ECON622 Notes: Graphical Models

Philip Solimine

November 2023

1 Introduction

Graphical models are a useful paradigm for visualizing and understanding complex probability distributions. They formalise the definition of joint probability distributions over sets of random variables. They encode *conditional independence relationships* between variables, which inform how the probability distribution can be factorized.

There are two primary types of graphical model; Bayesian graphical models, which are *directed*, and Markov Random Fields or MRFs, which are *undirected*. Methodologically, the primary difference is that direction defines a *topological ordering* on variables, while undirected graphs define no such ordering.

These notes contain information from [1] Probabilistic Machine Learning: Advanced Topics, Chapter 4. As well as slides from [2].

2 Bayesian graphical models (Directed Acyclic Graphs)

Bayesian graphical models satisfy the *ordered Markov property* – a variable is conditionally independent of all its predecessors, given its parents. The conditional independence assumptions are

$$x_i \perp \mathbf{x}_{\text{pred}(i)\setminus \text{pa}(i)}$$

where $\text{pa}(i)$ are the parents of node i and $\text{pred}(i)$ are the predecessors.

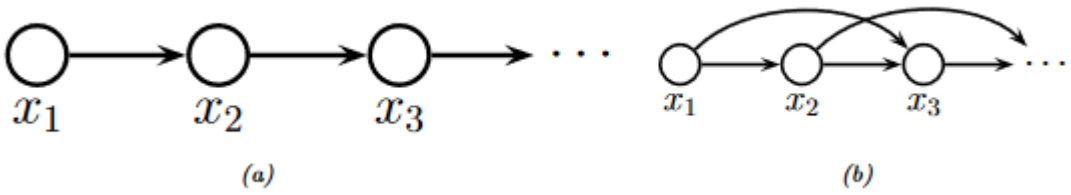


Figure 1: A first order (a) and second-order (b) Markov process

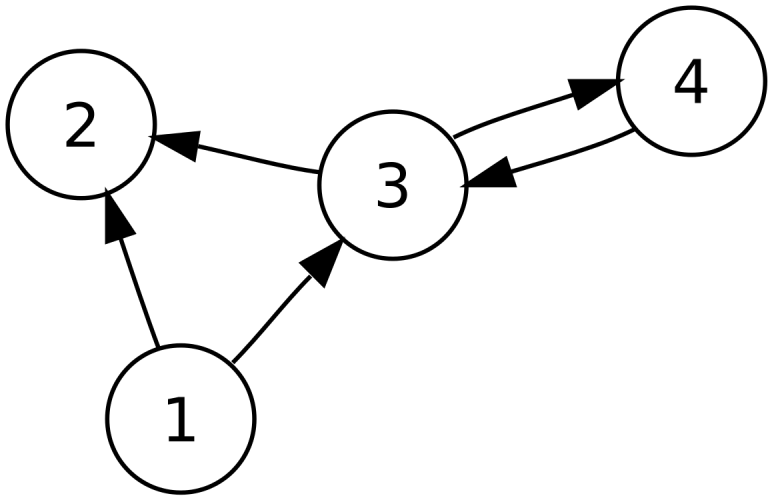


Figure 2: Since x_3 and x_4 are in a cycle, $x_3 \leq x_4$ and $x_4 \leq x_3$, but $x_3 \neq x_4$

2.1 Topological ordering

Bayesian models represent a set of *topological orderings* of the variables. A topological ordering is any *linear extension* of a partial order. Define the partial order $x_1 \leq x_2 \Leftrightarrow$ there is a directed path from x_1 to x_2.
A partial order must be

1. Reflexive: $x_1 \leq x_1$
2. Antisymmetric: $x_1 \leq x_2 \wedge x_2 \leq x_1 \Rightarrow x_1 = x_2$
3. Transitive: $x_1 \leq x_2 \wedge x_2 \leq x_3 \Rightarrow x_1 \leq x_3$

Q: Which one of the above is not trivial?
A: The restriction of Antisymmetry bites. It says that not any directed graph is a valid Bayesian graphical model; they cannot include any cycles. For this reason, we can think of the edges as being causal relationships between the variables. (This provides a framework to use DAGs to identify causal models. But it is not what we are covering today)

A *topological order* is a valid linear extension of this partial order. A *linear extension* of a partial order is any ordering $\leq^*$ that satisfies:

- $x_1 \leq^* x_2 \vee x_2 \leq^* x_1 \; \forall x_1, x_2$ ($\leq^*$ is a total order)
- $x_1 \leq x_2 \Rightarrow x_1 \leq^* x_2 \; \forall x_1, x_2$

Q: Are the topological orderings associated with a bayes-net unique?
A: Topological orderings are not generically unique. If a unique path exists, then it forms a *Hamiltonian path* (a path that visits each node exactly once).

While the Hamiltonian path problem is NP-Complete in general directed graphs, it is straightforward in DAGs. If there are two adjacent nodes in the ordering that are not connected by a path, then swapping their order will create another valid topological order.

However, counting all topological orderings is #P-complete, (and finding the best one is NP-complete).

