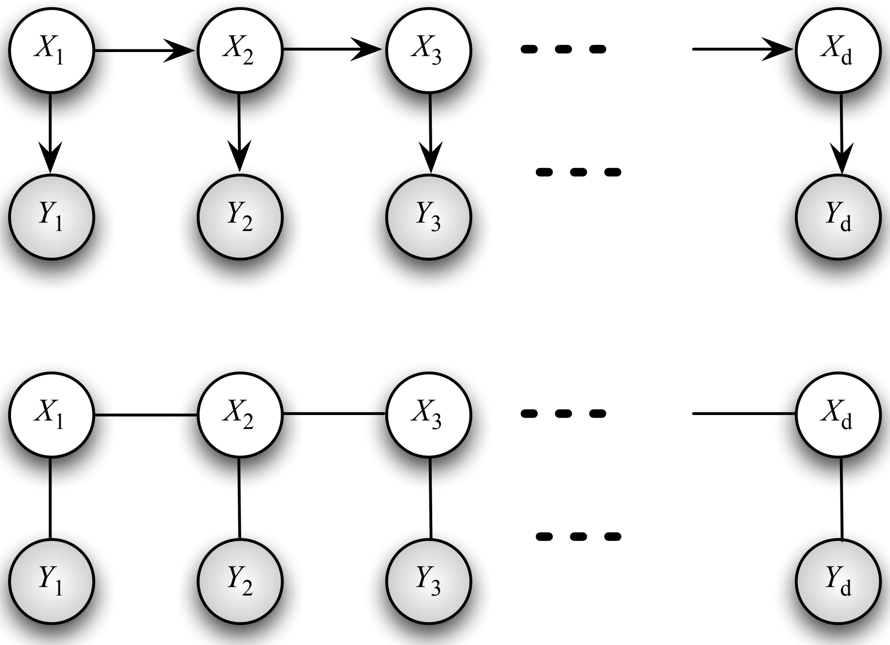


FIG 13. *The top graph is a directed graph representing a hidden Markov model. The shaded nodes are observed, but the unshaded nodes, representing states in a latent Markov chain, are unobserved. Replacing the directed edges by undirected edges (bottom) does not changed the independence relations.*

B are not separated by C. However, in the directed graph, X_A and X_B are marginally independent, such an independence relationship is lost in the moral graph. Conversely, Figure 12 provides an undirected graph over four variables. There is no directed graph over four variables that implies the same set of conditional independence properties.

The upper plot in Figure 13 shows the directed graph underlying a hidden Markov model. There are no colliders in this graph, and therefore the undirected skeleton represents an equivalent set of independence relations. Thus, hidden Markov models are equivalent to hidden Markov fields with an underlying tree graph.

4. Exponential Family Representation

By the Hammersley-Clifford-Besag theorem, we see that the essential property of undirected graphical models is the exponential family representation. In particular, a strictly positive distribution that is locally Markov with respect to a graph can be represented as

$$p(x) = \frac{1}{Z(f)} \exp\left(\sum_{C \in \mathcal{C}} f_C(x_C)\right) \tag{36}$$

where the sum is over all maximal cliques in the graph, and

$$Z(f) = \sum_x \exp\left(\sum_{C \in \mathcal{C}} f_C(x_C)\right) \text{ or } Z(f) = \int \exp\left(\sum_{C \in \mathcal{C}} f_C(x_C)\right) dx \tag{37}$$

in either discrete or continuous settings.

A sufficient condition for positivity is that the *potential functions* f_C satisfy

$$\inf_x \min_{C\in\mathcal{C}} f_C(x_C) > -\infty. \tag{38}$$

Thus, the Hammersley-Clifford-Besag theorem characterizes Markov fields in terms of arbitrary functions of the local configuration in the graph.

This family is often restricted to a linear functions of a set of pre-given *feature vector* $\{f_C(x_C)\}_{C\in\mathcal{C}}$:

$$p(x;\theta) = \frac{1}{Z(\theta)} \exp\left(\sum_{C\in\mathcal{C}} \theta_C f_C(x_C)\right) \tag{39}$$

where $f(x) = (f_C(x_C) : \ C \in \mathcal{C})$ is a vector of sufficient statistics and $\theta = (\theta_C : \ C \in \mathcal{C})$ is the parameter vector. Curved exponential families can also be employed, but these are used less often in practice. Note the difference between (36) and (39); in the latter, the function $f_C(x_C)$ are fixed, and not free parameters of the model as in (36).

5. Gaussian Random Fields for Continuous Data

We now represent multivariate Gaussian distributions using undirected graphical models. A *Gaussian random field* is another name for a multivariate Gaussian distribution, where the structure of the underlying undirected graph is emphasized. This is a simple, but important graphical model that is appropriate for real-valued data $X = (X_1, \ldots, X_d) \in \mathbf{R}^d$.

Let X be distributed as

$$\begin{aligned} p(x_1, x_2, \ldots, x_d) \quad &= \quad \frac{1}{(2\pi)^{d/2}|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)\right) && (40)\\ &\propto \quad \exp\left(-\frac{1}{2}x^T\Omega x + x^T\Omega\mu\right) && (41)\end{aligned}$$

where $\Omega = [\omega_{ij}] = \Sigma^{-1}$ is the *inverse covariance matrix*, also known as the *precision matrix*. Let $G = (V, E)$ be the graph with $V = \{1, 2, \ldots, d\}$, and $(i, j) \in E$ in case $\omega_{ij} \neq 0$. Thus, the inverse of the covariance matrix encodes the graph. The distribution can then be expressed as

$$p(x_1, x_2, \ldots, x_d) \quad \propto \quad \prod_{(i,j)\in E} \psi_{ij}(x_i, x_j) \prod_{i\in V} \psi_i(x_i) \tag{42}$$

where

$$\psi_{ij}(x_i, x_j) = \exp\left(-\frac{1}{2}\omega_{ij}x_ix_j\right) \ \text{ and } \ \psi_i(x_i) = \exp\left(x_i \sum_{j=1}^d \omega_{ij}\mu_j\right). \tag{43}$$

Writing this as a product of factors with respect to the edges, we have

$$p(x_1, x_2, \ldots, x_d) \quad \propto \quad \prod_{(i,j)\in E} \tilde{\psi}_{ij}(x_i, x_j) \tag{44}$$

where $\tilde{\psi}_{ij}(x_i, x_j) = \psi_{ij}(x_i, x_j)\psi_i(x_i)^{1/n(i)}\psi_j(x_j)^{1/n(j)}$ and $n(i) = |\{(i,j) : (i,j) \in E\}|$ is the number of neighbors of node i in the graph. Therefore, the distribution P is Markov to the graph G. One thing to note is that there can be many cycles in the graph.

The following summarizes the above discussion.

Theorem 5.11. *Let $X_1, \ldots, X_d \sim N(0, \Sigma)$ and $\Omega = \Sigma^{-1}$. Let $G = (V, E)$ be the graph defined above, and assume that the distribution P is faithful to G. Then*

$$(i,j) \notin E \iff \omega_{ij} = 0 \iff X_i \perp\!\!\!\perp X_j \mid X_{V\setminus\{i,j\}}. \tag{45}$$

Proof. The result directly follows from (44) and Theorem 2.2. □

6. Log-linear Models for Discrete Data

We now represent multivariate discrete distributions using undirected graphical models. The *log-linear models* are exponential parameterizations of multinomials. Suppose $X_j \in \{0, 1, \ldots, m-1\}$, for $j \in V$, with $V = \{1, \ldots, d\}$; thus each of the d variables takes one of m possible values.

