Graphical Models

Robin Evans
evans@stats.ox.ac.uk

Michaelmas 2016

This version: December 1, 2016

These notes will be updated as the course goes on. If you find any mistakes or omissions, I'd be very grateful to be informed.

Administration

The course webpage is at

http://www.stats.ox.ac.uk/~evans/gms/

Here you will find problem sheets, slides and any other materials.

Problem Sheets and Classes

There will be four problem sheets, each covering roughly four lectures' material. Part C students should sign up for classes in the usual way, these will be in weeks 3, 5, 7 and Hilary Week 0. MSc students will have classes in weeks 4, 5, 7 and 8:

Sheet	Day	Time	Place
1	Wednesday Week 4	11:00	LG.01
2	Tuesday Week 5	12:00	LG.01
3	Tuesday Week 7	12:00	LG.01
4	Thursday Week 8	11:00	LG.01

Resources

Books are useful, though not required. Here are the main ones this course is based on.

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

The 'bible' of graphical models, and much of the first half of this course is based on this. One complication is that the book makes a distinction between two different types of vertex, which can make some ideas look more complicated.

2. M.J. Wainwright and M.I. Jordan, *Graphical Models, Exponential Families, and Variational Inference*, Foundations and Trends in Machine Learning, 2008.

Relevant for the later part of the course, and for understanding much of the computational advantages of graphical models. Available for free at `https://people.eecs.berkeley.edu/~wainwrig/Papers/WaiJor08_FTML.pdf`.

3. J. Pearl, *Causality*, third edition, Cambridge, 2013.

Book dealing with the causal interpretation of directed models, which we will touch upon.

4. D. Koller and N. Friedman, *Probabilistic Graphical Models: Principles and Techniques*, MIT Press, 2009.

A complementary book, written from a machine learning perspective.

Contents

1 Introduction **5**

2 Conditional Independence **6**

2.1 Independence 6

2.2 Conditional Independence 6

2.3 Statistical Inference 9

3 Contingency Tables **11**

3.1 Computation 11

3.2 Log-linear models 12

3.3 Conditional Independence 12

4 Undirected Graphical Models **15**

4.1 Undirected Graphs 15

4.2 Markov Properties 15

4.3 Cliques and Factorization 17

4.4 Decomposability 18

4.5 Separator Sets 21

4.6 Non-Decomposable Models 22

5 Multivariate Gaussian Distributions **23**

5.1 Gaussian Graphical Models 23

5.2 Maximum Likelihood Estimation 25

5.3 Data Examples 25

6 Directed Graphical Models **26**

6.1 Markov Properties 27

6.2 Ancestrality 28

6.3 Statistical Inference 30

6.4 Markov Equivalence 31

6.5 Directed Graphs, Undirected Graphs, and Decomposability 32

7 Junction Trees and Message Passing **34**

7.1 Junction Trees 35

7.2 Message Passing and the Junction Tree Algorithm 38

7.3 Directed Graphs and Triangulation 41
7.4 Evidence 42

8 Causal Inference **44**
8.1 Interventions 45
8.2 Adjustment Sets and Back-Door Paths 47
8.3 Paths and d-separation 48
8.4 Back-door Adjustment 49
8.5 Example: HIV Treatment 50
8.6 Gaussian Causal Models 51

9 Variational and Monte Carlo Methods **52**
9.1 Exponential Families 52
9.2 Empirical Moment Matching 54
9.3 Maximum Entropy 56
9.4 Duality and Variation 57
9.5 The EM Algorithm 58
9.6 Conjugate Priors 59
9.7 Variational Bayes 60
9.8 Gibbs Sampling 62

1 Introduction

The modern world is replete with sources of massively multivariate data, sometimes called 'big data'. In many cases, the number of variables being measured (p) exceeds the number of samples available (n), and in almost all cases the number of possible ways of classifying individuals is greater than n.

Examples:

- There are around 25,000 human genes, which gives more possible human genomes than humans who have ever existed. Even if a gene is present, whether or not it is expressed depends upon other genes and also environmental factors. Good genetic data sets might have a few thousand individuals in, the best ones have one hundred thousand. How do we study what effect these genes have on diseases, or on each other's expression?
- A doctor has to diagnose one (or more) of hundreds of different possible diseases in a patient with a handful out of thousands of possible symptoms, and with a few pieces of information about his medical history. She can perhaps order some tests to provide evidence in favour of one condition or another. How should she decide whether the evidence is behind a particular condition?
- Photographs are typically made up of millions of pixels, each of which can take one of $256^3 \approx 17$ million colours. How do we train a computer to recognize the object in an image?

The nature of these data sets leads to two related challenges: a statistical challenge and a computational one. Both are features of the so-called *curse of dimensionality*. The statistical problems are easy to see: suppose I ask 1,000 people 10 questions each with two answers. This gives $2^{10} = 1024$ possible response patterns, so that it is impossible to observe all the response patterns, and in practice we won't observe most of them even once. How can we sensibly estimate the probability of those missing response patterns in future?

The computational problem is related. Suppose now that I *know* the distribution of outcomes, so I have $P(X_V = x_V)$ for every $x_V \in \mathcal{X}_V$. How can I compute the marginal probability of a particular variable? Well:

$$P(X_i = x_i) = \sum_{x_{V\setminus\{i\}}} P(X_V = x_V).$$

But notice that, if $p = |V|$ is large, say 1,000 variables, then this sum could easily involve $2^{1000} \approx 10^{301}$ terms! Even for a very fast computer this is totally infeasible, and of course we wouldn't be able to store all the probabilities in the first place.

Each of these examples—although theoretically massive—has a lot of underlying structure that makes the problem potentially tractable. Particular medical symptoms are closely tied to particular diseases, with probabilities that we understand. Adjacent pixels in photographs are often almost the same; if every pixel were completely different we would never discern an image.

Graphical models provide a convenient way of modelling this structure, and make it computationally feasible to perform calculations with the networks.

2 Conditional Independence

The primary tool we will use to provide statistical and computationally feasible models is conditional independence. This ensures that distributions factorize into smaller pieces that can be evaluated separately and quickly.

2.1 Independence

Recall that two discrete variables X and Y are *independent* if

$$P(X = x, Y = y) = P(X = x) \cdot P(Y = y) \qquad \forall x \in \mathcal{X}, y \in \mathcal{Y}.$$

Note that this is equivalent to

$$P(X = x \,|\, Y = y) = P(X = x) \qquad \text{whenever } P(Y = y) > 0, \forall x \in \mathcal{X}.$$

In other words, knowing the value of Y gives us no information about the distribution of X; we say that Y is **irrelevant** for X. Similarly, two variables with joint density f_{XY} are independent if

$$f_{XY}(x, y) = f_X(x) \cdot f_Y(y) \qquad \forall x \in \mathcal{X}, y \in \mathcal{Y}.$$

The qualification that these expressions hold for all $x \in \mathcal{X}$ and $y \in \mathcal{Y}$ is very important[1], and sometimes forgotten.

Example 2.1. Suppose that X, W are independent Exponential(λ) random variables. Define $Y = X + W$. Then the joint density of X and Y is

$$f_{XY}(x, y) = \begin{cases} \lambda^2 e^{-\lambda y} & \text{if } y > x > 0, \\ 0 & \text{otherwise} \end{cases}.$$

Note that the expression within the valid range for x, y factorizes, so when performing the usual change of variables one may mistakenly conclude that X and Y are independent.

2.2 Conditional Independence

Given random variables X, Y we denote the joint density $p(x, y)$, and call

$$p(y) = \int_{\mathcal{X}} p(x, y)\, dx.$$

the **marginal density** (of Y). The **conditional density** of X given Y is defined as any function $p(x \,|\, y)$ such that

$$p(x, y) = p(y) \cdot p(x \,|\, y).$$

Note that if $p(y) > 0$ then the solution is unique and given by the familiar expression

$$p(x \,|\, y) = \frac{p(x, y)}{p(y)}.$$

[1] Of course, for continuous random variables densities are only defined up to a set of measure zero, so the condition should really read 'almost everywhere'. We will ignore such measure theoretic niceties in this course.

Definition 2.2. Let X, Y be random variables defined on a product space $\mathcal{X} \times \mathcal{Y}$; let Z be a third random variable so that the joint density is $p(x, y, z)$. We say that X and Y are *conditionally independent* given Z if

$$p(x \,|\, y, z) = p(x \,|\, z), \qquad \forall x \in \mathcal{X}, y \in \mathcal{Y}, z \in \mathcal{Z} \text{ such that } p(y, z) > 0.$$

When this holds we write $X \perp\!\!\!\perp Y \mid Z \,[p]$, possibly omitting the p for brevity.

In other words, once $Z = z$ is known, the value of Y provides no *additional* information that would allow us to predict or model X. If Z is degenerate—that is, there is some z such that $P(Z = z) = 1$, then the definition above is the same as saying that X and Y are independent. This is called *marginal independence*, and denoted $X \perp\!\!\!\perp Y$.

Example 2.3. Let $X_1, \ldots, X_k$ be a Markov chain. Then X_k is independent of $X_1, \ldots, X_{k-2}$ conditional upon X_{k-1}:

$$P(X_k = x \mid X_{k-1} = x_{k-1}, \cdots, X_1 = x_1) = P(X_k = x \mid X_{k-1} = x_{k-1})$$

for all $x, x_{k-1}, \ldots, x_1$. That is, $X_k \perp\!\!\!\perp X_1, \ldots, X_{k-2} \mid X_{k-1}$. This is known as the *Markov property*, or memoryless property.

Although the definition of conditional independence appears to be asymmetric in X and Y, in fact it is not: if X gives no additional information about Y then the reverse is also true, as the following theorem shows.

Theorem 2.4. *Let X, Y, Z be random variables on a Cartesian product space. The following are equivalent.*

(i) $p(x \,|\, y, z) = p(x \,|\, z)$ *for all* x, y, z *such that* $p(y, z) > 0$*;*

(ii) $p(x, y \,|\, z) = p(x \,|\, z) \cdot p(y \,|\, z)$ *for all* x, y, z *such that* $p(z) > 0$*;*

(iii) $p(x, y, z) = p(y, z) \cdot p(x \,|\, z)$ *for all* x, y, z *such that* $p(z) > 0$*;*

(iv) $p(z) \cdot p(x, y, z) = p(x, z) \cdot p(y, z)$ *for all* x, y, z*;*

(v) $p(x, y, z) = f(x, z) \cdot g(y, z)$ *for some functions* f, g *and all* x, y, z*.*

Proof. Note that $p(y, z) > 0$ implies $p(z) > 0$, so (i) $\implies$ (ii) follows from multiplying by $p(y \,|\, z)$, and (ii) $\implies$ (iii) by multiplying by $p(z)$. (iii) $\implies$ (i) directly.

The equivalence of (iii) and (iv) is also clear, and (iii) implies (v). It remains to prove that (v) implies the others. Suppose that (v) holds. Then

$$p(y, z) = \int p(x, y, z)\, dx = g(y, z) \int f(x, z)\, dx = g(y, z) \cdot \tilde{f}(z).$$

If $\tilde{f}(z) > 0$ (which happens whenever $p(z) > 0$) we have

$$p(x, y, z) = \frac{f(x, z)}{\tilde{f}(z)} p(y, z).$$

But by definition $f(x, z)/\tilde{f}(z)$ is $p(x \,|\, y, z)$, and it does not depend upon y, so we obtain (iii). $\square$

Conditional independence is a complicated and often unintuitive notion, as the next example illustrates.

Example 2.5 (Simpson's Paradox)**.** Below is a famous data set that records the races of the victim and defendants in various murder cases in Florida between 1976 and 1987, and whether or not the death penalty was imposed upon the killer. The data are presented as counts, though we can turn this into an empirical probability distribution by dividing by the total, 674.

Victim	White	
Defendant	White	Black
Yes	53	11
No	414	37

Victim	Black	
Defendant	White	Black
Yes	0	4
No	16	139

The marginal table has

Defendant	White	Black
Yes	53	15
No	430	176

Here we see that the chance of receiving a death sentence is approximately independent of the defendant's race. $P(\text{Death} \mid \text{White}) = 53/(53 + 430) = 0.11$, $P(\text{Death} \mid \text{Black}) = 15/(15 + 176) = 0.08$. (One could fiddle the numbers to obtain exact independence.)

However, restricting only to cases where the victim is white we see that black defendants have nearly a 1/3 chance of receiving the death penalty, compared to about 1/8 for whites. And for black victims the story is the same, a handful of blacks were were sentenced to death while no white defendants were. (In fact we will see in Chapter 3 that this conditional dependence is not statistically significant either, but for the purposes of this discussion this doesn't matter: we could multiply all the numbers by 10 and get a data set in which the correlations *are* significant. For more on this data set, take a look at Example 2.3.2 in the book *Categorical Data Analysis* by Agresti).

The previous example teaches us the valuable lesson that marginal independence does not imply conditional independence (nor vice versa). More generally, conditioning on additional things may result in dependence being induced. However, there are properties that relate conditional independences, the most important of which are given in the next theorem.

Theorem 2.6 (Graphoid Axioms)**.** *Conditional independence satisfies the following properties, sometimes called the* graphoid axioms*.*

1. $X \perp\!\!\!\perp Y \mid Z \implies Y \perp\!\!\!\perp X \mid Z$;
2. $X \perp\!\!\!\perp Y, W \mid Z \implies X \perp\!\!\!\perp Y \mid Z$;
3. $X \perp\!\!\!\perp Y, W \mid Z \implies X \perp\!\!\!\perp W \mid Y, Z$;
4. $X \perp\!\!\!\perp W \mid Y, Z$ *and* $X \perp\!\!\!\perp Y \mid Z \implies X \perp\!\!\!\perp Y, W \mid Z$;
5. *if* $p(x, y, z, w) > 0$, *then* $X \perp\!\!\!\perp W \mid Y, Z$ *and* $X \perp\!\!\!\perp Y \mid W, Z \implies X \perp\!\!\!\perp Y, W \mid Z$.

These properties are sometimes referred to respectively as symmetry, decomposition, weak union, contraction and intersection.

Proof. 1. Symmetry follows from Theorem 2.4

2. Starting from $p(x, y, w \mid z) = p(x \mid z)p(y, w \mid z)$ and integrating out w gives $p(x, y \mid z) = p(x \mid z)p(y \mid z)$.

3. and 4: see Examples sheet.

5. By Theorem 2.4 we have $p(x, y, w, z) = f(x, y, z)g(y, w, z)$ and $p(x, y, w, z) = \tilde{f}(x, w, z)\tilde{g}(y, w, z)$. By positivity, taking ratios shows that

$$\begin{aligned} f(x, y, z) &= \frac{\tilde{f}(x, w, z)\tilde{g}(y, w, z)}{g(y, w, z)} \\ &= \frac{\tilde{f}(x, w_0, z)\tilde{g}(y, w_0, z)}{g(y, w_0, z)} \end{aligned}$$

for any w_0, since the LHS does not depend upon w; now we see that the right hand side is a function of x, z times a function of y, z, so

$$f(x, y, z) = a(x, z) \cdot b(y, z).$$

Plugging into the first expression gives the result. □

Remark 2.7. Properties 2–4 can be combined into a single 'chain rule':

$$X \perp\!\!\!\perp W \mid Y, Z \qquad \text{and} \qquad X \perp\!\!\!\perp Y \mid Z \qquad \iff \qquad X \perp\!\!\!\perp Y, W \mid Z.$$

The fifth property is often extremely useful (as we shall see), but doesn't generally hold if the distribution is not positive: see the Examples Sheet.

2.3 Statistical Inference

Conditional independence crops up in various areas of statistics; here is an example that should be familiar.

Example 2.8. Suppose that $X \sim f_\theta$ for some parameter $\theta \in \Theta$. We say that $T \equiv t(X)$ is a *sufficient statistic* for θ if the likelihood can be written as

$$L(\theta \mid X = x) = f_\theta(x) = g(t(x), \theta) \cdot h(x).$$

Note that under a Bayesian interpretation of θ, this is equivalent to saying that $X \perp\!\!\!\perp \theta \mid T$.

Conditional independence can also give huge computational advantages for dealing with complex distributions and large datasets. Take random variables X, Y, Z on a product space with joint density

$$p_\theta(x, y, z) = g_\eta(x, y) \cdot h_\zeta(y, z), \qquad \forall x, y, z, \theta,$$

for some functions g, h, where $\theta = (\eta, \zeta)$ is a Cartesian product.

Then suppose we wish to find the maximum likelihood estimate of θ; well this is just $\hat{\theta} = (\hat{\eta}, \hat{\zeta})$ where

$$\hat{\eta} = \arg\max_{\eta} \prod_{i=1}^{n} g_\eta(x_i, y_i), \qquad \hat{\zeta} = \arg\max_{\zeta} \prod_{i=1}^{n} h_\zeta(y_i, z_i).$$

So we can maximize these two pieces separately. Notice in particular that we don't need all the data in either case!

If in a Bayesian mood, we might impose a prior $\pi(\eta, \zeta) = \pi(\eta)\pi(\zeta)$. Then

$$\begin{aligned}
\pi(\eta, \zeta \mid \boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z}) &\propto \pi(\eta) \cdot \pi(\zeta) \cdot \prod_i g_\eta(x_i, y_i) \cdot h_\zeta(y_i, z_i) \\
&= \left\{ \pi(\eta) \prod_i g_\eta(x_i, y_i) \right\} \cdot \left\{ \pi(\zeta) \prod_i h_\zeta(y_i, z_i) \right\} \\
&= \pi(\eta \mid \boldsymbol{x}, \boldsymbol{y}) \cdot \pi(\zeta \mid \boldsymbol{y}, \boldsymbol{z}).
\end{aligned}$$

Applying Theorem 2.4(ii) we see that $\eta \perp\!\!\!\perp \zeta \mid \boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{Z}$, and so we can perform inference about this distribution for the two pieces separately (e.g. by running an MCMC procedure or finding the posterior mode).

Indeed, each piece only require part of the data, and for large problems this can be a tremendous computational saving.

3 Contingency Tables

For much of the rest of the course we will be dealing with large numbers of random variables, which we will denote by X_v, taking values in sets $\mathcal{X}_v$, for some index v. For sets of variables, $V = \{1, \ldots, p\}$, we write $X_V \equiv (X_1, \ldots, X_p)$.

In this section we will assume that our variables X_v are discrete with a finite set of levels $\mathcal{X}_v \equiv \{1, \ldots, d_v\}$. Though we use integers as labels, they can represent something completely arbitrary and unordered such as religion, social preference, or a car model.

Given a vector of these categories $X_V^{(i)} = (X_1^{(i)}, \ldots, X_p^{(i)})$ sampled over individuals $i = 1, \ldots, n$, it is helpful to cross-tabulate their responses. Define:

$$n(x_V) \equiv \sum_{i=1}^{n} \mathbb{1}\{X_1^{(i)} = x_1, \ldots, X_p^{(i)} = x_p\},$$

i.e. the number of individuals who have the response pattern x_V. These counts are the sufficient statistics for a multinomial model, whose log-likelihood is

$$l(p; \boldsymbol{n}) = \sum_{x_V} n(x_V) \log p(x_V), \qquad p(x_V) \geq 0, \quad \sum_{x_V} p(x_V) = 1.$$

Each possibility x_V is called a *cell* of the table. Given a subset of the responses $A \subseteq V$ we may be interested in the *marginal table*:

$$n(x_A) \equiv \sum_{x_B} n(x_A, x_B),$$

where $B = V \setminus A$.

Example 3.1. Consider the death penalty data again:

Victim	White	
Defendant	White	Black
Yes	53	11
No	414	37

Victim	Black	
Defendant	White	Black
Yes	0	4
No	16	139

The marginal table has

Defendant	White	Black
Yes	53	15
No	430	176

3.1 Computation

As noted in the introduction, even a moderately sized contingency table will cause statistical problems in practice due to the *curse of dimensionality*. If we have p binary variables, then the contingency table will have 2^p cells. Even for $p = 10$ we will have over a thousand possibilities, and for $p = 50$ there are too many to cells to store in a computer.

Conditional independence can help, however; suppose that $X_A \perp\!\!\!\perp X_B \mid X_S$ for some $A \cup B \cup S = V$, so that we have

$$p(x_V) = p(x_S) \cdot p(x_A \mid x_S) \cdot p(x_B \mid x_S).$$

Now we can store each of these factors in computer memory separately, which means $2^s + 2^{a+s} + 2^{b+s} = 2^s(1 + 2^a + 2^b)$ cells instead of 2^{s+a+b}. This is a considerable saving if s is small. For calculations, if we want to find $P(X_1 = 1)$ and $1 \in A$, then we need only sum over the 2^{s+a} entries in $p(x_S) \cdot p(x_A \mid x_S)$ rather than the 2^{a+b+s} entries in $p(x_V)$.

Of course, if there are several conditional independences then one might imagine that further computational savings are possible: indeed this is correct, and is the main idea behind graphical models.

3.2 Log-linear models

The *log-linear* parameters for $p(x_V) > 0$ are defined by the relation

$$\begin{aligned}\log p(x_V) &= \sum_{A \subseteq V} \lambda_A(x_A) \\ &= \lambda_\emptyset + \lambda_1(x_1) + \cdots + \lambda_V(x_V),\end{aligned}$$

and the identifiability constraint $\lambda_A(x_A) = 0$ whenever $x_a = 1$ for some $a \in A$. (Other identifiability constraints can also be used.)

In the case of binary variables (that is, each variable takes only two states, $d_v = 2$, $\mathcal{X}_v = \{1, 2\}$), there is only one possibly non-zero level for each log-linear parameter $\lambda_A(x_A)$, which is when $x_A = (2, \ldots, 2)$. In this case we will simply write $\lambda_A = \lambda_A(2, \ldots, 2)$. We will proceed under this assumption from now on.

Example 3.2. Consider a 2×2 table with probabilities $\pi_{ij} = P(X = i, Y = j)$. The log-linear parametrization has

$$\begin{aligned}\log \pi_{11} &= \lambda_\emptyset & \log \pi_{21} &= \lambda_\emptyset + \lambda_X \\ \log \pi_{12} &= \lambda_\emptyset + \lambda_Y & \log \pi_{22} &= \lambda_\emptyset + \lambda_X + \lambda_Y + \lambda_{XY}.\end{aligned}$$

From this we can deduce that

$$\lambda_{XY} = \log \frac{\pi_{11}\pi_{22}}{\pi_{21}\pi_{12}}.$$

The quantity λ_{XY} is called the *odds ratio* between X and Y, and is a fundamental quantity in statistical inference.

3.3 Conditional Independence

Log-linear parameters provide a convenient way of expressing conditional independence constraints, since factorization of a density is equivalent to an additive separation of the log-density.