2.2 Density factorization

From now on, choose a topological ordering $1 : N_G$. Bayesian networks provide a factorization of the joint density function

$$p\left(\mathbf{x}_{1:N_G}\right) = p\left(x_1\right) p\left(x_2 \mid x_1\right) p\left(x_3 \mid x_1, x_2\right) \ldots p\left(x_{N_G} \mid x_1, \ldots, x_{N_G-1}\right) = \prod_{i=1}^{N_G} p\left(x_i \mid \mathbf{x}_{\mathrm{pa}(i)}\right)$$

Suppose each variable is discrete, with a finite outcome space $|\mathcal{X}| = K < \infty$. Then to specify the full joint distribution without factorizing, we would need $O(K^{N_G})$ parameters (one probability for each combination of outcomes). On the other hand, if each node has at most N_P parents, the representational complexity is only $O(N_G K^{N_P+1})$.

2.3 Dependence flows and d-separation: Bayes-ball algorithm

Graphical models are primarily a way to encode conditional independence assumptions. We will write conditional independence as $x_a \perp x_b \mid x_c$, meaning that x_a is independent of x_b given x_c in the graph.

A simple way to figure out if two variables are d-separated is to play a game of *Bayes-ball*. Think about a ball that rolls through the graph according to a set of rules.

We say that an *undirected path* P is *blocked* by a conditioning set C if and only if at least one of the following holds:

1. P contains a chain or pipe $s \to m \to t$ or $s \leftarrow m \leftarrow t$ where $m \in C$
2. P contains a tent or fork $s \leftarrow m \to t$ where $m \in C$
3. P contains a collider $s \to m \leftarrow t$ and $m \notin C$

Then $x_a \perp x_b \mid \mathbf{x}_c \iff$ all undirected paths from x_a to x_b are blocked by $\mathbf{x}_c$.
These "rules of Bayes-ball" can be derived as follows:

1. Consider a chain $x \to y \to z$. This encodes the factorization $p(x, y, z) = p(x)p(y \mid x)p(z \mid y)$. Now condition on y.

   $$p(x, z \mid y) = \frac{p(x, y, z)}{p(y)} = \frac{p(x)p(y \mid x)p(z \mid y)}{p(y)} = \frac{p(x, y)p(z \mid y)}{p(y)} = p(x \mid y)p(z \mid y)$$

   Therefore conditioning on y separates x and z into independent factors.

2. Given a fork structure $x \leftarrow y \to z$, the joint is $p(x, y, z) = p(y)p(x \mid y)p(z \mid y)$. Conditioning on the confounder y gives

   $$p(x, z \mid y) = \frac{p(x, y, z)}{p(y)} = \frac{p(y)p(x \mid y)p(z \mid y)}{p(y)} = p(x \mid y)p(z \mid y)$$

   and therefore $x \perp z \mid y$.

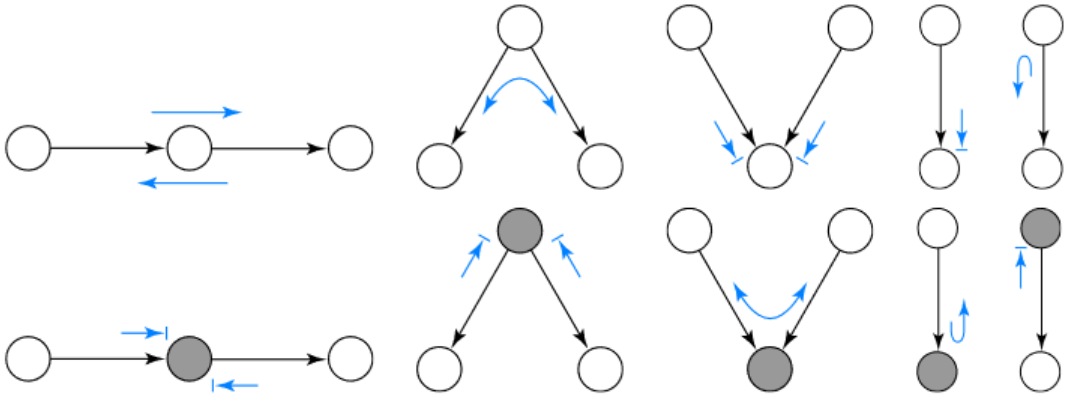


Figure 3: The rules of Bayes-ball (from Mark Paskin)

3. The final case is different; the collider, where $x \rightarrow y \leftarrow z$. In this case, $p(x, y, z) = p(x)p(z)p(y \mid x, z)$, and therefore $p(x, y, z) = \frac{p(x)p(z)p(y|x,z)}{p(y)}$, so $x \not\perp z \mid y$. On the other hand, the unconditional $p(x, z) = p(x)p(z)$, so $x \perp z$.

2.4 Example: Visualizing bias

In order to identify a treatment effect, we need to make sure that the treatment effect we want to identify is the only open path from the treatment to the outcome.