Definition 6.12. *Let $X = (X_1, \ldots, X_d)$ be a discrete random vector with probability function $p(x) = \mathbb{P}(X = x) = \mathbb{P}(X_1 = x_1, \ldots, X_d = x_d)$ where $x = (x_1, \ldots, x_d)$. The probability mass function $p(x)$ is in log-linear form in case*

$$\log p(x) = \sum_{A \subset V} \psi_A(x_A) \tag{46}$$

with the constraints that ψ_Φ is a constant , and if $j \in A$ and $x_j = 0$ then $\psi_A(x_A) = 0$. The formula in (46) is called the *log-linear expansion* of $p(x)$. Each $\psi_A(x_A)$ may depend on some unknown parameters θ_A. Note that the total number of parameters satisfies $\sum_{j=1}^{d} \binom{d}{j}(m-1)^j = m^d$, however one of the parameters is the normalizing constant, and is determined by the constraint that the sum of the probabilities is one. Thus, there are $m^d - 1$ *free parameters*, and this is a minimal exponential parameterization of the multinomial. Let $\theta = (\theta_A : A \subset V)$ be the set of all these parameters. We will write $p(x) = p(x; \theta)$ when we want to emphasize the dependence on the unknown parameters θ.

The next theorem provides an easy way to read out conditional independence in a log-linear model.

Theorem 6.13. *Let (X_A, X_B, X_C) be a partition of $X = (X_1, \ldots, X_d)$. Then $X_B \perp\!\!\!\perp X_C \mid X_A$ if and only if all the ψ-terms in the log-linear expansion that have at least one coordinate in B and one coordinate in C are zero.*

Proof. From the definition of conditional independence, we know that $X_B \perp\!\!\!\perp X_C \mid X_A$ if and only if $p(x_A, x_B, x_C) = f(x_A, x_B)g(x_A, x_C)$ for some functions f and g.

Suppose that ψ_t is 0 whenever t has coordinates in B and C. Hence, ψ_t is 0 if $t \not\subseteq A \cup B$ or $t \not\subseteq A \cup C$. Therefore

$$\log p(x) = \sum_{t \subset A \cup B} \psi_t(x_t) + \sum_{t \subset A \cup C} \psi_t(x_t) - \sum_{t \subset A} \psi_t(x_t). \tag{47}$$

Exponentiating, we see that the joint density is of the form $f(x_A, x_B)g(x_A, x_C)$. Therefore $X_B \perp\!\!\!\perp X_C \mid X_A$. The reverse follows by reversing the argument. □

A *graphical log-linear model* with respect to a graph G is a log-linear model for which the parameters ψ_A satisfy $\psi_A(x_A) \neq 0$ if and only if A is a clique of G. Thus, a graphical log-linear model has potential functions on each clique, both maximal and non-maximal, with the restriction that $\psi_A(x_A) = 0$ in case $x_j = 0$ for any $j \in A$. In a *hierarchical log-linear model*, if $\psi_A(x_A) = 0$ then $\psi_B(x_B) = 0$ whenever $A \subset B$. Thus, the parameters in a hierarchical model are nested, in the sense that if a parameter is identically zero for some subset of variables, the parameter for supersets of those variables must also be zero. Every graphical log-linear model is hierarchical, but a hierarchical model need not be graphical; Such a relationship is shown in Figure 14 and is characterized by the next lemma.

Corollary 6.14. *A graphical log-linear model is hierarchical but the reverse need not be true.*

Proof. We assume there exists a model that is graphical but not hierarchical. There must exist two sets A and B, such that $A \subset B$ with $\psi_A(x_A) = 0$ and $\psi_B(x_B) \neq 0$. Since the model is graphical, $\psi_B(x_B) \neq 0$ implies that B is a clique. We then know that A must also be a clique due to $A \subset B$, which implies that $\psi_A(x_A) \neq 0$. A contradiction.

To see that a hierarchical model does not have to be graphical. We consider the following example. Let

$$\log p(x) = \psi_\Phi + \sum_{i=1}^{3} \psi_i(x_i) + \sum_{1 \leq j < k \leq 3} \psi_{jk}(x_{jk}). \tag{48}$$

This model is hierarchical but not graphical. The graph corresponding to this model is a complete graph with three nodes X_1, X_2, X_3. It is not graphical since $\psi_{123}(x) = 0$, which is contradict with the fact that the graph is complete. □

6.1. Overcomplete Representations of log-linear Models

While log-linear models are, by convention, minimal exponential family models, it is often more convenient to work with *overcomplete representation* (or non-minimal representations). Consider

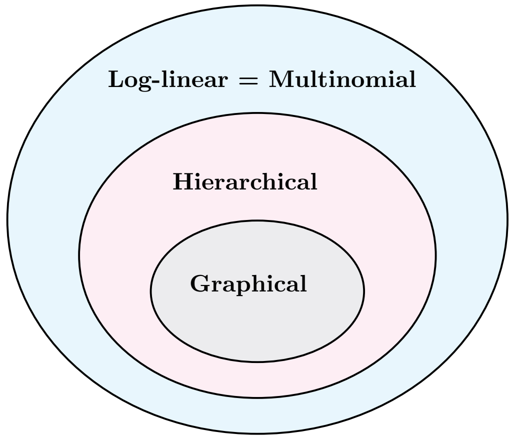


FIG 14. *Every graphical log-linear model is hierarchical but the reverse may not be true.*

a graph G, and consider the following exponential family model with parameters for each edge and vertex, where we assume that each random variable $X_s \in \{0, 1, \ldots, m-1\}$:

$$p_\theta(x) \propto \exp\left(\sum_{(s,t)\in E} \sum_{k,\ell=0}^{m-1} \theta_{s,t;k,\ell}\delta(x_s,k)\delta(x_t,\ell) + \sum_{s\in V}\sum_{k=0}^{m-1} \theta_{s;k}\delta(x_s,k)\right).$$

Here $\delta(x_s, k) = 1$ if $x_s = k$ and $\delta(x_s, k) = 0$ otherwise. This model has $m^2|E| + m|V|$ free parameters, where $|E|$ is the number of edges and $|V|$ is the number of vertices, compared with the number of free parameters in the corresponding hierarchical log-linear model of $(m-1)^2|E| + (m-1)|V|$. Thus, there are linear dependencies among the sufficient statistics; for instance,

$$\sum_{k,\ell} \delta(x_s, k)\, \delta(x_t, \ell) = 1 \tag{49}$$

for every edge $(s, t) \in E$. It is often more convenient to work in this overparameterized form. This can be viewed as a discrete version of the multivariate Gaussian.

6.2. Ising and Potts Models

The Ising and Potts models are special cases of hierarchical log-linear models that originated in statistical and solid state physics, but also arise naturally in image analysis, models of social networks, and other areas. The *Ising model* can be thought of as a discrete analogue of the Gaussian graphical model. In essence, an Ising model is a hierarchical model where the nodes have binary variables and the distribution has pairwise interactions. Thus, the model is hierarchical with nonzero parameters on vertices and cliques only. The *Potts model* is the extension of the Ising model to the case of $m \geq 2$ possible labels at each node.

In more detail, let $G = (V, E)$ be a graph, and let $X_s \in \{0, 1\}$ be a binary labeling of each node.