Theorem 3.3. *Let $p > 0$ be a discrete distribution on X_V with associated log-linear parameters λ_C, $C \subseteq V$. The conditional independence $X_a \perp\!\!\!\perp X_b \mid X_{V \setminus \{a,b\}}$ holds if and only if $\lambda_C = 0$ for all $\{a, b\} \subseteq C \subseteq V$.*

Proof. See examples sheet. □

If there is a conditional independence, then the log-linear parameters can be calculated by just looking at the distribution of each 'piece' of the conditional independence separately. For example, suppose that $X_A \perp\!\!\!\perp X_B \mid X_C$, where $A \cup B \cup C = V$. Then by Theorem 2.4, we have

$$p(x_C) \cdot p(x_A, x_B, x_C) = p(x_A, x_C) \cdot p(x_B, x_C),$$

and hence

$$\log p(x_A, x_B, x_C) = \log p(x_A, x_C) + \log p(x_B, x_C) - \log p(x_C).$$

Then applying the log-linear expansions to each term, we get

$$\sum_{W \subseteq V} \lambda_W(x_W) = \sum_{W \subseteq A \cup C} \lambda_W^{AC}(x_W) + \sum_{W \subseteq B \cup C} \lambda_W^{BC}(x_W) - \sum_{W \subseteq C} \lambda_W^{C}(x_W),$$

where λ_{BC} By equating terms we can see that

$$\begin{aligned}
\lambda_W(x_W) &= \lambda_W^{AC}(x_W) && \text{for any } W \subseteq A \cup C \text{ with } W \cap A \neq \emptyset \\
\lambda_W(x_W) &= \lambda_W^{BC}(x_W) && \text{for any } W \subseteq B \cup C \text{ with } W \cap B \neq \emptyset \\
\lambda_W(x_W) &= \lambda_W^{AC}(x_W) + \lambda_W^{BC}(x_W) - \lambda_W^{C}(x_W) && \text{for any } W \subseteq C.
\end{aligned}$$

So under this conditional independence, the log-linear parameters for $p(x_V)$ are easily obtainable from those for $p(x_A, x_C)$ and $p(x_B, x_C)$.

Example 3.4. Let us now try applying this to our death penalty dataset using R. The file `deathpen.txt` is available on the class website.

```
> df <- read.table("deathpen.txt", header=TRUE)
> df


  DeathPen Defendant Victim freq
1      Yes     White  White   53
2       No     White  White  414
3      Yes     Black  White   11
4       No     Black  White   37
5      Yes     White  Black    0
6       No     White  Black   16
7      Yes     Black  Black    4
8       No     Black  Black  139
```

We can fit log-linear models using the `glm()` command with a Poisson response. This gives the model DeathPen $\perp\!\!\!\perp$ Defendant | Victim.

```
> mod1 = glm(freq ~ DeathPen*Victim + Defendant*Victim,
+               family=poisson, data=df)
> summary(mod1)$coefficients
```

The output (edited for brevity) is:

```
Coefficients:
                           Estimate Std. Error z value Pr(>|z|)
(Intercept)                  4.0610     0.1258  32.283  < 2e-16 ***
DeathPenNo                   1.9526     0.1336  14.618  < 2e-16 ***
VictimBlack                 -4.9711     0.5675  -8.760  < 2e-16 ***
DefendantBlack              -2.2751     0.1516 -15.010  < 2e-16 ***
DeathPenNo:VictimBlack       1.7045     0.5237   3.255  0.00114 **
VictimBlack:DefendantBlack   4.4654     0.3041  14.685  < 2e-16 ***
```

We can verify that the coefficient of Victim-Defendant is the same as the marginal log odds-ratio between those two variables by fitting a model that ignores whether or not the death penalty was administered:

```
> mod2 = glm(freq ~ Defendant*Victim,
+               family=poisson, data=df)
> summary(mod2)$coefficients
```

```
                           Estimate Std. Error z value Pr(>|z|)
(Intercept)                 5.45318    0.04627  117.84   <2e-16 ***
DefendantBlack             -2.27513    0.15157  -15.01   <2e-16 ***
VictimBlack                -3.37374    0.25423  -13.27   <2e-16 ***
DefendantBlack:VictimBlack  4.46538    0.30407   14.69   <2e-16 ***
```

Note that the parameter estimates relating to the Defendant's race (and their standard errors) are the same as in the larger model.

It is perhaps easier just to recover the predicted counts under the model:

```
> count1 <- predict.glm(mod1, type="response")
> count1

      1       2       3       4       5       6       7       8
 58.035 408.965   5.965  42.035   0.403  15.597   3.597 139.403
```

Compare these to the actual counts: a goodness of fit test can be performed by using Pearson's χ^2 test or (almost equivalently) by looking at the residual deviance of the model.

4 Undirected Graphical Models

Conditional independence is, in general, a rather complicated object. In fact, one can derive a countably infinite number of properties like those in Theorem 2.6 to try to describe it. Graphical models are a class of conditional independence models with particularly nice properties. In this section we introduce *undirected graphical models*.

4.1 Undirected Graphs

Definition 4.1. Let V be a finite set. An *undirected graph* $\mathcal{G}$ is a pair (V, E) where:

- V are the *vertices*;
- $E \subseteq \{\{i, j\} : i, j \in V, i \neq j\}$ is a set of unordered distinct pairs of V, called *edges*.

We represent graphs by drawing the vertices (also called *nodes*) and then joining pairs of vertices by a line if there is an edge between them.

Example 4.2. The graph in Figure 1(a) has five vertices and six edges:

$$V = \{1, 2, 3, 4, 5\};$$
$$E = \{\{1, 2\}, \{1, 3\}, \{2, 3\}, \{3, 4\}, \{3, 5\}, \{4, 5\}\}.$$

We write $i \sim j$ if $\{i, j\} \in E$, and say that i and j are *adjacent* in the graph. The vertices adjacent to i are called the *neighbours* of i, and the set of neighbours is often called the *boundary* of i and denoted $\mathrm{bd}_{\mathcal{G}}(i)$.

A *path* in a graph is a sequence of adjacent vertices, without repetition. For example, $1 - 2 - 3 - 5$ is a path in the graph in Figure 1(a). However $3 - 1 - 2 - 3 - 4$ is not a path, since the vertex 3 appears twice. The *length* of a path is the number of edges in it. There is trivially a path of length zero from each vertex to itself.

Definition 4.3 (Separation)**.** Let $A, B, S \subseteq V$. We say that A and B are *separated by* S in $\mathcal{G}$ (and write $A \perp_s B \mid S\,[\mathcal{G}]$) if every path from any $a \in A$ to any $b \in B$ contains at least one vertex in S.

For example, $\{1, 2\}$ is separated from $\{5\}$ by $\{3\}$ in Figure 1(a).

Note that there is no need for A, B, S to be disjoint for the definition to make sense, though in practice this is usually assumed.

Given a subset of vertices $W \subseteq V$, we define the *induced subgraph* $\mathcal{G}_W$ of $\mathcal{G}$ to be the graph with vertices W, and all edges from $\mathcal{G}$ whose endpoints are contained in W. For example, the induced subgraph of Figure 1(a) over $\{2, 3, 5\}$ is the graph $2 - 3 - 5$.

We remark that A and B are separated by S (where $S \cap A = S \cap B = \emptyset$) if and only if A and B are separated by $\emptyset$ in $\mathcal{G}_{V \setminus S}$.

4.2 Markov Properties

A graphical model is a statistical model based on the structure of a graph. We associate each vertex v with a random variable X_v, and infer structure (a model) on the joint

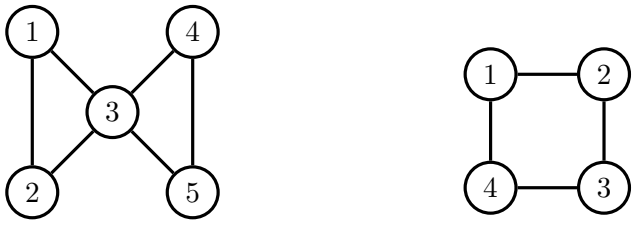


Figure 1: Two undirected graphs.

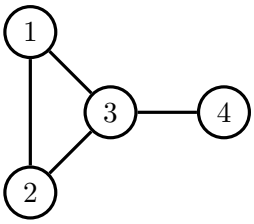


Figure 2: An undirected graph.

distribution of the random variables from the structure of the graph. In all the examples we consider, the model will be defined by conditional independences arising from missing edges in the graph.

Definition 4.4. Let $\mathcal{G}$ be a graph with vertices V, and let p be a probability distribution over the random variables X_V. We say that p satisfies the *pairwise Markov property* for $\mathcal{G}$ if

$$i \not\sim j \text{ in } \mathcal{G} \implies X_i \perp\!\!\!\perp X_j \mid X_{V\setminus\{i,j\}} \, [p].$$

In other words, whenever an edge is missing in $\mathcal{G}$ there is a corresponding conditional independence in p.

Example 4.5. Looking at the graph in Figure 2, we see that there are two missing edges, $\{1,4\}$ and $\{2,4\}$. Therefore a distribution obeys the pairwise Markov property for this graph if and only if $X_1 \perp\!\!\!\perp X_4 \mid X_2, X_3$ and $X_2 \perp\!\!\!\perp X_4 \mid X_1, X_3$.

Note that, if the distribution is positive then we can apply Property 5 of Theorem 2.6 to obtain that $X_1, X_2 \perp\!\!\!\perp X_4 \mid X_3$.

The word 'Markov' is used by analogy with Markov chains, in which a similar independence structure is observed. In fact, undirected graph models are often called *Markov random fields* or *Markov networks* in the machine learning literature.

Definition 4.6. We say that p satisfies the *global Markov property* for $\mathcal{G}$ if for any disjoint sets A, B, S

$$A \perp_s B \mid S \text{ in } \mathcal{G} \implies X_A \perp\!\!\!\perp X_B \mid X_S \, [p].$$

In other words, whenever a separation is present in $\mathcal{G}$ there is a corresponding conditional independence in p.

Proposition 4.7. *The global Markov property implies the pairwise Markov property.*

Proof. If $i \not\sim j$ then clearly any path from i to j first visits a vertex in $V \setminus \{i, j\}$. Hence $V \setminus \{i, j\}$ separates i and j. □

We will shortly see that the pairwise property 'almost' implies the global property.

It is common, though a pet peeve of your lecturer, to confuse a 'graph' with a 'graphical model'. A graph is—as should now be clear from the definitions above—a purely mathematical (as opposed to statistical) object; a graphical model is a statistical model that is based on the structure of a graph.

4.3 Cliques and Factorization

The pairwise Markov property implies a conditional independence involving all the variables represented in a graph for each edge that is missing from the graph; from Theorem 2.4 it is therefore a factorization on the joint distribution. A natural question is whether these separate factorizations can be combined into a single constraint on the joint distribution; in this section we show that they can, at least for positive distributions.

Definition 4.8. Let $\mathcal{G}$ be a graph with vertices V. We say C is *complete* if $i \sim j$ for every $i, j \in C$. A maximal complete set is called a *clique*. We will denote the set of cliques in a graph by $\mathcal{C}(\mathcal{G})$.

The cliques of Figure 1(a) are $\{1, 2, 3\}$ and $\{3, 4, 5\}$, and the complete sets are any subsets of these vertices. Note that $\{v\}$ is trivially complete in any graph.

The graph in Figure 1(b) has cliques $\{1, 2\}$, $\{2, 3\}$, $\{3, 4\}$ and $\{1, 4\}$.

Definition 4.9. Let $\mathcal{G}$ be a graph with vertices V. We say a distribution with density p *factorizes according to* $\mathcal{G}$ if

$$p(x_V) = \prod_{C \in \mathcal{C}(\mathcal{G})} \psi_C(x_C) \tag{1}$$

for some functions ψ_C. The functions ψ_C are called *potentials*.

Recalling Theorem 2.4, it is clear that this factorization implies conditional independence constraints. In fact, it implies those conditional independence statements given by the global Markov property.

Theorem 4.10. *If $p(x_V)$ factorizes according to $\mathcal{G}$, then p obeys the global Markov property with respect to p.*

Proof. Suppose that S separates A and B in $\mathcal{G}$. Let $\tilde{A}$ be the set of vertices that are connected to A by paths in $\mathcal{G}_{V \setminus S}$; in particular, $B \cap \tilde{A} = \emptyset$. Let $\tilde{B} = V \setminus (\tilde{A} \cup S)$, so that $\tilde{A}$ and $\tilde{B}$ are separated by S, $V = \tilde{A} \cup \tilde{B} \cup S$, and $A \subseteq \tilde{A}$, $B \subseteq \tilde{B}$.

Every clique in $\mathcal{G}$ must be a subset of either $\tilde{A} \cup S$ or $\tilde{B} \cup S$, since there are no edges between $\tilde{A}$ and $\tilde{B}$. Hence we can write

$$\begin{aligned} \prod_{C \in \mathcal{C}} \psi_C(x_C) &= \prod_{C \in \mathcal{C}_A} \psi_C(x_C) \cdot \prod_{C \in \mathcal{C}_B} \psi_C(x_C) \\ &= f(x_{\tilde{A}}, x_S) \cdot f(x_{\tilde{B}}, x_S). \end{aligned}$$

and hence $X_{\tilde{A}} \perp\!\!\!\perp X_{\tilde{B}} \mid X_S$. Then applying property 2 of Theorem 2.6 gives $X_A \perp\!\!\!\perp X_B \mid X_S$. □

Theorem 4.11 (Hammersley-Clifford Theorem)**.** *If $p(x_V) > 0$ obeys the pairwise Markov property with respect to $\mathcal{G}$, then p factorizes according to $\mathcal{G}$.*

The proof of this is omitted, but if of interest it can be found in Lauritzen's book.

We can now summarize our Markov properties as follows:

$$\text{factorization} \implies \text{global Markov property} \implies \text{pairwise Markov property},$$

and if p is positive, then we also have

$$\text{pairwise Markov property} \implies \text{factorization},$$

so all three are equivalent. The result is not true in general if p is not strictly positive.

Example 4.12. Let X_3 and X_4 be independent Bernoulli variables with $P(X_3 = 1) = P(X_4 = 1) = \frac{1}{2}$, and $P(X_1 = X_2 = X_4) = 1$. Then $X_4 \perp\!\!\!\perp X_1 \mid X_2, X_3$ and $X_4 \perp\!\!\!\perp X_2 \mid X_1, X_3$, but $X_4 \not\perp\!\!\!\perp X_1, X_2 \mid X_3$.

Hence, P satisfies the pairwise Markov property with respect to Figure 2, but not the global Markov property.

It is important to note that one can define models of the form (1) that are not graphical, if the sets $\mathcal{C}$ do not correspond to the cliques of a graph. See the Examples Sheet.

4.4 Decomposability

Given the discussion in Section 2.3 we might wonder whether we can always perform inference on cliques separately in graphical models? The answer turns out to be that, in general, we can't—at least not without being more careful. However, for a particularly important subclass known as *decomposable models*, we can.

Definition 4.13. Let $\mathcal{G}$ be an undirected graph with vertices $V = A \cup S \cup B$, where A, B, S are disjoint sets. We say that (A, S, B) constitutes a *decomposition* of $\mathcal{G}$ if:

- $\mathcal{G}_S$ is complete;
- A and B are separated by S in $\mathcal{G}$.

If A and B are both non-empty we say the decomposition is *proper*.

Example 4.14. Consider the graph in Figure 1(a). Here $\{1, 2\}$ is separated from $\{4, 5\}$ by $\{3\}$, and $\{3\}$ is trivially complete so $(\{1, 2\}, \{3\}, \{4, 5\})$ is a decomposition. Note that $(\{2\}, \{1, 3\}, \{4, 5\})$ is also a decomposition, for example. We say that a decomposition is *minimal* if there is no subset of S that can be used to separate A and B.

The graph in Figure 1(b) cannot be decomposed, since the only possible separating sets are $\{1, 3\}$ and $\{2, 4\}$, which are not complete. A graph which cannot be (properly) decomposed is called *prime*.

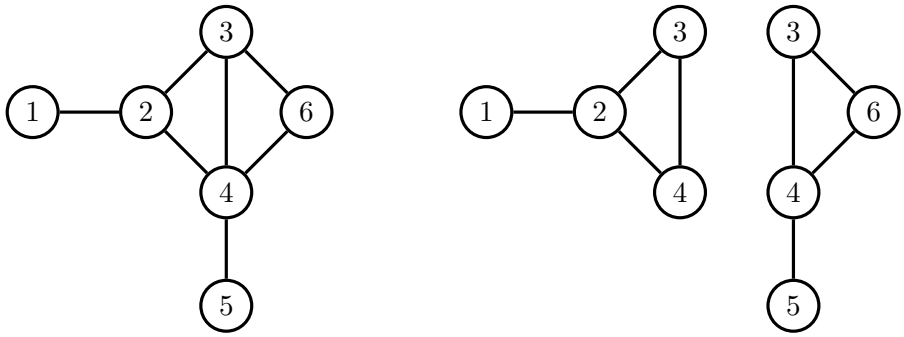


Figure 3: Left: a decomposable graph. Right: the results of a possible decomposition of the graph, $(\{1,2\},\{3,4\},\{5,6\})$.

Definition 4.15. Let $\mathcal{G}$ be a graph. We say that $\mathcal{G}$ is *decomposable* if it is complete, or there is a proper decomposition (A,S,B) and both $\mathcal{G}_{A\cup S}$ and $\mathcal{G}_{B\cup S}$ are also decomposable.

The graph in Figure 1(a) is decomposable, because using the decomposition $(\{1,2\},\{3\},\{4,5\})$ we can see that $\mathcal{G}_{\{1,2,3\}}$ and $\mathcal{G}_{\{3,4,5\}}$ are complete (and therefore decomposable by definition).

The graph in Figure 3 can be decomposed as shown, into $\mathcal{G}_{\{1,2,3,4\}}$ and $\mathcal{G}_{\{3,4,5,6\}}$, both of which are themselves decomposable.

Definition 4.16. Let $\mathcal{C}$ be a collection of subsets of V. We say that the sets $\mathcal{C}$ satisfy the *running intersection property* if there is an ordering $C_1,\dots,C_k$, such that for every $j=2,\dots,k$ there exists $\sigma(j)<j$ with

$$C_j \cap \bigcup_{i=1}^{j-1} C_i = C_j \cap C_{\sigma(j)}.$$

In other words, the intersection of each set with all the previously seen objects is contained in a single set.

Example 4.17. The sets $\{1,2,3\}$, $\{3,4\}$, $\{2,3,5\}$, $\{3,5,6\}$ satisfy the running intersection property, under that ordering.

The sets $\{1,2\}$, $\{2,3\}$, $\{3,4\}$, $\{1,4\}$ cannot be ordered in such a way.

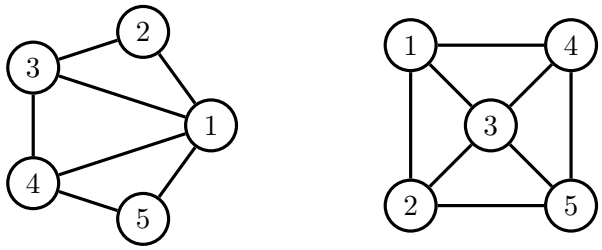
Proposition 4.18. *If $C_1,\dots,C_k$ satisfy the running intersection property, then there is a graph whose cliques are precisely (the inclusion maximal elements of) $\mathcal{C}=\{C_1,\dots,C_k\}$.*

Proof. This is left as an exercise for the interested reader. □

Definition 4.19. Let $\mathcal{G}$ be an undirected graph. A *cycle* is a sequence of vertices $\langle v_1,\dots,v_k\rangle$ for $k\geq 3$, such that there is a path $v_1-\cdots-v_k$ and an edge v_k-v_1.

A *chord* on a cycle is any edge between two vertices not adjacent on the cycle. We say that a graph is *chordal* or *triangulated* if whenever there is a cycle of length ≥ 4, it contains a chord.

Beware of taking the word 'triangulated' at face value: the graph in Figure 4(b) is *not* triangulated because of the cycle $1-2-5-4$, which contains no chords.



(a) (b)

Figure 4: Two undirected graphs: (a) is chordal, (b) is not.

Theorem 4.20. *Let $\mathcal{G}$ be an undirected graph. The following are equivalent:*

(i) $\mathcal{G}$ is decomposable;

(ii) $\mathcal{G}$ is triangulated;

(iii) every minimal a, b-separator is complete;

(iv) the cliques of $\mathcal{G}$ satisfy the running intersection property.

Proof. (i) $\implies$ (ii). We proceed by induction on p, the number of vertices in the graph. Let $\mathcal{G}$ be decomposable; if it is complete then it is clearly triangulated, so the result holds for $p = 1$. Otherwise, let (A, S, B) be a proper decomposition, so that $\mathcal{G}_{A\cup S}$ and $\mathcal{G}_{B\cup S}$ are both have strictly fewer vertices and are decomposable. By the induction hypothesis, there are no chordless cycles entirely contained in $A \cup S$ or $B \cup S$, so any such cycle must contain a vertex $a \in A$ and $b \in B$. Then the cycle must pass through S twice, and since S is complete this means there is a chord on the cycle.

(ii) $\implies$ (iii). Suppose there is a minimal a, b-separator, say S, which is not complete; let $s_1, s_2 \in S$ be non-adjacent. Since the separator is minimal there is a path π_1 from a to b via $s_1 \in S$, and another path π_2 from a to b via $s_2 \in S$, and neither of these paths intersects any other element of S. By concatenating the paths we obtain a closed walk; by shrinking the end of the paths to any vertices which are common to both we obtain a cycle. Make the cycle of minimal length by traversing chords, and we end up with a chordless cycle of length ≥ 4.

(iii) $\implies$ (iv). If the graph is complete there is nothing to prove, otherwise pick a, b not adjacent and let S be a minimal separator. As in Theorem 4.10, let $\tilde{A}$ be the connected component of a in $\mathcal{G}_{V\setminus S}$, and $\tilde{B}$ the rest. Then apply the result by induction to the strictly smaller graphs $\mathcal{G}_{\tilde{A}\cup S}$ and $\mathcal{G}_{\tilde{B}\cup S}$. Then claim that this gives a series of cliques that satisfies the RIP. [See Examples Sheet 2.]

(iv) $\implies$ (i). We proceed by induction, on the number of cliques. If $k = 1$ there is nothing to prove. Let $R_k = V \setminus C_k$, and $S_k = C_k \cap \bigcup_{i=1}^{k-1} C_i$; we claim that $(R_k, S_k, C_k \setminus S_k)$ is a proper decomposition, and that the graph $\mathcal{G}_{R_k\cup S_k}$ has $k - 1$ cliques that also satisfy the running intersection property. □

Corollary 4.21. *Let $\mathcal{G}$ be decomposable and (A, S, B) be a proper decomposition. Then $\mathcal{G}_{A\cup S}$ and $\mathcal{G}_{B\cup S}$ are also decomposable.*

Proof. If $\mathcal{G}$ is triangulated then so are any induced subgraphs of $\mathcal{G}$. □

This corollary reassures us that to check if a graph is decomposable we can just go ahead and start decomposing, and we will never have to 'back track'.

Definition 4.22. A *forest* is a graph that contains no cycles. If a forest is connected we call it a *tree*.

All forests (and hence trees) are decomposable, since they are clearly triangulated. In fact, the relationship between trees and connected decomposable graphs is more fundamental than this. Decomposable graphs are 'tree-like', in a sense we will make precise later in the course (Section 7). This turns out to be extremely useful for computational reasons.