This involves choosing which variables to control for. You want to control for all confounders (to avoid omitted variable bias), but not any colliders, mediators, or endogenous variables (to avoid selection bias)

2.5 Markov blankets and the full conditional

A variable's *Markov blanket* is defined as the smallest set of controls that renders a variable conditionally independent of all other variables in the model. We denote the Markov blanket of variable i as $\text{mb}(i)$

In a directed graphical model, the Markov blanket consists of all direct connections (ingoing or outgoing), plus all of the co-parents.

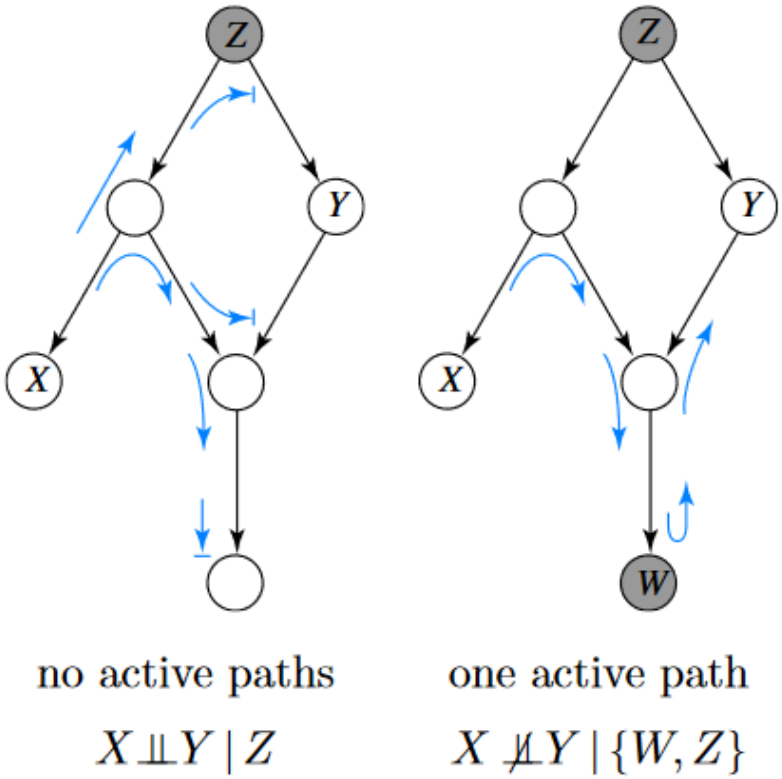


Figure 4: Two games of Bayes ball (from Mark Paskin)

probably skip this derivation and draw a graph with intuition?

$$
\begin{aligned}
p\left(X_i \mid X_{-i}\right) &= \frac{p\left(X_i, X_{-i}\right)}{\sum_x p\left(X_i = x, X_{-i}\right)} \\
&= \frac{p\left(X_i, U, Y, Z, O\right)}{\sum_x p\left(X_i = x, U, Y, Z, O\right)} \\
&= \frac{p\left(X_i \mid U\right)\left[\prod_j p\left(Y_j \mid X_i, Z_j\right)\right] P(U, Z, O)}{\sum_x p\left(X_i = x \mid U\right)\left[\prod_j p\left(Y_j \mid X_i = x, Z_j\right)\right] P(U, Z, O)} \\
&= \frac{p\left(X_i \mid U\right)\left[\prod_j p\left(Y_j \mid X_i, Z_j\right)\right]}{\sum_x p\left(X_i = x \mid U\right)\left[\prod_j p\left(Y_j \mid X_i = x, Z_j\right)\right]} \\
&\propto p\left(X_i \mid \mathrm{pa}\left(X_i\right)\right) \prod_{Y_j \in \mathrm{ch}(X_i)} p\left(Y_j \mid \mathrm{pa}\left(Y_j\right)\right)
\end{aligned}
$$

Therefore we can express the full conditional distribution of x_i, given $\mathbf{x}_{-i}$, as

$$
p(x_i \mid \mathbf{x}_i) \propto p(x_i \mid \mathbf{x}_{\mathrm{pa}(i)}) \prod_{k \in \mathrm{ch}(i)} p(x_k \mid \mathbf{x}_{\mathrm{pa}(k)})
$$

This concept will be useful in Gibbs samplers and MCMC.

2.6 Sampling and inference

Generation or sampling from a directed model is simple. The important part is that we visit the nodes in topological order, parents before children. Sample from the parents, then sample from the children given the parents' values. This is called *ancestral sampling*

Inference refers to the task of computing the posterior distribution over a set of *query* variables Q, given the observed values of a set of *visible* nodes V (the data), while marginalizing over the irrelevant *nuisance variables* R.

If Q is a singleton, then $p_\theta(Q \mid V)$ is called the *posterior marginal* for Q.

The goal of inference is to compute the posterior over the queries given the data, in a way that is invariant to the nuisance variables. (However, we can't avoid the nuisance variables entirely, we need to infer their values in order to get at the query.)

2.7 Learning

Learning parameters the parameters of a graphical model works by treating the parameters (θ) as variables in the joint density, and conducting inference on them.