The Ising model takes the form

$$p(x;\theta) \;\;\propto\;\; \exp\left(\sum_{s\in V}\theta_s x_s + \sum_{(s,t)\in E}\theta_{s,t}x_s x_t\right). \tag{50}$$

For a given node $s \in V$, the conditional distribution, by the local Markov property, is given by a logistic regression model

$$\begin{aligned} p_s(x;\theta) \;\;&\equiv\;\; p(x_s = 1 \,|\, x_t, t \neq s; \theta) = p(x_s = 1 \,|\, x_t, (s,t) \in E) && (51)\\ &=\;\; \frac{\exp\left(\theta_s + \sum_{t:\,(s,t)\in E}\theta_{s,t}x_t\right)}{1+\exp\left(\theta_s + \sum_{t:\,(s,t)\in E}\theta_{s,t}x_t\right)}. && (52)\end{aligned}$$

The model is identifiable, and the Fisher information matrix at node $s \in V$ is given by

$$Q_s(\theta) = \mathbb{E}\left[p_s(X;\theta)(1-p_s(X;\theta))X_{V\setminus\{s\}}X_{V\setminus\{s\}}^T\right]. \tag{53}$$

Note that the matrix $Q_s(\theta)$ is the Fisher information matrix associated with the local conditional probability distribution. Intuitively, it serves as the counterpart for discrete graphical models of the covariance matrix $\mathbb{E}[XX^T]$ of Gaussian graphical models.

The m-dimensional *Potts model* allows $X_s \in \{0, 1, \ldots, m-1\}$, and has potential functions that test whether neighboring states are the same:

$$p(x;\theta) \;\;\propto\;\; \exp\left(\sum_{(s,t)\in E}\theta_{s,t}\delta(x_s, x_t)\right) \tag{54}$$

An "external field" can be incorporated through vertex potentials, so

$$p(x;\theta) \;\;\propto\;\; \exp\left(\sum_{s\in V}\theta_s x_s + \sum_{(s,t)\in E}\theta_{s,t}\delta(x_s, x_t)\right). \tag{55}$$

A simplified version of the model is used to study the properties of random colorings of a graph; here $X_s \in \{0, 1, \ldots, m-1\}$ denotes the color of node s, and $S(X) = \sum_{s,t}\delta(X_s, X_t)$ is the number of edges in the graph whose endpoints are colored with the same color. The associated Gibbs distribution is then

$$p(x;\theta) \;\;\propto\;\; \exp\left(-\beta J S(x)\right) \tag{56}$$

where $J \in \mathbf{R}$ and $\beta > 0$ is the "inverse temperature." Thus, the parameters in (54) satisfy $\theta_{s,t} = -\beta J$. When J is negative, higher probability is assigned to colorings X where the endpoints of edges have the same color. This is referred to as the *ferromagnetic Potts model*. If J is positive, however, then colorings X where the endpoints are identically colored have low probability; this is known as the *antiferromagnetic Potts model*. The parameter β is interpreted as inverse temperature: as $\beta \to \infty$ the temperature cools to zero. As the temperature cools, the antiferromagnetic model becomes more likely to sample a proper coloring of the graph, where no two adjacent nodes have the same color.

The Ising model corresponds to a Potts model with two colors, and the configurations can be thought of as weighted cuts in the graph, which divide the graph into two parts corresponding to the color of each node. When the temperature is low, and $J < 0$, with high probability the cut will be large. Since determining the largest cut in a graph is an NP-complete problem, it is (believed to be) not possible to efficiently sample from the Ising model in this regime. An interesting simulation of the Potts model with four states on a square two-dimensional lattice is shown in Figure 15.

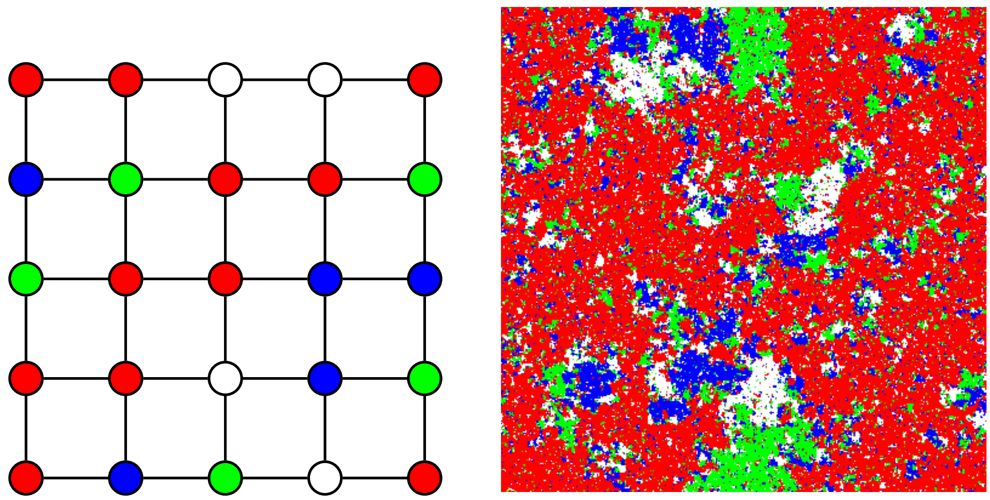


FIG 15. *An important example of an undirected graphical model is the simple 2-dimensional grid graph. This arises in image processing, for instance, where each node may represent a pixel in an image. Random fields (or probability distributions) on such graphs were first studied in statistical physics. The right plot shows a sample from a Potts model over four states on a 2-dimensional lattice at the critical temperature (D. Wilson).*

7. Structure Learning

In this section we describe methods that could estimate the undirected graphs based on observational samples. We are especially interested in the high dimensional settings. To handle high dimensions, the most common assumption is that G is sparse. This means that $\Omega = \Sigma^{-1}$ has many zeroes. (Alternatively, you can take the point of view that we are finding a sparse approximation to the true graph G.)

7.1. Structure Learning for Gaussian and Ising Models

In the high dimensional setting, we cannot estimate G unless we make some extra assumption. The most popular method for estimating G when the dimension d is large is to assume that X has a multivariate Gaussian distribution, $X \in N(\mu, \Sigma)$. From Theorem 5, we know that the graph is encoded by the sparsity pattern of the inverse covariance matrix $\Omega = [\omega_{ij}] = \Sigma^{-1}$.

Inspired by the success of the lasso, Meinshausen and Bühlmann (2006) proposed a parallel lasso approach to estimate the graph. They use the lasso to regress X_i on $(X_j : j \neq i)$. This is repeated for each X_i. Let

$$\beta^i = \text{argmin}_\beta \mathbb{E}(X_i - \sum_{j \neq i} \beta_j X_j)^2$$

and define the neighborhood of i, $N_i = \{j : \beta^i_j \neq 0\}$. The lasso gives estimates $\widehat{\beta}^i$ for all i. Let $\widehat{N}_i = \{j : \widehat{\beta}^i_j \neq 0\}$ and $\widehat{E}$ be the set of edges (i, j) such that $i \in \widehat{N}_j$ and $j \in \widehat{N}_i$. Under suitable sparsity assumptions they prove that $\mathbb{P}(N_i = \widehat{N}_i) \to 1$ as $n \to \infty$ even if $d = n^\gamma$ for some $\gamma > 0$. Similarly, $\mathbb{P}(E = \widehat{E}) \to 1$ as $n \to \infty$.