4.5 Separator Sets

Let $\mathcal{G}$ be a decomposable graph, and let $C_1, \ldots, C_k$ be an ordering of the cliques which satisfies running intersection. Define the jth *separator set* for $j \geq 2$ as

$$S_j \equiv C_j \cap \bigcup_{i=1}^{j-1} C_i = C_j \cap C_{\sigma(j)}.$$

By convention $S_1 = \emptyset$.

Lemma 4.23. *Let $\mathcal{G}$ be a graph with decomposition (A, S, B), and let p be a distribution; then p factorizes with respect to $\mathcal{G}$ if and only if its marginals $p(x_{A\cup S})$ and $p(x_{B\cup S})$ factorize according to $\mathcal{G}_{A\cup S}$ and $\mathcal{G}_{B\cup S}$ respectively, and*

$$p(x_V) \cdot p(x_S) = p(x_{A\cup S}) \cdot p(x_{B\cup S}). \tag{2}$$

Proof. If (2) and the other factorizations hold then the fact that p factorizes with respect to $\mathcal{G}$ is clear.

Now suppose the converse. From the decomposition, we have $A \perp_s B \mid S$ in $\mathcal{G}$, and so the equation (2) must hold by Theorem 2.4. Since this is a decomposition, all cliques of $\mathcal{G}$ are contained either within $A\cup S$ or $B\cup S$ (or both). Let $\mathcal{A}$ be the cliques contained in $A\cup S$, and $\mathcal{B}$ the rest.

Then $p(x_V) = \prod_{C\in\mathcal{A}} \psi_C(x_C) \cdot \prod_{C\in\mathcal{B}} \psi_C(x_C) = h(x_A, x_S) \cdot k(x_B, x_S)$. By Theorem 2.4, we have that (2) holds; substituting $p(x_V)$ into (2) and integrating both sides with respect to x_A gives

$$p(x_S) \cdot k(x_B, x_S) \int h(x_A, x_S)\, dx_A = p(x_S) \cdot p(x_B, x_S)$$
$$p(x_S) \cdot k(x_B, x_S) \cdot \tilde{h}(x_S) = p(x_S) \cdot p(x_B, x_S),$$

which shows that $p(x_B, x_S) = \psi'_S(x_S) \prod_{C\in\mathcal{B}} \psi_C$ as required. □

Theorem 4.24. *Let $\mathcal{G}$ be a decomposable graph with cliques $C_1, \ldots, C_k$. Then p factorizes with respect to $\mathcal{G}$ if and only if*

$$p(x_V) = \prod_{i=1}^{k} p(x_{C_i\setminus S_i} \mid x_{S_i}) = \prod_{i=1}^{k} \frac{p(x_{C_i})}{p(x_{S_i})}.$$

Further, these quantities are variation independent, so inference for $p(x_V)$ can be based on separate inferences for each $p(x_{C_i})$.

Proof. If p factorizes in the manner suggested then it satisfies the factorization property for $\mathcal{G}$.

For the converse we proceed by induction on k. If $k = 1$ the result is trivial. Otherwise, note that $C_k \setminus S_k$ is separated from $H_k \equiv \left(\bigcup_{i<k} C_i\right) \setminus S_k$ by S_k, so we have a decomposition $(H_k, S_k, C_k \setminus S_k)$, and hence applying Lemma 4.23,

$$p(x_{S_k}) \cdot p(x_V) = p(x_{C_k}) \cdot p(x_{H_k}, x_{S_k})$$

where $p(x_{H_k}, x_{S_k})$ factorizes according to $\mathcal{G}_{H_k \cup S_k}$. This is the graph with cliques $C_1, \ldots, C_{k-1}$, which trivially also satisfy running intersection. Hence, by the induction hypothesis

$$p(x_{S_k}) \cdot p(x_V) = p(x_{C_k}) \cdot \prod_{i=1}^{k-1} \frac{p(x_{C_i})}{p(x_{S_i})},$$

giving the required result.

The variation independence follows from the fact that $p(x_{C_k \setminus S_k} \mid x_{S_k})$ can take the form of any valid probability distribution. □

This result is extremely useful for statistical inference, since we only need to consider the margins of variables corresponding to cliques. Suppose we have a contingency table with counts $n(x_V)$. The likelihood for a decomposable graph is

$$\begin{aligned} l(p; n) &= \sum_{x_V} n(x_V) \log p(x_V) \\ &= \sum_{x_V} n(x_V) \sum_{i=1}^{k} \log p(x_{C_i \setminus S_i} \mid x_{S_i}) \\ &= \sum_{i=1}^{k} \sum_{x_{C_i}} n(x_{C_i}) \log p(x_{C_i \setminus S_i} \mid x_{S_i}), \end{aligned}$$

so inference about $p(x_{C_i \setminus S_i} \mid x_{S_i})$ should be based entirely upon $n(x_{C_i})$. Using Lagrange multipliers (see also Sheet 0, Question 4) we can see that the likelihood is maximized by choosing

$$\hat{p}(x_{C_i \setminus S_i} \mid x_{S_i}) = \frac{n(x_{C_i})}{n(x_{S_i})}, \qquad \text{i.e. } \hat{p}(x_{C_i}) = \frac{n(x_{C_i})}{n},$$

using the empirical distribution for each clique.

4.6 Non-Decomposable Models

It would be natural to ask at this point whether the closed-form results for decomposable models also hold for general undirected graph models; unfortunately they do not. However, we can say the following:

Theorem 4.25. *Let $\mathcal{G}$ be an undirected graph, and suppose we have counts $n(x_V)$. Then the maximum likelihood estimate $\hat{p}$ under the set of distributions that are Markov to $\mathcal{G}$ is the unique element in which*

$$n \cdot \hat{p}(x_C) = n(x_C).$$

In fact, more is true, but we will leave this point until we have studied exponential families in Section 9.

5 Multivariate Gaussian Distributions

Let $X_V = (X_1, \ldots, X_p)^T \in \mathbb{R}^p$ be a random vector. Let $\mu \in \mathbb{R}^p$ and $\Sigma \in \mathbb{R}^{p \times p}$ be a positive definite symmetric matrix. We say that X_V has a **multivariate Gaussian distribution** with parameters μ and Σ if the joint density is

$$f(x_V) = \frac{1}{(2\pi)^{p/2}|\Sigma|^{1/2}} \exp\left\{-\frac{1}{2}(x_V - \mu)^T \Sigma^{-1}(x_V - \mu)\right\}, \qquad x_V \in \mathbb{R}^p.$$

This is also called the multivariate normal distribution.

Exercise: show that $\mathbb{E}X = \mu$ and $\operatorname{Cov} X = \Sigma$ (hint: using the Cholesky decomposition write $\Sigma = LL^T$, where L is a lower triangular invertible matrix, and use a change of variables).

The **concentration matrix** is $K \equiv \Sigma^{-1}$.

Proposition 5.1. *Let $X_V \sim N_p(\mu, \Sigma)$, and let A be a $q \times p$ matrix of full rank q. Then*

$$AX_V \sim N_q(A\mu, A\Sigma A^T).$$

In particular, for any $U \subseteq V$ we have $X_U \sim N_q(\mu_U, \Sigma_{UU})$.

Proof sketch (you should fill in the gaps). For $q = p$ this just follows from applying the transformation $Z = AX_V$ to the density of X_V. If $q < p$ then since Σ is positive definite we can write $\Sigma = LL^T$ for a non-singular lower triangular matrix L; then construct a non-singular $p \times p$ matrix

$$\tilde{A} = \begin{pmatrix} A \\ B \end{pmatrix}$$

whose first q rows are A, and such that $\tilde{A}L$ has its first q rows orthogonal to its last $p - q$ rows. Then

$$\tilde{A}\Sigma\tilde{A}^T = \begin{pmatrix} A\Sigma A^T & 0 \\ 0 & B\Sigma B^T \end{pmatrix}$$

and the first q components have the desired marginal distribution. □

For simplicity of notation, we will assume throughout that $\mu = 0$.

5.1 Gaussian Graphical Models

We only consider the case in which Σ is positive definite, so all our density functions are strictly positive. Hence, by the Hammersley-Clifford Theorem, the pairwise and global Markov properties, and the factorization criterion all lead to the same conditional independence restrictions. If any of these hold, we will say that Σ 'is Markov with respect to' a graph, without ambiguity.

Proposition 5.2. *Let X_V have a multivariate Gaussian distribution with concentration matrix $K = \Sigma^{-1}$. Then $X_A \perp\!\!\!\perp X_B \mid X_{V \setminus (A \cup B)}$ if and only if $K_{AB} = 0$.*

In addition, $X_A \perp\!\!\!\perp X_B$ if and only if $\Sigma_{AB} = 0$.

Proof. Looking at the log-likelihood, it is clear that the only term involving both x_a and x_b is $-\frac{1}{2}x_a x_b k_{ab}$. Hence, if $k_{ab} = 0$ then the log-likelihood has separate terms for each x_a, x_b, and the likelihood factorizes.

The second claim follows from Proposition 5.1. □

Note that as a corollary of this, $X \perp\!\!\!\perp Y$ and $X \perp\!\!\!\perp Z$ *does* imply $X \perp\!\!\!\perp Y, Z$ for jointly Gaussian random variables.

Theorem 5.3. *Let $X_V \sim N_p(\mu, \Sigma)$ for positive definite Σ, with $K = \Sigma^{-1}$. Then the distribution of X_V is Markov with respect to $\mathcal{G}$ if and only if $k_{ab} = 0$ whenever $a \not\sim b$ in $\mathcal{G}$.*

Proof. This follows immediately from Proposition 5.2. □

We introduce some notation for convenience. If M is a matrix whose rows and columns are indexed by $A \subseteq V$, we write $\{M\}_{A,A}$ to indicate the matrix indexed by V whose A, A-entries are M and with zeroes elsewhere.

Lemma 5.4. *Let $\mathcal{G}$ be a graph with decomposition (A, S, B), and $X_V \sim N_p(0, \Sigma)$. Then $p(x_V)$ is Markov with respect to $\mathcal{G}$ if and only if*

$$\Sigma^{-1} = \left\{(\Sigma_{A\cup S, A\cup S})^{-1}\right\}_{A\cup S, A\cup S} + \left\{(\Sigma_{B\cup S, B\cup S})^{-1}\right\}_{B\cup S, B\cup S} - \left\{(\Sigma_{S,S})^{-1}\right\}_{S,S},$$

and $\Sigma_{A\cup S, A\cup S}$ and $\Sigma_{B\cup S, B\cup S}$ are Markov with respect to $\mathcal{G}_{A\cup S}$ and $\mathcal{G}_{B\cup S}$ respectively.

Proof. We know from Lemma 4.23 that

$$p(x_V) \cdot p(x_S) = p(x_A, x_S) \cdot p(x_B, x_S).$$

where $p(x_A, x_S)$ and $p(x_B, x_S)$ are Markov with respect to $\mathcal{G}_{A\cup S}$ and $\mathcal{G}_{B\cup S}$ respectively. Since margins of multivariate Gaussians are also multivariate Gaussian, we can insert the appropriate density for each term, take logs and rearrange to see that:

$$x_V^T \Sigma^{-1} x_V + x_S^T (\Sigma_{SS})^{-1} x_S = x_{A\cup S}^T (\Sigma_{A\cup S, A\cup S})^{-1} x_{A\cup S} + x_{B\cup S}^T (\Sigma_{B\cup S, B\cup S})^{-1} x_{B\cup S} + \text{const.}$$

Now, comparing coefficients for each term we have

$$\Sigma^{-1} = \left\{(\Sigma_{A\cup S, A\cup S})^{-1}\right\}_{A\cup S, A\cup S} + \left\{(\Sigma_{B\cup S, B\cup S})^{-1}\right\}_{B\cup S, B\cup S} - \left\{(\Sigma_{S,S})^{-1}\right\}_{S,S},$$

This gives the result. □

Applying the previous result to a decomposable graph repeatedly we see that X_V is Markov with respect to $\mathcal{G}$ if and only if

$$\Sigma^{-1} = \sum_{i=1}^{k} \left\{(\Sigma_{C_i, C_i})^{-1}\right\}_{C_i, C_i} - \sum_{i=2}^{k} \left\{(\Sigma_{S_i, S_i})^{-1}\right\}_{S_i, S_i}.$$

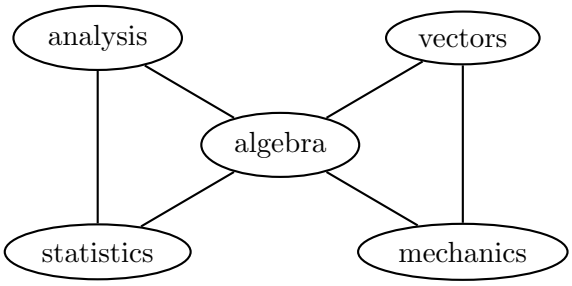


Figure 5: A graph for the maths test data.

5.2 Maximum Likelihood Estimation

Let $X_V^{(1)}, \ldots, X_V^{(n)}$ be i.i.d. $N_p(0, \Sigma)$; then the maximum likelihood estimator of Σ is the sample covariance matrix

$$\hat{\Sigma} = W \equiv \frac{1}{n} \sum_{i=1}^{n} X_V^{(i)} X_V^{(i)T}.$$

Let $\hat{\Sigma}^{\mathcal{G}}$ denote the MLE for Σ under the restriction that the distribution satisfies the Markov property for $\mathcal{G}$, and $\hat{K}^{\mathcal{G}}$ its inverse.

Then, since the MLE of a saturated Gaussian is to have $\Sigma = W$, the maximum likelihood estimate for $K = \Sigma^{-1}$ is given by

$$\left(\hat{\Sigma}^{\mathcal{G}}\right)^{-1} = \sum_{i=1}^{k} \left\{(W_{C_i,C_i})^{-1}\right\}_{C_i,C_i} - \sum_{i=2}^{k} \left\{(W_{S_i,S_i})^{-1}\right\}_{S_i,S_i}.$$

5.3 Data Examples

Example 5.5. Whittaker (1990) analyses data on five maths test results administered to 88 students, in analysis, algebra, vectors, mechanics and statistics. The empirical concentration matrix (i.e. S^{-1}) is given by the following table (entries multiplied by 10^3)

	mechanics	vectors	algebra	analysis	statistics
mechanics	5.24	-2.43	-2.72	0.01	-0.15
vectors	-2.43	10.42	-4.72	-0.79	-0.16
algebra	-2.72	-4.72	26.94	-7.05	-4.70
analysis	0.01	-0.79	-7.05	9.88	-2.02
statistics	-0.15	-0.16	-4.70	-2.02	6.45

Notice that some of the entries in the concentration matrix are quite small, suggesting that conditional independence holds. Indeed, fitting the graphical model in Figure 5 gives an excellent fit (see Example Sheet 2). The model suggests that ability in analysis and statistics is independent of ability in mechanics and vector calculus, conditional on one's fundamental abilities in algebra.

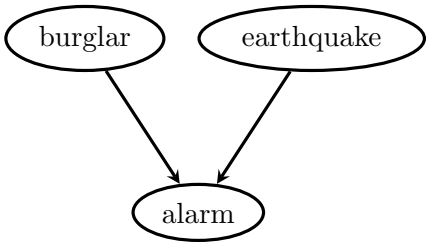


Figure 6: A directed graph representing a burglar alarm and the reasons it might go off.

6 Directed Graphical Models

Undirected graphs represent symmetrical relationships between random variables: the vertices in an undirected graph are typically unordered. In many realistic situations the relationships we wish to model are not symmetric: for example, in regression we have a outcome that is modelled as a function of covariates, and implicitly this suggests that the covariates 'come before' the outcome (in a temporal sense or otherwise).

A further limitation of undirected graphs is that they are only able to represent conditional independences; marginal independences arise very naturally. For example, suppose that we have independent inputs to a system, and an output that is a (random) function of the inputs. An example is given in Figure 6.

Such situations are naturally represented by a directed graph.

Definition 6.1. A *directed graph* $\mathcal{G}$ is a pair (V, D), where

- V is a finite set of *vertices*; and
- $D \subseteq V \times V$ is a collection of *edges*, which are *ordered* pairs of vertices. Loops (i.e. edges of the form (v, v)) are not allowed.

If $(v, w) \in D$ we write $v \to w$, and say that v is a *parent* of w, and conversely w a *child* of v. Examples are given in Figures 6 and 7(a).

We still say that v and w are *adjacent* if $v \to w$ or $w \to v$. A *path* in $\mathcal{G}$ is a sequence of distinct vertices such that each adjacent pair in the sequence is adjacent in $\mathcal{G}$. The path is *directed* if all the edges point away from the beginning of the path.

For example, in the graph in Figure 7(a), 1 and 2 are parents of 3. There is a path $1 \to 3 \leftarrow 2 \to 5$, and there is a directed path $1 \to 3 \to 5$ from 1 to 5.

The set of parents of w is $\mathrm{pa}_{\mathcal{G}}(w)$, and the set of children of v is $\mathrm{ch}_{\mathcal{G}}(v)$.

Definition 6.2. A graph contains a *directed cycle* if there is a directed path from v to w together with an edge $w \to v$. A directed graph is *acyclic* if it contains no directed cycles. We call such graphs *directed acyclic graphs* (DAGs).

All the directed graphs considered in this course are acyclic.

A *topological ordering* of the vertices of the graph is an ordering $1, \ldots, k$ such that $i \in \mathrm{pa}_{\mathcal{G}}(j)$ implies that $i < j$. That is, vertices at the 'top' of the graph come earlier in the

ordering. Acyclicity ensures that a topological ordering always exists (see the Examples Sheet).

We say that a is an *ancestor* of v if *either* $a = v$, or there is a directed path $a \to \cdots \to v$. The set of ancestors of v is denoted by $\mathrm{an}_{\mathcal{G}}(v)$. The ancestors of 4 in the DAG in Figure 7(a) are $\mathrm{an}_{\mathcal{G}}(4) = \{2, 4\}$. The *descendants* of v are defined analogously and denoted $\mathrm{de}_{\mathcal{G}}(v)$; the *non-descendants* of v are $\mathrm{nd}_{\mathcal{G}}(v) \equiv V \setminus \mathrm{de}_{\mathcal{G}}(v)$. The non-descendants of 4 in Figure 7(a) are $\{1, 2, 3\}$.

6.1 Markov Properties

As with undirected graphs, we will associate a model with each DAG via various Markov properties. The most natural way to describe the model associated with a DAG is via a factorization criterion, so this is where we begin.

For any multivariate probability distribution $p(x_V)$, given an arbitrary ordering of the variables $x_1, \ldots, x_k$, we can iteratively use the definition of conditional distributions to see that

$$p(x_V) = \prod_{i=1}^{k} p(x_i \mid x_1, \ldots, x_{i-1}).$$

A directed acyclic graph model uses this form with a topological ordering of the graph, and states that the right-hand side of each factor only depends upon the parents of i.

Definition 6.3 (Factorization Property)**.** Let $\mathcal{G}$ be a directed acyclic graph with vertices V. We say that a probability distribution $p(x_V)$ *factorizes with respect to* $\mathcal{G}$ if

$$p(x_V) = \prod_{v \in V} p(x_v \mid x_{\mathrm{pa}_{\mathcal{G}}(v)}), \qquad\qquad x_V \in \mathcal{X}_V.$$

This is clearly a conditional independence model; given a total ordering on the vertices V, let $\mathrm{pre}_<(v) = \{w \mid w < v\}$ denote all the vertices that precede v according to the ordering. It is not hard to see that we are requiring

$$p(x_v \mid x_{\mathrm{pre}_<(v)}) = p(x_v \mid x_{\mathrm{pa}_{\mathcal{G}}(v)}), \qquad\qquad v \in V$$

for an arbitrary topological ordering of the vertices $<$. That is,

$$X_v \perp\!\!\!\perp X_{\mathrm{pre}_<(v) \setminus \mathrm{pa}_{\mathcal{G}}(v)} \mid X_{\mathrm{pa}_{\mathcal{G}}(v)} \, [p]. \tag{3}$$

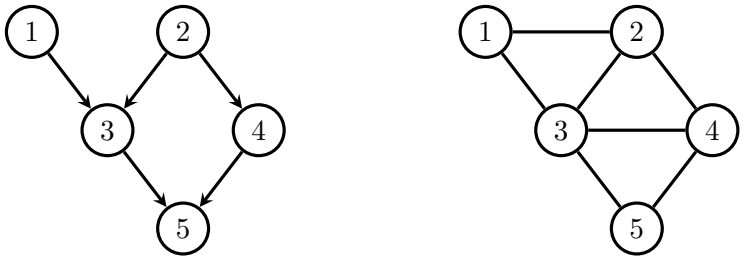
Since the ordering is arbitrary provided that it is topological, we can pick $<$ so that as many vertices come before v as possible; then we see that (3) implies

$$X_v \perp\!\!\!\perp X_{\mathrm{nd}_{\mathcal{G}}(v) \setminus \mathrm{pa}_{\mathcal{G}}(v)} \mid X_{\mathrm{pa}_{\mathcal{G}}(v)} \, [p]. \tag{4}$$

Distributions are said to obey the *local Markov property* with respect to $\mathcal{G}$ if they satisfy (4) for every $v \in V$.

For example, the local Markov property applied to each vertex in Figure 7(a) would require that

$$X_1 \perp\!\!\!\perp X_2, X_4 \qquad X_2 \perp\!\!\!\perp X_1 \qquad X_3 \perp\!\!\!\perp X_4 \mid X_1, X_2$$
$$X_4 \perp\!\!\!\perp X_1, X_3 \mid X_2 \qquad X_5 \perp\!\!\!\perp X_1, X_2 \mid X_3, X_4$$



(a) (b)

Figure 7: (a) A directed graph and (b) its moral graph.

There is some redundancy here, but not all independences that hold are given directly. For example, using Theorem 2.6 we can deduce that $X_4, X_5 \perp\!\!\!\perp X_1 \mid X_2, X_3$, but we might wonder if there is a way to tell this immediately from the graph. For such a 'global Markov property' we need to do a bit more work.

6.2 Ancestrality

We say that a set of vertices A is *ancestral* if it contains all its own ancestors. So, for example, the set $\{1, 2, 4\}$ is ancestral in Figure 7(a); however $\{1, 3\}$ is not, because $\{2\}$ is an ancestor of $\{3\}$ but it not included.

Ancestral sets play an important role in directed graphs because of the following proposition.

Proposition 6.4. *Let A be an ancestral set in $\mathcal{G}$. Then $p(x_V)$ factorizes with respect to $\mathcal{G}$ only if $p(x_A)$ factorizes with respect to $\mathcal{G}_A$.*

Proof. See Examples Sheet 2. □

Now suppose we wish to interrogate whether a conditional independence $X_A \perp\!\!\!\perp X_B \mid X_C$ holds under a DAG model. From the previous result, we can restrict ourselves to asking if this independence holds in the induced subgraph over the ancestral set $\mathrm{an}_{\mathcal{G}}(A \cup B \cup C)$.

Definition 6.5. A *v-structure* is a triple $i \to k \leftarrow j$ such that $i \not\sim j$.

Let $\mathcal{G}$ be a directed acyclic graph; the *moral graph* $\mathcal{G}^m$ is formed from $\mathcal{G}$ by joining any non-adjacent parents and dropping the direction of edges.

In other words, the moral graph removes any 'v-structures' by filling in the missing edge, and then drops the direction of edges. An example is given in Figure 7.