We then use conditional densities to infer something about the distribution of parameters. These are usually

1. A likelihood function; $p(X, Z \mid \theta)$
2. A posterior distribution of the parameter; $p(\theta \mid X)$

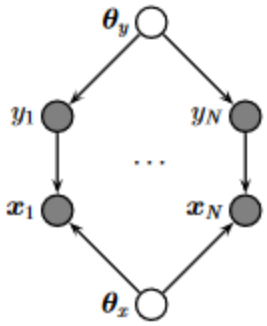


Figure 5: A fully observed Bayes net for parameter inference

$$p_\theta(z \mid x) = \sum_r p_\theta(z, r \mid x) = \sum_r \frac{p_\theta(z, r, x)}{p_\theta(x)} = \sum_r \frac{p_\theta(z, r, x)}{\sum_{z', r'} p_\theta(z', r', x)}$$

The joint distribution for the model in Figure 5 is

$$\begin{aligned} p(\boldsymbol{\theta}, \mathcal{D}) &= p\left(\boldsymbol{\theta}_x\right) p\left(\boldsymbol{\theta}_y\right) \left[\prod_{n=1}^{N} p\left(y_n \mid \boldsymbol{\theta}_y\right) p\left(\boldsymbol{x}_n \mid y_n, \boldsymbol{\theta}_x\right)\right] \\ &= \left[p\left(\boldsymbol{\theta}_y\right) \prod_{n=1}^{N} p\left(y_n \mid \boldsymbol{\theta}_y\right)\right] \left[p\left(\boldsymbol{\theta}_x\right) \prod_{n=1}^{N} p\left(\boldsymbol{x}_n \mid y_n, \boldsymbol{\theta}_x\right)\right] \\ &= \left[p\left(\boldsymbol{\theta}_y\right) p\left(\mathcal{D}_y \mid \boldsymbol{\theta}_y\right)\right] \left[p\left(\boldsymbol{\theta}_x\right) p\left(\mathcal{D}_x \mid \boldsymbol{\theta}_x\right)\right] \end{aligned}$$

Where $\mathcal{D}_y = \{y\}$ is all that is needed to estimate θ_y and $\mathcal{D}_x = \{x, y\}$

In this formulation, the prior, likelihood, and posterior all factorize nicely. In particular,

$$p(\theta \mid \mathcal{D}) = \prod_{i=1}^{N_G} p(\theta_i) p(\mathcal{D}_i \mid \theta_i)$$

It doesn't get better than that! This means that we can solve for each node independently:

$$\hat{\theta} = \arg\max_\theta \prod_{i=1}^{N_G} p(\mathcal{D}_i \mid \theta_i)$$

In essence, we can think of each parameter as the subject of its own, independent estimation problem. After a log transform, this becomes a simple sum.

With discrete data and multinomial distributions, this often means that we can use simple estimators like empirical frequencies.

$$\begin{aligned} p(\mathcal{D} \mid \boldsymbol{\theta}) &= \prod_{n=1}^{N} \prod_{i=1}^{N_G} p\left(x_{ni} \mid \boldsymbol{x}_{n,\mathrm{pa}(i)}, \boldsymbol{\theta}_i\right) \\ &= \prod_{n=1}^{N} \prod_{i=1}^{N_G} \prod_{j=1}^{J_i} \prod_{k=1}^{K_i} \theta_{ijk}^{\mathbf{I}\left(x_{ni}=k, \boldsymbol{x}_{n,\mathrm{pa}(i)}=j\right)} \end{aligned}$$

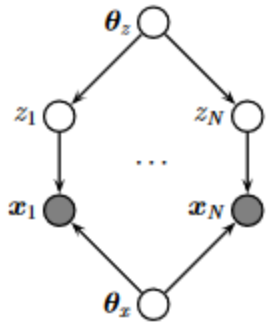


Figure 6: A partially-observed Bayesian network with latent features z

Where $\theta_{ijk} = p(x_i = k \mid \mathbf{x}_{n,\text{pa}(i)} = j)$ is a parameter of the density. Define the *sufficient statistics* for a variable to be the number of times it appears

$$N_{ijk} = \sum_{n=1}^{N} \mathbf{I}(x_{n,i} = k, x_{n,\text{pa}(i)} = j)$$

Then the MLE is simply the empirical frequiencies:

$$\hat{\theta}_{ijk} = \frac{N_{ijk}}{\sum_{k'} N_{ijk'}}$$

2.8 Learning from incomplete data

Often we are not lucky enough to observe every relevant feature in the data. For example, see Figure 6. This is the same estimation problem as before, but now we have an additional set of latent states z that must be inferred.

