Undirected Graphical Models
36-708

Contents

Graphical models are a way of representing the relationships between features (variables). There are two main brands: directed and undirected. We shall focus on undirected graphical models. See Figure 1 for an example of an undirected graph.

Undirected graphs come in different flavors, such as:

1. Marginal Correlation Graphs.
2. Partial Correlation Graphs.
3. Conditional Independence Graphs.

In each case, there are parametric and nonparametric versions.

Let $X_1, \ldots, X_n \sim P$ where $X_i = (X_i(1), \ldots, X_i(d))^T \in \mathbb{R}^d$. The vertices (nodes) of the graph refer to the d features. Each node of the graph corresponds to one feature. Edges represent relationships between the features. The graph is represented by $G = (V, E)$ where $V = (V_1, \ldots, V_d)$ are the vertices and E are the edges. We can regard the edges E as a $d \times d$ matrix where $E(j, k) = 1$ if there is an edge between feature j and feature k and 0 otherwise. Alternatively, you can regard E as a list of pairs where $(j, k) \in E$ if there is an edge between j and k. We write

$$X \amalg Y$$

to mean that X and Y are independent. In other words, $p(x, y) = p(x)p(y)$. We write

$$X \amalg Y \mid Z$$

to mean that X and Y are independent given Z. In other words, $p(x, y|z) = p(x|z)p(y|z)$.

1 Marginal Correlation Graphs

In a marginal correlation graph (or association graph) we put an edge between V_j and V_k if $|\rho(j, k)| \geq \epsilon$ where $\rho(j, k)$ is some measure of association. Often we use $\epsilon = 0$ in which case there is an edge iff $\rho(j, k) \neq 0$. We also write $\rho(X_j, X_k)$ to mean the same as $\rho(j, k)$.

The parameter $\rho(j, k)$ is required to have the following property:

$$X \amalg Y \quad \text{implies that} \quad \rho(X, Y) = 0.$$

In general, the reverse may not be true. We will say that ρ is *strong* if

$$X \amalg Y \quad \text{if and only if} \quad \rho(X, Y) = 0.$$

We would lile ρ to have several properties: easy to compute, robust to outliers and there is some way to calculate a confidence interval for the parameter. Here is a summary of the association measures we will consider:

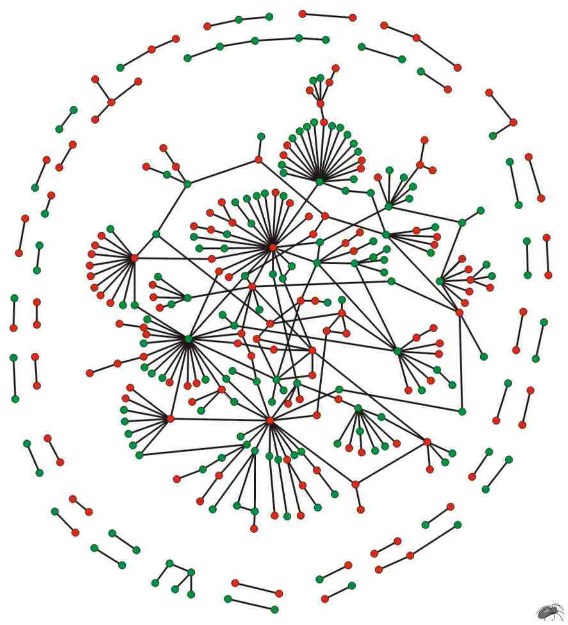


Figure 1: *A Protein network. From: Maslov and Sneppen (2002). Specificity and Stability in Topology of Protein Networks.* Science, *296, 910-913.*

	Strong	Robust	Fast	Confidence Interval
Pearson	$\times$	$\times$	$\surd$	$\surd$
Kendall	$\times$	$\surd$	$\surd$	$\surd$
Dcorr	$\surd$	$\times$	$\times$	sort of
τ^*	$\surd$	$\surd$	$\times$	$\surd$

Pearson Correlation. A common choice of ρ is the Pearson correlation. For two variables X and Y te Pearson correlation is

$$\rho(X, Y) = \frac{\mathrm{Cov}(X, Y)}{\sigma_X \sigma_Y} \tag{1}$$

. The sample estimate is

$$r(X, Y) = \frac{\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{s_X s_Y}.$$

When dealing with d feaures $X(1), \ldots, X(d)$, we write $\rho(j, k) \equiv \rho(X(j), X(k))$. The sample correlation is denoted by r_{jk}.

To test $H_0 : \rho(j, k) = 0$ versus $H_1 : \rho(j, k) \neq 0$ we can use an asymptotic test or an exact test.

The asymptotic test works like this. Define

$$Z_{jk} = \frac{1}{2}\log\left(\frac{1+r_{jk}}{1-r_{jk}}\right).$$

Fisher proved that

$$Z_{jk} \approx N\left(\theta_{jk}, \frac{1}{n-3}\right)$$

where

$$\theta_{jk} = \frac{1}{2}\log\left(\frac{1+\rho_{jk}}{1-\rho_{jk}}\right).$$

We reject H_0 if $|Z_{jk}| > z_{\alpha/2}/\sqrt{n-3}$. In fact, to control for multiple testing, we should reject when $|Z_{jk}| > z_{\alpha/(2m)}/\sqrt{n-3}$ where $m = \binom{d}{2}$. The confidence interval is $C_n = [a, b]$ where $a = \exp(Z_{jk} - z_{\alpha/2}/\sqrt{n-3})$ and $b = \exp(Z_{jk} + z_{\alpha/2}/\sqrt{n-3})$. A simultaneous confidence set for all the correlations can be obtained using the high dimensional bootstrap which we decribe later.

An exact test can be obtained by using a permutation test. Permute one of the variables and recompute the correlation. Repeat B times to get $r_{jk}^1, \ldots, r_{jk}^B$. The p-value is

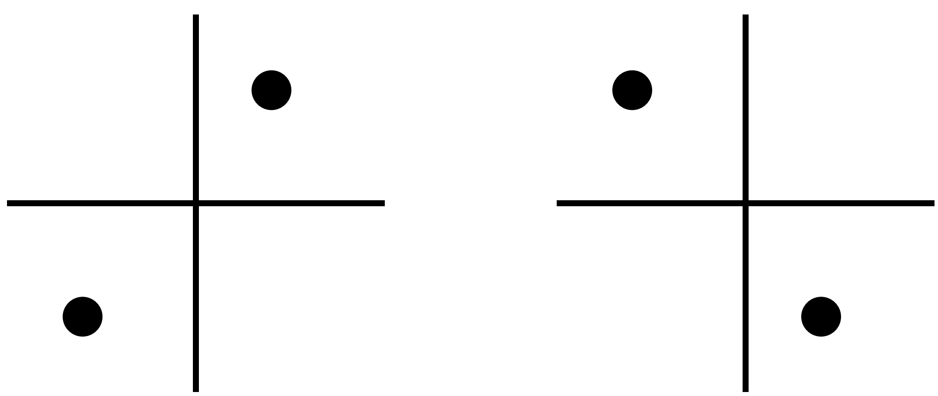
$$p = \frac{1}{B}\sum_s I(|r_{jk}^s| \geq |r_{jk}|).$$

Reject if $p \leq \alpha/m$.

Kendall's τ. The Pearson correlation is not very robust to outliers. A more robust measure of association is Kendall's tau defined by

$$\tau(X, Y) = \mathbb{E}\Bigg[\text{sign}\big[(X_1 - X_2)(Y_1 - Y_2)\big]\Bigg].$$

Kendall's τ can be intepreted as: probabilty(concordant) - probabilty(disconcordant). See this plot:



Concordant Pair　　Discordant Pair

τ can be estimated by

$$\widehat{\tau}(X,Y) = \frac{1}{\binom{n}{2}} \sum_{s\neq t} \Bigg[\text{sign}[(X_s - X_t)(Y_s - Y_t)] \Bigg].$$

A statistic of this form is called a U-statistic. Under H_0, $\widehat{\tau}_{jk} \approx N(0, 4/(9n))$ so we reject when $\sqrt{9n/4}|\widehat{\tau}_{jk}| > z_{\alpha/2m}$. Alternatively, use the permutation test.

Distance Correlation. There are various nonparametric measures of association. The most common are the distance correlation and the RKHS correlation. The squared *distance covariance* between two random vectors X and Y is defined by (Szekely et al 2007)

$$\gamma^2(X,Y) = \text{Cov}(||X - X'||, ||Y - Y'||) - 2\text{Cov}(||X - X'||, ||Y - Y''||) \tag{2}$$

where $(X,Y), (X',Y')$ and (X'',Y'') are independent pairs. We can write this as

$$\gamma^2(X,Y) = \frac{1}{4}\mathbb{E}[b(X_1,X_2,X_3,X_4)b(Y_1,Y_2,Y_3,Y_4)]$$

where

$$b(z_1,z_2,z_3,z_4) = |z_1 - z_2| + |z_3 - z_4| - |z_1 - z_3| - |z_2 - z_4|$$

The distance correlation is

$$\rho^2(X,Y) = \frac{\gamma^2(X,Y)}{\sqrt{\gamma^2(X,X)\gamma^2(Y,Y)}}.$$

It can be shown that

$$\gamma^2(X,Y) = \frac{1}{c_1 c_2} \int \frac{|\phi_{X,Y}(s,t) - \phi_X(s)\phi_Y(t)|^2}{||s||^{1+d}||t||^{1+d}}\, ds\ dt \tag{3}$$

where c_1, c_2 are constants and ϕ denotes the characteristic function. Another expression (Lyons 2013) for γ is

$$\gamma^2(X,Y) = \mathbb{E}[\delta(X,X')\delta(Y,Y')]$$

where

$$\delta(X,X') = d(X,X') - 2\int d(X,u)dP(u) + \int\int d(u,v)dP(u)dP(v)$$

and $d(x,y) = ||x - y||$. In fact, other metrics d can be used.

Lemma 1 *We have that* $0 \leq \rho(X,Y) \leq 1$ *and* $\rho(X,Y) = 0$ *if and only if* $X \amalg Y$.

An estimate of γ is

$$\widehat{\gamma}^2(X,Y) = \frac{1}{n^2}\sum_{j,k} A_{jk}B_{jk}$$

where

$$A_{jk} = a_{jk} - a_{j\cdot} - a_{\cdot k} + a_{\cdot\cdot}, \quad B_{jk} = b_{jk} - b_{j\cdot} - b_{\cdot k} + b_{\cdot\cdot}.$$

Here, $a_{jk} = ||X_j - X_k||$ and $a_{j\cdot}, a_{\cdot k}, a_{\cdot\cdot}$ are the row, column and grand means of the matrix $\{a_{jk}\}$. The limiting distribution of $\widehat{\gamma}^2(X,Y)$ is complicated. But we can easily test $H_0 : \gamma(X,Y) = 0$ using a permutation test.