Alternatively, Yuan and Lin (2007) suggested a penalized likelihood estimator

$$\widehat{\Omega} = \arg\max_{\Omega \succ 0} \{\text{loglikelihood}(\Omega) - \lambda \sum_{j,k} |\omega_{jk}|\},$$

where the loglikelihood of Ω is evaluated under the Gaussian model. The estimator $\widehat{\Omega}$ can be efficiently computed using the glasso algorithm (Friedman et al., 2007; Banerjee et al., 2008), which is a block coordinate descent procedure that uses the standard lasso to estimate a single row and column of Ω in each iteration. In this subsection, we focus on introducing the glasso algorithm, which provides a good example of convex duality, and in particular how efficient algorithms can be derived by formulating the dual problem.

Suppose that we have n observations $x_1, \ldots, x_n$ where $x_i \in \mathbf{R}^d$. The maximum likelihood estimate of μ and Ω are obtained in terms of the sample mean and sample covariance,

$$\widehat{\mu}_n = \frac{1}{n} \sum_{i=1}^n x_i, \tag{57}$$

$$\widehat{\Omega}_n = S_n^{-1} \text{ where } S_n = \frac{1}{n} \sum_{i=1}^n (x_i - \widehat{\mu}_n)(x_i - \widehat{\mu}_n)^T. \tag{58}$$

In the following, we will simply assume that the mean μ is known. For $d < n$, the maximum likelihood estimate is easily obtained by noting that the log-likelihood of the data is

$$\ell(\Omega) = \frac{n}{2} \log |\Omega| - \frac{n}{2} \text{tr}(\Omega S_n) - \frac{nd}{2} \log(2\pi) \tag{59}$$

and the gradient of $\log |\mathcal{X}|$ as a function of $\mathcal{X} \in \mathcal{S}^d_{++}$ is $\nabla \log |\mathcal{X}| = \mathcal{X}^{-1}$.

If $d > n$, the maximum likelihood estimate is no longer valid because the matrix S_n is singular. We need to use regularized estimator. The negative log-likelihood, rescaled for convenience, is

$$f(\Omega) = \text{tr}(\Omega S_n) - \log |\Omega|. \tag{60}$$

Just as for the lasso, we can look for a sparse matrix Ω by imposing an L_1 penalty, leading to the optimization problem

$$\min_{\Omega \succ 0} \quad \text{tr}(\Omega S_n) - \log |\Omega| \tag{61}$$

$$\text{such that} \quad \|\Omega\|_1 \leq L, \tag{62}$$

where $\|\Omega\|_1 = \sum_{i,j} |\omega_{ij}|$. As is often the case, the dual form leads to insight into the optimization problem and its properties. To derive the dual, first note that we can write

$$\|\mathcal{X}\|_1 = \max_{\|U\|_\infty \leq 1} \text{tr}(\mathcal{X}U), \tag{63}$$

where $U = [u_{ij}]$ and $\|U\|_\infty = \max_{i,j} |u_{ij}|$. Therefore, the Lagrangian of the problem can be written as

$$\begin{aligned} \mathcal{L}(\Omega, \lambda) &= \text{tr}(\Omega S_n) - \log|\Omega| + \lambda\|\Omega\|_1 & (64)\\ &= \max_{\|U\|_\infty \leq \lambda} \Big\{\text{tr}(\Omega S_n) - \log|\Omega| + \text{tr}(\Omega U)\Big\} & (65)\\ &= \max_{\|U\|_\infty \leq \lambda} \Big\{\text{tr}(\Omega(S_n + U)) - \log|\Omega|\Big\}. & (66) \end{aligned}$$

The dual is thus

$$h(\lambda) = \min_{\Omega \succ 0} \max_{\|U\|_\infty \leq \lambda} \Big\{\text{tr}(\Omega(S_n + U)) - \log|\Omega|\Big\}. \tag{67}$$

Interchanging the min and max (which is justified by strong duality), we can then perform the minimization analytically, to obtain $\Omega = (S_n + U)^{-1}$, and thus

$$h(\lambda) = \max_{\|U\|_\infty \leq \lambda} \log|S_n + U| + d. \tag{68}$$

The log determinant acts as a barrier function that ensures $S_n + U$ is positive definite. Letting $\Sigma = S_n + U$, we can then re-express the dual optimization as estimating the covariance W according to

$$\widehat{\Sigma}_n = \mathop{\text{argmax}}_{W\,:\,\|W - S_n\|_\infty \leq \lambda} \log|W|. \tag{69}$$

Thus, the dual estimates the covariance matrix while the primal estimates the inverse covariance.

To carry out this dual optimization, Banerjee et al. (2008) propose a block-coordinate ascent algorithm, optimizing over one row and column of W at a time. To derive the algorithm, first note that

$$\widehat{\Sigma}_{jj} = S_{jj} + \lambda \tag{70}$$

(where we drop the sample size subscript on S_n for notational clarity). Thus, the diagonal is fixed in closed form. Suppose that we are optimizing over the jth row and column; then after reordering rows by moving the jth row into the first position, we can write the current estimate as

$$\widehat{\Sigma} = \begin{pmatrix} S_{jj} + \lambda & y^T \\ y & W_{\backslash jj} \end{pmatrix} \tag{71}$$

where $W_{\backslash jj}$ is the matrix W with the jth row and column removed. By *Schur complement*s, we have

$$\log|\widehat{\Sigma}| = \log|W_{\backslash jj}| + \log\left(S_{jj} + \lambda - y^T W_{\backslash jj}^{-1} y\right). \tag{72}$$

Therefore, optimizing over y is equivalent to the following quadratic program:

$$\begin{aligned} \min_{y} \quad & y^T W_{\setminus jj}^{-1} y && (73)\\ \text{such that} \quad & \|y - S_j\|_\infty \le \lambda && (74) \end{aligned}$$

where S_j is the jth column of S.

Now, we can recast this inner optimization problem into a more convenient form by forming the dual quadratic program, which is given as follows:

$$\min_x x^T W_{\setminus jj} x - S_j^T x + \lambda \|x\|_1, \tag{75}$$

where the mapping between these primal and dual subproblems is $y = W_{\setminus jj} x$. This is equivalent to a lasso regression problem. To see this, let $\mathbf{Q}$ be the square root $W_{\setminus jj}^{1/2}$ and let $b = \frac{1}{2}\mathbf{Q}^{-1} S_j$. Then the problem above is equivalent to

$$\min_x \|\mathbf{Q}x - b\|_2^2 + \lambda \|x\|_1. \tag{76}$$

Thus, we estimate the covariance by an iterated sequence of lasso problems. This can be implemented in an efficient manner using iterative soft thresholding. In particular, Equation (75) is solved iteratively by

$$x \leftarrow \mathcal{S}_\lambda^{(1)}\left(x + S_j - W_{\setminus jj} x\right) \tag{77}$$

where $\mathcal{S}_\lambda^{(1)}$ is the soft thresholding operator (assuming that $W_{\setminus jj}$ has norm less than or equal to one; the problem can be rescaled to ensure this is the case). After convergence, y in Equation (71) is updated with $W_{\setminus jj} x$. At each step, the inverse W^{-1} can be efficiently updated using Schur complements, to yield the estimate $\widehat{\Omega}_n$.

Figure 16 shows a gene-gene interaction graph estimated using this method. The data are based on Affymetrix GeneChip microarrays for the plant *Arabidopsis thaliana*, (Wille et al., 2004), with sample size is $n = 118$. The expression levels for each chip are pre-processed by log-transformation and standardization. A subset of 40 genes from the isoprenoid pathway are chosen, and the graphs for two different values of the regularization parameter λ are shown.

To handle discrete data, Ravikumar et al. (2010) proposed a parallel lasso based graph estimation procedure for high dimensional Ising models (or discrete Markov random fields). Their idea is similar to that of Meinshausen and Bühlmann (2006) but replacing the lasso with L_1-regularized logistic regression.