Now, we will have to marginalize out over z.

$$p(\mathcal{D} \mid \theta) = \sum_{z_{1:N}} \prod_{n=1}^{N} p(z_n \mid \theta_z) p(x_n \mid z_n, \theta_x) = \prod_{n=1}^{N} \sum_{z_{1:N}} p(z_n \mid \theta_z) p(x_n \mid z_n, \theta_x)$$

This leaves us with the log-likelihood

$$\ell(\theta) = \sum_{n} \log \sum_{z_n} p(z_n \mid \theta_z) p(x_n \mid z_n, \theta_x)$$

Since log doesn't distribute over the sum, this is the best we can do. If everything is discrete and tabular, then

$$\log p(\mathcal{D} \mid \theta) = \sum_{i=1}^{N_G} \sum_{j=1}^{J_i} \sum_{k=1}^{K_i} N_{ijk} \log \theta_{ijk}$$

and thus the expected "complete data" log likelihood would be:

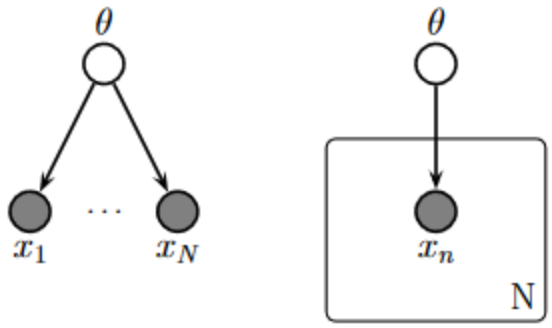


Figure 7: Plates represent repetition

$$\mathbb{E}[\log p(\mathcal{D} \mid \boldsymbol{\theta})] = \sum_i \sum_j \sum_k \bar{N}_{ijk} \log \theta_{ijk}$$

where

$$\bar{N}_{ijk} = \sum_{n=1}^{N} \mathbb{E}\left[\mathbb{I}\left(x_{ni} = k, \boldsymbol{x}_{n,\text{pa}(i)} = j\right)\right] = \sum_{n=1}^{N} p\left(x_{ni} = k, \boldsymbol{x}_{n,\text{pa}(i)} = j \mid \mathcal{D}_n, \boldsymbol{\theta}^{\text{old}}\right)$$

Notice that θ became θ^{old}.

That's because this is the first step of an algorithm called the E-M Algorithm (expectation-maximization algorithm). In each "expectation" step, the probabilities $p\left(x_{ni} = k, \boldsymbol{x}_{n,\text{pa}(i)} = j \mid \mathcal{D}_n, \boldsymbol{\theta}^{\text{old}}\right)$ are inferred from the data and parameters. In the subsequent "maximization" step, the new MLE, given these *expected sufficient statistics*, would then be found as

$$\hat{\theta}_{ijk} = \frac{\bar{N}_{ijk}}{\sum_{k'} \bar{N}_{ijk}}$$

The E-M algorithm consists of executing these two steps repeatedly until convergence.

2.9 Plating

Plating is a visual aid to simplify *repetition* in graphical models. Plating is a powerful abstraction and is used in the background to speed up computations in some PPLs, like pyro/numpyro. Here are some examples:

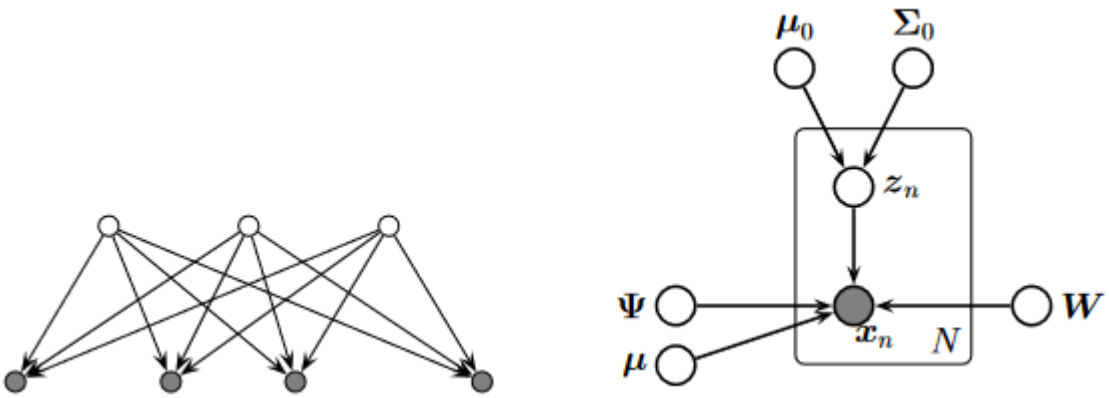


Figure 8: A factor analysis model with 3 latent factors and 4 outcomes, drawn with plating

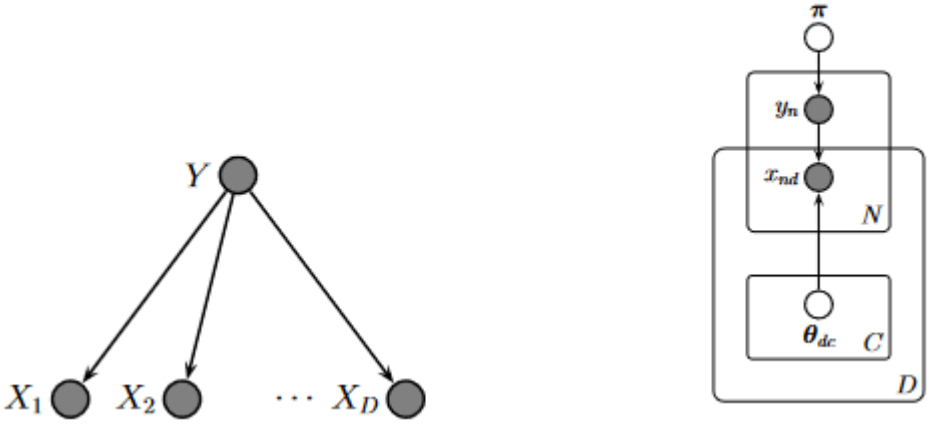


Figure 9: Naive Bayes classifier, drawn with plating; plates can overlap!