Proposition 6.6. *If p_V factorizes with respect to a DAG $\mathcal{G}$, then it also factorizes with respect to the undirected graph $\mathcal{G}^m$.*

Proof. This follows from an inspection of the factorization and checking the cliques from $\mathcal{G}^m$. See the Examples Sheet. □

Using this proposition, we see that the DAG in Figure 7(a) implies $X_1 \perp\!\!\!\perp X_4, X_5 \mid X_2, X_3$, by using the global Markov property applied to the moral graph in Figure 7(b). In fact, moral graphs are used to define the global Markov property for DAGs.

Definition 6.7. We say that $p(x_V)$ satisfies the *global Markov property* with respect to $\mathcal{G}$ if whenever A and B are separated by C in $(\mathcal{G}_{\mathrm{an}(A\cup B\cup C)})^m$ we have $X_A \perp\!\!\!\perp X_B \mid X_C\,[p]$.

The global Markov property is *complete* in the sense that any independence not exhibited by a separation will not generally hold in distributions Markov to $\mathcal{G}$. We state the result formally here, but the proof is not given in this course.

Theorem 6.8 (Completeness of global Markov property.)**.** *Let $\mathcal{G}$ be a DAG. There exists a probability distribution p such that $X_A \perp\!\!\!\perp X_B \mid X_C\ [p]$ if and only if $A \perp_s B \mid C$ in $(\mathcal{G}_{\mathrm{an}(A\cup B\cup C)})^m$.*

In other words, the global Markov property gives all conditional independences that are implied by the DAG model.

We now give the main result concerning Markov equivalence, which says that each of our three properties give equivalent models.

Theorem 6.9. *Let $\mathcal{G}$ be a DAG and p a probability distribution. Then the following are equivalent:*

(i) p factorizes according to $\mathcal{G}$;

(ii) p is globally Markov with respect to $\mathcal{G}$;

(iii) p is locally Markov with respect to $\mathcal{G}$.

Notice that, unlike for undirected graphs, there is no requirement of positivity on p: it is true even for degenerate distributions. There is also a 'pairwise' Markov property for directed graphs, which we will not cover; see Lauritzen's book for interest.

Proof. (i) $\implies$ (ii). Let $W = \mathrm{an}_{\mathcal{G}}(A \cup B \cup C)$, and suppose that there is a separation between A and B given C in $(\mathcal{G}_W)^m$. The distribution $p(x_W)$ can be written as

$$p(x_W) = \prod_{v\in W} p(x_v \mid x_{\mathrm{pa}(v)}),$$

so in other words it is Markov w.r.t. $\mathcal{G}_W$ and hence to $(\mathcal{G}_W)^m$ (see Propositions 6.6 and 6.4). But if p factorizes according to the undirected graph $(\mathcal{G}_W)^m$ then it is also globally Markov with respect to it by Theorem 4.10, and hence the separation implies $X_A \perp\!\!\!\perp X_B \mid X_C\ [p]$.

(ii) $\implies$ (iii). Note that moralizing only adds edges adjacent to vertices that have a child in the graph, and also that $\{v\} \cup \mathrm{nd}_{\mathcal{G}}(v)$ is an ancestral set. It follows that in the moral graph $(\mathcal{G}_{\{v\}\cup\mathrm{nd}_{\mathcal{G}}(v)})^m$, there is a separation between v and $\mathrm{nd}_{\mathcal{G}}(v) \setminus \mathrm{pa}_{\mathcal{G}}(v)$ given $\mathrm{pa}_{\mathcal{G}}(v)$.

(iii) $\implies$ (i). Let $<$ be a topological ordering of the vertices in $\mathcal{G}$. The local Markov property implies that X_v is independent of $X_{\mathrm{nd}(v)\setminus\mathrm{pa}(v)}$ given $X_{\mathrm{pa}(v)}$, so in particular it is independent of $X_{\mathrm{pre}_<(v)\setminus\mathrm{pa}(v)}$ given $X_{\mathrm{pa}(v)}$. Hence

$$p(x_V) = \prod_v p(x_v \mid x_{\mathrm{pre}_<(v)}) = \prod_v p(x_v \mid x_{\mathrm{pa}(v)})$$

as required. □

6.3 Statistical Inference

The factorization of distributions that are Markov with respect to a DAG is particularly attractive statistically because, as with the decomposable models in Theorem 4.24, the conditional distributions can all be dealt with entirely separately.

Consider again the example of a contingency table with counts $n(x_V)$. The likelihood for a DAG model is

$$\begin{aligned} l(p; n) &= \sum_{x_V} n(x_V) \log p(x_V) \\ &= \sum_{x_V} n(x_V) \sum_{v \in V} \log p(x_v \mid x_{\mathrm{pa}(v)}) \\ &= \sum_{v \in V} \sum_{x_v, x_{\mathrm{pa}(v)}} n(x_v, x_{\mathrm{pa}(v)}) \log p(x_v \mid x_{\mathrm{pa}(v)}) \\ &= \sum_{v \in V} \sum_{x_{\mathrm{pa}(v)}} \sum_{x_v} n(x_v, x_{\mathrm{pa}(v)}) \log p(x_v \mid x_{\mathrm{pa}(v)}), \end{aligned}$$

where each of the conditional distributions $p(x_v \mid x_{\mathrm{pa}(v)})$ can be dealt with entirely separately. That is, we can separately maximize each inner sum $\sum_{x_v} n(x_v, x_{\mathrm{pa}(v)}) \log p(x_v \mid x_{\mathrm{pa}(v)})$ subject to the restriction that $\sum_{x_v} p(x_v \mid x_{\mathrm{pa}(v)}) = 1$, and hence obtain the MLE

$$\begin{aligned} &\hat{p}(x_v \mid x_{\mathrm{pa}(v)}) = \frac{n(x_v, x_{\mathrm{pa}(v)})}{n(x_{\mathrm{pa}(v)})}; \\ \text{hence} \quad &\hat{p}(x_V) = \prod_{v \in V} \hat{p}(x_v \mid x_{\mathrm{pa}(v)}) = \prod_{v \in V} \frac{n(x_v, x_{\mathrm{pa}(v)})}{n(x_{\mathrm{pa}(v)})}. \end{aligned}$$

This looks rather like the result we obtained for decomposable models, and indeed we will see that there is an important connection.

A slightly more general result is to say that if we have a separate parametric model defined by some parameter θ_v for each conditional distribution $p(x_v \mid x_{\mathrm{pa}(v)}; \theta_v)$, then we can perform our inference on each θ_v separately.

Formally: the MLE for θ satisfies

$$p(x_V; \hat{\theta}) = \prod_{v \in V} p(x_v \mid x_{\mathrm{pa}(v)}; \hat{\theta}_v), \qquad\qquad x_V \in \mathcal{X}_V.$$

In addition, if we have independent priors $\pi(\theta) = \prod_v \pi(\theta_v)$, then

$$\begin{aligned} \pi(\theta \mid x_V) &\propto \pi(\theta) \cdot p(x_V \mid \theta) \\ &= \prod_v \pi(\theta_v) \cdot p(x_v \mid x_{\mathrm{pa}(v)}, \theta_v), \end{aligned}$$

which factorizes into separate functions for each θ_v, showing that the θ_v are independent conditional on X_V. Hence

$$\pi(\theta_v \mid x_V) \propto \pi(\theta_v) \cdot p(x_v \mid x_{\mathrm{pa}(v)}, \theta_v),$$

so $\pi(\theta_v \mid x_V) = \pi(\theta_v \mid x_v, x_{\mathrm{pa}(v)})$, and θ_v only depends upon X_v and $X_{\mathrm{pa}(v)}$.

In other words, the data from $X_v, X_{\mathrm{pa}(v)}$ are sufficient for each θ_v. This means that if no vertex has many parents, even very large graphs represent manageable models. For

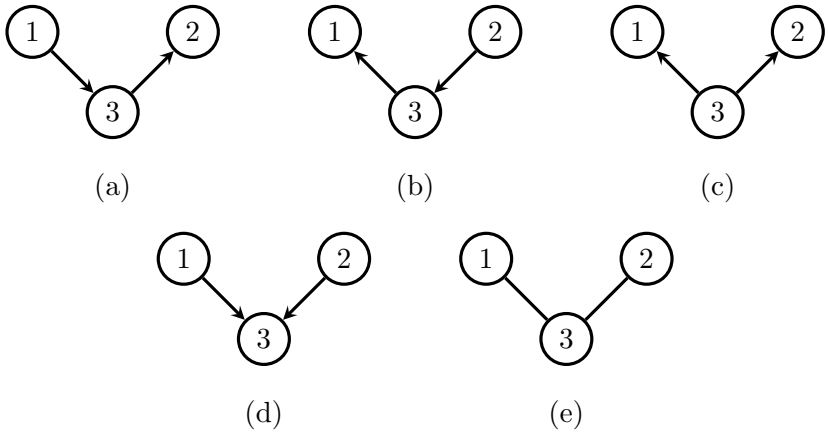


Figure 8: (a)-(c) Three directed graphs, and (e) an undirected graph to which they are all Markov equivalent; (d) a graph which is not Markov equivalent to the others.

a Gaussian distribution we can use our results about conditional distributions to obtain closed form expressions for the covariance matrices that are Markov with respect to a graph (see Example Sheet 2).

6.4 Markov Equivalence

For undirected graphs, the independence $X_a \perp\!\!\!\perp X_b \mid X_{V\setminus\{a,b\}}$ is implied by the graphical model if and only if the edge $a - b$ is not present in the graph. This shows that (under any choice of Markov property) each undirected graphical model is distinct.

For directed graphs this is not the case. The graphs in Figures 8 (a), (b) and (c) are all different, but all imply precisely the independence $X_1 \perp\!\!\!\perp X_2 \mid X_3$.

We say that two graphs $\mathcal{G}$ and $\mathcal{G}'$ are *Markov equivalent* if any p which is Markov with respect to $\mathcal{G}$ is also Markov with respect to $\mathcal{G}'$, and vice-versa. This is an equivalence relation, so we can partition graphs into sets we call *Markov equivalence classes*.

In model selection problems we are not trying to learn the graph itself, but rather the Markov equivalence class of indistinguishable models. The presence or absence of edges induces all conditional independences, so unsurprisingly the graph of adjacencies is very important.

Given a DAG $\mathcal{G} = (V, D)$, define the *skeleton* of $\mathcal{G}$ as the undirected graph $\mathrm{skel}(\mathcal{G}) = (V, E)$, where $\{i, j\} \in E$ if and only if either $(i, j) \in D$ or $(j, i) \in D$. In other words, we drop the orientations of edges in $\mathcal{G}$.

For example, the skeleton of the graphs in Figures 8(a)–(d) is the graph in Figure 8(e).

Lemma 6.10. *Let $\mathcal{G}$ and $\mathcal{G}'$ be graphs with different skeletons. Then $\mathcal{G}$ and $\mathcal{G}'$ are not Markov equivalent.*

Proof. Suppose without loss of generality that $i \to j$ in $\mathcal{G}$ but that $i \not\sim j$ in $\mathcal{G}$. Then let p be any distribution in which $X_v \perp\!\!\!\perp X_{V\setminus\{v\}}$ for each $v \in V \setminus \{i, j\}$, but that X_i and X_j are dependent.

The local Markov property for $\mathcal{G}$ is clearly satisfied, since each variable is independent of its non-descendants given its parents. For $\mathcal{G}'$, however, we claim that the global Markov property is not satisfied. By Sheet 2 Question 5, there is some set C such that the GMP requires $X_i \perp\!\!\!\perp X_j \mid X_C$.

Let $c \in C$; under p we have $X_c \perp\!\!\!\perp X_{V\setminus\{c\}}$, so by applying property 2 of the graphoid axioms, $X_c \perp\!\!\!\perp X_j, X_{C\setminus\{c\}}$. Then using properties 3 and 4 we see that $X_i \perp\!\!\!\perp X_j \mid X_C$ is equivalent to $X_i \perp\!\!\!\perp X_j \mid X_{C\setminus\{c\}}$. Repeating this we end up with a requirement that $X_i \perp\!\!\!\perp X_j$, which does not hold by construction. Hence p is not Markov with respect to $\mathcal{G}'$, and the graphs are not Markov equivalent. □

Theorem 6.11. *Directed graphs $\mathcal{G}$ and $\mathcal{G}'$ are Markov equivalent if and only if they have the same skeletons and v-structures.*

Proof. We will prove the 'only if' direction for now: the converse is harder.

If $\mathcal{G}$ and $\mathcal{G}'$ have different skeletons then the induced models are different by the previous Lemma. Otherwise, suppose that $a \to c \leftarrow b$ is a v-structure in $\mathcal{G}$ but not in $\mathcal{G}'$.

Let p be a distribution in which all variables other than X_a, X_b, X_c are independent of all other variables. By the factorization property, we can then pick an arbitrary

$$p(x_V) = p(x_c \mid x_a, x_b) \prod_{v\in V\setminus\{c\}} p(x_v)$$

and obtain a distribution that is Markov with respect to $\mathcal{G}$.

In $\mathcal{G}'$ there is no v-structure, so either $a \to c \to b$, $a \leftarrow c \to b$, or $a \leftarrow c \leftarrow b$. In particular, either a or b is a child of c. Now let $A = \mathrm{an}_{\mathcal{G}}(\{a, b, c\})$; we claim that there is no $d \in A$ such that $a \to d \leftarrow b$. To see this, note that if this is true, then d is a descendant of each of a, b and c, and if $d \in A$ it is also an ancestor of one a, b and c, so the graph is cyclic.

Now, it follows that in the moral graph $(\mathcal{G}'_A)^m$, there is no edge between a and b, so $a \perp_s b \mid A \setminus \{a, b\}$ in $(\mathcal{G}'_A)^m$. But by a similar argument to the previous Lemma, the corresponding independence does not hold in p, and therefore p is not Markov with respect to $\mathcal{G}'$ if $p(x_c \mid x_a, x_b)$ is chosen not to factorize. □

6.5 Directed Graphs, Undirected Graphs, and Decomposability

Closely related to the previous point is whether an undirected graph can represent the same conditional independences as a directed one. The undirected graph in Figure 8(e) represents the same model as each of the directed graphs in Figures 8(a)–(c), so clearly in some cases this occurs.

However the graph in Figure 8(d) does not induce the same model as any undirected graph. Indeed, it is again this 'v-structure' that is the important factor in determining whether the models are the same.

Theorem 6.12. *A directed graph is Markov equivalent to an undirected graph if and only if it contains no v-structures.*

Proof. We proceed by induction on p; the result is clearly true for graphs of size $p \leq 2$. We have already established that if $\mathcal{G}$ is a DAG, then p being Markov with respect to $\mathcal{G}$ implies that it is also Markov with respect to $\mathcal{G}^m$.

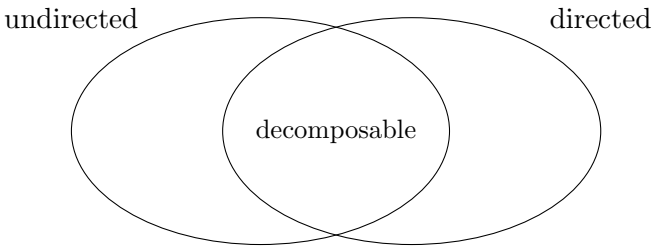


Figure 9: Venn diagram of model classes introduced by directed and undirected graphs.

Now suppose that p is Markov with respect to $\mathcal{G}^m$. Let v be a vertex in $\mathcal{G}$ without children. We will attempt to show that $p(x_{V\setminus\{v\}})$ is Markov with respect to $G_{V\setminus\{v\}}$ and that $X_v \perp\!\!\!\perp X_{V\setminus(\mathrm{pa}(v)\cup\{v\})} \mid X_{\mathrm{pa}(v)}$ under p, and hence that p satisfies the local Markov property with respect to $\mathcal{G}$.

The neighbours of v in $\mathcal{G}^m$ are its parents in $\mathcal{G}$, and in the moral graph $\mathcal{G}^m$ these are all adjacent, so there is a decomposition $(\{v\}, \mathrm{pa}_{\mathcal{G}}(v), W)$ in $\mathcal{G}^m$, where $W = V\setminus(\{v\}\cup\mathrm{pa}_{\mathcal{G}}(v))$. By Lemma 4.23, we have $X_v \perp\!\!\!\perp X_W \mid X_{\mathrm{pa}(v)}$, and that $p(x_{V\setminus\{v\}})$ is Markov with respect to $(\mathcal{G}^m)_{V\setminus\{v\}}$. Now, since $\mathcal{G}$ has no v-structures, $(\mathcal{G}^m)_{V\setminus\{v\}} = (\mathcal{G}_{V\setminus\{v\}})^m$, so by the induction hypothesis, $p(x_{V\setminus\{v\}})$ is Markov with respect to $\mathcal{G}_{V\setminus\{v\}}$. □

Corollary 6.13. *A undirected graph is Markov equivalent to a directed graph if and only if it is decomposable.*

Proof. This can be seen by the same decomposition and induction as in the proof of the Theorem above. □

This shows that decomposable models represent the intersection of undirected and directed graphical models.

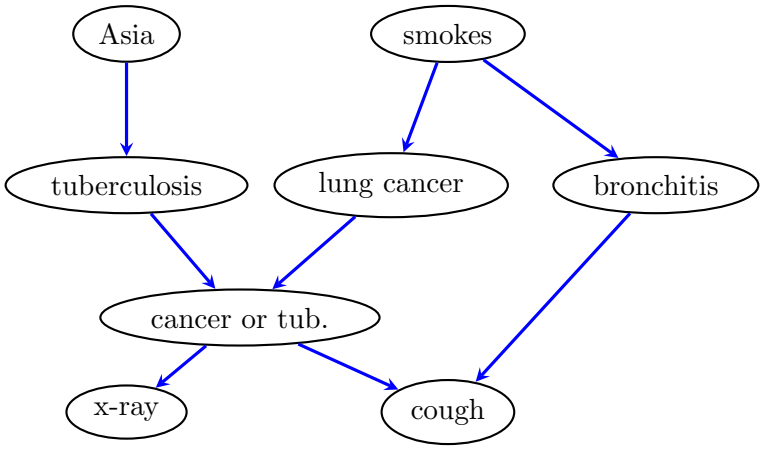
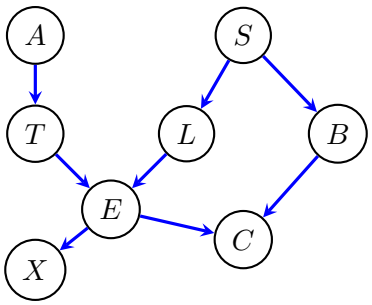


Figure 10: The 'Chest Clinic' network, a fictitious diagnostic model.

7 Junction Trees and Message Passing

In this chapter we answer some of the problems mentioned in the introduction: given a large network of variables, how can we efficiently evaluate conditional and marginal probabilities? And how should we update our beliefs given new information?

Consider the graph in Figure 10, which is a simplified diagnostic model, containing patient background, diseases, and symptoms. In practice, we observe the background and symptoms and with to infer the probability of disease given this 'evidence'. Of course, to calculate the updated probability we just need to use Bayes' formula, but for large networks this is computationally infeasible. Instead we will develop an algorithm that exploits the structure of the graph to simplify the calculations.



We abbreviate the variable names as indicated in the graph below. From the DAG factorization, we have

$$p(a, s, t, l, b, e, x, c) = p(a) \cdot p(s) \cdot p(t \mid a) \cdot p(l \mid s) \cdot p(b \mid s) \cdot p(e \mid t, l) \cdot p(x \mid e) \cdot p(c \mid e, b).$$

Suppose a patient smokes, has not visited Asia (tuberculosis is endemic in several Asian countries), has a negative x-ray, and a cough. Then to work out the probability of lung cancer:

$$p(l \mid x, c, a, s) = \frac{p(l, x, c \mid a, s)}{\sum_l p(l, x, c \mid a, s)}$$

The quantity we need can be obtained from the factorization of the directed graph as

$$p(l, x, c \mid a, s) = \sum_{t,e,b} p(t \mid a) \cdot p(l \mid s) \cdot p(b \mid s) \cdot p(e \mid t, l) \cdot p(x \mid e) \cdot p(c \mid e, b).$$

There is more than one way to evaluate this quantity, because some of the summations can be 'pushed in' past terms that do not depend upon them. So, for example,

$$\begin{aligned} &p(l, x, c \mid a, s) \\ &= p(l \mid s) \sum_{e} p(x \mid e) \left(\sum_{b} p(b \mid s) \cdot p(c \mid e, b) \right) \left(\sum_{t} p(t \mid a) \cdot p(e \mid t, l) \right). \end{aligned}$$

If calculated in this way we require 144 multiplications and additions. The naïve way implied by the first expression requires 816 multiplications and additions. Over larger networks with dozens or hundreds of variables these differences are very substantial.

This section provides a method for systematically arranging calculations of this sort in an efficient way, using the structure of a graph.

7.1 Junction Trees

We have already seen that we can write distributions that are Markov with respect to an undirected graph as a product of 'potentials', which are functions only of a few variables. A junction tree is a way of arranging these potentials that is computationally convenient.

Let $\mathcal{T}$ be a tree (i.e. a connected, undirected graph without any cycles) with vertices $\mathcal{V}$ contained in the power set of V; that is, each vertex of $\mathcal{T}$ is a subset of V. We say that $\mathcal{T}$ is a *junction tree* if whenever we have $C_i, C_j \in \mathcal{V}$ with $C_i \cap C_j \neq \emptyset$, there is a (unique) path π in $\mathcal{T}$ from C_i to C_j such that for every vertex C on the path, $C_i \cap C_j \subseteq C$.

The graph in Figure 11(b) is a junction tree. Note that, for example, $\{2, 4, 5\}$ and $\{4, 6\}$ have a non-zero intersection $\{4\}$, and that indeed 4 is contained on the intermediate vertex $\{2, 3, 4\}$.

The graph in Figure 12 is not a junction tree, because the sets $\{1, 2\}$ and $\{1, 3\}$ have the non-empty intersection $\{1\}$, but the intermediate sets in the tree (i.e. $\{2, 3\}$) do not contain $\{1\}$; this more general object is sometimes called a *clique tree*. The fact that these sets cannot be arranged in a junction tree is a consequence of these sets not satisfying the running intersection property (under any ordering), as the next result shows.

Proposition 7.1. *If $\mathcal{T}$ is a junction tree then its vertices $\mathcal{V}$ can be ordered to satisfy the running intersection property. Conversely, if a collection of sets satisfies the running intersection property they can be arranged into a junction tree.*

Proof. We proceed by induction on $k = |\mathcal{V}|$. If $k \leq 2$ then both the junction tree and running intersection conditions are always satisfied. Otherwise, since $\mathcal{T}$ is a tree it contains a leaf (i.e. a vertex joined to exactly one other), say C_k which is adjacent to $C_{\sigma(k)}$.

Consider $\mathcal{T}^{-k}$, the graph obtained by removing C_k from $\mathcal{T}$. The set of paths between C_i and C_j vertices in $\mathcal{T}^{-k}$ is the same as the set of such paths in $\mathcal{T}$: we cannot have paths via C_k because it would require repetition of $C_{\sigma(k)}$. Hence $\mathcal{T}^{-k}$ is still a junction tree, and by induction its elements $C_1, \ldots, C_{k-1}$ satisfy the RIP.