7.2. Learning Forest Graphical Models

Besides Gaussian and Ising models, another tractable family for structure learning is forest graphical models. If F is a d-node undirected forest with vertex set $V_F = \{1, \ldots, d\}$ and edge set $E_F \subset \{1, \ldots, d\} \times \{1, \ldots, d\}$, the number of edges satisfies $|E_F| < d$, noting that we do not

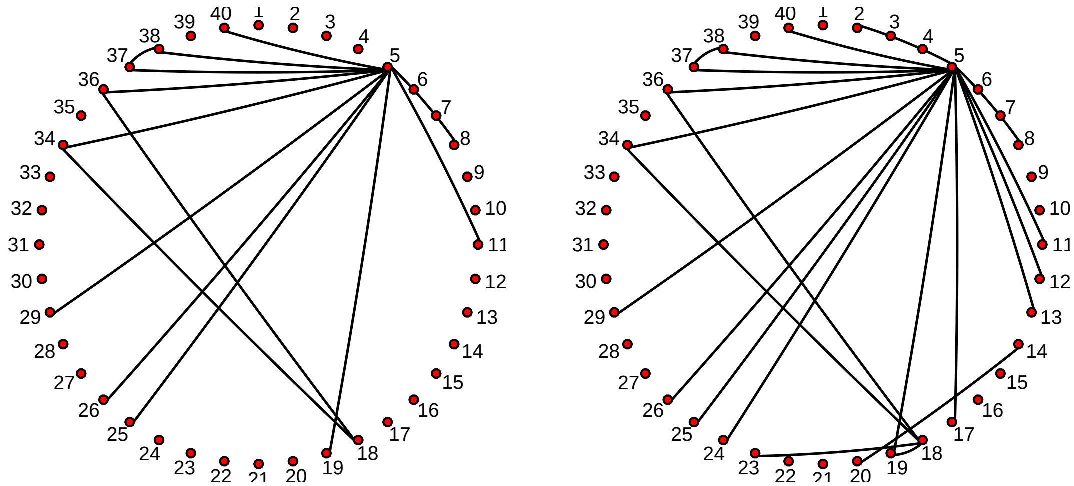


FIG 16. *Estimated gene-gene interaction graphs for the Arabidopsis thaliana data, for the regularization levels* $\lambda = 0.2448, 0.30857$.

restrict the graph to be connected. A probability density function $p(x)$ that is Markov to F can be written as

$$p(x) = \prod_{(i,j)\in E_F} \frac{p(x_i, x_j)}{p(x_i)\, p(x_j)} \prod_{k\in V_F} p(x_k), \tag{78}$$

where each $p(x_i, x_j)$ is a bivariate density, and each $p(x_k)$ is a univariate density. Using (78), we have

$$\begin{aligned}
&\mathbb{E}\log p(X) &\quad (79)\\
&= -\int p(x)\Big(\sum_{(i,j)\in E_F} \log\frac{p(x_i,x_j)}{p(x_i)p(x_j)} + \sum_{k\in V_F}\log\left(p(x_k)\right)\Big)dx \\
&= -\sum_{(i,j)\in E_F}\int p(x_i,x_j)\log\frac{p(x_i,x_j)}{p(x_i)p(x_j)}dx_i dx_j - \sum_{k\in V_F}\int p(x_k)\log p(x_k) dx_k \\
&= -\sum_{(i,j)\in E_F} I(X_i;X_j) + \sum_{k\in V_F} H(X_k), &\quad (80)
\end{aligned}$$

where

$$I(X_i; X_j) \equiv \int p(x_i, x_j)\log\frac{p(x_i,x_j)}{p(x_i)\,p(x_j)}\, dx_i dx_j \tag{81}$$

is the mutual information between the pair of variables X_i, X_j and

$$H(X_k) \equiv -\int p(x_k)\log p(x_k)\, dx_k \tag{82}$$

is the entropy.

The optimal forest F^* can be found by minimizing the right hand side of (80). Since the entropy term $H(X) = \sum_k H(X_k)$ is constant across all forests, this can be recast as the problem of finding the maximum weight spanning forest for a weighted graph, where the weight $w(i, j)$ of the edge connecting nodes i and j is $I(X_i; X_j)$. Kruskal's algorithm (Kruskal, 1956) is a greedy algorithm that is guaranteed to find a maximum weight spanning tree of a weighted graph. In the setting of density estimation, this procedure was proposed by Chow and Liu (1968) as a way of constructing a tree approximation to a distribution. At each stage the algorithm adds an edge connecting that pair of variables with maximum mutual information among all pairs not yet visited by the algorithm, if doing so does not form a cycle. When stopped early, after $k < d$ edges have been added, it yields the best k-edge weighted forest.

Of course, the above procedure is not practical since the true density $p(x)$ is unknown. In applications, we parameterize bivariate and univariate distributions to be $p_{\theta_{ij}}(x_i, x_j)$ and $p_{\theta_k}(x_k)$. We replace the population mutual information $I(X_i; X_j)$ in (80) by the plug-in estimate $\widehat{I}_n(X_i, X_j)$, defined as

$$\widehat{I}_n(X_i, X_j) = \int p_{\widehat{\theta}_{ij}}(x_i, x_j) \log \frac{p_{\widehat{\theta}_{ij}}(x_i, x_j)}{p_{\widehat{\theta}_i}(x_i)\, p_{\widehat{\theta}_j}(x_j)}\, dx_i dx_j \tag{83}$$

where $\widehat{\theta}_{ij}$ and $\widehat{\theta}_k$ are maximum likelihood estimates. Given this estimated mutual information matrix $\widehat{M} = \left[\widehat{I}_n(X_i, X_j)\right]$, we can apply Kruskal's algorithm (equivalently, the Chow-Liu algorithm) to find the best forest structure $\widehat{F}$. The detailed algorithm is described in the following:

Chow-Liu Algorithm for Learning Forest Graphs

Initialize $E^{(0)} = \emptyset$ and the desired forest size $K < d$.
Calculate the mutual information matrix $\widehat{M} = \left[\widehat{I}_n(X_i, X_j)\right]$ according to (83).
For $k = 1, \ldots, K$

(a) $(i^{(k)}, j^{(k)}) \leftarrow \text{argmax}_{(i,j)}\, \widehat{M}(i, j)$ such that $E^{(k-1)} \cup \{(i^{(k)}, j^{(k)})\}$ does not contain a cycle.
(b) $E^{(k)} \leftarrow E^{(k-1)} \cup \{(i^{(k)}, j^{(k)})\}$.

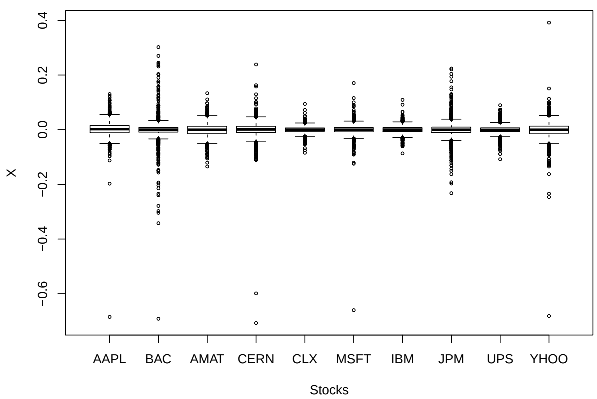
Output the obtained edge set $E^{(K)}$.