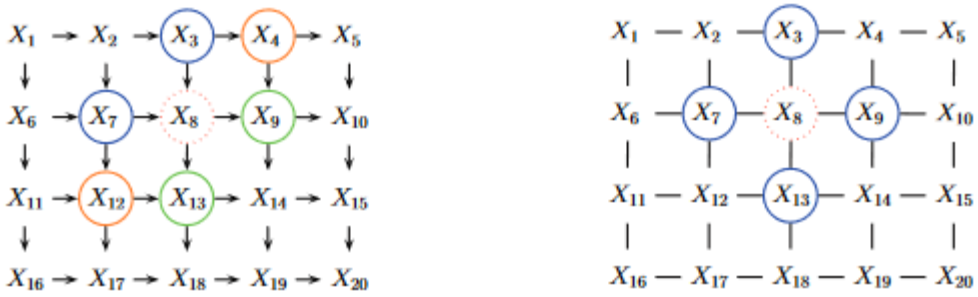


Figure 10: Common applications of MRFs arise in image processing and statistical physics

3 Markov random fields (undirected graphical models)

3.1 Motivation

Sometimes, relationships between variables aren't best thought of as directed. The most obvious example in econ is in the outcomes of a game. In the definition of a Nash equilibrium, both players strategies should inherently depend on each other. There are basically two ways around this problem, when estimating games.

The first way is to specify a structural behavioral process for beliefs and learning. This can be nice if you are studying a particular class of strategies or learning behavior. But different models of learning lead to different predictions, and there is no "one-size fits all" behavioral model.

On the other hand, you could write down a model that actually describes the situation. In a game with rational players, we think that player 1's strategy affects player 2's, and player 2's also affects player 1's. However, this type of dependence is explicitly banned from the Bayesian network paradigm, since it is a cyclic dependence. Therefore, it would be nice to describe a set of models that do not characterize an ordering.

3.2 Joint distribution

An undirected graphical model (Markov Random Field or MRF) does not encode any topological ordering. Therefore, we can't simply employ the chain rule of probabilities to express the joint probability as a product of factors. Instead, we will make use of some nonnegative functions we call *potential functions*.

The *Hammersley-Clifford theorem* explains that an MRF joint density can be factorized as a product of "clique potentials". The MRF clique factorization looks like this

$$p(\mathbf{x} \mid \theta) = \frac{1}{Z(\theta)} \prod_{c \in C} \phi_c(\mathbf{x}_c; \theta_c)$$

where the *partition function* or *normalizing constant* $Z(\theta)$ is defined as

$$Z(\theta) = \sum_{\mathbf{x}} \prod_{c \in C} \phi_c(\mathbf{x}_c; \theta_c)$$

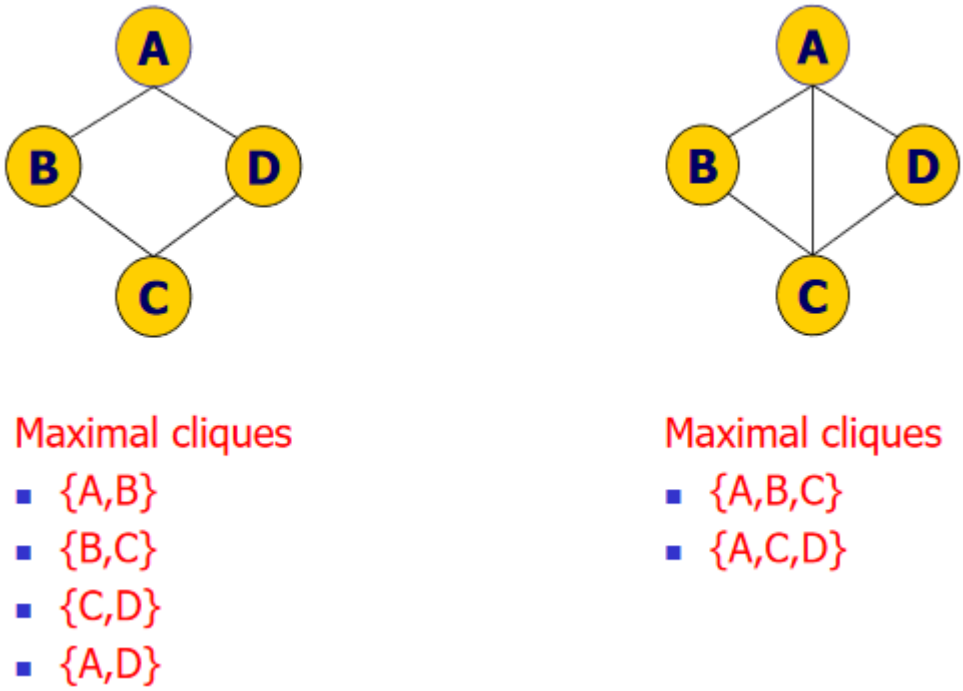


Figure 11: Maximal clique sets in some simple graphs (from Su-In Lee)