Another nonparametric measure of independence based on RKHS is

$$\begin{aligned}\gamma^2(X,Y) &= \mathbb{E}[K_h(X,X')K_h(Y,Y')] + \mathbb{E}[K_h(X,X')]\mathbb{E}[K_h(Y,Y')] \\ &\quad - 2\mathbb{E}\left[\int K_h(X,u)dP(u)\int K_h(Y,v)dP(v)\right]\end{aligned}$$

for a kernel K_h. See Gretton et al (2008).

To apply any of these methods to graphs, we need to test all $\binom{d}{2}$ correlations.

The Bergsma-Dassios τ^* Correlation. Bergsma and Dassios (2014) extended Kendall's τ into a strong correlation. The definition is

$$\tau^*(X,Y) = \mathbb{E}[a(X_1,X_2,X_3,X_4)a(Y_1,Y_2,Y_3,Y_4)] \tag{4}$$

where

$$a(z_1,z_2,z_3,z_4) = \text{sign}(|z_1 - z_2| + |z_3 - z_4| - |z_1 - z_3 - |z_2 - z_4|).$$

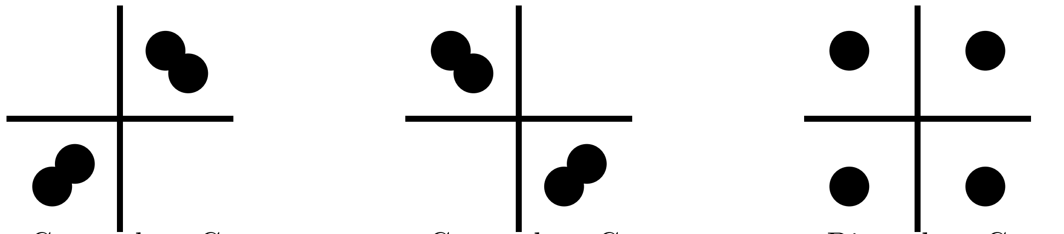
Lemma 2 $\tau^*(X,Y) \geq 0$. *Further,* $\tau^*(X,Y) = 0$ *if and only if* $X \amalg Y$.

An estimate of τ^* is

$$\widehat{\tau}^* = \frac{1}{\binom{n}{4}}\sum a(X_i,X_j,X_k,X_\ell)a(Y_i,Y_j,Y_k,Y_\ell) \tag{5}$$

where the sum is over all distinct quadruples.

The τ^* parameter can also be given an interpretation in terms of concordant and discordant points if we define them as follows:



Concordant Group Concordant Group Discordant Group

Then

$$\tau^* = \frac{2P(\text{concordant}) - P(\text{discordant})}{3}.$$

This statistic is related to the distance covariance as follows:

$$\tau^*(X,Y) = \mathbb{E}[a(X_1,X_2,X_3,X_4)a(Y_1,Y_2,Y_3,Y_4)]$$
$$\gamma^2(X,Y) = \frac{1}{4}\mathbb{E}[b(X_1,X_2,X_3,X_4)b(Y_1,Y_2,Y_3,Y_4)]$$

where

$$b(z_1,z_2,z_3,z_4) = |z_1 - z_2| + |z_3 - z_4| - |z_1 - z_3| - |z_2 - z_4|$$

and $a(z_1,z_2,z_3,z_4) = \text{sign}(b(z_1,z_2,z_3,z_4))$.

To test $H_0 : X \amalg Y$ we use a permutation test. Recently Dhar, Dassios and Bergsma (2016) showed that $\widehat{\tau}^*$ has good power and is quite robust.

Confidence Intervals. Constructing confidence intervals for γ is not easy. The problem is that the statistics have different limiting distributions depending on whether the the null $H_0 : X \amalg Y$ is true or not. For example, if H_0 is true then

$$n\widehat{\gamma}^2 \rightsquigarrow \sum_{j=1}^{\infty} \lambda_j[(Z_j + a_j)^2 - 1]$$

where $Z_1, Z_2, \ldots, N(0,1)$ and $\{\lambda_j, a_j\}_1^\infty$ are (unknown) constants. A similar result holds for $\widehat{\gamma}$. This is called a Gaussian chaos. On the other hand, when H_0 is false, the limiting distribution is different.

Since the limiting distribution varies, we cannot really use it to construct a confidence interval. On way to solce this problem is to use blocking. Instead of using a U-statistics based on all subsets of size 4, we can break the dataset into non-overlapping blocks of size 4. We construct $Q = a(X_i, X_j, X_k, X_\ell)a(Y_i, Y_j, Y_k, Y_\ell)$ on each block. Then we define

$$g = \frac{1}{m}\sum_j Q_j$$

where $m = n/4$ is the number of blocks. Since this is an average, it will have a limiting Normal distribution. We can then use the Normal approximation or the bootstrap to get confidence intervals. However, g uses less information then $\widehat{\gamma}$ so we will get larger confidence intervals than necessary.

For τ^* the situation is better. Note that $\widehat{\tau}*$ is a U-statistic of order 4. That is

$$\widehat{\tau}^* = \widehat{\tau}^* = \frac{1}{\binom{n}{4}} \sum K(Z_i, Z_j, Z_k, Z_\ell)$$

where the sum is over all distinct quadruples and $Z_i = (X_i, Y_i)$. Also, $-1 \leq K \leq 1$. By Hoeffding's inequalty for U-statsitics of order b we have

$$\mathbb{P}(|\widehat{\tau}^* - \tau^*| > t) \leq 2e^{-2(n/b)t^2/r^2}$$

where r is the range of K. In our case, $r = 2$ and $b = 4$ so

$$\mathbb{P}(|\widehat{\tau}^* - \tau^*| > t) \leq 2e^{-nt^2/8}.$$

If we set $t_n = \sqrt{(8/n)\log(2/\alpha)}$ then $C_n = \widehat{\tau}^* \pm t_n$ is a $1 - \alpha$ confidence interval. However, it may not be shortest possible confidence interval. Find the shortest valid confidence interval is an open question.

Example. Figure 2 shows some Pearson graphs. These are: two Markov chains, a hub, four clusters, and a band. Technically, these should be best discovered using conditional independence graphs (discussed later). But correlation graphs are easy to estimate and often reveal the salient structure.

Figure 3 shows a graph from highly non-Normal data. The data have the structure of two Markov chains. I used the distance correlation on all pairs with permutation tests. Nice!

High Dimensional Bootstrap For Pearson Correlations We can also get simultaneous confidence intervals for many Pearson correlations. This is especially important if we want to put an edge when $|\rho(j,k)| \geq \epsilon$. If we have a confidence interval C then we can put an edge whenever $[-\epsilon, \epsilon] \cap C = \emptyset$.

The easiest way to get simultaneous confidence intervals is to use the bootstrap. Let R be the $d \times d$ matrix of true correlations and let $\widehat{R}$ be the $d \times d$ matrix of sample correlations. (Actually, it is probably better to use the Fisher transformed correlations.) Let $X_1^*, \ldots, X_n^*$ denote a bootstrap sample and let $\widehat{R}^*$ be the $d \times d$ matrix of correlations from the bootstrap sample. After taking B bootstrap samples we have $\widehat{R}_1^*, \ldots, \widehat{R}_B^*$. Let $\delta_j = \sqrt{n}\max_{s,t}|\widehat{R}_j^*(s,t) - \widehat{R}(s,t)|$ and define

$$\widehat{F}_n(w) = \frac{1}{B}\sum_{j=1}^{B} I(\delta_j \leq w)$$

which approximates

$$F_n(w) = \mathbb{P}(\sqrt{n}\max_{s,t}|\widehat{R}(s,t) - R(s,t)| \leq w).$$

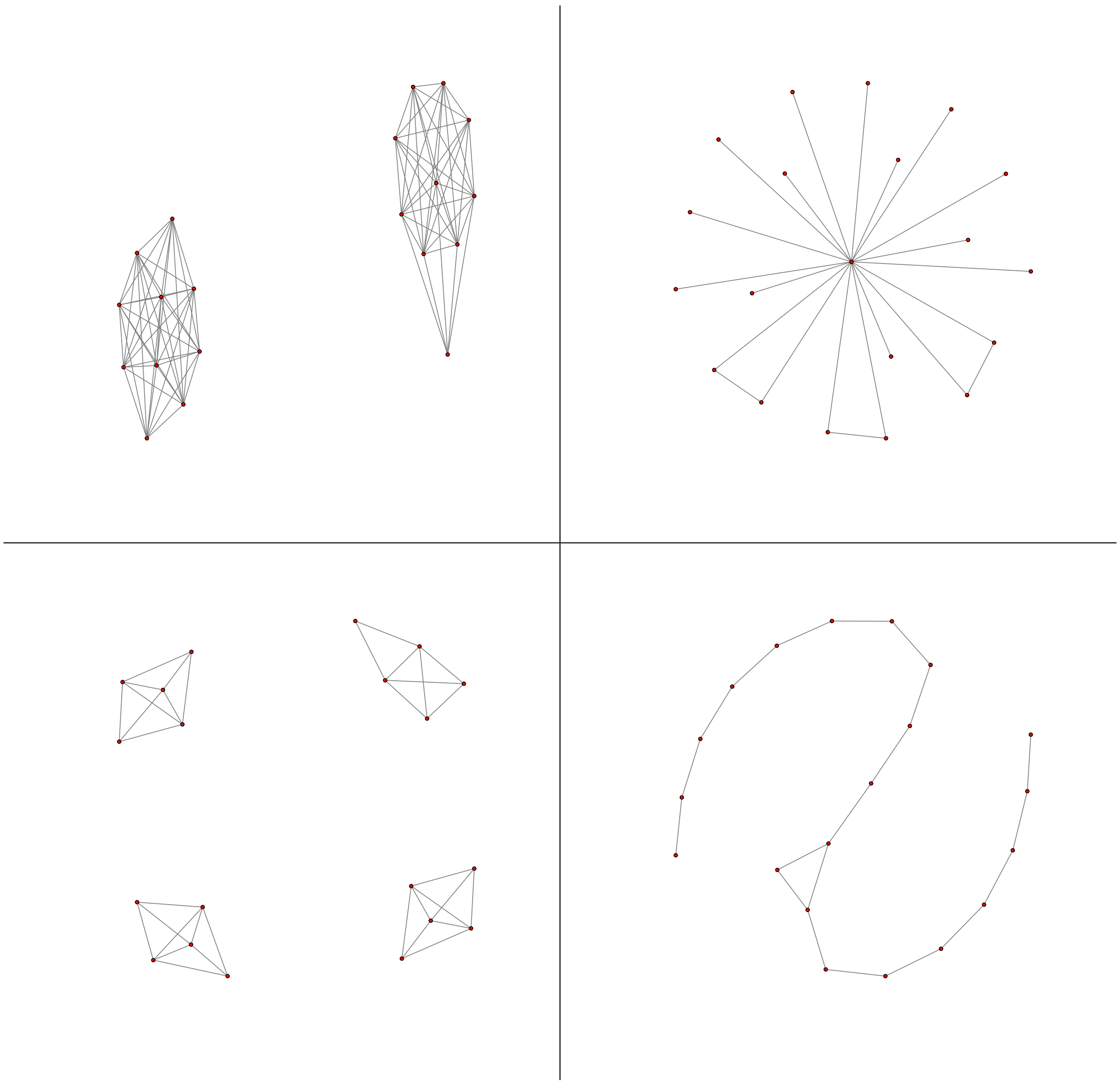


Figure 2: Pearson correlation graphs. Top left: two Markov-chains. Top right: a Hub. Bottom left: 4 clusters. Bottom right: banded.