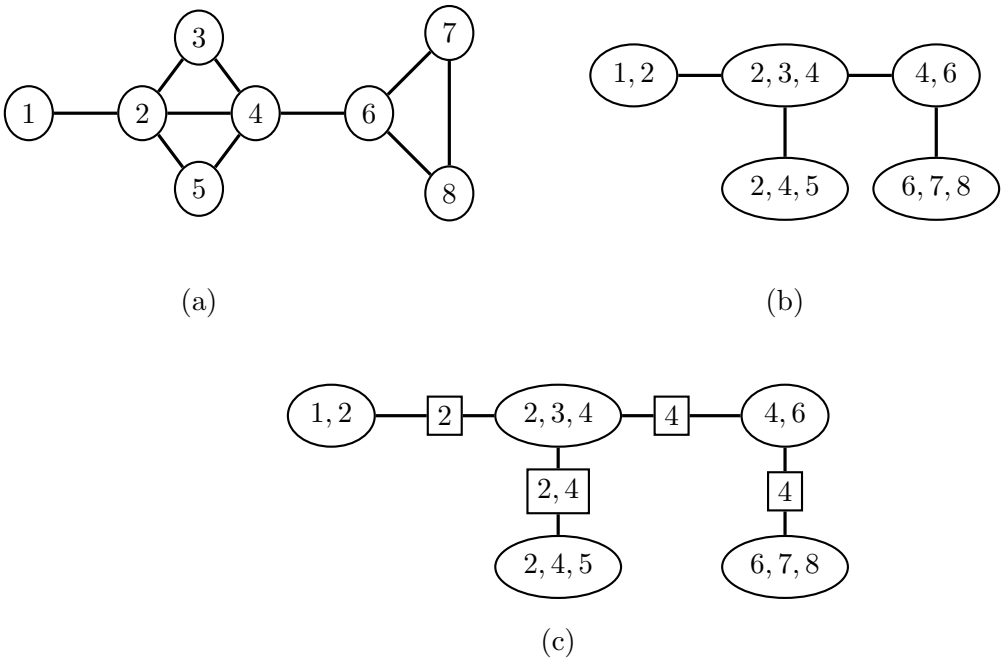


Figure 11: (a) A decomposable graph and (b) a possible junction tree of its cliques. (c) The same junction tree with separator sets explicitly marked.

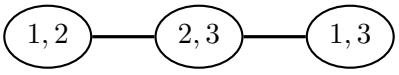


Figure 12: A tree of sets that is not a junction tree.

But then by the definition of a junction tree, $C_k \cap \bigcup_{i<k} C_i = C_k \cap C_{\sigma(k)}$, so $C_1, \ldots, C_k$ satisfies the RIP.

For the converse result, again by induction just join the final set C_k to $C_{\sigma(k)}$ and it is clear that we obtain a junction tree by definition of running intersection. □

In other words, this result shows that junction trees are available for the cliques of decomposable graphs. The graph in Figure 11(a) for example has cliques $\{1,2\}$, $\{2,3,4\}$, $\{2,4,5\}$, $\{4,6\}$ and $\{6,7,8\}$. Since it is a decomposable graph, these satisfy the running intersection property, and can be arranged in a junction tree such as the one in Figure 11(b). Notice that this is not unique, since we could join either (or both) of $\{1,2\}$ or $\{4,6\}$ to $\{2,4,5\}$ instead of $\{2,3,4\}$.

We can explicitly add in the separator sets as nodes in our tree, so that each edge contains an additional node, as shown in Figure 11(c).

We will associate each node C in our junction tree with a *potential* $\psi_C(x_C) \geq 0$, which is a function over the variables in the corresponding set. We say that two potentials ψ_C, ψ_D are *consistent* if

$$\sum_{x_{C \setminus D}} \psi_C(x_C) = f(x_{C \cap D}) = \sum_{x_{D \setminus C}} \psi_D(x_D).$$

That is, the margins of ψ_C and ψ_D over $C \cap D$ are the same.

Of course, the standard example of when we would have consistent margins comes when each potential is the margin of a probability distribution. Indeed, this relationship turns out to be quite fundamental.

Proposition 7.2. *Let $C_1, \ldots, C_k$ satisfy the running intersection property with separator sets $S_2, \ldots, S_k$, and let*

$$p(x_V) = \prod_{i=1}^{k} \frac{\psi_{C_i}(x_{C_i})}{\psi_{S_i}(x_{S_i})}$$

(where $\psi_\emptyset = 1$ by convention). Then each $\psi_{C_i}(x_{C_i}) = p(x_{C_i})$ and $\psi_{S_i}(x_{S_i}) = p(x_{S_i})$ if (and only if) each pair of potentials is consistent.

Proof. The only if is clear, since margins of probabilities are indeed consistent in this way.

For the converse we proceed by induction on k; for $k = 1$ there is nothing to prove. Otherwise, let $R_k = C_k \setminus S_k \left(= C_k \setminus \bigcup_{i<k} C_i\right)$, so

$$p(x_{V \setminus R_k}) = \sum_{x_{R_k}} p(x_V) = \prod_{i=1}^{k-1} \frac{\psi_{C_i}(x_{C_i})}{\psi_{S_i}(x_{S_i})} \times \frac{1}{\psi_{S_k}(x_{S_k})} \sum_{x_{R_k}} \psi_{C_k}(x_{C_k})$$

Since the cliques are consistent, we have

$$\frac{\sum_{x_{R_k}} \psi_{C_k}(x_{C_k})}{\psi_{S_k}(x_{S_k})} = \frac{\psi_{S_k}(x_{S_k})}{\psi_{S_k}(x_{S_k})} = 1,$$

so

$$p(x_{V \setminus R_k}) = \prod_{i=1}^{k-1} \frac{\psi_{C_i}(x_{C_i})}{\psi_{S_i}(x_{S_i})}. \tag{5}$$

By the induction hypothesis, we have that $\psi_{C_i}(x_{C_i}) = p(x_{C_i})$ for $i \leq k-1$. In addition, by the RIP $S_k = C_k \cap C_j$ for some $j < k$, and hence by consistency

$$\psi_{S_k}(x_{S_k}) = \sum_{x_{C_j \setminus S_k}} \psi_{C_j}(x_{C_j}) = \sum_{x_{C_j \setminus S_k}} p(x_{C_j}) = p(x_{S_k}).$$

Finally, substituting (5) into our original expression, we have

$$p(x_V) = p(x_{V \setminus R_k}) \frac{\psi_{C_k}(x_{C_k})}{\psi_{S_k}(x_{S_k})} = p(x_{V \setminus R_k}) \frac{\psi_{C_k}(x_{C_k})}{p(x_{S_k})},$$

and so $p(x_{R_k} \mid x_{V \setminus R_k}) = \frac{\psi_{C_k}(x_{C_k})}{p(x_{S_k})}$ by definition of conditional probabilities. Since this only depends upon x_{C_k}, this is also $p(x_{R_k} \mid x_{S_k})$. Hence,

$$\psi_{C_k}(x_{C_k}) = p(x_{R_k} \mid x_{S_k}) \cdot p(x_{S_k}) = p(x_{C_k})$$

as required. □

If a graph is not decomposable then we can *triangulate* it by adding edges. We discuss will this further later on.

7.2 Message Passing and the Junction Tree Algorithm

We have seen that having locally consistent potentials is enough to deduce that we have correctly calculated marginal probabilities. The obvious question now is how we arrive at consistent margins in the first place. In fact we shall do this with 'local' update steps, that alter potentials to become consistent without altering the overall distribution. We will show that this leads to consistency in a finite number of steps.

Suppose that two cliques C and D are adjacent in the junction tree, with a separator set $S = C \cap D$. An *update* from C to D consists of replacing ψ_S and ψ_D with the following:

$$\psi'_S(x_S) = \sum_{x_{C \setminus S}} \psi_C(x_C), \qquad \psi'_D(x_D) = \frac{\psi'_S(x_S)}{\psi_S(x_S)} \psi_D(x_D).$$

This operation is also known as *message passing*, with the 'message' $\psi'_S(x_S)$ being passed from C to D. We note three important points about this updating step:

- after updating, ψ_C and ψ'_S are consistent;
- if ψ_D and ψ_S are consistent, then so are ψ'_D and ψ'_S: to see this, note that

$$\begin{aligned} \sum_{x_{D \setminus S}} \psi'_D(x_D) &= \sum_{x_{D \setminus S}} \frac{\psi'_S(x_S)}{\psi_S(x_S)} \psi_D(x_D) \\ &= \frac{\psi'_S(x_S)}{\psi_S(x_S)} \sum_{x_{D \setminus S}} \psi_D(x_D), \end{aligned}$$

  so if ψ_S and ψ_D are consistent then $\psi_S(x_S) = \sum_{x_{D \setminus S}} \psi_D(x_D)$ and we are left with ψ'_S.

- the product

$$\frac{\prod_{C\in\mathcal{C}}\psi_C(x_C)}{\prod_{S\in\mathcal{S}}\psi_S(x_S)}=\frac{\prod_{C\in\mathcal{C}}\psi'_C(x_C)}{\prod_{S\in\mathcal{S}}\psi'_S(x_S)}.$$

  is unchanged: the only altered terms are ψ_D and ψ_S, and by definition of ψ'_D we have

$$\frac{\psi'_D(x_D)}{\psi'_S(x_S)}=\frac{\psi_D(x_D)}{\psi_S(x_S)}.$$

Hence, updating preserves the joint distribution and does not upset margins that are already consistent. The junction tree algorithm is a way of updating all the margins such that, when it is complete, they are all consistent.

Let $\mathcal{T}$ be a tree. Given any node $t\in\mathcal{T}$, we can 'root' the tree at t, and replace it with a directed graph in which all the edges point away from t.[2] The *junction tree algorithm* involves messages being passed from the edge of the junction tree (the leaves) towards a chosen root (the *collection* phase), and then being sent away from that root back down to the leaves (the *distribution* phase). Once these steps are completed, the potentials will all be consistent. This process is also called *belief propagation*.

Algorithm 1 Collect and distribute steps of the junction tree algorithm.

```
function COLLECT(rooted tree T, potentials ψ_t)
    let 1 < ... < k be a topological ordering of T
    for t in k, ..., 2 do
        send message from ψ_t to ψ_σ(t);
    end for
    return updated potentials ψ_t
end function
function DISTRIBUTE(rooted tree T, potentials ψ_t)
    let 1 < ... < k be a topological ordering of T
    for t in 2, ..., k do
        send message from ψ_σ(t) to ψ_t;
    end for
    return updated potentials ψ_t
end function
```

The *junction tree algorithm* consists of running $\textsc{Collect}(\mathcal{T},\psi_t)$ and $\textsc{Distribute}(\mathcal{T},\psi'_t)$, as given in Algorithm 1.

Theorem 7.3. *Let $\mathcal{T}$ be a junction tree with potentials $\psi_{C_i}(x_{C_i})$. After running the junction tree algorithm, all pairs of potentials will be consistent.*

Proof. We have already seen that each message passing step will make the separator node consistent with the child node. It follows that each pair ψ_{C_i} and ψ_{S_i} are consistent after the collection step. We also know that this consistency will be preserved after future updates from $\psi_{C_{\sigma(i)}}$. Hence, after the distribution step, each ψ_{C_i} and ψ_{S_i} remain consistent, and

[2]This process always gives a Markov equivalent graph although, of course, we are not really applying the Markov property to our junction tree. The directions are just for convenience.

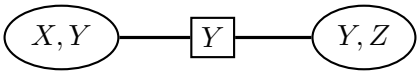
Figure 13: Illustration of the junction tree algorithm with $\{2, 3, 4\}$ chosen as the root. (a) Collect steps towards the root: note that the $\{4, 6\}$ to $\{2, 3, 4\}$ step must happen after the $\{6, 7, 8\}$ to $\{4, 6\}$ update. (b) Distribute steps away from the root and towards the leaves: this time the constraint on the ordering is reversed.

$\psi_{C_{\sigma(i)}}$ and ψ_{S_i} become consistent for each i. Hence, every adjacent pair of cliques is now consistent.

But whenever $C_i \cap C_j \neq \emptyset$ there is a path in the junction tree such that every intermediate clique also contains $C_i \cap C_j$, so this local consistency implies global consistency of the tree. □

Remark 7.4. In practice, message passing is often done in parallel, and it is not hard to prove that if all potentials update simultaneously then the potentials will converge to a consistent solution in at most d steps, where d is the width of the tree.

Example 7.5. Suppose we have just two tables, ψ_{XY} and ψ_{YZ} arranged in the junction tree:



representing a distribution in which $X \perp\!\!\!\perp Z \mid Y$. We can initialize by setting

$$\psi_{XY}(x, y) = p(x \mid y) \qquad \psi_{YZ}(y, z) = p(y, z) \qquad \psi_Y(y) = 1,$$

so that $p(x, y, z) = p(y, z) \cdot p(x \mid y) = \psi_{YZ}\psi_{XY}/\psi_Y$.

Now, we could pick YZ as the root node of our tree, so the collection step consists of replacing

$$\psi'_Y(y) = \sum_x \psi_{XY}(x, y) = \sum_x p(x \mid y) = 1;$$

so ψ'_Y and ψ_Y are the same; hence the collection step leaves ψ_Y and ψ_{YZ} unchanged.

The distribution step consists of

$$\psi''_Y(y) = \sum_z \psi_{YZ}(y, z) = \sum_z p(y, z) = p(y);$$

$$\psi'_{XY}(x, y) = \frac{\psi''_Y(y)}{\psi_Y(y)}\psi_{XY}(x, y) = \frac{p(y)}{1}p(x \mid y) = p(x, y);$$

Hence, after performing both steps, each potential is the marginal distribution corresponding to those variables.

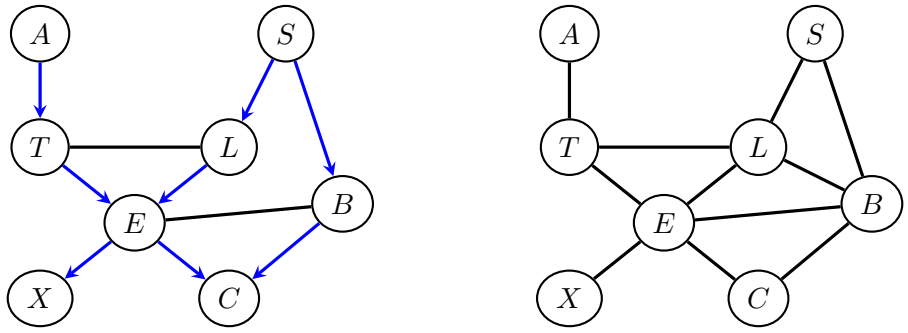


Figure 14: The moral graph of the Chest Clinic network, and a possible triangulation.

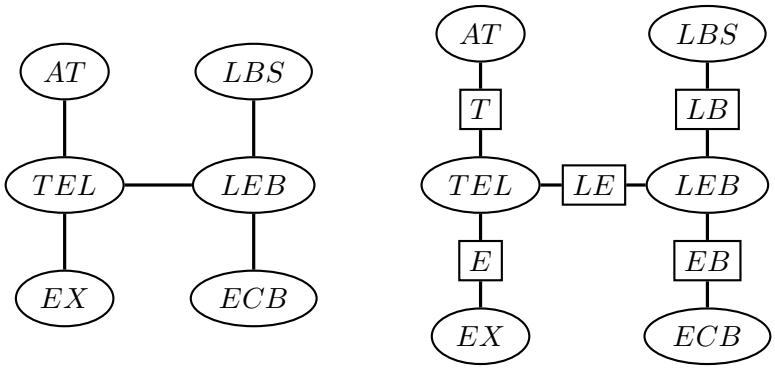


Figure 15: A possible junction tree for the Chest Clinic network, and (right) with separator sets drawn on.

In junction graphs that are not trees it is still possible to perform message passing, but convergence is not guaranteed. This is known as 'loopy belief propagation, and is a topic of current research.

7.3 Directed Graphs and Triangulation

How does any of this relate to directed graphs? And what should we do if our model is *not* decomposable? In this case we cannot immediately form a junction tree. However, all is not lost, since we can always embed our model in a larger model which *is* decomposable.

For a directed graph, we start by taking the moral graph, so that we obtain an undirected model. If the directed model is decomposable then so is the moral graph. If the moral graph is still not decomposable, then we can *triangulate* it by adding edges to obtain a decomposable graph. Figure 14(b) contains a triangulation of the moral graph of Figure 10. We can arrange the cliques as

$$\{L, E, B\}, \qquad \{T, E, L\}, \qquad \{L, B, S\}, \qquad \{E, C, B\}, \qquad \{A, T\}, \qquad \{E, X\},$$

giving rise to the junction tree in Figure 15

Taking the 4-cycle in Figure 16(a) as an example, we can add chords to the cycle until we obtain a graph that is triangulated; a resulting graph is called a *triangulation*. This process is not unique, as is obvious from this example. Given the new graph we can form a junction tree for the larger model.

(a) (b)

Figure 16: A non-decomposable graph, and a possible triangulation of it.

Naturally, to keep our computations efficient we want the cliques in the model to remain small when we triangulate: after all, we could always embed our graph in the complete model! Finding a triangulation that is 'optimal'—in the sense of giving the smallest cliques—is a very hard problem in general. Some approximate and heuristic methods exist. A simple method, *Tarjan elimination*, is given on Examples Sheet 3.

Suppose we have a directed graphical model embedded within a decomposable model $C_1, \ldots, C_k$. For each vertex v, the set $\{v\} \cup \mathrm{pa}_{\mathcal{G}}(v)$ is contained within at least one of these cliques. Assigning each vertex arbitrarily to one such clique, let $v(C)$ be the vertices assigned to C. Then we can set $\psi_C(x_C) = \prod_{v \in v(C)} p(x_v \mid x_{\mathrm{pa}(v)})$ and $\psi_S(x_S) = 1$, and we have

$$\prod_{i=1}^{k} \frac{\psi_{C_i}(x_{C_i})}{\psi_{S_i}(x_{S_i})} = \prod_{v \in V} p(x_v \mid x_{\mathrm{pa}(v)}) = p(x_V).$$

This is called *initialization*. Now if we run the junction tree algorithm to obtain consistent potentials, then these will just be the marginal probabilities for each clique.

7.4 Evidence

The junction tree gives us a mechanism for calculating marginal distributions for quantities that are contained in the same clique. How should we deal with queries about conditional distributions for quantities that may not be adjacent? For example, what difference does it make to our chest clinic network if a patient smokes?

We can answer this by introducing 'evidence' into our tables, and then propagating it through the tree. The new evidence corresponds to replacing an existing marginal table with one in which the event that occurred has probability 1: for example,

$$p(s) = \begin{array}{c|c} \text{smokes} & \text{doesn't smoke} \\ \hline 0.25 & 0.75 \end{array} \quad \text{becomes} \quad \tilde{p}(s) = \begin{array}{c|c} \text{smokes} & \text{doesn't smoke} \\ \hline 1 & 0 \end{array}.$$

Let our evidence be the event $\{X_e = y_e\}$ for some relevant node e; we can write the new joint distribution as

$$p(x_V \mid X_e = y_e) = p(x_V, x_e) \frac{\mathbb{1}_{\{x_e = y_e\}}}{p(x_e)}.$$

Thus, replacing

$$\psi'_C(x_C) \leftarrow \psi_C(x_C) \cdot \frac{\mathbb{1}_{\{x_e = y_e\}}}{p(y_e)}$$

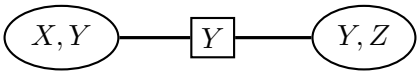
for any potential with $C \ni e$ will alter the joint distribution in the required way. If the potentials are already consistent then $p(y_e)$ can be calculated from ψ_C directly.

Of course, after replacing ψ_C the potentials will no longer be consistent, and therefore the junction tree algorithm needs to be run again. In fact, only a distribution step with ψ_C chosen as the root node is needed.

Proposition 7.6. *Suppose that potentials Ψ are all consistent except for ψ_C. Then after running* DISTRIBUTE*$(\mathcal{T}, \Psi)$, all potentials are consistent.*

Proof. Each separator set potential is already consistent with the clique potential(s) 'away' from C in the graph. This consistency is preserved, and distribution will ensure that each separator set is consistent with the clique potentials 'towards' C. Hence, all clique potentials and separator sets are now consistent. □

If we try to introduce evidence in two different places without propagating in between then we may not obtain the conditional distribution that we want. To see this, consider again our very simple example with two cliques:



If the potentials are already consistent, then $\psi_{XY} = p(x, y)$ and $\psi_{YZ} = p(y, z)$ with $\psi_Y = p(y)$. Now suppose we want to introduce two pieces of evidence: $\{X = x^*\}$ and $\{Z = z^*\}$. To introduce the first, we replace ψ_{XY} with

$$\psi'_{XY} = \psi_{XY} \frac{\mathbb{1}_{\{x=x^*\}}}{p(x^*)} = p(y \mid x^*)\mathbb{1}_{\{x=x^*\}}.$$

This means that the potentials are jointly representing the distribution q in which

$$q(x, y, z) = \frac{\psi'_{XY}(x, y)\psi_{YZ}(y, z)}{\psi_Y(y)} = \frac{p(y \mid x^*) \cdot p(y, z)}{p(y)} \mathbb{1}_{\{x=x^*\}} = p(y, z \mid x^*)\mathbb{1}_{\{x=x^*\}},$$

as required.

Now, the second would be introduced by replacing ψ_{YZ} with

$$\psi'_{YZ} = p(y \mid z^*)\mathbb{1}_{\{z=z^*\}}.$$

But now this gives

$$\begin{aligned} r(x, y, z) = \frac{\psi'_{XY}(x, y)\psi'_{YZ}(y, z)}{\psi_Y(y)} &= \frac{p(y \mid x^*) \cdot p(y \mid z^*)}{p(y)} \mathbb{1}_{\{x=x^*, z=z^*\}} \\ &= \frac{p(y, x^*) \cdot p(y, z^*)}{p(y)p(x^*)p(y^*)} \mathbb{1}_{\{x=x^*, z=z^*\}} \\ &= p(y \mid x^*, z^*) \frac{p(x^*, z^*)}{p(x^*)p(z^*)} \mathbb{1}_{\{x=x^*, z=z^*\}}, \end{aligned}$$

where the last equality holds from applying Theorem 2.4(iv) to $X \perp\!\!\!\perp Z \mid Y$. Now since $X \not\perp\!\!\!\perp Z$ in general, this final expression is not equal to $p(y \mid x^*, z^*)$.

8 Causal Inference

Causal inference, at its heart, asks what would happen if we were to perform an experiment in a system. This is different to usual 'prediction' with conditional distributions, as the following example illustrates.

Example 8.1. Suppose that a health and safety inspector is interested in the safety of a set of outdoor steps. She commissions a study that monitors the weather conditions each day, and whether anyone slips on the steps.