Note that in direct calculation of $Z(\theta)$ requires *enumeration* over all values of $\mathbf{x}$. Therefore, it is typically intractable.

3.3 Inspiration from statistical physics: Gibbs distribution

Virtually all MRFs are formulated as examples of a Gibbs distribution. A Gibbs distribution models the probabilities of the particles in a thermodynamic system being in any particular state. The Gibbs density is

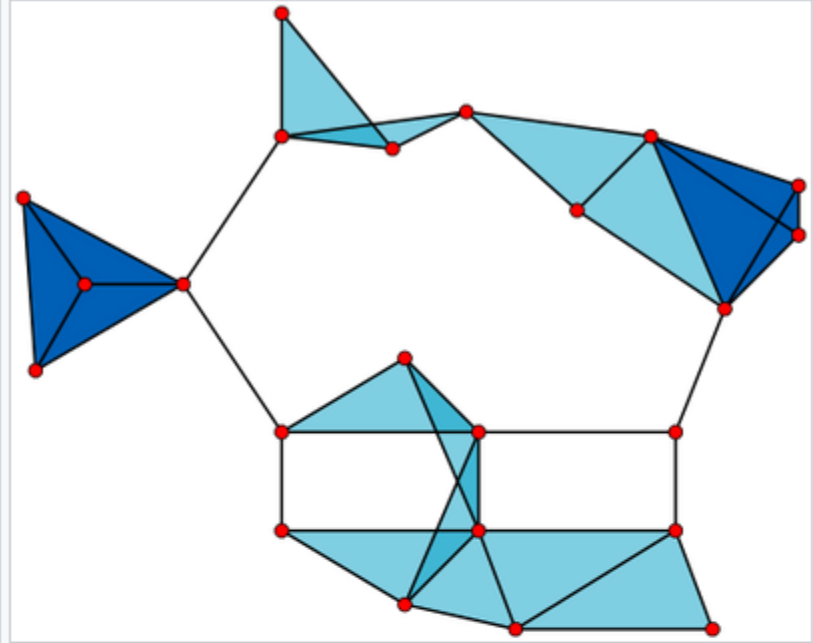
$$p(\mathbf{x} \mid \theta) = \frac{1}{Z(\theta)} \exp(-\mathcal{E}(\mathbf{x}; \theta))$$

Where $\mathcal{E}(\mathbf{x}) > 0$ is the energy of the state $\mathbf{x}$, defined as $\mathcal{E}(\mathbf{x}; \theta) = \sum_{c \in C} \mathcal{E}(\mathbf{x}_c; \theta_c)$. In other words, the system is less likely to be in a high-energy state. Set the clique potentials to $\phi_c(\mathbf{x}_c; \theta_c) = \exp(-\mathcal{E}(\mathbf{x}_c; \theta_c)$ and the two models are equivalent.

This model is where the partition function $Z(\theta)$ gets its name. It encodes information about the cliques or physical partitions of a thermodynamic system.

3.4 Examples

Ising models are a particular variant of MRF that is popular across many fields of science. They are characterized by a lattice structure and binary outcomes (which originally represented particle spins).



A graph with

- 23 × 1-vertex cliques (the vertices),
- 42 × 2-vertex cliques (the edges),
- 19 × 3-vertex cliques (light and dark blue triangles), and
- 2 × 4-vertex cliques (dark blue areas).

The 11 light blue triangles form maximal cliques. The two dark blue 4-cliques are both maximum and maximal, and the clique number of the graph is 4.

Figure 12: Another example of a clique factorization (from Wikipedia)

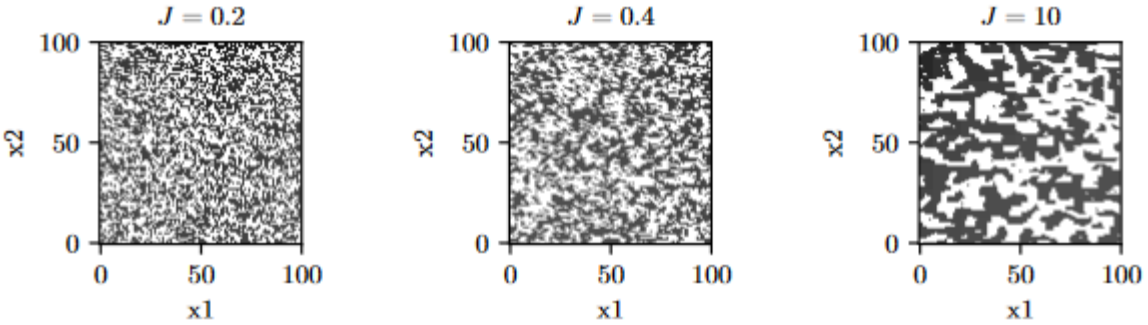


Figure 13: Ising models with varying $J > 0$

In an Ising model, the clique potentials are

$$\psi_{ij}\left(x_i, x_j; \boldsymbol{\theta}\right) = \begin{cases} e^{J_{ij}} & \text{if } x_i = x_j \\ e^{-J_{ij}} & \text{if } x_i \neq x_j \end{cases}$$