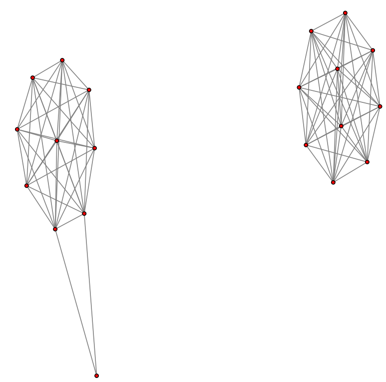


Figure 3: Graph based on nonparametric distance correlation. Two Markov chains.

Let $w_\alpha = \widehat{F}_n^{-1}(\alpha)$. Finally, we set

$$C_{st} = \left[\widehat{R}(s,t) - \frac{w_\alpha}{\sqrt{n}}, \quad \widehat{R}(s,t) + \frac{w_\alpha}{\sqrt{n}}\right].$$

Theorem 3 *Suppose that* $d = o(e^{n^{1/6}})$. *Then*

$$\mathbb{P}\left(R(s,t) \in C_{st} \text{ for all } s,t\right) \to 1 - \alpha$$

as $n \to \infty$.

2 Partial Correlation Graphs

Let $X, Y \in \mathbb{R}$ and Z be a random vector. The partial correlation between X and Y, given Z, is a measure of association between X and Y after removing the effect of Z.

Specifically, $\rho(X, Y|Z)$ is the correlation between ϵ_X and ϵ_Y where

$$\epsilon_X = X - \Pi_Z X, \qquad \epsilon_Y = Y - \Pi_Z Y.$$

Here, $\Pi_Z X$ is the projection of X onto the linear space spanned by Z. That is $\Pi_Z X = \beta^T X$ where β minimizes $\mathbb{E}[Y - \beta^T X]^2$. In other words, $\Pi_Z X$ is the linear regression of X on Z. Similarly, for $\Pi_Z Y$. We'll give an explicit formula for the partial correlation shortly.

Now let's go back to graphs. Let $X = (X(1), \ldots, X(d))$ and let ρ_{jk} denote the partial correlation between $X(j)$ and $X(k)$ given all the other variables. Let $R = \{\rho_{jk}\}$ be the $d \times d$ matrix of partial correlations.

Lemma 4 *The matrix R is given by $R(j,k) = -\Omega_{jk}/\sqrt{\Omega_{jj}\Omega_{kk}}$ where $\Omega = \Sigma^{-1}$ and Σ is the covariance matrix of X.*

The partial correlation graph G has an edge between j and k when $\rho_{jk} \neq 0$.

In the low-dimensional setting, we can estimate R as follows. Let S_n be the sample covariance. Let $\widehat{\Omega} = S_n^{-1}$ and define $\widehat{R}(j,k) = -\widehat{\Omega}_{jk}/\sqrt{\widehat{\Omega}_{jj}\widehat{\Omega}_{kk}}$. The easiest way to construct the graph is to use get simultaneous confidence intervals C_{jk} using the bootstrap. Then we put an edge if $0 \notin C_{jk}$.

There is also a Normal approximation similar to correlations. Define

$$Z_{jk} = \frac{1}{2}\log\left(\frac{1+r_{jk}}{1-r_{jk}}\right)$$

where $r_{jk} = \widehat{R}(j,k)$. Then

$$Z_{jk} \approx N\left(\theta_{jk}, \frac{1}{n-g-3}\right)$$

where $g = d - 2$ and

$$\theta_{jk} = \frac{1}{2}\log\left(\frac{1+\rho_{jk}}{1-\rho_{jk}}\right).$$

We reject H_0 if $|Z_{jk}| > z_{\alpha/(2m)}/\sqrt{n-g-3}$.

In high dimensions, this won't work since S_n is not invertible. In fact,

$$\mathrm{Var}(\widehat{R}(j,k)) \approx \frac{1}{n-d}$$

which shows that we cannot reliably estimate the partial correlation when d is large. You can do three things:

1. Compute a correlation graph instead. This is easy, works well, and often reveals similar structure that is in the partial correlation graph.
2. Shrinkage: let $\widehat{\Omega} = [(1-\epsilon)S_n + \epsilon D]^{-1}$ where $0 \leq \epsilon \leq 1$ and D is a diagonal matrix with $D_{jj} = S_{jj}$. Then we use the bootstrap to test the entries of the matrix. Based on calculations in Schafer and Strimmer (2005) amd Ledoit and Wolf (2004), a good choice of ϵ is
$$\epsilon = \frac{\sum_{j\neq k}\widehat{\mathrm{Var}}(s_{jk})}{\sum_{j\neq j} s_{jk}^2}$$

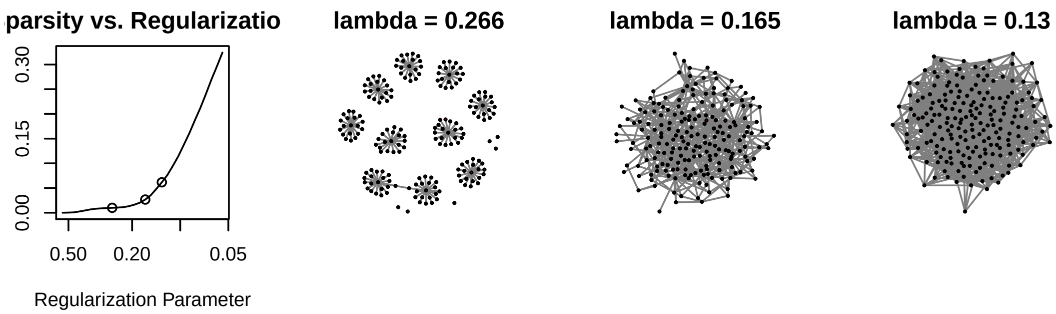


Figure 4: Graphical Lasso for a hub graph.

where

$$\widehat{\mathrm{Var}}(s_{jk}) = \frac{n}{(n-1)^3}\sum_{i=1}^{n}(s_{ijk} - \overline{w}_{jk})^2,$$

$w_{ijk} = (X_i(j) - \overline{X}(j))(X_i(k) - \overline{X}(k))$ and $\overline{w}_{jk} = n^{-1}\sum_i w_{ijk}$. This choice is based on minimizing the estimated risk.

3. Use the graphical lasso (described below). Warning! The reliability of the graphical lasso depends on lots of non-trivial, uncheckable assumptions.

For the graphical lasso we proceed as follows. We assume that $X_i \sim N(\mu, \Sigma)$. Then we estimated Σ (and hence $\Omega = \Sigma^{-1}$) using the penalized log-likelihood,

$$\widehat{\Omega} = \underset{\Omega \succ 0}{\arg\max}\Big[\ell(\Omega) - \lambda \sum_{j \neq k} |\omega_{jk}|\Big]$$

where the log-likelihood (after maximizing over μ) is

$$\ell(\Omega) = \frac{n}{2}\log|\Omega| - \frac{n}{2}\,\mathrm{tr}(\Omega S_n) - \frac{nd}{2}\log(2\pi). \tag{6}$$

Node-wise regression. A related but different approach due to Meinshausen and Buhlmann (2006). The idea is to regress each variable on all the others using the lasso.

Example: Figure 4 shows a hub graph. The graphs was estimated by Meinshausen and Buhlmann (2006) method using the R package `huge`.

How To Choose λ. If we use the graphical lasso, how do we choose λ? One idea is to use cross-validation based on the Normal log-likelihood. In this case we fit the model on part of the data and evaluate the log-likehood on the held-out data. This is not very reliable since it depends heavily on the Normality assumption. Currently, I do not think there is a rigorously justified, robust method for choosing λ. Perhaps the best strategy is to plot the graph for many values of λ.

More Robust Approach. Here is a more robust method. For each pair of variables, regress them on all the other variables (using your favorate regression method). Now compute the Kendal correlation on the residuals. This seems like a good idea but I have not seen anyone try it.

Nonparametric Partial Correlation Graphs. There are various ways to create a non-parametric partial correlation. Let us write

$$X = g(Z) + \epsilon_X$$
$$Y = h(Z) + \epsilon_Y.$$

Thus, $\epsilon_X = X - g(Z)$ and $\epsilon_Y = Y - h(Z)$ where $g(z) = \mathbb{E}[X|Z = x]$ and $h(z) = \mathbb{E}[Y|Z = z]$. Now define

$$\rho(X, Y|Z) = \rho(\epsilon_X, \epsilon_Y).$$

We can estimate ρ by using nonparametric regression to estimate $g(z)$ and $h(z)$. Then we take the correlation between the residuals $\widehat{\epsilon}_{X,i} = X_i - \widehat{g}(X_i)$ and $\widehat{\epsilon}_{Y,i} = Y_i - \widehat{h}(Y_i)$. When Z is high-dimensional, we can use SpAM to estimate g and h.

3 Conditional Independence Graphs

The strongest type of undirected graph is a *conditional independence graph*. In this case, we omit the edge between j and k if $X(j)$ is independent of $X(k)$ given the rest of the variables. We write this as

$$X(j) \amalg X(k) \,|\, \text{rest}. \tag{7}$$

Conditional independence graphs are the most informative undirected graphs but they are also the hardest to estimate.

3.1 Gaussian

In the special case of Normality, they are equivalent to partial correlation graphs.

Theorem 5 *Suppose that $X = (X(1), \ldots, X(d)) \sim N(\mu, \Sigma)$. Let $\Omega = \Sigma^{-1}$. Then $X(j)$ is independent of $X(k)$ given the rest, if and only if $\Omega_{jk} = 0$.*

So, in the Normal case, we are back to partial correlations.

3.2 Multinomials and Log-Linear Models

When all the variables are discrete, the joint distribution is multinomial. It is convenient to reparameterize the multinomial in a form known as a *log-linear model*.

Let's start with a simple example. Suppose $X = (X(1), X(2))$ and that each variable is binary. Let

$$p(x_1, x_2) = \mathbb{P}(X_1 = x_1, X_2 = x_2).$$

So, for example, $p(0,1) = \mathbb{P}(X_1 = 0, X_2 = 1)$. There are four unknown parameters: $p(0,0), p(0,1), p(1,0), p(1,1)$. Actually, these have to add up to 1, so there are really only three free parameters.