She finds that it rains (making the steps wet, $W = 1$) on 40% of days, and that the temperature is below freezing (making the steps icy, $I = 1$) on one day out of 10, and that these happen independently. Given the four possible conditions, she finds the probability of someone slipping each day is:

		$P(S = s \mid W, I)$	
W	I	$s = 0$	$s = 1$
0	0	0.95	0.05
1	0	0.9	0.1
0	1	0.8	0.2
1	1	0.5	0.5

So, for example, if it is icy but not wet then the probability of someone slipping is 0.2.

Now, suppose we know that someone has slipped on the steps: what is the probability that the steps were wet? Using Bayes' formula,

$$\begin{aligned} P(W = 1 \mid S = 1) &= \frac{\sum_i P(W = 1) \cdot P(I = i) \cdot P(S = 1 \mid W = 1, I = i)}{\sum_{f,w} P(W = w) \cdot P(I = i) \cdot P(S = 1 \mid W = w, I = i)} \\ &= \frac{0.06}{0.1} = 0.59. \end{aligned}$$

Unsurprisingly, if someone slips then this is predictive of wet steps: since the probability increases from 0.4 to 0.59. Similarly, if there is no slip then the probability decreases slightly to $P(W = 1 \mid S = 0) = 0.38$.

Now suppose the health and safety inspector insists that salt and grit be placed on the steps, so that they never get icy. Given this event, how would we estimate the probability of slipping? Well, this should just depend on whether the steps are wet as before, but always with $I = 0$. So our new distribution is $P(R = r, S = s \mid I = 0)$. In particular, this means that the overall probability of someone slipping is $P(S = 1 \mid I = 0) = 0.07$, down from $P(S = 1) = 0.1$.

Consider a third scenario: suppose that the health and safety inspector shuts the steps, so that no-one can slip ($S = 0$). What happens to the probability of the steps being icy? Following the same approach as above, we would look at $P(I = 1 \mid S = 0) = 0.34$, which is higher than $P(I = 1) = 0.1$ But this is surely absurd: health and safety inspector's actions will have no affect on the local climate! Indeed, we would expect that the probability of icy steps remains at $P(I = 1) = 0.1$, regardless of the action taken to fix $S = 0$.

The asymmetry in the previous example is an example of a *causal relationship*. Ordinary prediction is, in some sense, symmetric: if the steps being icy increase the chance of a slip, then a slip makes it more likely that it was icy. However, causal prediction is not

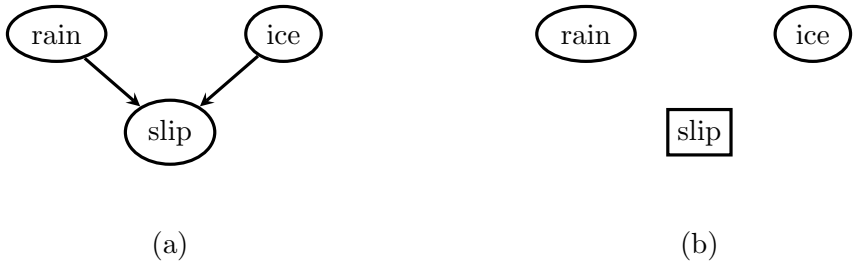


Figure 17: (a) A causal DAG on three vertices; (b) after intervening on 'slip' none of the variables are correlated.

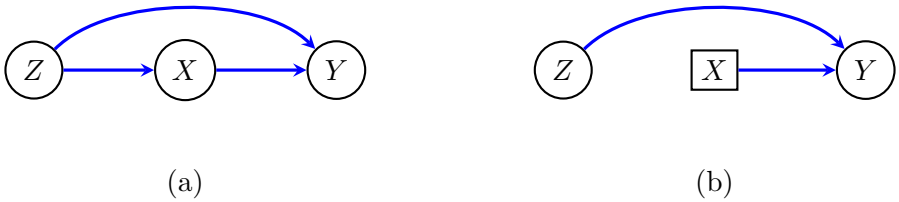


Figure 18: (a) A causal DAG on three vertices, and (b) after intervening on X.

symmetric: if I make it rain then that will makes the steps wet, but if I make the steps wet then it will not cause it to start raining.

The scenarios of adding grit to prevent ice, or of closing the steps are examples of *interventions* or *treatments* that affect the variables in the system and the relationships between them. If we intervene in a system in such a way as to set a variable such as $S = s$, we denote the resulting distribution of other variables as

$$P(R = r, I = i \mid do(S = s)).$$

The example above shows that in some cases this is the same as the relevant conditional distribution, but not always:

$$P(S = s \mid do(I = i)) = P(S = s \mid I = i)$$
$$P(I = i \mid do(S = s)) = P(I = i).$$

Directed graphs provide a convenient framework for representing the structural assumptions underlying a causal system, and the asymmetry in interventions. We can think of each edge $v \to w$ as saying that X_v is a 'direct cause' of X_w; i.e. that it affects it in a way that is not mediated by any of the other variables. In our example, the system could be represented by the graph in Figure 17(a).

8.1 Interventions

Let $\mathcal{G}$ be a directed acyclic graph representing a causal system, and let p be a probability distribution over the variables X_V. An *intervention* on a variable $w \in V$ does two things:

- *graphically* we represent this by removing edges pointing into w (i.e. of the form $v \to w$);

- *probabilistically*, we replace our usual factorization

$$p(x_V) = \prod_{v \in V} p(x_v \mid x_{\mathrm{pa}(v)})$$

with

$$\begin{aligned} p(x_{V \setminus \{w\}} \mid do(x_w^*)) &= p(x_V) \frac{\mathbb{1}_{\{x_w = x_w^*\}}}{p(x_w^* \mid x_{\mathrm{pa}(w)})} \\ &= \mathbb{1}_{\{x_w = x_w^*\}} \prod_{v \in V \setminus \{w\}} p(x_v \mid x_{\mathrm{pa}(v)}). \end{aligned}$$

In words, we are assuming that w no longer depends upon its parents, but has been fixed to x_w^*; hence the $p(x_w \mid x_{\mathrm{pa}(w)})$ factor is replaced with the indicator function that assigns probability 1 to the event that $\{X_w = x_w^*\}$. Other variables will continue to depend upon their parents according to the same conditionals $p(x_v \mid x_{\mathrm{pa}(v)})$.

When we say a graph and its associated probability distribution is *causal*, we mean that we are making the assumption that, if we were to intervene on a variable X_v via some experiment, then the distribution would change in the way described above. This assumption is something that has to be justified in specific applied examples.

Example 8.2 (Confounding)**.** Consider the graph in Figure 18(a); here Z causally affects both X and Y, so some of the observed correlation between X and Y will be due to this 'common cause' Z. We say that X and Y are 'confounded' by Z. Suppose we intervene to fix $X = x$, so that it is no longer causally affected by Z. Hence, we go from

$$p(z, x, y) = p(z) \cdot p(x \mid z) \cdot p(y \mid z, x)$$

to

$$p(z, y \mid do(x^*)) = p(z) \cdot p(y \mid z, x^*).$$

Note that this last object is not generally the same as the ordinary conditional distribution:

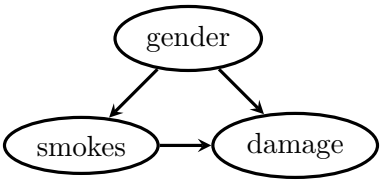
$$\begin{aligned} p(z, y \mid x^*) &= p(z \mid x^*) \cdot p(y \mid z, x^*) \\ p(z, y \mid do(x^*)) &= p(z) \cdot p(y \mid z, x^*). \end{aligned}$$

Example 8.3. Suppose we have a group of 64 people, half men and half women. We ask them whether they smoke, and test them for lung damage. The results are given by the following table.

	women			men	
	not smoke	smoke		not smoke	smoke
no damage	21	6		6	6
damage	3	2		2	18

Given that a person smokes, the probability that they have lung damage is $P(D = 1 \mid S = 1) = \frac{20}{32} = \frac{5}{8}$. If someone doesn't smoke the probability is $P(D = 1 \mid S = 0) = \frac{5}{32}$.

What happens if we had prevented everyone from smoking? Would this mean that only $\frac{5}{32} \times 64 = 10$ of our participants showed lung damage? If we assume the following causal model, then the answer is no.



We have (taking $G = 0$ to represent male) that

$$\begin{aligned} P(D = 1 \mid do(S = 0)) &= \sum_g P(D = 1 \mid S = 0, G = g) \cdot P(G = g) \\ &= P(D = 1 \mid S = 0, G = 0) \cdot P(G = 0) + P(D = 1 \mid S = 0, G = 1) \cdot P(G = 1) \\ &= \frac{2}{8} \cdot \frac{1}{2} + \frac{3}{24} \cdot \frac{1}{2} \\ &= \frac{3}{16} > \frac{5}{32}. \end{aligned}$$

So in fact, we would expect $\frac{3}{16} \times 64 = 12$ people to have damage if no-one was able to smoke.

The difference can be accounted for by the fact that some of the chance of getting lung damage is determined by gender. If we 'observe' that someone does not smoke then they are more likely to be female; but forcing someone not to smoke does *not* make them more likely to be female!

8.2 Adjustment Sets and Back-Door Paths

For this section we will assume we are interested in the distribution of Y after intervening on Z. The method given above for finding $p(y \mid do(z))$ appears to involve summing over all the variables in the graph:

$$p(y \mid do(z)) = \sum_{x_W} \frac{p(y, z, x_W)}{p(z \mid x_{\mathrm{pa}(z)})}$$

Here we present some methods for 'adjusting' by only a small number of variables.

Lemma 8.4. *Let $\mathcal{G}$ be a causal DAG. Then*

$$p(y \mid do(z)) = \sum_{x_{\mathrm{pa}(z)}} p(y \mid z, x_{\mathrm{pa}(z)}) \cdot p(x_{\mathrm{pa}(z)}).$$

Proof. Let X_V be divided into $Y, Z, X_{\mathrm{pa}(z)}$ and X_W, where X_W is any other variable (that is, not Y, Z, nor a parent of Z). Then

$$p(y, x_{\mathrm{pa}(z)}, x_W \mid do(z)) = \frac{p(y, z, x_{\mathrm{pa}(z)}, x_W)}{p(z \mid x_{\mathrm{pa}(z)})} = p(y, x_W \mid z, x_{\mathrm{pa}(z)}) \cdot p(x_{\mathrm{pa}(z)}).$$

Then

$$\begin{aligned} p(y \mid do(z)) &= \sum_{x_W, x_{\mathrm{pa}(z)}} p(y, x_W \mid z, x_{\mathrm{pa}(z)}) \cdot p(x_{\mathrm{pa}(z)}) \\ &= \sum_{x_{\mathrm{pa}(z)}} p(x_{\mathrm{pa}(z)}) \sum_{x_W} p(y, x_W \mid z, x_{\mathrm{pa}(z)}) \\ &= \sum_{x_{\mathrm{pa}(z)}} p(x_{\mathrm{pa}(z)}) p(y \mid z, x_{\mathrm{pa}(z)}) \end{aligned}$$

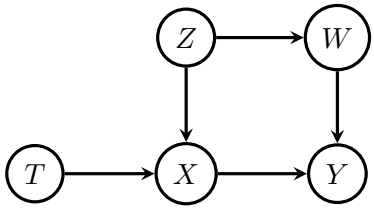


Figure 19: A causal directed graph.

as required. □

This result is called an 'adjustment' formula. Applied to the graph in Figure 19, for example, it would tell us that $p(y \mid do(x)) = \sum_{z,t} p(y \mid x, z, t) \cdot p(z, t)$, so, for example, we do not need to consider W. In fact, though, you might notice that $Y \perp\!\!\!\perp T \mid X, Z$, so we can write

$$\begin{aligned} p(y \mid do(x)) &= \sum_{z,t} p(y \mid x, z) \cdot p(z, t) \\ &= \sum_{z} p(y \mid x, z) \cdot p(z), \end{aligned}$$

and we only need to adjust for Z! Further,

$$\begin{aligned} p(y \mid do(x)) = \sum_{z} p(y \mid x, z) \cdot p(z) &= \sum_{z,w} p(y, w \mid x, z) \cdot p(z) \\ &= \sum_{z,w} p(y \mid x, w, z) \cdot p(w \mid x, z) \cdot p(z) \\ &= \sum_{z,w} p(y \mid x, w) \cdot p(w \mid z) \cdot p(z) \\ &= \sum_{z,w} p(y \mid x, w) \cdot p(w, z) \\ &= \sum_{w} p(y \mid x, w) \cdot p(w); \end{aligned}$$

the fourth equality here uses the fact that $W \perp\!\!\!\perp X \mid Z$ and $Y \perp\!\!\!\perp Z \mid W, X$, which can be seen from the graph.

So, in other words, we could adjust by W instead of Z! This illustrates that there are often multiple equivalent ways of obtaining the same causal quantity. We will give a criterion for valid adjustment sets, but we first need an extra definition and theorem to prove this criterion correct.

8.3 Paths and d-separation

Let $\mathcal{G}$ be a directed graph and π a path in $\mathcal{G}$. We say that an internal vertex t on π is a *collider* if the edges adjacent to t meet as $\rightarrow t \leftarrow$. Otherwise ($\rightarrow t \rightarrow$, $\leftarrow t \leftarrow$, or $\leftarrow t \rightarrow$) we say t is a *non-collider*.

Let π be a path from a to b. We say that π is *open* given (or conditional on) $C \subseteq V \setminus \{a, b\}$ if

- all colliders on π are in $\text{an}_{\mathcal{G}}(C)$;
- all non-colliders are outside C.

(Recall that $C \subseteq \text{an}_{\mathcal{G}}(C)$.) A path which is not open given C is said to be *blocked* by C.

Example 8.5. Consider the graph in Figure 19. There are two paths from T to W:

$$T \to X \leftarrow Z \to W \qquad\qquad T \to X \to Y \leftarrow W.$$

Without conditioning on any variable, both these paths are both blocked, since they contain colliders. Given $\{Y\}$, however, both paths are open, because Y is the only collider on the second path, and the only collider on the first is X, which is an ancestor of Y. Given $\{Z, Y\}$, the first path is blocked because Z is a non-collider, but the second is open.

Definition 8.6. Let A, B, C be disjoint sets of vertices in $\mathcal{G}$ (C may be empty). We say that A and B are *d-separated* given C if every path from $a \in A$ to $b \in B$ is blocked by C.

Theorem 8.7. *Let $\mathcal{G}$ be a DAG and let A, B, C be disjoint subsets of $\mathcal{G}$. Then A is d-separated from B by C in $\mathcal{G}$ if and only if A is separated from B by C in $(\mathcal{G}_{\text{an}(A\cup B\cup C)})^m$.*

In other words, this gives us an alternative version of the global Markov property for DAGs: instead of being based on paths in moral graphs, we can use paths in the original DAG.

Proof (not examinable). Suppose A is not d-separated from B by C in $\mathcal{G}$, so there is an open path π in $\mathcal{G}$ from some $a \in A$ to some $b \in B$. Dividing the path up into sections of the form $\leftarrow \cdots \leftarrow\!\!\rightarrow \cdots \rightarrow$, we see that π must lie within $\text{an}_{\mathcal{G}}(A \cup B \cup C)$, because every collider must be an ancestor of C, and the extreme vertices are in A and B. Each of the colliders $i \to k \leftarrow j$ gives an additional edge $i - j$ in the moral graph and so can be avoided; all the other vertices are not in C since the path is open. Hence we obtain a path from $a \in A$ to $b \in B$ in the moral graph that avoids C.

Conversely, suppose A is not separated from B by C in $(\mathcal{G}_{\text{an}(A\cup B\cup C)})^m$, so there is a path π in $(\mathcal{G}_{\text{an}(A\cup B\cup C)})^m$ from some $a \in A$ to some $b \in B$ that does not traverse any element of C. Each such path is made up of edges in the original graph and edges added over v-structures. Suppose an edge corresponds to a v-structure over k; then k is in $\text{an}_{\mathcal{G}}(A \cup B \cup C)$. If k is an ancestor of C then the path remains open; otherwise, if k is an ancestor of A then there is a directed path from k to $a' \in A$, and every vertex on it is a non-collider that is not contained in C. Hence we can obtain a path with fewer edges over v-structures from a' to b. Repeating this process we obtain a path from A to B in which every edge is either in $\mathcal{G}$ or is a v-structure over an ancestor of C. Hence the path is open. □

8.4 Back-door Adjustment

We say that C is a *back-door* adjustment set for the ordered pair (v, w) if

- no vertex in C is a descendant of v.
- every path from v to w with an arrow into v (i.e. starting $v \leftarrow \cdots$) is blocked by C;

Theorem 8.8. *Let C be a back-door adjustment set for (v, w). Then*

$$p(x_w \mid do(x_v)) = \sum_{x_C} p(x_C) \cdot p(x_w \mid x_v, x_C).$$

That is, C is a valid adjustment set for the causal distribution.

Proof. Since no vertex in C is a descendant of v, we have that $X_v \perp\!\!\!\perp X_C \mid X_{\mathrm{pa}(v)}$ using the local Markov property. We also claim that w is d-separated from $\mathrm{pa}_{\mathcal{G}}(v)$ by $C \cup \{v\}$. To see this, note that if there is an open path from w to some $t \in \mathrm{pa}_{\mathcal{G}}(v)$ given $C \cup \{v\}$, then there is an open path given C obtained by adding the edge $t \to v$. Hence the global Markov property implies that $X_w \perp\!\!\!\perp X_{\mathrm{pa}(v)} \mid X_v, X_C$. Then:

$$\begin{aligned}
p(x_w \mid do(x_v)) &= \sum_{x_{\mathrm{pa}(v)}} p(x_{\mathrm{pa}(v)}) \cdot p(x_w \mid x_v, x_{\mathrm{pa}(v)}) \\
&= \sum_{x_{\mathrm{pa}(v)}} p(x_{\mathrm{pa}(v)}) \sum_{x_C} p(x_w, x_C \mid x_v, x_{\mathrm{pa}(v)}) \\
&= \sum_{x_{\mathrm{pa}(v)}} p(x_{\mathrm{pa}(v)}) \sum_{x_C} p(x_w \mid x_C, x_v, x_{\mathrm{pa}(v)}) \cdot p(x_C \mid x_v, x_{\mathrm{pa}(v)}) \\
&= \sum_{x_{\mathrm{pa}(v)}} p(x_{\mathrm{pa}(v)}) \sum_{x_C} p(x_w \mid x_C, x_v) \cdot p(x_C \mid x_{\mathrm{pa}(v)}) \\
&= \sum_{x_C} p(x_w \mid x_C, x_v) \sum_{x_{\mathrm{pa}(v)}} p(x_{\mathrm{pa}(v)}) \cdot p(x_C \mid x_{\mathrm{pa}(v)}) \\
&= \sum_{x_C} p(x_C) \cdot p(x_w \mid x_v, x_C).
\end{aligned}$$

□

Proposition 8.9. *Let $\mathcal{G}$ be a causal DAG. The set $\mathrm{pa}_{\mathcal{G}}(v)$ is a back-door adjustment set for (v, w).*

Proof. Any (v, w) back-door path starts with an edge $v \leftarrow t$, so clearly $t \in \mathrm{pa}_G(v)$ is a non-collider on the path, which is therefore blocked. □

8.5 Example: HIV Treatment

Figure 20 depicts a situation that arises in HIV treatment, and more generally in the treatment of chronic diseases. A doctor prescribes patients with AZT (A), which is known to reduce AIDS related mortality, but also harms the immune system of the patient, increasing the risk of opportunistic infections such as pneumonia (L). If pneumonia arises, patients are generally treated with antibiotics (B), and the outcome of interest is 5 year survival (Y).

An epidemiologist might ask what the effect on survival would be if we treated all patients with antibiotics and AZT from the start, without waiting for an infection to present. What would this do to survival?

Well,

$$p(y \mid do(a, b)) = \sum_l p(y \mid a, l, b) p(l \mid a),$$

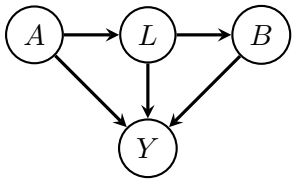


Figure 20: Causal diagram representing treatment for HIV patients. A is treatment with AZT (an anti-retroviral drug), L represents infection with pneumonia, B treatment with antibiotics, and Y survival.

so the answer can be determined directly from observed data without having to perform an experiment.

$$P(Y = 1 \mid do(A = 1, B = 1)) = \sum_{l=0}^{1} P(Y = 1 \mid A = 1, L = l, B = 1) \cdot P(L = l \mid A = 1).$$

Note that, in this case, there is no 'back-door' like solution, because L is a descendant of A so cannot form part of a back-door set, but without including on L the back-door path $B \leftarrow L \rightarrow Y$ will introduce spurious (i.e. non-causal) correlations.

8.6 Gaussian Causal Models

The adjustment formula can be thought of as averaging the conditional distribution over a portion of the population:

$$p(y \mid do(z)) = \sum_{x_C} p(x_C) \cdot p(y \mid z, x_C).$$

If the variables we are dealing with are multivariate Gaussian, then conditional distributions such as $p(y \mid z, x_C)$ are determined by regressing Y on Z, X_C using a simple linear model.

The regression coefficient between Z and Y in such a model is the same for all values of $X_C = x_C$, and therefore in this case we can forget the averaging and just look at the regression to obtain the causal effect. Consider the example in Figure 19: if we regress Y on X then the estimate we obtain is biased because of the back-door path $X \leftarrow Z \rightarrow W \rightarrow Y$; but if we add in Z or W (or both), then the estimate will be unbiased. See slides for an example.

9 Variational and Monte Carlo Methods

Not all interesting graphs can be triangulated in such a way as to give a junction tree with small cliques. This makes operations such as marginalization and updating with evidence intractable for large graphs. Alternative approaches to inference are based on either Markov chain Monte Carlo or, more recently, approximations to the model with *variational methods*.

9.1 Exponential Families

Let $p(\cdot;\theta)$ be a collection of probability densities over $\mathcal{X}$ indexed by θ. We say that p is an *exponential family* if it can be written as

$$p(x;\theta) = \exp\left\{\sum_i \theta_i \phi_i(x) - A(\theta)\right\}.$$

The functions ϕ_i are the *sufficient statistics*, and the components θ_i are called the *canonical parameters*. We can replace the sum with an inner product of vectors $\theta = (\theta_i)$ and $\phi = (\phi_i(x))$:

$$p(x;\theta) = \exp\left\{\langle \theta, \phi\rangle - A(\theta)\right\}.$$

You may be used to seeing an additional function of x in the definition of an exponential family. However, one can incorporate this quantity into the 'base measure' used for integrating over x, so we do not write it explicitly.

The function $A(\theta)$ is the *cumulant function*, and must be chosen so that the distribution normalizes, i.e.

$$A(\theta) = \log \int \exp\left\{\langle \theta, \phi\rangle\right\}\, dx.$$

$Z(\theta) \equiv e^{A(\theta)}$ is also called the *partition function*.

Lemma 9.1. *We have*

$$\nabla A(\theta) = \mathbb{E}_\theta \phi(X), \qquad \nabla\nabla^T A(\theta) = \mathrm{Cov}_\theta\, \phi(X).$$

Consequently we call the function $\mu(\theta) \equiv \nabla A(\theta)$ the mean function.