Example 7.15. [Learning Gaussian maximum weight spanning tree] For Gaussian data $X \sim N(\mu, \Sigma)$, we know that the mutual information between two variables are

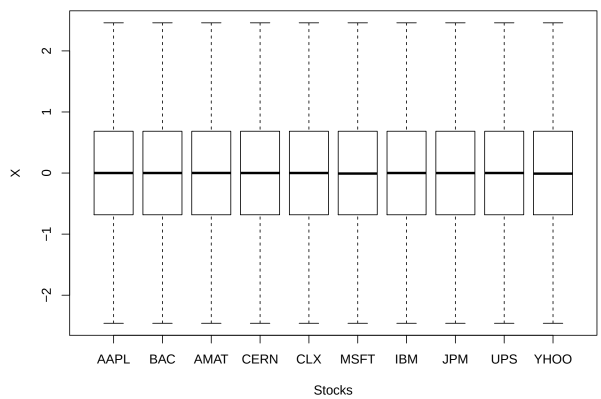
$$I(X_i; X_j) = -\frac{1}{2} \log\left(1 - \rho_{ij}^2\right), \tag{84}$$

where ρ_{ij} is the correlation between X_i and X_j. To obtain an empirical estimator, we simply plug-in the sample correlation $\widehat{\rho}_{ij}$. Once the mutual information matrix is calculated, we could apply the Chow-Liu algorithm to get the maximum weight spanning tree.

Example 7.16. [Graphs for Equities Data] We collect the daily closing prices were obtained for 452 stocks that were consistently in the S&P 500 index between January 1, 2003 through January 1, 2011. This gave us altogether 2,015 data points, each data point corresponds to the vector of closing prices on a trading day. With $S_{t,j}$ denoting the closing price of stock j on day t, we consider the variables $X_{tj} = \log(S_{t,j}/S_{t-1,j})$ and build graphs over the indices j. We simply treat the instances X_t as independent replicates, even though they form a time series. We truncate every stock so that its data points are within six times the mean absolute deviation from the sample average. In Figure 17(a) we show boxplots for 10 randomly chosen stocks. It can be seen that the data contains outliers even after truncation; the reasons for these outliers includes splits in a stock, which increases the number of shares. In Figure 17(b) we show the boxplots of the data after the nonparanormal transformation (the details of nonparanormal transformation will be explained in the nonparametric graphical model chapter). In this analysis, we use the subset of the data between January 1, 2003 to January 1, 2008, before the onset of the "financial crisis." There are altogether $n = 1,257$ data points and $d = 452$ dimensions.



(a) original data



(b) after nonparanormal transformation

FIG 17. *Boxplots of $X_t = \log(S_t/S_{t-1})$ for 10 stocks. As can be seen, the original data has many outliers, which is addressed by the nonparanormal transformation on the re-scaled data (right).*

The 452 stocks are categorized into 10 Global Industry Classification Standard (GICS) sectors, including `Consumer Discretionary` (70 stocks), `Energy` (37 stocks), `Financials` (74 stocks), `Consumer Staples` (35 stocks), `Telecommunications Services` (6 stocks), `Health Care` (46 stocks), `Industrials` (59 stocks), `Information Technology` (64 stocks), `Materials` (29 stocks), and `Utilities` (32 stocks). It is expected that stocks from the same GICS sectors should tend to be clustered together, since stocks from the same GICS sector tend to interact more with each other. In the graphs shown below, the nodes are colored according to the GICS sector of the corresponding stock.

With Gaussian assumption, we directly apply Chow-Liu algorithm to obtain a full spanning tree of $d - 1 = 451$ edges. The resulting graph is shown in Figure 18. We see that the stocks from the same GICS sector are clustered very well.

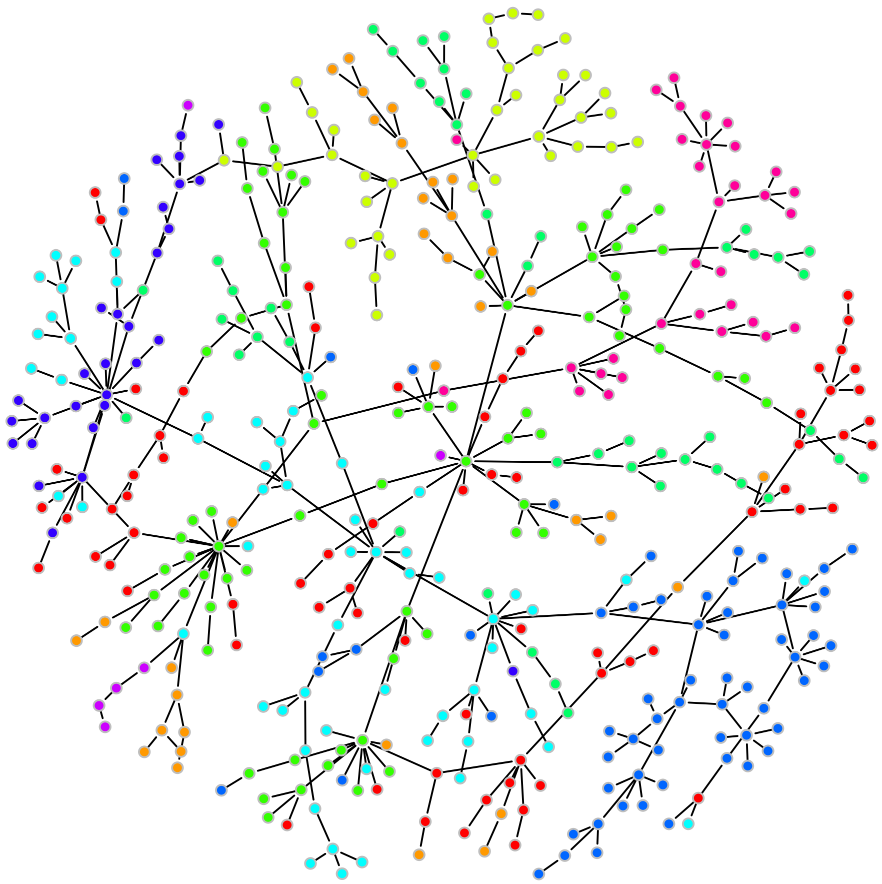


FIG 18. *Tree graph learned from S&P 500 stock data from Jan. 1, 2003 to Jan. 1, 2008. The graph is estimated using the Chow-Liu algorithm under the Gaussian model. The nodes are colored according to their GICS sector categories.*

8. Bibliographic Remarks

Textbooks on undirected graphical models include Whittaker (1990), Edwards (1995), Lauritzen (1996), and Jordan (1999). Nice treatments on undirected graphical model inference appears in Bishop (2007). More details about the graphical lasso algorithm can be found in Hastie et al. (2009). A thorough treatment of undirected graphical models can be found in Koller and Friedman (2009). Some discussions on chordal graphs and junction trees are drawn from an unpublished notes from Peter Bartlett.

9. Exercises

1. Prove Theorem 2.2.

2. Prove Theorem 3.1.

3. State and prove the continuous version of Hammersley-Clifford theorem.

4. Prove that the undirected graph (the square) of Figure 12 represents a family of probability distributions that cannot be represented by a directed graph on the same set of vertices.

5. Consider random variables (X_1, X_2, X_3, X_4). Suppose the log-density is

$$\log p(x) = \psi_\Phi + \psi_{12}(x_1, x_2) + \psi_{13}(x_1, x_3) + \psi_{24}(x_2, x_4) + \psi_{34}(x_3, x_4). \tag{85}$$

   (a) Draw the graph G for these variables.