We can now write

$$\log p(x_1, x_2) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{12} x_1 x_2.$$

This is the log-linear representation of the multinomial. The parameters $\beta = (\beta_0, \beta_1, \beta_2, \beta_{12})$ are functions of $p(0,0), p(0,1), p(1,0), p(1,1)$. Conversely, $\beta = (\beta_0, \beta_1, \beta_2, \beta_{12})$ are functions of $p(0,0), p(0,1), p(1,0), p(1,1)$. In fact we can solve and get:

$$\beta_0 = \log p(0,0), \quad \beta_1 = \log\left(\frac{p(1,0)}{p(0,0)}\right), \quad \beta_2 = \log\left(\frac{p(0,1)}{p(0,0)}\right), \quad \beta_{12} = \log\left(\frac{p(1,1)p(0,0)}{p(0,1)p(1,0)}\right).$$

So why should be bother writing the model this way? The answer is:

Lemma 6 *In the above model, $X_1 \amalg X_2$ if and only if $\beta_{12} = 0$.*

The log-linear representation converts statements about independence and conditional independence into statements about parameters being 0.

Now suppose that $X = (X(1), \ldots, X(d))$. Let's continue to assume that each variable is binary. The log-linear representation is:

$$\log p(x_1, \ldots, x_d) = \beta_0 + \sum_j \beta_j x_j + \sum_{j<k} \beta_{jk} x_j x_k + \cdots + \beta_{12\cdots d}\, x_1 \cdots x_d.$$

Theorem 7 *We have that $X(j) \amalg X(k) | \mathrm{rest}$ if and only if every $\beta_A = 0$ if $(j,k) \in A$.*

Here is an example. Suppose that $d = 3$ and suppose that

$$\log p(x_1, x_2, x_3) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_{12} x_1 x_2 + \beta_{13} x_1 x_3.$$

In this model, $\beta_{23} = \beta_{123} = 0$. We conclude that $X(2) \amalg X(3)|X(1)$. Hence, we can omit the edge between $X(2)$ and $X(3)$.

Log-linear models thus make a nice connection between conditional independence graphs and parameters. There is a simple one-line command in R for fitting log-linear models. The function gives the parameter estimates as well as tests that each parameter is 0.

When the variables are not binary, the model is a bit more complicated. Each variable is now represented by a vector of parameters rather than one parameter. But conceptually, it is the same. 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 8 *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 log-linear representation of $p(x)$ is*

$$\log p(x) = \sum_{A \subset V} \psi_A(x_A) \tag{8}$$

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 (8) 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 9 *Let (X_A, X_B, X_C) be a partition of $X = (X_1, \ldots, X_d)$. Then $X_B \perp\!\!\!\perp X_C \,|\, X_A$ 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 $X_B \perp\!\!\!\perp X_C \,|\, X_A$ 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 $t \nsubseteq A \cup B$ or $t \nsubseteq A \cup C$. Therefore

$$\log p(x) = \sum_{t \subset A \cup B} \psi_t(x_t) + \sum_{t \subset A \cup C} \psi_t(x_t) - \sum_{t \subset A} \psi_t(x_t). \tag{9}$$

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

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 5 and is characterized by the next lemma.

Lemma 10 *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

$$\log p(x) = \psi_\Phi + \sum_{i=1}^{3} \psi_i(x_i) + \sum_{1 \le j < k \le 3} \psi_{jk}(x_{jk}). \tag{10}$$

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. $\square$

3.3 The Nonparametric Case

In real life, nothing has a Normal distribution. What should we do? We could just use a correlation graph or partial correlation graph. That's what I recommend. But if you really want a nonparametric conditional independence graph, there are some possible approaches.

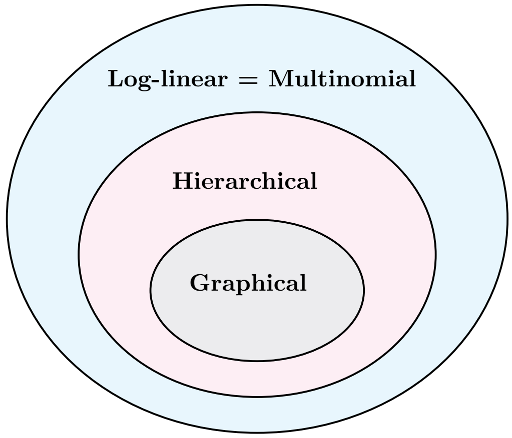


Figure 5: *Every graphical log-linear model is hierarchical but the reverse may not be true.*

Conditional cdf Method. Let X and Y be real and let Z be a random vector. Define

$$U = F(X|Z), \quad V = G(Y|Z)$$

where

$$F(x|z) = \mathbb{P}(X \leq x|Z = z), \quad G(y|z) = \mathbb{P}(Y \leq y|Z = z).$$

Lemma 11 *If $X \amalg Y \mid Z$ then $U \amalg V$.*

If we knew F and G, we could compute $U_i = F(X_i|Z_i)$ and $V_i = G(Y_i|Z_i)$ and then test for independence between U and V.

In practice, we estimate U_i and V_i using smoothing:

$$\widehat{U}_i = \widehat{F}(X_i|Z_i), \; \widehat{V}_i = \widehat{G}(Y_i|Z_i)$$

where

$$\widehat{F}(x|z) = \frac{\sum_{s=1}^n K(||z - Z_s||/h) I(X_s \leq x)}{\sum_{s=1}^n K(||z - Z_s||/h)}$$
$$\widehat{G}(y|z) = \frac{\sum_{s=1}^n K(||z - Z_s||/b) I(Y_s \leq y)}{\sum_{s=1}^n K(||z - Z_s||/b)}.$$

We have the following result from Bergsma (2011).

Theorem 12 (Bergsma) *Let θ be the Pearson or Kendall measure of association between U and V. Let $\widetilde{\theta}$ be the sample version based on (U_i, V_i), $i = 1, \ldots, n$. Let $\widehat{\theta}$ be the sample version based on $(\widehat{U}_i, \widehat{V}_i)$, $i = 1, \ldots, n$. Suppose that*

$$n^{1/2}(\widetilde{\theta} - \theta) = O_P(1), \quad n^{\beta_1}(\widehat{F}(x|z) - F(x|z)) = O_P(1), \quad n^{\beta_2}(\widehat{F}(y|z) - F(y|z)) = O_P(1).$$

Then

$$\sqrt{n}(\widehat{\theta} - \theta) = \sqrt{n}(\widetilde{\theta} - \theta) + O_P(n^{-\gamma})$$

where $\gamma = \min\{\beta_1, \beta_2\}$.

This means that, asymptotically, we can treat $(\widehat{U}_i, \widehat{V}_i)$ as if they were (U_i, V_i). Of course, for graphs, the whole procedure needs to be repeated for each pair of variables.

There are some caveats. First, we are essentially doing high dimensional regression. In high dimensions, the convergence will be very slow. Second, we have to choose the bandwidths h and b. It is not obvious how to do this in practice.

Challenge: Can you think of a way to do sparse estimation of $F(x|z)$ and $F(y|z)$?

Nonparanormal. Another approach is to use a Gaussian copula, also known as a *Nonparanormal* (Liu, Lafferty and Wasserman 2009). Recall that, in high dimnsional nonparametric regression, we replaced the linear model $Y = \sum_j \beta_j X_j + \epsilon$ with the *sparse additive model*:

$$Y = \sum_j f_j(X_j) + \epsilon \quad \text{ where most } ||f_j|| = 0.$$

We can take a similar strategy for graphs.

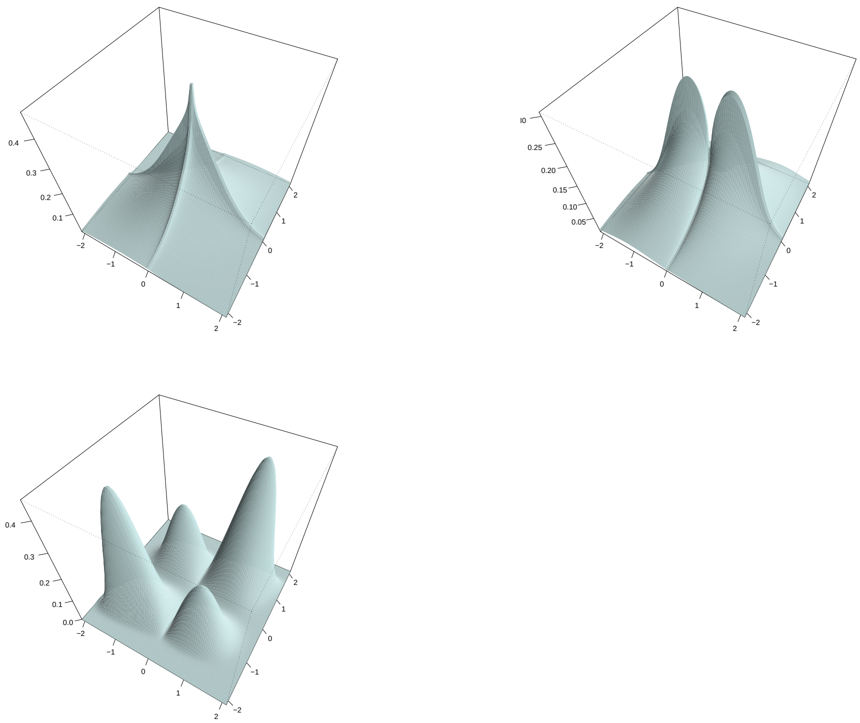
Assumptions	Dimension	Regression	Graphical Models
parametric	low	linear model	multivariate normal
	high	lasso	graphical lasso
nonparametric	low	additive model	nonparanormal
	high	sparse additive model	ℓ_1-regularized nonparanormal

One idea is to do node-wise regression using SpAM (see Voorman, Shojaie and Witten 2007). An alternative is as follows. Let $f(X) = (f_1(X_1), \ldots, f_p(X_p))$. Assume that $f(X) \sim N(\mu, \Sigma)$. Write $X \sim \mathrm{NPN}(\mu, \Sigma, f)$. If each f_j is monotone then this is just a Gaussian copula, that is,

$$F(x_1, \ldots, x_p) = \Phi_{\mu,\Sigma}\Big(\Phi^{-1}(F_1(x_1)), \ldots, \Phi^{-1}(F_p(x_p))\Big).$$

Lemma 13 $X_j \amalg X_k | \mathsf{rest}$ *iff* $\Sigma^{-1}_{jk} = 0$ *where* $\Sigma = \mathsf{cov}(f(X))$, $f(x) = (f_1(x_1), \ldots, f_d(x_d))$ *and* $f_j(x_j) = \Phi^{-1}(F_j(x_j))$.

The marginal means and variances μ_j and σ_j are not identifiable but this does not affect the graph G.



Three examples of nonparanormals.

We can estimate G using a two stage procedure:

1. Estimate each $Z_j = f_j(x_j) = \Phi^{-1}(F_j(x_j))$.
2. Apply the glasso to the Z_j's.

Let $\widehat{f}_j(x_j) = \Phi^{-1}(\widehat{F}_j(x_j))$. The usual empirical $\widehat{F}_j(x_j)$ will not work if d increases with n. We use a Winsorized version:

$$\widetilde{F}_j(x) = \begin{cases} \delta_n & \text{if } \widehat{F}_j(x) < \delta_n \\ \widehat{F}_j(x) & \text{if } \delta_n \leq \widehat{F}_j(x) \leq 1 - \delta_n \\ (1 - \delta_n) & \text{if } \widehat{F}_j(x) > 1 - \delta_n, \end{cases}$$

where

$$\delta_n \equiv \frac{1}{4n^{1/4}\sqrt{\pi \log n}}.$$

This choice of δ_n provides the right bias-variance balance so that we can achieve the desired rate of convergence in our estimate of Ω and the associated undirected graph G. Now compute the sample covariance S_n of the Normalized variables: $Z_j = \widehat{f}_j(X_j) = \Phi^{-1}(\widetilde{F}_j(X_j))$. Finally, apply the glasso to S_n. Let S_n^* be the covariance using the true f_j's.