Proof. For the first part,

$$\begin{aligned} e^{A(\theta)} \frac{\partial}{\partial \theta_i} A(\theta) &= \frac{\partial}{\partial \theta_i} e^{A(\theta)} \\ &= \frac{\partial}{\partial \theta_i} \int \exp\left\{\langle \theta, \phi\rangle\right\}\, dx \\ &= \int \frac{\partial}{\partial \theta_i} \exp\left\{\langle \theta, \phi\rangle\right\}\, dx \\ &= \int \phi_i(x) \exp\left\{\langle \theta, \phi\rangle\right\}\, dx \\ &= e^{A(\theta)} \int \phi_i(x) \exp\left\{\langle \theta, \phi\rangle - A(\theta)\right\}\, dx \\ &= e^{A(\theta)} \mathbb{E}_\theta \phi_i(X). \end{aligned}$$

The second part follows similarly, see Examples Sheet 4. □

The property of convexity plays an important role in the computational advantages of exponential families. Recall that a real valued function f is convex if $f((x+y)/2) \leq (f(x)+f(y))/2$ for all x, y. Convex functions are easy to work with for the purposes of optimization: in particular, they do not contain multiple local minima.

As a consequence of the fact that the Hessian matrix of A is a covariance matrix and therefore positive definite, it follows that A is a convex function. Hence $\log p(x;\theta)$ is a concave function of θ for each x.

Example 9.2. Let $X \sim \text{Poisson}(\lambda)$. We have

$$p_\lambda(x) = e^{-\lambda}\frac{\lambda^x}{x!} = \frac{1}{x!}\exp\{x\log\lambda - \lambda\}.$$

Clearly the canonical parameter is $\theta = \log\lambda$, so we can rewrite as

$$p_\theta(x) = \frac{1}{x!}\exp\left\{\theta x - e^\theta\right\},$$

giving $A(\theta) = e^\theta$ (which is convex, as expected). Note that $A'(\theta) = A''(\theta) = e^\theta = \lambda$, which is indeed the mean and variance of a Poisson distribution. We can incorporate $1/x!$ into the measure so that expectations over a density $p(x)$ are calculated as

$$\mathbb{E}_\theta f(X) = \sum_x \frac{p(x)}{x!} f(x).$$

Examples: Binary Graphical Model

Let $\mathcal{G}(V, E)$ be an undirected graph with cliques $\mathcal{C}$. Let

$$\tilde{\mathcal{C}} = \{C : \emptyset \neq C \subseteq D \text{ for some } D \in \mathcal{C}\};$$

i.e., $\tilde{\mathcal{C}}$ is the collection of (non-empty) complete sets in $\mathcal{G}$.

Suppose that $X_V \in \{0,1\}^{|V|}$; we already know that $p(x_V) > 0$ is Markov with respect to $\mathcal{G}$ if and only if

$$\log p(x_V) = \lambda_\emptyset + \sum_{A\in\tilde{\mathcal{C}}} \lambda_A(x_A)$$

for functions λ_A that are zero unless all entries in x_A are 1. We can rewrite this as

$$\log p(x_V) = \lambda_\emptyset + \sum_{A\in\tilde{\mathcal{C}}} \theta_A \phi_A(x_A),$$

where $\phi_A(x_A) = \prod_{a\in A} x_a$, and θ_A are constants. This is an exponential family form with cumulant function $A(\theta) = -\lambda_\emptyset$, and the sufficient statistics $\phi_A(x_A)$ are the 'raw moments' of X_A. Note that $\mathbb{E}\phi_A(x_A) = \mathbb{E}\prod_{a\in A} X_a = P(X_A = 1)$.

Since

$$P(X_A = 1) = \sum_{x_{C\setminus A}} P(X_A = 1, X_{C\setminus A} = x_{C\setminus A}) \qquad \forall A \subseteq C,$$

it follows from the Möbius transformation that (Examples Sheet 1)

$$P(X_A = 1, X_{C\setminus A} = 0) = \sum_{B: A\subseteq B\subseteq C} (-1)^{|B\setminus A|} P(X_B = 1) \qquad \forall A \subseteq C$$

Hence the distribution of each clique C is determined by the raw moments of X_A for $A \subseteq C$.

Examples: Gaussian

Let $\mathcal{G}(V, E)$ be an undirected graph, and take

$$\begin{aligned} p(x_V) &= \exp\left\{\sum_{i\in V}\left(\theta_i x_i + \theta_{ii}x_i^2\right) + \sum_{\{i,j\}\in E}\theta_{ij}x_i x_j - A(\theta)\right\}, \qquad x_V \in \mathbb{R}^{|V|} \\ &= \exp\left\{\theta_V^T x_V - \frac{1}{2}x_V^T K(\theta_E)x_V - A(\theta)\right\} \end{aligned}$$

where $K_{ij} = K_{ji} = -\theta_{ij}$. Then, assuming that $K(\theta_E)$ is positive definite, we have

$$p(x_V) = \exp\left\{-\frac{1}{2}(x_V - K^{-1}\theta_V)^T K(x_V - K^{-1}\theta_V) + \frac{1}{2}\theta_V^T K^{-1}\theta_V - A(\theta)\right\},$$

which proportional to a multivariate Gaussian distribution with mean $\mu = K^{-1}\theta_V$ and covariance matrix K^{-1}. It follows that

$$A(\theta) = \frac{1}{2}\theta_V^T K^{-1}\theta_V + \frac{1}{2}\log\det K^{-1}.$$

So the Gaussian graphical model corresponds to the exponential family whose sufficient statistics are $(x_i, x_i^2, i \in V)$ and $(x_i x_j, \{i,j\} \in E)$.

9.2 Empirical Moment Matching

To find the maximum likelihood estimate in an exponential family, we maximize the log-likelihood

$$\begin{aligned} l(\theta; x^{(1)}, \ldots, x^{(n)}) &= \left\langle \sum_{i=1}^n \phi(x^{(i)}), \theta \right\rangle - nA(\theta) \\ n^{-1} l(\theta; x^{(1)}, \ldots, x^{(n)}) &= \langle \overline{\phi(x)}, \theta\rangle - A(\theta) \end{aligned}$$

where $\overline{\phi(x)} = n^{-1}\sum_i \phi(x^{(i)})$ is the sample mean of the sufficient statistics. To maximize this, we can differentiate and set to zero, obtaining

$$\overline{\phi(x)} - \nabla A(\theta) = 0,$$

so in other words when we choose θ so that the mean of the sufficient statistics matches the empirical mean from the data.

Note also that if we differentiate just with respect to θ_i, we obtain the same result for each sufficient statistic separately; hence if we update the parameters to match the moment $\overline{\phi_i(x)} = \mathbb{E}_\theta \phi_i(X)$, then we increase the log-likelihood. If we iterate this over i, we will converge to the global maximum likelihood estimate, because the log-likelihood is a concave function.

Iterative Proportional Fitting

For graphical models, this update process has a particularly nice interpretation. Suppose that we have an undirected model with cliques $\mathcal{C}$. We have seen above that this corresponds to the exponential family with sufficient statistics given by the relevant marginal

distributions $\{p(x_C), C \in \mathcal{C}\}$. The previous commentary shows that we can obtain the maximum likelihood estimator by finding the unique distribution in the model with those margins. Formally this result is given by the following theorem, also given as Theorem 4.25.

Theorem 9.3. *Let $\mathcal{G}$ be an undirected graph with cliques $\mathcal{C}$, and let $n(x_V)$ be a table of counts. Then the MLE under $\mathcal{G}$ is the unique distribution $\hat{p}(x_V)$ that is Markov with respect to $\mathcal{G}$ and such that*

$$n\hat{p}(x_C) = n(x_C) \qquad \text{for all } C \in \mathcal{C}.$$

The *iterative proportional fitting* (IPF) algorithm, also sometimes called the *iterative proportional scaling* (IPS) algorithm, starts with a discrete distribution that satisfies the Markov property for the graph $\mathcal{G}$ (usually we pick the uniform distribution, so that everything is independent), and then iteratively fixes each margin $p(x_C)$ to match the required distribution using the update step:

$$\begin{aligned} p^{(t+1)}(x_V) &= p^{(t)}(x_V) \cdot \frac{p(x_C)}{p^{(t)}(x_C)} \\ &= p^{(t)}(x_{V \setminus C} \mid x_C) \cdot p(x_C). \end{aligned} \tag{6}$$

Note that this is closely related to the message passing algorithm in Section 7.

Algorithm 2 Iterative Proportional Fitting (IPF) algorithm.

```
function IPF(collection of consistent margins q(x_{C_i}) for sets C_1, ..., C_k)
    set p(x_V) to uniform distribution;
    while max_i max_{x_{C_i}} |p(x_{C_i}) − q(x_{C_i})| > tol do
        for i in 1, ..., k do
            update p(x_V) to p(x_{V\C_i} | x_{C_i}) · q(x_{C_i});
        end for
    end while
    return distribution p with margins p(x_{C_i}) = q(x_{C_i}).
end function
```

The sequence of distributions in IPF converges to the MLE $\hat{p}(x_V)$. To see this, first note that the update (6) ensures that the moments for the sufficient statistics involving the clique C are matched. Second, after each update step the joint distribution remains Markov with respect to $\mathcal{G}$: this can be seen easily by considering the factorization. Performing each step increases the likelihood, and since the log-likelihood is concave, this sort of co-ordinate based iterative updating scheme will converge to the global maximum.

Example 9.4. Consider the 4-cycle in Figure 16(a), with cliques $\{1,2\}, \{2,3\}, \{3,4\}, \{1,4\}$.

Suppose we have data from $n = 96$ observations as shown in the table below (the column count).

X_1	X_2	X_3	X_4	count	step 0	step 1	step 2	step 3	step 4	$\hat{n}$
0	0	0	0	5	6	7.5	13	13	12.59	12.6
1	0	0	0	10	6	3.75	6.5	6.5	6.97	6.95
0	1	0	0	20	6	9.25	11.97	11.97	11.59	11.58
1	1	0	0	1	6	3.5	4.53	4.53	4.86	4.87
0	0	1	0	0	6	7.5	2	1.17	1.13	1.13
1	0	1	0	3	6	3.75	1	0.58	0.63	0.63
0	1	1	0	4	6	9.25	6.53	3.81	3.69	3.69
1	1	1	0	0	6	3.5	2.47	1.44	1.55	1.55
0	0	0	1	24	6	7.5	13	13	13.33	13.35
1	0	0	1	0	6	3.75	6.5	6.5	6.11	6.1
0	1	0	1	9	6	9.25	11.97	11.97	12.28	12.27
1	1	0	1	3	6	3.5	4.53	4.53	4.26	4.28
0	0	1	1	1	6	7.5	2	2.83	2.91	2.91
1	0	1	1	2	6	3.75	1	1.42	1.33	1.33
0	1	1	1	4	6	9.25	6.53	9.25	9.49	9.46
1	1	1	1	10	6	3.5	2.47	3.5	3.29	3.3

The marginals over the cliques are:

$n(x_{12})$	$X_2=0$	1
$X_1=0$	30	37
1	15	14

$n(x_{23})$	$X_3=0$	1
$X_2=0$	39	6
1	33	18

$n(x_{34})$	$X_4=0$	1
$X_3=0$	36	36
1	7	17

$n(x_{14})$	$X_4=0$	1
$X_1=0$	29	38
1	14	15

To implement IPF, we start with a uniform table, given in the column 'step 0'. We then scale the entries so as to match the X_1, X_2 margin above. For instance, the four entries corresponding to $X_1 = X_2 = 0$ are scaled to add up to 30; this gives the column 'step 1'. This is repeated for each of the other cliques, giving steps 2–4. By the fourth step the distribution of all cliques has been updated, but note that the margin over X_1, X_2 is now 29.96, 15.04, 37.04, 13.96. We keep cycling until the process converges to the final column, which matches all four margins.

9.3 Maximum Entropy

Let $p(x)$, $x \in \mathcal{X}$ be a probability distribution. The *entropy* (or *Shannon entropy*) of p is defined as

$$H(p) = -\mathbb{E} \log p(X) = -\int_{\mathcal{X}} p(x) \log p(x)\, dx.$$

This is the expected 'self-information' provided by a draw from p.

The *principle of maximum entropy* says that, given some knowledge about a model, any remaining aspects should be chosen so as to maximize the uncertainty about the model in the form of the entropy.

Example 9.5. The maximum entropy distribution such that $\mathbb{E}X = \mu$ and $\operatorname{Var} X = \sigma^2$ is the Gaussian. (See exercise sheet.)

Now suppose we have some general set of statistics $\mathbb{E}\phi_i(X) = \mu_i$ to match. The optimization problem is:

$$\text{maximize } H(p) = -\mathbb{E}_p \log p \qquad \text{subject to } \mathbb{E}_p\phi_i(X) = \mu_i, i = 1, \ldots, k.$$

We can write this as a Lagrange multiplier problem, so that the solution is a stationary point of

$$\begin{aligned} L(p, \theta) &= H(p) + \sum_{i=1}^{k} \theta_i(\mu_i - \mathbb{E}\phi_i(X)) + \theta_0(1 - \sum_x p(x)) \\ &= -\sum_x p(x) \log p(x) + \sum_{i=1}^{k} \theta_i(\mu_i - \sum_x p(x)\phi_i(x)) + \theta_0(1 - \sum_x p(x)) \\ &= \sum_x \left\{ -p(x) \log p(x) + \sum_{i=1}^{k} \theta_i p(x)(\mu_i - \phi_i(x)) + \theta_0(1 - p(x)) \right\} \end{aligned}$$

Differentiating with respect to $p(x)$ and looking for a stationary point, we obtain

$$\frac{\partial L}{\partial p(x)} = -\log p(x) - 1 + \sum_{i=1}^{k} \theta_i(\mu_i - \phi_i(x)) - \theta_0 = 0$$

and rearranging this,

$$p(x) \propto \exp\left\{ \sum_{i=1}^{k} \theta_i \phi(x) \right\}.$$

The distribution which maximizes the entropy is therefore of the form

$$p(x) = \exp\left\{ \sum_i \theta_i \phi_i(x) - A(\theta) \right\},$$

which is the *exponential family* with sufficient statistics ϕ_i and canonical parameters θ_i! Note that the Lagrange multiplier θ_0 does not appear in the final result directly, since A is chosen to ensure normalization. The same result can also be obtained for the continuous case using techniques from Calculus of Variations.

9.4 Duality and Variation

We saw above that $\mu = \nabla A(\theta) = \mathbb{E}_\theta \phi(X)$ is a map from the canonical parameter θ to the mean of the sufficient statistics; the derivative of this is positive definite, from which it follows that this map $\mu(\theta)$ is smooth and bijective with inverse $\theta(\mu)$. Thus we can equally well describe a particular distribution $p_\theta(x)$ using the canonical parameter θ or the *mean parameter* $\mu(\theta) = \mathbb{E}_\theta \phi(x)$.

The *conjugate dual* function to $A(\theta)$ is defined as

$$A^*(\mu) = \langle \mu, \theta(\mu) \rangle - A(\theta(\mu)).$$

The mappings $\mu(\theta)$ and $\theta(\mu)$ are not necessarily easy to evaluate (think about the Möbius transformation). However, when we consider the right hand side as a function of *separate*

parameters μ and θ we have $\langle \mu, \theta \rangle - A(\theta)$, and we have already seen that this expression is maximized by choosing the θ such that $\mathbb{E}_\theta \phi(X) = \mu$, i.e. by $\theta = \theta(\mu)$. Hence

$$A^*(\mu) \geq \langle \mu, \theta \rangle - A(\theta) \tag{7}$$

with equality if and only if $\theta = \theta(\mu)$. This inequality is called the *variation bound*, and plays a key part in variational inference.

9.5 The EM Algorithm

Suppose we have random variables X, Y from an exponential family

$$p_\theta(x, y) = \exp\left\{ \langle \theta, \phi(x, y) \rangle - A(\theta) \right\}.$$

If the observations X are unobserved, then this computationally nice family becomes intractable to deal with. The log-likelihood is

$$\begin{aligned} \log p_\theta(y) &= \log \int_{\mathcal{X}} \exp\left\{ \langle \theta, \phi(x, y) \rangle - A(\theta) \right\} \, dx \\ &= \log \int_{\mathcal{X}} \exp\left\{ \langle \theta, \phi(x, y) \rangle \right\} \, dx - A(\theta) \\ &\equiv A_y(\theta) - A(\theta). \end{aligned}$$

Notice that for fixed y,

$$p_\theta(x \mid y) = \exp\left\{ \langle \theta, \phi(x, y) \rangle - A_y(\theta) \right\}$$

is an exponential family with sufficient statistics $\phi(X, y)$. Applying the variation bound (7) we see that

$$A_y(\theta) \geq \langle \mu_y, \theta \rangle - A_y^*(\mu_y),$$

where A_y^* is the conjugate dual for this exponential family on $X \mid Y = y$.

Using this dual representation we have

$$\log p_\theta(y) \geq \langle \theta, \mu_y \rangle - A_y^*(\mu_y) - A(\theta).$$

The *EM algorithm* (Expectation-Maximization) arises from iteratively maximizing the right hand side of this expression with respect to μ_y and θ. In the first case we want to maximize

$$\langle \theta, \mu_y \rangle - A_y^*(\mu_y)$$

with respect to μ_y, which, by the variation bound discussion is achieved at $\mu_y = \mathbb{E}_\theta[\phi(X, y) \mid Y = y]$. So, in other words, we pick μ_y to be the mean of our sufficient statistics given the current parameter values. This is the **E-step**. For the second case we maximize

$$\langle \mu_y, \theta \rangle - A(\theta)$$

with respect to θ. This corresponds to just finding the MLE of the original full exponential family, but with the unknown $\phi(X, y)$ replaced by its conditional mean $\mathbb{E}_\theta[\phi(X, y) \mid Y = y]$. This is the **M-step**.

Example 9.6. Suppose that we have $X \sim \text{Bernoulli}(\pi)$ and $Y_i \mid X = x \sim \text{Bernoulli}(\theta_x)$ independently for $i = 1, \ldots, I$. If we observe only $Y_1, \ldots, Y_I$, what should we conclude about π, θ_x? The log-likelihood for one observation is

$$\begin{aligned} l(\pi, \theta_1, \theta_0; X, \boldsymbol{Y}) &= X \log \pi + (1 - X) \log(1 - \pi) + X \sum_{i=1}^{I} \{Y_i \log \theta_1 + (1 - Y_i) \log(1 - \theta_1)\} \\ &\quad + (1 - X) \sum_{i=1}^{I} \{Y_i \log \theta_0 + (1 - Y_i) \log(1 - \theta_0)\} \\ &= X \left\{ \log \pi + \log \theta_1 \sum_i Y_i + \log(1 - \theta_1) \sum_i (1 - Y_i) \right\} \\ &\quad + (1 - X) \left\{ \log(1 - \pi) + \log \theta_0 \sum_i Y_i + \log(1 - \theta_0) \sum_i (1 - Y_i) \right\}, \\ &= \alpha X + \beta(1 - X), \end{aligned}$$

and hence, conditional on $Y_1, \ldots, Y_I$, the distribution of X is still Bernoulli with probability

$$\mu_Y = \frac{e^\alpha}{e^\alpha + e^\beta} = \frac{\pi \theta_1^{\sum_i Y_i} (1 - \theta_1)^{\sum_i (1 - Y_i)}}{\pi \theta_1^{\sum_i Y_i} (1 - \theta_1)^{\sum_i (1 - Y_i)} + (1 - \pi) \theta_0^{\sum_i Y_i} (1 - \theta_0)^{\sum_i (1 - Y_i)}}. \tag{8}$$

This is the E-step. Now, returning to the complete data log-likelihood, the maximization with respect to the parameters occurs after replacing each X with $\mu_y = \mathbb{E}_\theta[X \mid Y = y]$. This gives

$$\begin{aligned} l(\pi, \theta_1, \theta_0; X, \boldsymbol{Y}) &= \sum_j \mu_Y^j \left\{ \log \pi + \log \theta_1 \sum_i Y_i^j + \log(1 - \theta_1) \sum_i (1 - Y_i^j) \right\} \\ &\quad + \sum_j (1 - \mu_Y^j) \left\{ \log(1 - \pi) + \log \theta_0 \sum_i Y_i^j + \log(1 - \theta_0) \sum_i (1 - Y_i^j) \right\}. \end{aligned}$$

But this is straightforward: differentiating the log-likelihood and solving gives

$$\hat{\pi} = n^{-1} \sum_j \mu_Y^j \qquad \hat{\theta}_1 = \frac{\sum_j \mu_Y^j \sum_i Y_i^j}{\sum_j \mu_Y^j} \qquad \hat{\theta}_0 = \frac{\sum_j (1 - \mu_Y^j) \sum_i Y_i^j}{n - \sum_j \mu_Y^j}.$$

These updates are the M-step. Iterating between the E-step and the M-step leads to a local maximum in the marginal likelihood $p(\boldsymbol{y}; \pi, \theta_0, \theta_1)$; convergence to the global maximum is not guaranteed (and often fail in practice), though this can be overcome by restarting the algorithm from several different places and comparing the likelihood values for the different solutions.

9.6 Conjugate Priors

Now suppose we take a Bayesian approach and include a conjugate prior distribution for our likelihood, of the form

$$p_{\xi,\lambda}(\theta) = \exp\left\{ \langle \xi, \theta \rangle - \lambda A(\theta) - B(\lambda, \xi) \right\}.$$

Note that this is an exponential family over θ with sufficient statistics $(\theta, -A(\theta))$ and natural parameters (ξ, λ). This means that

$$\begin{aligned} p(\theta \mid x, y, \xi, \lambda) &\propto p_{\xi,\lambda}(\theta) p_\theta(x, y) \\ &\propto \exp\left\{\langle \phi(x, y) + \xi, \theta\rangle - (1+\lambda)A(\theta) - B(\lambda, \xi)\right\} \\ &\propto \exp\left\{\langle \phi(x, y) + \xi, \theta\rangle - (1+\lambda)A(\theta)\right\}, \end{aligned}$$

i.e. the same exponential family with the statistics (ξ, λ) replaced by $(\xi + \phi(x, y), \lambda + 1)$. In the case with n i.i.d. samples this becomes $(\xi + \sum_i \phi(x_i, y_i), \lambda + n)$.

Example 9.7. Suppose $X \sim \text{Bernoulli}(\pi)$, so that

$$\log p(x; \pi) = x \log \pi + (1 - x) \log(1 - \pi) = x \operatorname{logit} \pi + \log(1 - \pi)$$

which is in exponential family form with canonical parameter $\operatorname{logit} \pi$. A conjugate prior for π is one with the form

$$\log p_{\xi,\lambda}(\pi) = \xi \operatorname{logit} \pi + \lambda \log(1 - \pi) = \xi \log \pi + (\lambda - \xi) \log(1 - \pi).$$

This is a Beta distribution with parameters ξ and $\lambda - \xi$, so the posterior distribution is the Beta distribution with parameters $\xi + x$ and $\lambda - \xi + 1 - x$. One can verify that this is correct using the ordinary parameterizations of the Bernoulli and Beta distribution.