   (b) Write down all independence and conditional independence relationships implied by the graph.

   (c) Is this model graphical? Is it hierarchical?

6. Suppose that the parameters $p(x_1, x_2, x_3)$ are proportional to the following values:

$$\mathbb{P}(0,0,0) = 2,\ \ \mathbb{P}(0,0,1) = 8,\ \ \mathbb{P}(0,1,0) = 4,\ \ \mathbb{P}(0,1,1) = 16, \tag{86}$$
$$\mathbb{P}(1,0,0) = 16,\ \ \mathbb{P}(1,0,1) = 128,\ \ \mathbb{P}(1,1,0) = 32,\ \ \mathbb{P}(1,1,1) = 256. \tag{87}$$

   Find the ψ-terms for the log-linear expansion. Comment on the model.

7. Let $X_1, \ldots, X_4$ be binary. Draw the independence graphs corresponding to the following log-linear models (where $\alpha > 0$). Also, identify whether each is graphical and/or hierarchical (or neither).

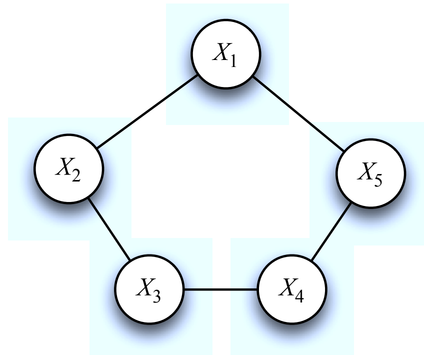
   (a) $\log p(x) = \alpha + 11x_1 + 2x_2 + 1.5x_3 + 17x_4$

   (b) $\log p(x) = \alpha + 11x_1 + 2x_2 + 1.5x_3 + 17x_4 + 12x_2x_3 + 78x_2x_4 + 3x_3x_4 + 32x_2x_3x_4$

   (c) $\log p(x) = \alpha + 11x_1 + 2x_2 + 1.5x_3 + 17x_4 + 12x_2x_3 + 3x_3x_4 + x_1x_4 + 2x_1x_2$

   (d) $\log p(x) = \alpha + 5055x_1x_2x_3x_4$.

8. This problem is based on the following graph:



   (a) Can you construct a DAG that has the same conditional independencies of the form $p(x_i \,|\, x_j,\ j \neq i) = p(x_i \,|\, x_{i_1}, x_{i_2})$ as those implied by the above graph?

   (b) For each of the following families of distributions, list any independence relations, if any, that are implied *in addition to* the independence relations for a general distribution that is Markov to the above graph.

      (i) $p(x) = \frac{1}{Z}\psi_{12}(x_1, x_2)\,\psi_{23}(x_2, x_3)\,\psi_{34}(x_3, x_4)\,\psi_{45}(x_4, x_5)\,\psi_{51}(x_5, x_1)$

      (ii) $p(x) = \frac{1}{Z}\psi_{12}(x_1, x_2)\,\psi_{23}(x_2, x_3)\,\psi_{34}(x_3, x_4)\,\psi_{45}(x_4, x_5)$

      (iii) $p(x) = \frac{1}{Z}\psi_{12}(x_1, x_2)\,\psi_{34}(x_3, x_4)\,\psi_5(x_5)$

   (c) Now suppose that each X_i is *binary*, so that $X_i \in \{0, 1\}$, and consider the following family of distributions:

$$p(x) \ \propto\ \psi_\theta(x_1, x_2)\,\psi_\theta(x_2, x_3)\,\psi_\theta(x_3, x_4)\,\psi_\theta(x_4, x_5)\,\psi_\theta(x_5, x_1)$$

   where $\psi_\theta(x, y) = \theta^{xy}$ with $\theta > 0$. For this family of models, calculate the conditional probability $P(X_1 = 1 \,|\, X_3 = 1, X_4 = 1)$ as a function of θ.

9. Let $X_1, \ldots, X_6$ be random variables with joint distribution of the form

$$p(x) \ \propto\ f(x_1, x_2, x_3)\, g(x_3, x_4)\, g(x_1, x_5)\, g(x_4, x_5)\, f(x_1, x_5, x_6)$$

   where $f : \mathbf{R}^3 \longrightarrow \mathbf{R}_+$ and $g : \mathbf{R}^2 \longrightarrow \mathbf{R}_+$ are arbitrary non-negative functions.

   (a) What is the graph with the fewest edges that represents the independence relations for this family of distributions?

   (b) For each of the following sets of variables $C \subset \{X_1, \ldots, X_6\}$, give non-empty sets of variables A and B such that $A \perp\!\!\!\perp B \,|\, C$.

      (i) $C = \{X_1, X_4\}$

      (ii) $C = \{X_1, X_3\}$

      (ii) $C = \{X_1, X_3, X_4\}$

   (c) Now suppose that $X_i \in \{0, 1\}$ are binary and that

$$f(x, y, z) = \alpha^{xyz} \quad g(x, y) = \alpha^{xy}$$

   Find the conditional probability $\mathbb{P}(X_4 = 1 \,|\, X_1 = 1, X_3 = 1)$.

10. Let $X = (X_1, X_2, X_3, X_4, X_5)$ be a random vector distributed as $X \sim N(0, \Sigma)$ where the covariance matrix Σ is given by

$$\Sigma = \frac{1}{15}\begin{pmatrix} 9 & -3 & -3 & -3 & -3 \\ -3 & 6 & 1 & 1 & 1 \\ -3 & 1 & 6 & 1 & 1 \\ -3 & 1 & 1 & 6 & 1 \\ -3 & 1 & 1 & 1 & 6 \end{pmatrix} \quad \text{with inverse} \quad \Sigma^{-1} = \begin{pmatrix} 3 & 1 & 1 & 1 & 1 \\ 1 & 3 & 0 & 0 & 0 \\ 1 & 0 & 3 & 0 & 0 \\ 1 & 0 & 0 & 3 & 0 \\ 1 & 0 & 0 & 0 & 3 \end{pmatrix}.$$

(a) What is the graph for X, viewed as an undirected graphical model?

(b) Which of the following independence statements are true?

i. $X_2 \perp\!\!\!\perp X_3 \mid X_1$

ii. $X_3 \perp\!\!\!\perp X_4$

iii. $X_1 \perp\!\!\!\perp X_3 \mid X_2$

iv. $X_1 \perp\!\!\!\perp X_5$

(c) List the local Markov properties for this graphical model.

(d) Find the conditional density $p(x_2 \mid X_1 = -3)$.

11. Consider a chain graph X_1—X_2—X_3—X_4—X_5, we assume all variables are binary. One log-linear model that is consistent with this graph is

$$\log p(x) = \beta_0 + 5\big(x_1x_2 + x_2x_3 + x_3x_4 + x_4x_5\big).$$

Simulate $n = 100$ random vectors from this distribution. Fit the model

$$\log p(x) = \beta_0 + \sum_j \beta_j x_j + \sum_{k<\ell} \beta_{k\ell} x_k x_\ell$$

using maximum likelihood. Report your estimators. Use forward model selection with BIC to choose a submodel. Compare the selected model to the true model.

12. Let $X \sim N(0, \Sigma)$ where $X = (X_1, \ldots, X_d)^T$, $d = 10$. Let $\Omega = \Sigma^{-1}$ and suppose that $\Omega(i,i) = 1$, $\Omega(i, i-1) = .5$, $\Omega(i-1, i) = .5$ and $\Omega(i,j) = 0$ otherwise. Simulate 50 random vectors and use the glasso method to estimate the covariance matrix. Compare your estimated graph to the true graph.