Suppose that $d \leq n^{\xi}$. For large n,

$$\mathbb{P}\left(\max_{jk} |S_n(j,k) - S_n^*(j,k)| > \epsilon\right) \leq \frac{c_1 d}{(n\epsilon^2)^{2\xi}} + \frac{c_1 d}{(n\epsilon^2)^{c_5 \xi - 1}} + c_3 \exp\left(-\frac{c_4 n^{1/2} \epsilon^2}{\log d \, \log^2 n}\right)$$

and hence

$$\max_{jk} |S_n(j,k) - S_n^*(j,k)| = O_P\left(\sqrt{\frac{\log d \log^2 n}{n^{1/2}}}\right).$$

Suppose (unrealistically) that $X^{(i)} \sim \text{NPN}(\mu_0, \Sigma_0, f_0)$, and let $\Omega_0 = \Sigma_0^{-1}$. If

$$\lambda_n \asymp \sqrt{\frac{\log d \log^2 n}{n^{1/2}}}$$

then $\|\widehat{\Omega}_n - \Omega_0\|_{\text{F}} = O_P\left(\sqrt{\frac{(s+d)\log d \log^2 n}{n^{1/2}}}\right)$ and
$\|\widehat{\Omega}_n - \Omega_0\|_2 = O_P\left(\sqrt{\frac{s \log d \log^2 n}{n^{1/2}}}\right)$ where s is the sparsity level. Under extra conditions we get sparsistency:

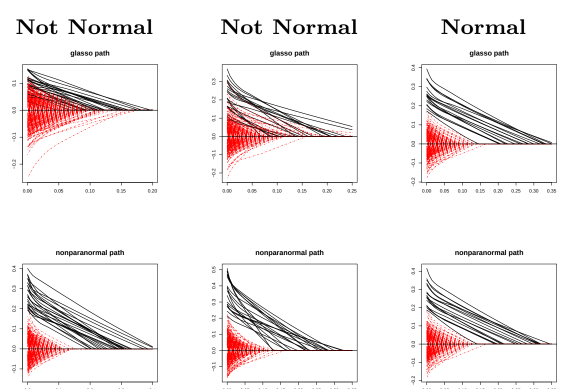
$$\mathbb{P}\Big(\text{sign}(\widehat{\Sigma}_n(j,k)) = \text{sign}(\Sigma_0(j,k) \quad \text{for all } j,k)\Big) \to 1.$$

Now suppose (more realistically) P is not NPN. Let $R(\widehat{f}, \widehat{\Sigma})$ denote risk (expected log-likelihood). Let $d \leq e^{n^{\xi}}$ for $\xi < 1$ and let

$$\mathcal{M}_n = \{f : \ f_j \text{ monotone}, \ ||f_j||_\infty \leq C\sqrt{\log n}\},$$
$$\mathcal{C}_n = \{\Omega : \ ||\Omega^{-1}||_1 \leq L_n\}$$

with $L_n = o(n^{(1-\xi)/2}/\sqrt{\log n})$. Then

$$R(\widehat{f}, \widehat{\Omega}) - \inf_{f,\Omega} R(f,\Omega) = o_P(1).$$



$n = 500$, $p = 40$

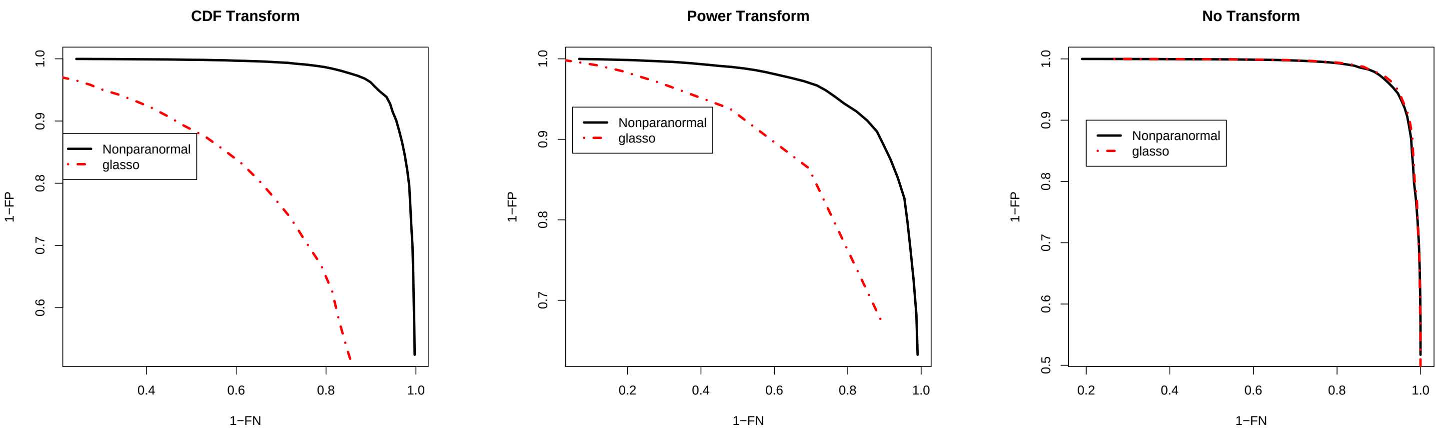


Figure 6: ROC curves for sample sizes $n = 200$.

The *Skeptic* is a more robst version (Spearman/Kendall Estimators Pre-empt Transformations to Infer Correlation). Set $\widehat{S}_{jk} = \sin\left(\frac{\pi}{2}\widehat{\tau}_{jk}\right)$ where

$$\widehat{\tau}_{jk} = \frac{2}{n(n-1)} \sum_{1 \le s < t \le n} \text{sign}\Big((X_s(j) - X_t(j))(X_s(k) - X_t(k))\Big).$$

Then (with $d \ge n$)

$$\mathbb{P}\left(\max_{jk} |\widehat{S}_{jk} - \Sigma_{jk}| > 2.45\pi\sqrt{\frac{\log d}{n}}\right) \le \frac{1}{d}.$$

As in Yuan (2010), let

$$\mathcal{M} = \left\{\Omega:\ \Omega \succ 0,\ ||\Omega||_1 \le \kappa,\ \frac{1}{c} \le \lambda_{\min}(\Omega) \le \lambda_{\max}(\Omega) < c,\ \deg(\Omega) \le M\right\}.$$

Then, for all $1 \le q \le \infty$,

$$\sup_{\Omega \in \mathcal{M}} ||\widehat{\Omega} - \Omega||_q = O_P\left(M\sqrt{\frac{\log d}{n}}\right).$$

From Yuan, this implies that the Skepic is minimax rate optimal. Now plug $\widehat{S}$ into glasso. See Liu, Han, Yuan, Lafferty and Wasserman arXiv:1202.2169 for numerical experiments and theoretical results.

Forests. Yet another approach is based on *forests.* A tractable family of graphs are *forests.* A graph F is a forest if it contains no cycles. 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 $|E_F| < d$. Suppose that P is Markov to F and has density p. Then p can be written as

$$p(x) = \prod_{(i,j) \in E_F} \frac{p(x_i, x_j)}{p(x_i)\, p(x_j)} \prod_{k \in V_F} p(x_k), \tag{11}$$

where each $p(x_i, x_j)$ is a bivariate density, and each $p(x_k)$ is a univariate density. Using (11), we have

$$\begin{aligned}
&\mathbb{E} \log p(X) && (12)\\
&= -\int p(x) \Bigg(\sum_{(i,j) \in E_F} \log \frac{p(x_i, x_j)}{p(x_i) p(x_j)} + \sum_{k \in V_F} \log \left(p(x_k)\right) \Bigg) dx \\
&= -\sum_{(i,j) \in E_F} \int p(x_i, x_j) \log \frac{p(x_i, x_j)}{p(x_i) p(x_j)} dx_i dx_j - \sum_{k \in V_F} \int p(x_k) \log p(x_k) dx_k \\
&= -\sum_{(i,j) \in E_F} I(X_i; X_j) + \sum_{k \in V_F} H(X_k), && (13)
\end{aligned}$$

where

$$I(X_i; X_j) \equiv \int p(x_i, x_j) \log \frac{p(x_i, x_j)}{p(x_i)\, p(x_j)}\, dx_i dx_j \tag{14}$$

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

$$H(X_k) \equiv -ds \int p(x_k) \log p(x_k)\, dx_k \tag{15}$$

is the entropy.

The optimal forest F^* can be found by minimizing the right hand side of (13). 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

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 (16).
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 $E^{(K)}$.

$p_{\theta_k}(x_k)$. We replace the population mutual information $I(X_i; X_j)$ in (13) by the plug-in estimate $\widehat{I}_n(X_i, X_j)$, defined as

$$\widehat{I}_n(X_i, X_j) = \int p_{\widehat{\theta}_{ij}}(x_i, x_j) \log \frac{p_{\widehat{\theta}_{ij}}(x_i, x_j)}{p_{\widehat{\theta}_i}(x_i)\, p_{\widehat{\theta}_j}(x_j)}\, dx_i dx_j \tag{16}$$

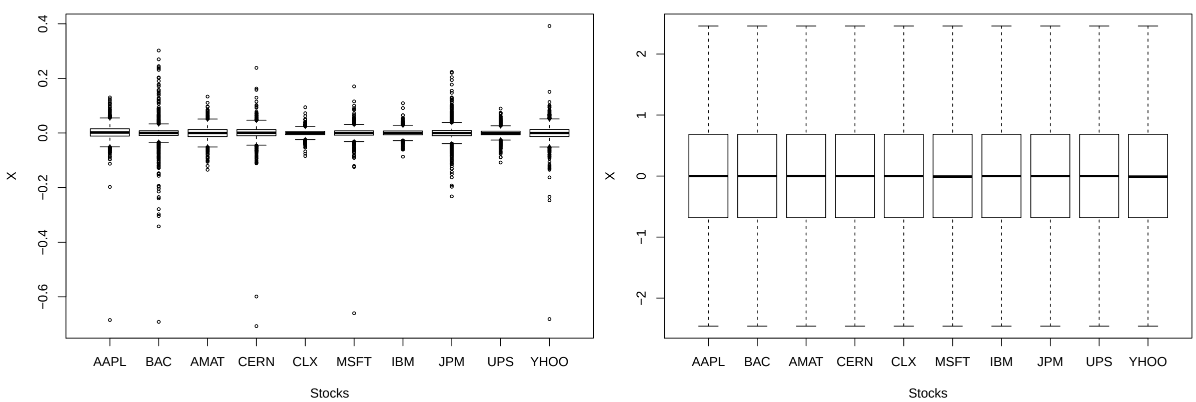
where $\widehat{\theta}_{ij}$ and $\widehat{\theta}_k$ are maximum likelihood estimates. Given this estimated mutual information matrix $\widehat{M} = \left[\widehat{I}_n(X_i, X_j)\right]$, we can apply Kruskal's algorithm (equivalently, the Chow-Liu algorithm) to find the best forest structure $\widehat{F}$.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