Example 9.8. Suppose $X \sim N(\eta, \tau^{-1})$, so that τ is the inverse variance. We have seen that

$$\log p(x; \eta, \tau) = \frac{1}{2} \log \tau - \frac{\tau}{2} x^2 + \tau \eta x - \frac{\tau}{2} \eta^2.$$

which is in exponential family form with canonical parameters τ and $\tau\eta$. A conjugate prior is of the form

$$\begin{aligned} \log p_{\xi,\lambda}(\eta, \tau) &= \xi_1 \tau + \xi_2 \tau \eta - \lambda \frac{\tau}{2} \eta^2 + \lambda \frac{1}{2} \log \tau + B(\xi_1, \xi_2, \lambda) \\ &= \xi_1 \tau - \lambda \frac{\tau}{2} (\eta - \lambda^{-1} \xi_2)^2 + \lambda \frac{1}{2} \log \tau + B'(\xi_1, \xi_2, \lambda). \end{aligned}$$

This is a Gamma distribution for τ with parameters $\lambda/2, -\xi_1$, and a conditional normal distribution for $\eta \mid \tau$ with mean $\lambda^{-1}\xi_2$ and variance $(\tau\lambda)^{-1}$.

9.7 Variational Bayes

Now suppose we are interested in finding the marginal likelihood, that is

$$p_{\xi^*,\lambda^*}(y) = \int_{\mathcal{X}} \int p_\theta(x, y) \cdot p_{\xi^*,\lambda^*}(\theta) \, d\theta \, dx.$$

This is a fundamental problem in Bayesian inference, since it is used in the computation of the Bayes factor for a model.

Taking logs, multiplying and dividing the right-hand side by $p_{\xi,\lambda}(\theta)$, and rearranging we obtain

$$\log p_{\xi^*,\lambda^*}(y) = \log \int \int_{\mathcal{X}} p_\theta(x, y) \, dx \cdot p_{\xi,\lambda}(\theta) \frac{p_{\xi^*,\lambda^*}(\theta)}{p_{\xi,\lambda}(\theta)} \, d\theta.$$

Now, for a concave function f, Jensen's inequality says that $f(\mathbb{E}X) \geq \mathbb{E}f(X)$; applying this to the density $p_{\xi,\lambda}(\theta)$ and the function $f(x) = \log(x)$ Jensen's inequality, we can write

$$\begin{aligned}
\log p_{\xi^*,\lambda^*}(y) &\geq \int_\theta p_{\xi,\lambda}(\theta) \cdot \log \left\{ \int_{\mathcal{X}} p_\theta(x,y)\, dx \cdot \frac{p_{\xi^*,\lambda^*}(\theta)}{p_{\xi,\lambda}(\theta)} \right\} d\theta \\
&\geq \int_\theta p_{\xi,\lambda}(\theta) \cdot \log \int_{\mathcal{X}} p_\theta(x,y)\, dx\, d\theta + \int_\theta p_{\xi,\lambda}(\theta) \log \frac{p_{\xi^*,\lambda^*}(\theta)}{p_{\xi,\lambda}(\theta)}\, d\theta \\
&\geq \int_\theta p_{\xi,\lambda}(\theta) \cdot \{A_y(\theta) - A(\theta)\}\, d\theta + \int_\theta p_{\xi,\lambda}(\theta) \log \frac{p_{\xi^*,\lambda^*}(\theta)}{p_{\xi,\lambda}(\theta)}\, d\theta \qquad (9)
\end{aligned}$$

by the same method as used for the EM algorithm. Using the dual bound, we can lower-bound the first term by

$$\int_\theta p_{\xi,\lambda}(\theta) \cdot \left\{ \langle \mu, \theta \rangle - A_y^*(\mu) - A(\theta) \right\} d\theta = \langle \mu, \mathbb{E}_{\xi,\lambda}\theta \rangle - A_y^*(\mu) - \mathbb{E}_{\xi,\lambda}A(\theta).$$

However, note that the value of μ at which we obtain equality is dependent upon θ, so it should be different at each point in the integral. Thus, this is certainly an approximation to the original problem and, unlike EM, we will not obtain the exact value of the marginal likelihood. Instead we will get a lower-bound.

The second term, meanwhile, is

$$\begin{aligned}
&\int_\theta p_{\xi,\lambda}(\theta) \left\{ \langle \theta, \xi^* - \xi \rangle - (\lambda^* - \lambda)A(\theta) - B(\xi^*, \lambda^*) + B(\xi, \lambda) \right\} d\theta \\
&= \langle \mathbb{E}_{\xi,\lambda}\theta, \xi^* - \xi \rangle - (\lambda^* - \lambda)\mathbb{E}_{\xi,\lambda}A(\theta) - B(\xi^*, \lambda^*) + B(\xi, \lambda)
\end{aligned}$$

By definition of the conjugate dual B^* we have

$$B(\xi, \lambda) = \langle \mathbb{E}_{\xi,\lambda}\theta, \xi \rangle + \langle -\mathbb{E}_{\xi,\lambda}A(\theta), \lambda \rangle - B^*(\mathbb{E}_{\xi,\lambda}\theta, \mathbb{E}_{\xi,\lambda}A(\theta)),$$

so substituting this in gives

$$\langle \mathbb{E}_{\xi,\lambda}\theta, \xi^* \rangle - \lambda^* \mathbb{E}_{\xi,\lambda}A(\theta) - B(\xi^*, \lambda^*) - B^*(\mathbb{E}_{\xi,\lambda}\theta, \mathbb{E}_{\xi,\lambda}A(\theta)).$$

Since ξ^*, λ^* are fixed, we can ignore $B(\xi^*, \lambda^*)$, and combining with the first term again we want to maximize

$$\langle \mu + \xi^*, \mathbb{E}_{\xi,\lambda}\theta \rangle - (1 + \lambda^*)\mathbb{E}_{\xi,\lambda}A(\theta) - A_y^*(\mu) - B^*(\mathbb{E}_{\xi,\lambda}\theta, \mathbb{E}_{\xi,\lambda}A(\theta))$$

With respect to μ, this involves setting

$$\hat{\mu} = \mathbb{E}_{\hat{\theta}}[\phi(X, y) \mid Y = y],$$

similar to the E-step in EM. With respect to θ we note that

$$\langle \mu + \xi^*, \mathbb{E}_{\xi,\lambda}\theta \rangle - (1 + \lambda^*)\mathbb{E}_{\xi,\lambda}A(\theta) - B^*(\mathbb{E}_{\xi,\lambda}\theta, \mathbb{E}_{\xi,\lambda}A(\theta))$$

looks like the conjugate dual for the original family with hyperparameters $(\mu + \xi^*, 1 + \lambda^*)$, so it will be maximized at the mean parameter for this distribution. Hence, the 'M-step' is to set

$$(\xi^{(t)}, \lambda^{(t)}) = (\mu^{(t)} + \xi^*, 1 + \lambda^*) \qquad \theta^{(t+1)} = \int \theta p_{\xi^{(t)},\lambda^{(t)}}(\theta)\, d\theta.$$

Example 9.9. Suppose $X_j \sim \text{Bernoulli}(\pi)$ and that $Y_{ij} \mid X_j = x \sim \text{Bernoulli}(\theta_x)$, with priors $\pi \sim \text{Beta}(a, b)$ and $\theta_x \sim \text{Beta}(c, d)$.

Then the posterior distribution for θ given the Y_{ij} can be approximated using variational Bayes. The full data likelihood is

$$l(\mu_x, \pi; x, y) = \sum_{j=1}^{n} x_j \log \pi + (1 - x_j) \log(1 - \pi) + \sum_{x=0}^{1} \mathbb{1}_{\{x_j = x\}} \sum_i \{y_{ij} \log \theta_x + (1 - y_{ij}) \log(1 - \theta_x)\}.$$

As in EM, the E-step is to replace x_i with

$$\mu_y = \frac{\pi \theta_1^{\sum_i Y_i} (1 - \theta_1)^{\sum_i (1 - Y_i)}}{\pi \theta_1^{\sum_i Y_i} (1 - \theta_1)^{\sum_i (1 - Y_i)} + (1 - \pi) \theta_0^{\sum_i Y_i} (1 - \theta_0)^{\sum_i (1 - Y_i)}}.$$

the updated hyperparameters are

$$\begin{aligned} a &= a^* + \sum_j \mu_y^j & b &= b^* + n - \sum_j \mu_y^j \\ c_1 &= c^* + \sum_j \mu_y^j \sum_i Y_{ij} & d_1 &= d^* + \sum_j \mu_y^j \sum_i (1 - Y_{ij}) \\ c_0 &= c^* + \sum_j (1 - \mu_y^j) \sum_i Y_{ij} & d_0 &= d^* + \sum_j (1 - \mu_y^j) \sum_i (1 - Y_{ij}). \end{aligned}$$

Under the distributions with these hyper-parameters, we want the expected natural parameters, i.e. $\text{logit}\, \pi = \log \pi - \log(1 - \pi)$ and $\text{logit}\, \theta_x$. These are computed using

$$\mathbb{E}_{a,b} \log \pi = \frac{1}{B(a, b)} \int_0^1 \pi^{a-1} (1 - \pi)^{b-1} \log \pi \, d\pi = \psi(a) - \psi(a + b),$$

where ψ is the digamma function ($\psi(x) = \frac{\partial}{\partial x} \log \Gamma(x)$). Then

$$\text{logit}\, \hat{\pi} = \psi(a) - \psi(b) \qquad \text{logit}\, \hat{\theta}_x = \psi(c_x) - \psi(d_x).$$

These steps are computationally easy to iterate. Upon convergence (which is guaranteed), we can evaluate the lower bound (9) from above. One can verify that

$$\begin{aligned} A(\mu, \theta_0, \theta_1) &= -\log(1 - \pi) - \log(1 - \theta_1) \sum_j \mu_Y^j - \log(1 - \theta_0) \sum_j (1 - \mu_Y^j) \\ \mathbb{E}_{a,b,c,d} A(\mu, \theta_0, \theta_1) &= \psi(a + b) - \psi(b) + \{\psi(c_1 + d_1) - \psi(d_1)\} \sum_j \mu_Y^j \\ &\quad + \{\psi(c_0 + d_0) - \psi(d_0)\} \sum_j (1 - \mu_Y^j), \end{aligned}$$

and we are approximating $A_y(\theta)$ by $A_y(\hat{\theta}) + \langle \mu, \theta - \hat{\theta} \rangle$, where $A_y(\hat{\theta}) = -\log(1 - \mu_Y)$. These calculations are relatively straightforward.

9.8 Gibbs Sampling

Finally, we discuss Gibbs Sampling, which gives a method of sampling from complicated joint distibutions, even if we are unable to obtain the normalizing constant. The idea is that we divide our distribution into simpler univariate conditional distributions, which often have nice parametric forms. Even if the univariate conditional is not a simple model, computing the normalizing constant only requires a one-dimensional integral (or sum), which is relatively easy to evaulate.

Example 9.10 (Ising Model)**.** Let $X_v \in \{-1, +1\}$, and let $\mathcal{G}(V, E)$ be a graph. The *Ising model* assumes that

$$p(\boldsymbol{x}; \theta) \propto \exp\left\{\theta \sum_{\{i,j\} \in E} x_i x_j\right\};$$

this is commonly used in the case of variables arranged on a $p \times p$ grid such as in a black and white image.

Note that

$$p(\boldsymbol{x}; \theta) = \frac{\exp\left\{\theta \sum_{\{i,j\} \in E} x_i x_j\right\}}{\sum_{x_V} \exp\left\{\theta \sum_{\{i,j\} \in E} x_i x_j\right\}} = \frac{\exp\left\{\theta \sum_{\{i,j\} \in E} x_i x_j\right\}}{Z(\theta)};$$

cannot generally be computed because the sum to obtain $Z(\theta)$ is intractable. The quantity $Z(\theta)$ is sometimes called the *partition function.*

Of course, if $\mathcal{G}$ is decomposable then we *can* perform these calculations using the expression from Theorem 4.24. But for non-decomposable graphs, there is generally no such nice expression.

To avoid this problem we introduce a very useful Markov chain Monte Carlo method called *Gibbs Sampling.* Suppose we wish to simulate from a distribution $p(x_1, \ldots, x_k)$, and are given each of the conditionals:

$$p(x_i \mid \boldsymbol{x}_{-i}) = p(x_i \mid x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_k).$$

The Gibbs Sampler is an algorithm that samples as follows:

Algorithm 3 Single iteration of the Gibbs sampler.

function GIBBS(current state $x_1, \ldots, x_k$, conditional distributions $p(x_i \mid x_{-i})$)
  for i in $1, \ldots, k$ **do**
    sample x_i^* from $p(x_i \mid x_1^*, \ldots, x_{i-1}^*, x_{i+1}, \ldots, x_k)$;
  end for
  return new state $(x_1^*, \ldots, x_k^*)$
end function

Repeatedly applying this algorithm gives a Markov chain on $x_1, \ldots, x_k$. Under mild conditions, the unique stationary distribution of the Markov chain is the joint distribution p, and the distribution of the state of the chain will converge to p. Hence, the state can be used as a sample from p. A potential disadvantage is that it may take a long time for the Markov chain to converge. We will not attempt to prove these facts, but see the Advanced Simulation course next term for more details.

Example 9.11. Suppose that p is a bivariate normal distribution with mean $\mu = (\mu_1, \mu_2)$

and covariance matrix K^{-1}. Then

$$\begin{aligned}\log p(x_1 \mid x_2) &= -\frac{1}{2}(x-\mu)^T K(x-\mu) + \text{const} \\ &= -\frac{1}{2}k_{11}(x_1-\mu_1)^2 - k_{12}(x_1-\mu_1)(x_2-\mu_2) - \frac{1}{2}k_{22}(x_2-\mu_2)^2 + \text{const} \\ &= -\frac{1}{2}k_{11}x_1^2 + (k_{11}\mu_1 - k_{12}(x_2-\mu_2))x_1 + \text{const} \\ &= -\frac{1}{2}k_{11}\left(x_1 - \mu_1 + \frac{k_{12}}{k_{11}}(x_2-\mu_2)\right)^2 + \text{const}.\end{aligned}$$

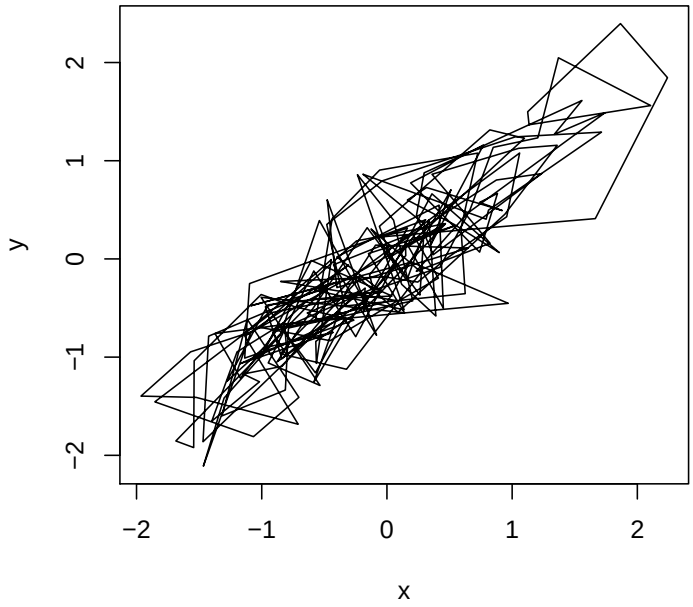
Hence, the conditional distribution of $X_1 \mid X_2 = x_2$ is Gaussian with mean $\mu_1 - \frac{k_{12}}{k_{11}}(x_2-\mu_2)$ and variance k_{11}^{-1}.

A similar result holds for $X_2 \mid X_1 = x_1$. Hence the Gibbs sampler consists of the following steps, starting from some initial $(x_1^{(0)}, x_2^{(0)})$, for $t = 1, 2, \ldots$,

- draw $x_1^{(t)}$ from $N(\mu_1 - \frac{k_{12}}{k_{11}}(x_2^{(t-1)} - \mu_2), k_{11}^{-1})$;
- draw $x_2^{(t)}$ from $N(\mu_2 - \frac{k_{12}}{k_{22}}(x_1^{(t)} - \mu_1), k_{22}^{-1})$.

In the case where $\mu_1 = \mu_2 = 0$ and the variances are 1, one can show that $K = \frac{1}{1-\rho^2}\begin{pmatrix} 1 & -\rho \\ -\rho & 1 \end{pmatrix}$, where ρ is the correlation. This gives updates of the form $x_1 \mid x_2 \sim N(\rho x_2, (1-\rho^2)^{-1})$. We implement this in the R code below.

```
> ## Gaussian Gibbs sampler
> rho <- 0.9  ## correlation
> N <- 200  ## number of samples
> x <- y <- numeric(N)
> x[1] <- y[1] <- 0
>
> for (i in 2:N) {
+   x[i] <- rnorm(1, mean=rho*y[i-1], sd=sqrt(1-rho^2))
+   y[i] <- rnorm(1, mean=rho*x[i], sd=sqrt(1-rho^2))
+ }
>
> plot(x,y, type="l")
```



Notice that there is correlation between the samples, as evidenced by the fact that the Markov chain tends to stay close to the previous step.

Applying Gibbs sampling to the Ising model makes it much easier to obtain samples from. Even though the joint distribution is hard to evaluate, the full conditional distribution of each variable is simple, because we have:

$$p(x_i \mid x_{V\setminus\{i\}}, \theta) \propto \exp\left\{\theta x_i \sum_{j \in \mathrm{bd}_{\mathcal{G}}(i)} x_j\right\}$$

$$p(x_i \mid x_{V\setminus\{i\}}, \theta) = \frac{\exp\left\{\theta x_i \sum_{j\in\mathrm{bd}_{\mathcal{G}}(i)} x_j\right\}}{\sum_{x_i} \exp\left\{\theta x_i \sum_{j\in\mathrm{bd}_{\mathcal{G}}(i)} x_j\right\}};$$

this is much easier to compute, because we only have to sum over a single variable x_i.

Example 9.12. Consider a model in which $X \sim \text{Bernoulli}(\pi)$ and $Y \mid X = x \sim N(\theta_x, 1)$ independently for $i = 1, \ldots, n$. We place priors $\pi \sim \text{Beta}(a, b)$ and $\theta_x \sim N(0, 1)$ for $x = 0, 1$.

Now suppose we observe Y but not X. The posterior distribution for the parameters is

$$p(\pi, \theta_0, \theta_1 \mid y) \propto p(y \mid \pi, \theta_0, \theta_1) p(\pi, \theta_0, \theta_1).$$

Unfortunately this is (relatively) hard to evaluate because $p(Y \mid \pi, \theta_0, \theta_1)$ does not have a simple form. However, we can instead consider

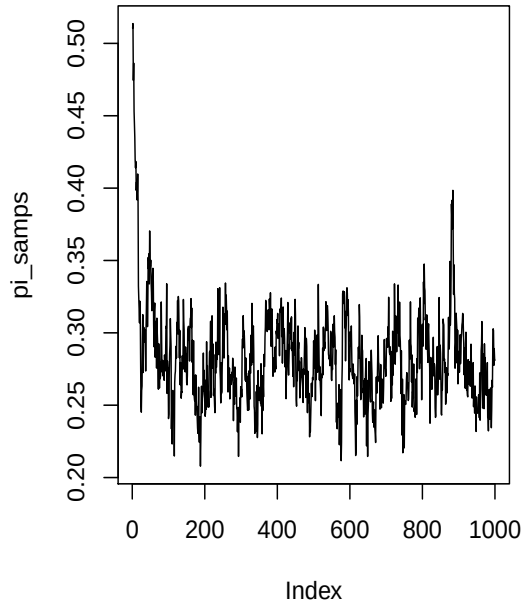
$$p(x, \pi, \theta_0, \theta_1 \mid y) \propto p(y \mid x, \theta_0, \theta_1) \cdot p(x \mid \pi) \cdot p(\pi, \theta_0, \theta_1).$$

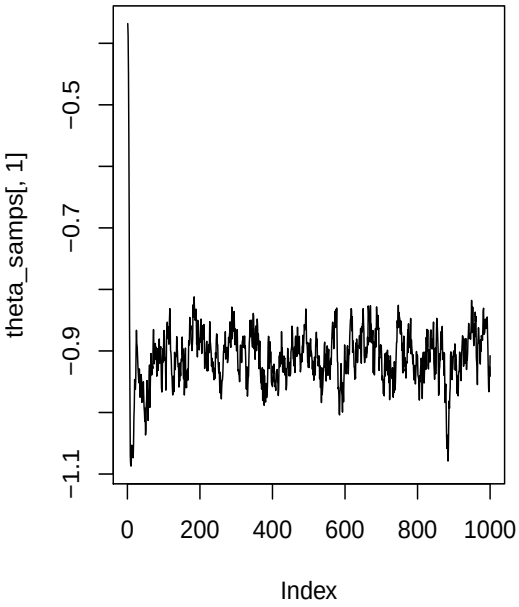
This is easy to work with because each factor on the right hand side has a simple closed form. In particular,

$$
\begin{aligned}
P(X = 1 \mid Y, \theta_0, \theta_1, \pi) &= \frac{\pi \exp(-(Y-\theta_1)^2/2)}{\pi \exp(-(Y-\theta_1)^2/2) + (1-\pi)\exp(-(Y-\theta_0)^2/2)} \\
\pi \mid X &\sim \text{Beta}(a + X, b + (1 - X)) \\
\theta_1 \mid X, Y &\sim N\left(\frac{1}{2}XY, \frac{1}{1+X}\right) \\
\theta_0 \mid X, Y &\sim N\left(\frac{1}{2}(1-X)Y, \frac{1}{1+(1-X)}\right).
\end{aligned}
$$

Since the full conditionals are easy to evaluate, we can just run a Gibbs sampler to obtain a sample from the joint posterior distribution of X and the parameters given Y. We can (if we choose) then just ignore the X samples and keep the sample of parameters.

```
> set.seed(674)
> ## generate data
> n <- 1000
> pi <- 0.3
> theta <- c(-1,1)
> X <- rbinom(n, 1, pi)
> Y <- rnorm(n, mean=theta[X+1], sd=1)
>
> ## initial states
> N <- 1000
> pi_samps <- numeric(N)
> theta_samps <- matrix(0, N, 2)
> X_samp <- rbinom(n, 1, 0.5)  # random starting point
>
> a <- b <- 1  ## prior for pi
>
> ## run Gibbs sampler
> for (i in 1:N) {
+   sumX <- sum(X_samp)
+   pi_samps[i] = rbeta(1, a+sumX, b+n-sumX)
+   theta_samps[i,1] = rnorm(1, sum(Y*(1-X_samp))/(1+n-sumX), 1/(1+n-sumX))
+   theta_samps[i,2] = rnorm(1, sum(Y*X_samp)/(1+sumX), 1/(1+sumX))
+   p <- pi_samps[i]*dnorm(Y, theta_samps[i,2])
+   p <- p/(p + (1-pi_samps[i])*dnorm(Y, theta_samps[i,1]))
+   X_samp <- rbinom(n, 1, p)
+ }
>
> par(mfrow=c(1,2))
> plot(pi_samps, type="l")
> plot(theta_samps[,1], type="l")
```






```
> par(mfrow=c(1,2))
> hist(pi_samps, col=2, breaks=50)
> hist(theta_samps[,1], col=4, breaks=50)
```