13. Let $X = (X_1, X_2, X_3, X_4)$ be a random vector satisfying

$$f(X) \sim N(0, \Sigma)$$

where $f(x) = (x_1^3, x_2^3, x_3^3, x_4^3)$ and the covariance matrix Σ is given by

$$\Sigma = \frac{1}{6}\begin{pmatrix} -4 & -2 & -2 & -2 \\ -2 & 4 & 1 & 1 \\ -2 & 1 & 4 & 1 \\ -2 & 1 & 1 & 4 \end{pmatrix}, \quad \text{with inverse} \quad \Sigma^{-1} = \begin{pmatrix} 3 & 1 & 1 & 1 \\ 1 & 2 & 0 & 0 \\ 1 & 0 & 2 & 0 \\ 1 & 0 & 0 & 2 \end{pmatrix}.$$

(a) Give an expression for the density of X.

(b) What is the graph for X, viewed as a graphical model?

(c) List the local Markov properties for this graphical model.

14. Let $X = (X_1, \ldots, X_d)$ where each $X_j \in \{0, 1\}$. Consider the log-linear model

$$\log p(x) = \beta_0 + \sum_{j=1}^{d} \beta_j x_j + \sum_{j<k}^{d} \beta_{jk} x_j x_k + \cdots + \sum_{j<k<\ell}^{d} \beta_{jk\ell} x_j x_k x_\ell + \cdots +$$

Suppose that $\beta_A = 0$ whenever $\{1, 2\} \subset A$. Show that $X_1 \perp\!\!\!\perp X_2 \,|\, X_3, \ldots, X_d$.

15. Let $X = (X_1, \ldots, X_d) \in \mathbb{R}^d$ be a random vector, with bivariate marginal densities $p_{ij}(x_i, x_j) > 0$ and univariate marginal densities $p_i(x_i)$. Let $G = (V, E)$ be a tree graph on $\{1, \ldots, d\}$, so that G does not contain any cycles. Consider the family of functions

$$f_m(x_1, \ldots, x_d) = \prod_{i=1}^{d} \theta_i(x_i)^{m_i} \prod_{(i,j) \in E} \theta_{ij}(x_i, x_j),$$

where $m_i \in \mathbb{Z}$ are integers. Find a set of integers $m_1, \ldots, m_d \in \mathbb{Z}$ for which the function f_m is a probability density; i.e., f_m is nonnegative and integrates to one.

16. Given $X = (Y, Z) \in \mathbb{R}^6$, let $X \sim N(0, \Sigma)$ be a random Gaussian vector where $Y = (Y_1, Y_2) \in \mathbf{R}^2$ and $Z = (Z_1, Z_2, Z_3, Z_4) \in \mathbf{R}^4$, with $\Sigma^{-1} = \Omega = \begin{pmatrix} A & B \\ B^T & C \end{pmatrix}$ where

$$A = \begin{bmatrix} 2 & 0 \\ 0 & 2 \end{bmatrix} \quad B = \begin{bmatrix} 1 & \frac{1}{2} & \frac{1}{3} & \frac{1}{4} \\ -1 & \frac{1}{2} & -\frac{1}{3} & \frac{1}{4} \end{bmatrix} \quad C = \begin{bmatrix} 2 & \frac{1}{2} & 0 & 0 \\ \frac{1}{2} & 2 & 0 & 0 \\ 0 & 0 & 2 & \frac{1}{2} \\ 0 & 0 & \frac{1}{2} & 2 \end{bmatrix}.$$

(a) Draw the undirected graph of X.

(b) Draw the undirected graph of Z. Hint: Recall that

$$\begin{bmatrix} A & B \\ B^T & C \end{bmatrix}^{-1} = \begin{bmatrix} A^{-1} + A^{-1} B S^{-1} B^T A^{-1} & -A^{-1} B S^{-1} \\ -S^{-1} B^T A^{-1} & S^{-1} \end{bmatrix}$$

where $S = C - B^T A^{-1} B$ is the Schur complement.

(c) Which of the following independence statements hold?

1. $Y_1 \perp\!\!\!\perp Y_2 \,|\, Z$
2. $Z_1 \perp\!\!\!\perp Z_4 \,|\, Z_2$
3. $Z_1 \perp\!\!\!\perp Z_4 \,|\, Y_1$
4. $Z_1 \perp\!\!\!\perp Z_2$.

References

Akkoyunlu, E. A. (1973). The enumeration of maximal cliques of large graphs. *SIAM Journal on Computing*, 2:1–6.

Banerjee, O., Ghaoui, L. E., and d'Aspremont, A. (2008). Model selection through sparse maximum likelihood estimation. *Journal of Machine Learning Research*, 9:485–516.

Besag, J. (1974). Spatial interaction and the statistical analysis of lattice systems. *Journal of the Royal Statistical Society, Series B (Methodological)*, 36(2):192–236.

Bishop, C. M. (2007). *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer, 1st ed. 2006. corr. 2nd printing edition.

Bomze, I. M., Budinich, M., Pardalos, P. M., and Pelillo, M. (1999). The maximum clique problem. *Handbook of Combinatorial Optimization*, 4:1–74.

Chow, C. and Liu, C. (1968). Approximating discrete probability distributions with dependence trees. *Information Theory, IEEE Transactions on*, 14(3):462–467.

Edwards, D. (1995). *Introduction to graphical modelling*. Springer-Verlag Inc.

Friedman, J. H., Hastie, T., and Tibshirani, R. (2007). Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, 9(3):432–441.

Hastie, T., Tibshirani, R., and Friedman, J. H. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction (Second Edition)*. Springer-Verlag.

Jordan, M. I. E. (1999). *Learning in Graphical Models*. MIT Press.

Koller, D. and Friedman, N. (2009). *Probabilistic Graphical Models - Principles and Techniques*. MIT Press.

Kruskal, J. B. (1956). On the shortest spanning subtree of a graph and the traveling salesman problem. *Proceedings of the American Mathematical Society*, 7(1):48–50.

Lauritzen, S. L. (1996). *Graphical Models*. Oxford University Press.

Maslov, S. and Sneppen, K. (2002). Specificity and stability in topology of protein networks. *Science*, 296:910–913.

Meinshausen, N. and Bühlmann, P. (2006). High dimensional graphs and variable selection with the lasso. *Annals of Statistics*, 34(3).

Moody, J. (2001). Race, school integration and friendship segregation in America. *American Journal of Sociology*, 107:679–716.

Pearl, J. and Paz, A. (1986). Graphoids: Graph-based logic for reasoning about relevance relations or when would x tell you more about y if you already know z? In *ECAI*, pages 357–363.

Ravikumar, P., Wainwright, M. J., and Lafferty, J. (2010). High-dimensional ising model selection using l1-regularized logistic regression. *Annals of Statistics*, 38:1287–1319.

Stern, D., Graepel, T., and MacKay, D. (2004). Modeling uncertainty in the game of Go. In *Advances in Neural Information Processing Systems*, volume 16.

Whittaker, J. (1990). *Graphical Models in Applied Multivariate Statistics*. Wiley.

Wille, A., Zimmermann, P., Vranová, E., Fürholz, A., Laule, O., Bleuler, S., Hennig, L., Prelić, A., von Rohr, P., Thiele, L., Zitzler, E., Gruissem, W., and Bühlmann, P. (2004). Sparse Gaussian graphical modelling of the isoprenoid gene network in *Arabidopsis thaliana*. *Genome Biology*, 5:R92.

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