Example 14 (Learning Gaussian maximum weight spanning tree) *For Gaussian data $X \sim N(\mu, \Sigma)$, we know that the mutual information between two variables are*

$$I(X_i; X_j) = -\frac{1}{2} \log\left(1 - \rho_{ij}^2\right), \tag{17}$$

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 15 (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 $X_{tj} = \log\left(S_{t,j}/S_{t-1,j}\right)$ 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 7(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 7(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

Figure 7: 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 the 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 8. We see that the stocks from the same GICS sector are clustered very well.*

To get a nonparametric version, we can just an nonparametric estimate of the mutual information. But for that matter, we might as well put in any measure of association such as

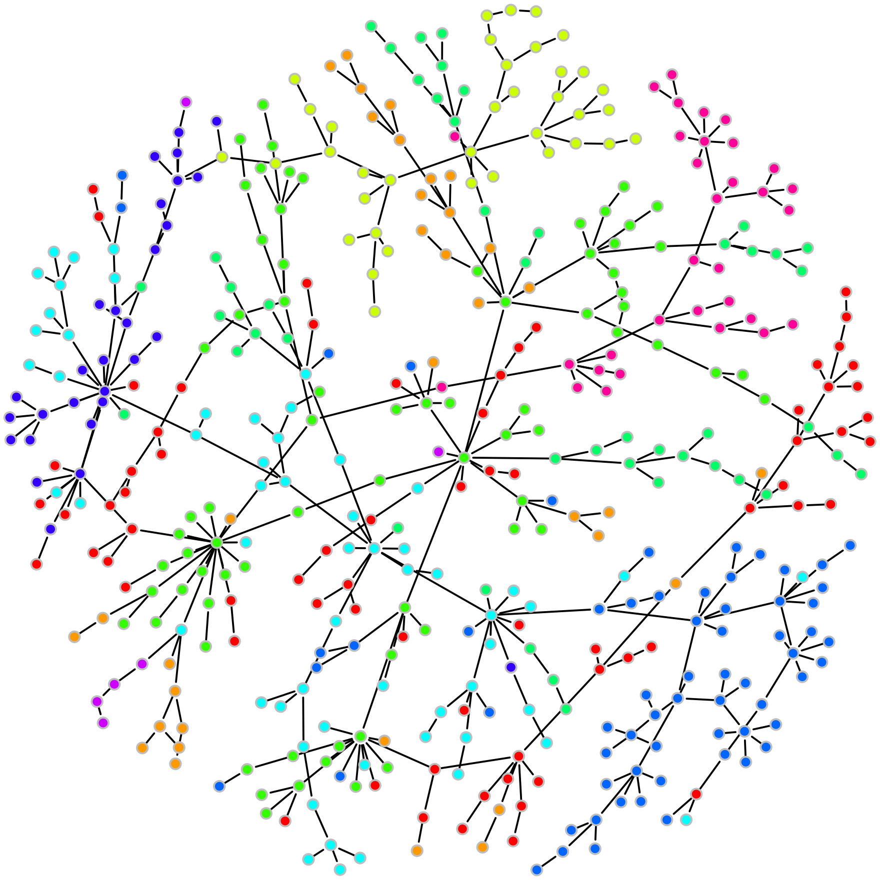


Figure 8: Tree graph learned from S&P 500 stock data from Jan. 1, 2003 to Jan. 1, 2008. The graph is estimated using the Chow-Liu algorithm under the Gaussian model. The nodes are colored according to their GICS sector categories.

distance correlation. In that case, we see that a forest is just a correlation graph without cycles.

4 A Deeper Look At Conditional Independence Graphs

In this section, we will take a closer look at conditional independence 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 $X_A = (X(j) : \; j \in A)$.

4.1 Markov Properties

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 $G = (V, E)$:

Definition 16 (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.*

The set of distributions that is globally Markov with respect to G is denoted by $\mathcal{P}(G)$.

Definition 17 (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*

$$p(x_s \mid x_t, \; t \neq s) = p\left(x_s \mid x_t, t \in N(s)\right) \tag{18}$$

for each node s.

Definition 18 (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*

$$X_s \perp\!\!\!\perp X_t \mid X_{V \setminus \{s,t\}}. \tag{19}$$

Consider for example the graph in Figure 9. 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 10.

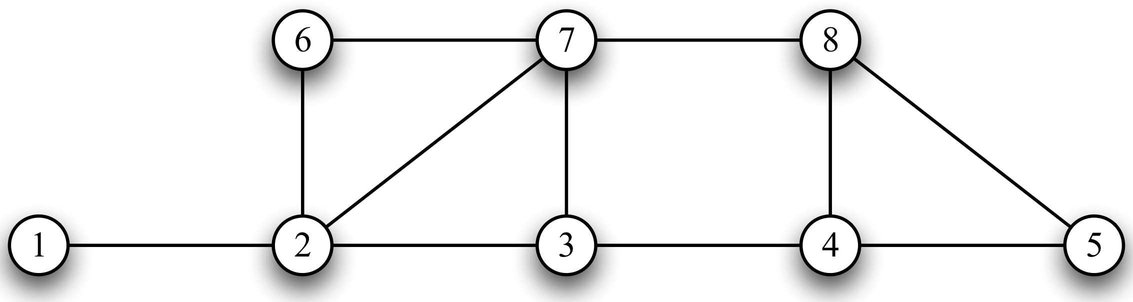


Figure 9: An undirected graph. $C = \{3, 7\}$ separates $A = \{1, 2\}$ and $B = \{4, 8\}$.

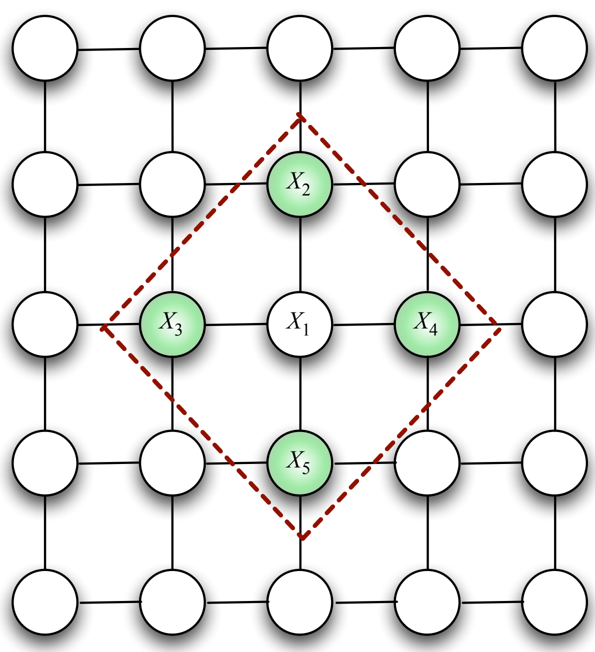


Figure 10: The local Markov property: Conditioned on its four neighbors X_2, X_3, X_4, and X_5, node X_1 is independent of the remaining nodes in the graph.

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

Proposition 19 *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

$$N(s) \cup [(V \setminus \{N(s) \cup \{s\}\}) \setminus \{t\}] = V \setminus \{s, t\}, \tag{20}$$

and it follows from the local Markov property that

$$X_s \perp\!\!\!\perp X_{V \setminus \{N(s) \cup \{s\}\}} \,|\, X_{V \setminus \{s,t\}}. \tag{21}$$

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

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

Theorem 20 *Suppose that, for all disjoint subsets $A, B, C, D \subset V$,*

$$\text{if } X_A \perp\!\!\!\perp X_B \,|\, X_{C \cup D} \text{ and } X_A \perp\!\!\!\perp X_C \,|\, X_{B \cup D}, \text{ then } X_A \perp\!\!\!\perp X_{B \cup C} \,|\, X_D, \tag{22}$$

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 $S, A, B \subset V$ with 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

$$X_{A \setminus \{s\}} \perp\!\!\!\perp X_B \,|\, X_{S \cup \{s\}} \text{ and } X_s \perp\!\!\!\perp X_B \,|\, S \cup (A \setminus \{s\}). \tag{23}$$

Now the condition (22) 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 (22) implies that $A \perp\!\!\!\perp B \,|\, S$. $\square$

The next proposition provides a stronger condition that implies (22).

Proposition 21 *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* (22) *holds.*

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

$$\text{if } X_1 \perp\!\!\!\perp X_2 \,|\, X_3 \text{ and } X_1 \perp\!\!\!\perp X_3 \,|\, X_2 \text{ then } X_1 \perp\!\!\!\perp \{X_2, X_3\}. \tag{24}$$

Since the density is positive and $X_1 \perp\!\!\!\perp X_2 \,|\, X_3$ and $X_1 \perp\!\!\!\perp X_3 \,|\, X_2$, 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:

$$p(x_1, x_2, x_3) = f_{13}(x_1, x_3) f_{23}(x_2, x_3) = g_{12}(x_1, x_2) g_{23}(x_2, x_3). \tag{25}$$

Since the density is continuous and positive, we have

$$g_{12}(x_1, x_2) = \frac{f_{13}(x_1, x_3) f_{23}(x_2, x_3)}{g_{23}(x_2, x_3)}. \tag{26}$$

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

$$p(x_1, x_2, x_3) = h(x_1)\ell(x_2) g_{23}(x_2, x_3) \tag{27}$$

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

From Proposition 21, 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*

4.2 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 11. 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 $\{X_1, X_2, X_3\}$.

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:

$$p(x_1, \ldots, x_{|V|}) = \frac{1}{Z} \prod_{C \in \mathcal{C}} \psi_C(x_C). \tag{28}$$

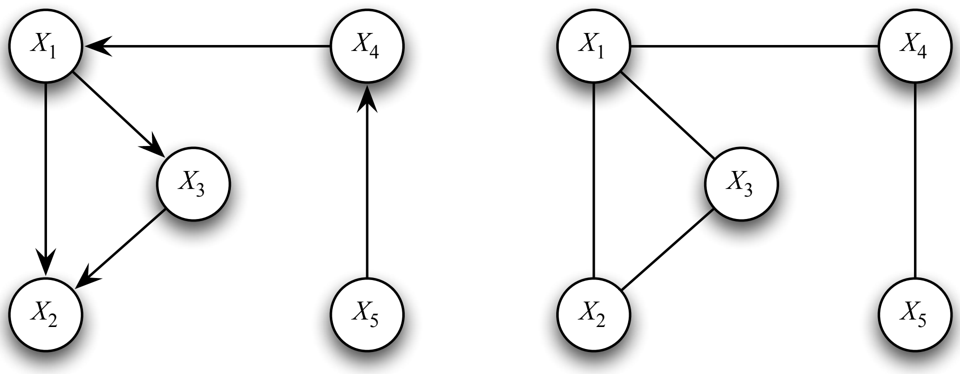


Figure 11: A directed graph encodes a factorization of the joint probability distribution in terms of conditional probability distributions. An undirected graph encodes a factorization of the joint probability distribution in terms of clique potentials.