Or from the Gibbs perspective, $\mathcal{E}(\mathbf{x}; J) = -\sum_{i \sim j} J_{ij} x_i x_j$

Where J is a symmetric matrix of weights that describes the strength of the interaction between any two variables (e.g. a network). In thermodynamics, it is common to write $J_{ij} = J_{ji} = J$ for all i, j. In this setting, $J > 0$ implies attraction of the variables to similar states and $J < 0$ implies that the states are likely to be different.

3.5 Gaussian graphical models

The multivariate Gaussian distribution is another example of a MRF.

$$\begin{aligned} p(\boldsymbol{x}) &= \frac{1}{Z(\boldsymbol{\theta})} \prod_{i \sim j} \psi_{ij}\left(x_i, x_j\right) \prod_i \psi_i\left(x_i\right) \\ \psi_{ij}\left(x_i, x_j\right) &= \exp\left(-\frac{1}{2} x_i \Lambda_{ij} x_j\right) \\ \psi_i\left(x_i\right) &= \exp\left(-\frac{1}{2} \Lambda_{ii} x_i^2 + \eta_i x_i\right) \\ Z(\boldsymbol{\theta}) &= (2\pi)^{D/2} |\boldsymbol{\Lambda}|^{-\frac{1}{2}} \end{aligned}$$

This means that any estimation of a Gaussian covariance can be seen as inference of a Markov random field. This includes sample covariance matrices $\hat{\Sigma} = \frac{1}{n} X X'$, and hard and soft thresholding based estimators. This tells us that in many cases, the network itself can be inferred as a parameter.

3.6 Moralization

Directed and undirected graphical models encode many similar types of conditional independence relationships. In fact, we can take any Bayesian network and convert it to a MRF

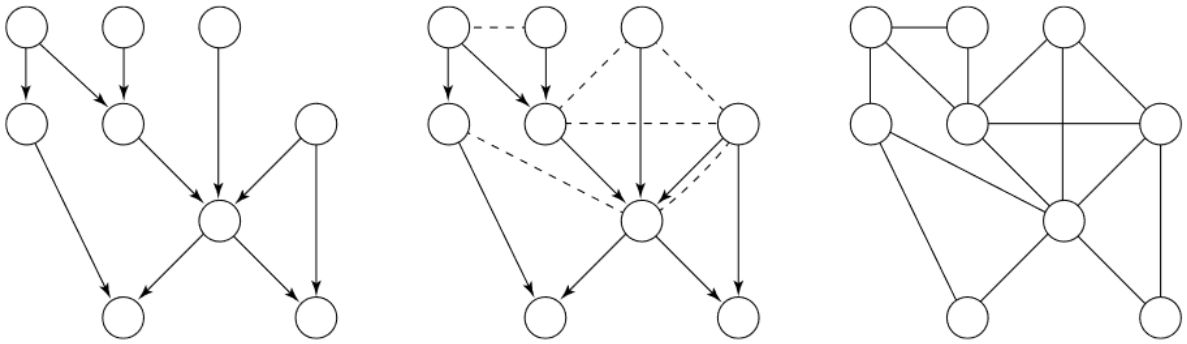


Figure 14: Moralization of a directed graph (from Mark Paskin)

by following a process called *moralization*.
Moralization consists of two steps.

1. Turn all directed edges into undirected edges

2. Connect all common parents

Remember that in a directed graphical model, controlling for a collider opens a path of dependence between the parent nodes. In an undirected model, we have no concept of a collider. Therefore, the fact that when controlling for the collider the parents are not independent of each other, in an undirected network would have to be represented by a link between the parents. In other words, a node's set of neighbors in an moralized graph is the same as that node's Markov blanket in the original model.

In an undirected model, the node's Markov blanket is simply the set of other nodes they are connected to. Conditional on those, they are independent of all other variables in the model.

A byproduct of this process is the realization that Bayesian models can encode more types of conditional independence relationships than undirected models. In fact, the conditional independence relationships of any MRF can be represented as a Bayesian model, but not the other way around. (Simple example, a collider converts to a triangle, but can't be converted back).

References

[1] Kevin P Murphy. *Probabilistic machine learning: Advanced topics*. MIT press, 2023.

[2] Mark Paskin. *A Short Course on Graphical Models. Section 2: Structured Representations*. Tech. rep. Intel Berkeley Research Center, 2003.