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

$$Z = \int_{x_1,\ldots,x_{|V|}} \prod_{C \in \mathcal{C}} \psi_C(x_C) dx_1 \ldots dx_{|V|}. \tag{29}$$

Thus, the family of distributions represented by the undirected graph in Figure 11 can be written as

$$p(x_1, x_2, x_3, x_4, x_5) = \psi_{1,2,3}(x_1, x_2, x_3)\, \psi_{1,4}(x_1, x_4)\, \psi_{4,5}(x_4, x_5). \tag{30}$$

In contrast, the family of distributions represented by the directed graph in Figure 11 can be factored into conditional distributions according to

$$p(x_1, x_2, x_3, x_4, x_5) = p(x_5)\, p(x_4 \,|\, x_5)\, p(x_1 \,|\, x_4)\, p(x_3 \,|\, x_1)\, p(x_2 \,|\, x_1, x_3). \tag{31}$$

Theorem 22 *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 $\widetilde{A}$ to be the connectivity components in $G_{V \setminus S}$ which contain A and $\widetilde{B} = V \setminus (\widetilde{A} \cup S)$. 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 $\widetilde{A} \cup S$ or $\widetilde{B} \cup S$. Let $\mathcal{C}_A$ be the set of cliques contained in $\widetilde{A} \cup S$, the joint density $p(x)$ takes the following factorization

$$p(x) = \prod_{C \in \mathcal{C}} \psi_C(x_C) = \prod_{C \in \mathcal{C}_A} \psi_C(x_C) \prod_{C \in \mathcal{C} \setminus \mathcal{C}_A} \psi_C(x_C). \tag{32}$$

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

It is worth remembering that while we think of the set of *maximal clique*s 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. However, many graphs of interest in statistical analysis are sparse, with the number of cliques of size $O(|V|)$.

Theorem 22 shows that factoring with respect to a graph implies global Markov property. The next question is, under what conditions the Markov properties imply 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.

Theorem 23 (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 $p(x) > 0$ 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 $X_{0\backslash i}$ denote the vector $X_{0\backslash i} = (X_1, X_2, \ldots, X_{i-1}, 0, X_{i+1}, \ldots, X_d)$ obtained by setting $X_i = 0$, and let $X_{\backslash i} = (X_1, X_2, \ldots, X_{i-1}, X_{i+1}, \ldots, X_d)$ denote the vector of all components except X_i. Then

$$\frac{\mathbb{P}(x)}{\mathbb{P}(x_{i\backslash 0})} = \frac{\mathbb{P}(x_i \,|\, x_{\backslash i})}{\mathbb{P}(0 \,|\, x_{\backslash i})}. \tag{33}$$

Now, let

$$Q(x) = \log\left(\frac{\mathbb{P}(x)}{\mathbb{P}(0)}\right). \tag{34}$$

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

$$\begin{aligned}
Q(x) &= \log\left(\frac{\mathbb{P}(x)}{\mathbb{P}(0)}\right) &(35)\\
&= \log\left(\frac{\mathbb{P}(0, \ldots, x_i, 0, \ldots, 0)}{\mathbb{P}(0)}\right) + \log\left(\frac{\mathbb{P}(x)}{\mathbb{P}(0, \ldots, x_i, 0, \ldots, 0)}\right) &(36)\\
&= \frac{1}{d}\sum_{i=1}^{d}\left\{\log\left(\frac{\mathbb{P}(0, \ldots, x_i, 0, \ldots, 0)}{\mathbb{P}(0)}\right) + \log\left(\frac{\mathbb{P}(x)}{\mathbb{P}(0, \ldots, x_i, 0, \ldots, 0)}\right)\right\}. &(37)
\end{aligned}$$

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

$$\begin{aligned} Q(x) - Q(x_{0\setminus i}) &= \log\left(\frac{\mathbb{P}(x_i \mid x_{\setminus i})}{\mathbb{P}(0 \mid x_{\setminus i})}\right) \\ &= \phi_1(x_1) + \sum_{i>1} \phi_{1i}(x_1, x_i) + \sum_{j>i>1} \phi_{1ij}(x_1, x_i, x_j) + \cdots + \phi_{12\cdots d}(x) \end{aligned} \tag{38}$$

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 $\Leftrightarrow$ local Markov $\Leftrightarrow$ factored

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

Thus we can write:

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

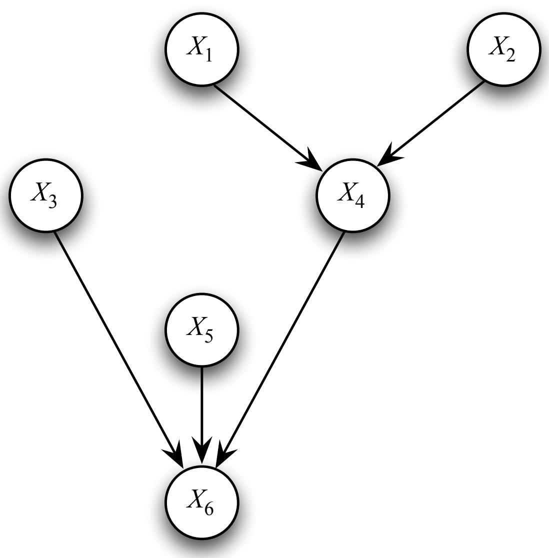
where $\mathcal{C}$ is the set of all (maximal) cliques in G and Z is the normalization constant. This is called the *Gibbs representation*.

4.3 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 12 can be sampled as follows:

$$
\begin{aligned}
X_1 &\sim P(X_1) && (39)\\
X_2 &\sim P(X_2) && (40)\\
X_3 &\sim P(X_3) && (41)\\
X_5 &\sim P(X_5) && (42)\\
X_4 \mid X_1, X_2 &\sim P(X_4 \mid X_1, X_2) && (43)\\
X_6 \mid X_3, X_4, X_5 &\sim P(X_6 \mid X_3, X_4, X_5). && (44)
\end{aligned}
$$

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. Instead one need something like MCMC.



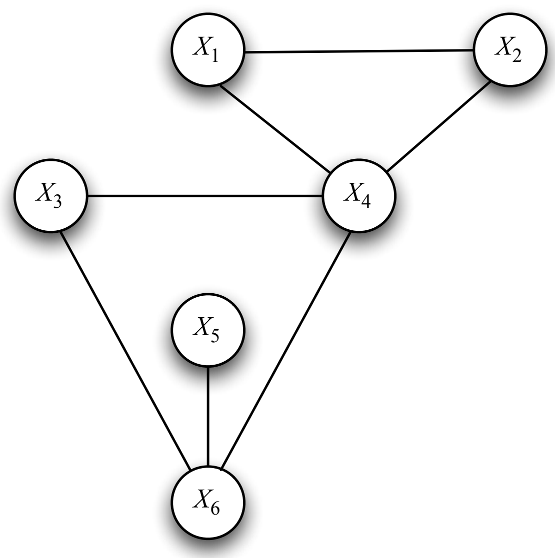


Figure 12: A DAG and its corresponding moral graph. A probability distribution that factors according to a DAG obeys the global Markov property on the undirected moral graph.

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 12. Here the directed model has a distribution

$$p(x_1)\, p(x_2)\, p(x_3)\, p(x_5)\, p(x_4 \mid x_1, x_2)\, p(x_6 \mid x_3, x_4, x_5). \tag{45}$$

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

$$\psi_{1,2,4}(x_1, x_2, x_4)\, \psi_{3,4,5,6}(x_3, x_4, x_5, x_6). \tag{46}$$

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 24 (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 25 *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 22. □

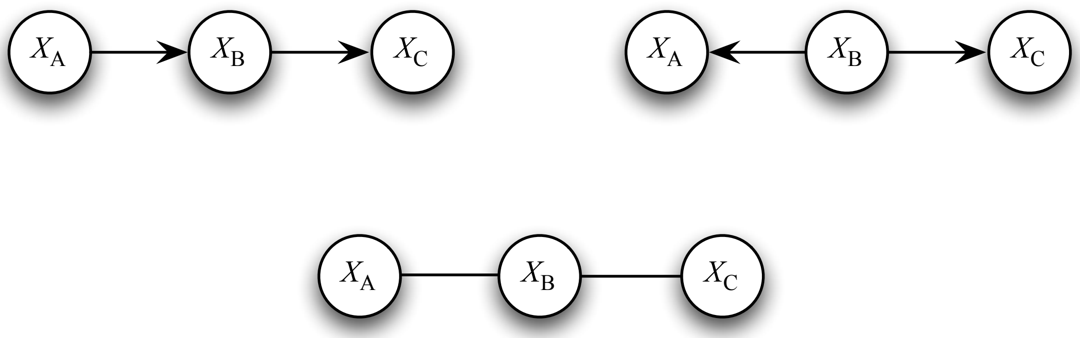


Figure 13: These three graphs encode distributions with identical independence relations. Conditioned on variable X_C, the variables X_A and X_B are independent; thus C separates A and B in the undirected graph.

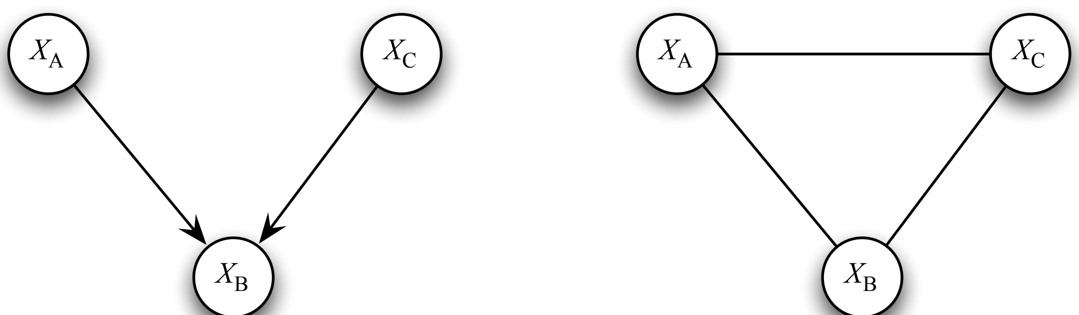


Figure 14: A directed graph whose conditional independence properties can not be perfectly expressed by its undirected moral graph. In the directed graph, the node C is a collider; therefore, X_A and X_B are not independent conditioned on X_C. In the corresponding moral graph, A and B are not separated 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.

Example 26 (Basic Directed and Undirected Graphs) *To illustrate some basic cases, consider the graphs in Figure 13. Each of the top three graphs encodes the same family of probability*

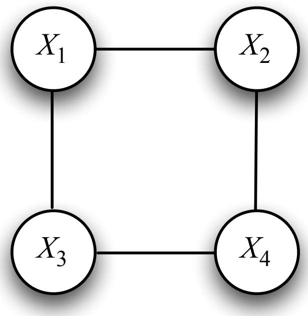


Figure 15: This undirected graph encodes a family of distributions that cannot be represented by a directed graph on the same set of nodes.

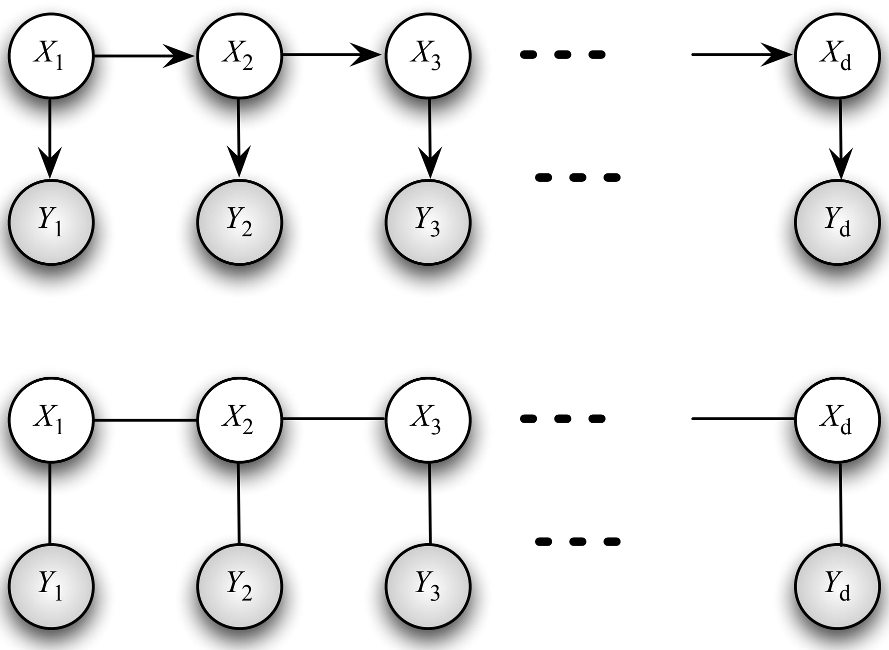


Figure 16: The top graph is a directed graph representing a *hidden Markov model*. The shaded nodes are observed, but the unshaded nodes, representing states in a latent Markov chain, are unobserved. Replacing the directed edges by undirected edges (bottom) does not changed the independence relations.

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 14 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 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 15 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 16 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.*

4.4 Faithfulness

The set of all distributions that are Markov to a graph G is denoted by $\mathcal{P}(G)$ To understand $\mathcal{P}(G)$ more clearly, we introduce some more notation. Given a distribution P let $\mathcal{I}(P)$ denote all conditional independence statements that are true for P. For example, if P has density p and $p(x_1, x_2, x_3) = p(x_1)p(x_2)p(x_3|x_1, x_2)$ then $\mathcal{I}(P) = \{X_1 \perp\!\!\!\perp X_2\}$. On the other hand, if $p(x_1, x_2, x_3) = p(x_1)p(x_2, x_3)$ then

$$\mathcal{I}(P) = \Big\{X_1 \perp\!\!\!\perp X_2,\ X_1 \perp\!\!\!\perp X_3,\ X_1 \perp\!\!\!\perp X_2|X_3,\ X_1 \perp\!\!\!\perp X_3|X_2\Big\}.$$

Similarly, given a graph G let $\mathcal{I}(G)$ denote all independence statements implied by the graph. For example, if G is the graph in Figure 15, then

$$\mathcal{I}(G) = \Big\{X_1 \perp\!\!\!\perp X_4 \mid \{X_2, X_3\}, \quad X_2 \perp\!\!\!\perp X_3 \mid \{X_1, X_4\}\Big\}.$$

From definition, we could write $\mathcal{P}(G)$ as

$$\mathcal{P}(G) = \Big\{P : \ \mathcal{I}(G) \subseteq \mathcal{I}(P)\Big\}. \tag{47}$$

This result often leads to confusion since you might have expected that $\mathcal{P}(G)$ would be equal to $\big\{P : \ \mathcal{I}(G) = \mathcal{I}(P)\big\}$. But this is incorrect. For example, consider the undirected graph X_1 — X_2, in this case, $\mathcal{I}(G) = \emptyset$ and $\mathcal{P}(G)$ consists of all distributions $p(x_1, x_2)$. Since, $\mathcal{P}(G)$ consists of all distributions, it also includes distributions of the form $p_0(x_1, x_2) = p_0(x_1)p_0(x_2)$. For such a distribution we have $\mathcal{I}(P_0) = \{X_1 \perp\!\!\!\perp X_2\}$. Hence, $\mathcal{I}(G)$ is a strict subset of $I(P_0)$.

In fact, you can think of $\mathcal{I}(G)$ as the set of independence statements that are common to all $P \in \mathcal{P}(G)$. In other words,

$$\mathcal{I}(G) = \bigcap \Big\{ \mathcal{I}(P) : \; P \in \mathcal{P}(G) \Big\}. \tag{48}$$

Every $P \in \mathcal{P}(G)$ has the independence properties in $\mathcal{I}(G)$. But some P's in $\mathcal{P}(G)$ might have extra independence properties.

We say that P is *faithful* to G if $\mathcal{I}(P) = \mathcal{I}(G)$. We define

$$\mathcal{F}(G) = \Big\{ P : \; \mathcal{I}(G) = \mathcal{I}(P) \Big\} \tag{49}$$

and we note that $\mathcal{F}(G) \subset \mathcal{P}(G)$. A distribution P that is in $\mathcal{P}(G)$ but is not in $\mathcal{F}(G)$ is said to be *unfaithful with respect to* G. The independence relation expressed by G are correct for such a P. It's just that P has extra independence relations not expressed by G. A distribution P is also Markov to some graph. For example, any distribution is Markov to the complete graph. But there exist distributions P that are not faithful to any graph. This means that there will be some independence relations of P that cannot be expressed using a graph.

Example 27 *The directed graph in Figure 14 implies that $X_A \perp\!\!\!\perp X_B$ but that X_A and X_B are not independent given X_C. There is no undirected graph G for (X_A, X_B, X_C) such that $\mathcal{I}(G)$ contains $X_A \perp\!\!\!\perp X_B$ but excludes $X_A \perp\!\!\!\perp X_B \,|\, X_C$. The only way to represent P is to use the complete graph. Then P is Markov to G since $\mathcal{I}(G) = \emptyset \subset \mathcal{I}(P) = \{X_A \perp\!\!\!\perp X_B\}$ but P is unfaithful to G since it has an independence relation not represented by G, namely, $\{X_A \perp\!\!\!\perp X_B\}$.*

Example 28 (Unfaithful Gaussian distribution) . *Let $\xi_1, \xi_2, \xi_3 \sim N(0,1)$ be independent.*

$$\begin{aligned} X_1 &= \xi_1 & (50)\\ X_2 &= aX_1 + \xi_2 & (51)\\ X_3 &= bX_2 + cX_1 + \xi_3 & (52) \end{aligned}$$

where a, b, c are nonzero. See Figure 17. Now suppose that $c = -\frac{b(a^2+1)}{a}$. Then

$$\begin{aligned} \mathrm{Cov}\,(X_2, X_3) &= \mathbb{E}(X_2 X_3) - \mathbb{E}X_2 \mathbb{E}X_3 & (53)\\ &= \mathbb{E}(X_2 X_3) & (54)\\ &= \mathbb{E}\left[(aX_1 + \xi_2)(bX_2 + cX_1 + \xi_3)\right] & (55)\\ &= \mathbb{E}\left[(a\xi_1 + \xi_2)(b(a\xi_1 + \xi_2) + cX_1 + \xi_3)\right] & (56)\\ &= (a^2 b + ac)\mathbb{E}\xi_1^2 + b\mathbb{E}\xi_2^2. & (57)\\ &= a^2 b + ac + b = 0. & (58) \end{aligned}$$

Thus, we know that $X_2 \perp\!\!\!\perp X_3$. We would like to drop the edge between X_2 and X_3. But this would imply that $X_2 \perp\!\!\!\perp X_3 \,|\, X_1$ which is not true.

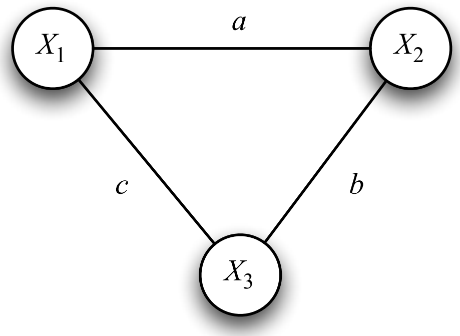


Figure 17: An unfaithful Gaussian distribution.

Generally, the set of unfaithful distributions $\mathcal{P}(G) \setminus \mathcal{F}(G)$ is a small set. In fact, it has Lebesgue measure zero if we restrict ourselves to nice parametric families. However, these unfaithful distributions are scattered throughout $\mathcal{P}(G)$ in a complex way; see Figure 18.

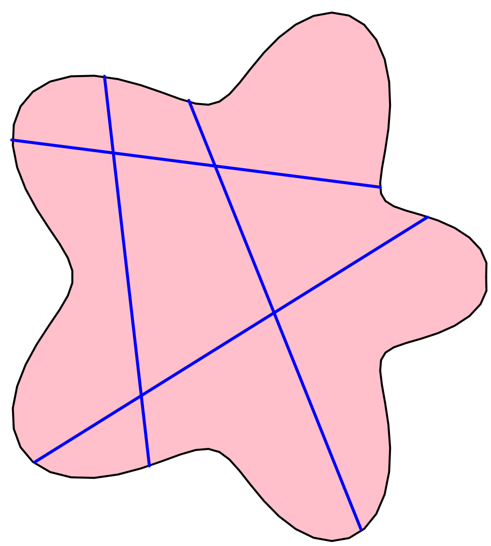


Figure 18: The blob represents the set $\mathcal{P}(G)$ of distributions that are Markov to some graph G. The lines are a stylized representation of the members of $\mathcal{P}(G)$ that are not faithful to G. Hence the lines represent the set $\mathcal{P}(G) \setminus \mathcal{F}(G)$. These distributions have extra independence relations not captured by the graph G. The set $\mathcal{P}(G) \setminus \mathcal{F}(G)$ is small but these distributions are scattered throughout $\mathcal{P}(G)$.

References

Gretton, Song, Fukumizu, Scholkopf, Smola (2008). A Kernel Statistical Test of Independence. NIPS.

Lyons, Russell. (2013). Distance covariance in metric spaces. *The Annals of Probability*, 41, 3284-3305.

Szekely, Gabor J., Maria L. Rizzo, and Nail K. Bakirov. (2007). Measuring and testing dependence by correlation of distances. *The Annals of Statistics*, 35, 2769-2794.