Graphical Models

Robin Evans
evans@stats.ox.ac.uk

Michaelmas 2018

This version: November 4, 2018

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, corresponding to classes in weeks 2, 5, 8 and HT week 1.

Part C and OMMS students should sign-up for classes with the online system.

Resources

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

   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.

   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.

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

A complementary book, written from a machine learning perspective.


As the name suggests, covers most of the material we will use for discussing contingency tables and log-linear models, as well as some data examples. Available for free at [https://mathdept.iut.ac.ir/sites/mathdept.iut.ac.ir/files/AGRESTI.PDF](https://mathdept.iut.ac.ir/sites/mathdept.iut.ac.ir/files/AGRESTI.PDF). There is also an updated third edition, but I will be using this second one.
Contents

1 Introduction 5

2 Conditional Independence 6
   2.1 Independence .................................................. 6
   2.2 Conditional Independence ..................................... 6
   2.3 Statistical Inference .......................................... 9

3 Exponential Families and Contingency Tables 11
   3.1 Exponential Families ........................................ 11
   3.2 Empirical Moment Matching .................................. 12
   3.3 Multivariate Gaussian Distribution .......................... 12
   3.4 Contingency Tables .......................................... 13
   3.5 Computation ..................................................... 14
   3.6 Log-linear models ............................................. 15
   3.7 Conditional Independence ................................. 15

4 Undirected Graphical Models 18
   4.1 Undirected Graphs ........................................... 18
   4.2 Markov Properties .......................................... 18
   4.3 Cliques and Factorization .................................... 20
   4.4 Decomposability .............................................. 21
   4.5 Separator Sets .................................................. 24
   4.6 Non-Decomposable Models .................................... 25

5 Gaussian Graphical Models 28
   5.1 Gaussian Graphical Models ................................. 28
   5.2 Maximum Likelihood Estimation .............................. 29
   5.3 Data Examples .................................................. 30

6 Directed Graphical Models 31
   6.1 Markov Properties ........................................... 32
   6.2 Ancestrality .................................................... 33
   6.3 Statistical Inference .......................................... 35
   6.4 Markov Equivalence ......................................... 36
   6.5 Directed Graphs, Undirected Graphs, and Decomposability 37
Recommended Prerequisites


Aims and Objectives

This course will give an overview of the use of graphical models as a tool for statistical inference. Graphical models relate the structure of a graph to the structure of a multivariate probability distribution, usually via conditional independence constraints. This has two broad uses: first, conditional independence can provide vast savings in computational effort, both in terms of the representation of large multivariate models and in performing inference with them; this makes graphical models very popular for dealing with big data problems. Second, conditional independence can be used as a tool to discover hidden structure in data, such as that relating to the direction of causality or to unobserved processes. As such, graphical models are widely used in genetics, medicine, epidemiology, statistical physics, economics, the social sciences and elsewhere.

Students will develop an understanding of the use of conditional independence and graphical structures for dealing with multivariate statistical models. They will appreciate how this is applied to causal modelling, and to computation in large-scale statistical problems.

Syllabus

Independence, conditional independence, graphoid axioms [1]

Exponential families, mean and canonical parameterisations, moment matching; contingency tables, log-linear models. [2]

Undirected graphs, cliques, paths; factorization and global Markov property, Hammersley-Clifford Theorem (statement only) [1]


The multivariate Gaussian distribution and Gaussian graphical models. [2]


Running intersection property, Junction trees; message passing, computation of marginal and conditional probabilities, introduction of evidence. Gibbs sampling. [2]

Causal models, structural equations, interventions; constraint-based learning, faithfulness. The trek rule and the backdoor criterion. [3]
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 \(2^{256^3} \approx 17\) 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 \(x_V \in X_V\). 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

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

Note that this is equivalent to

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

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

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

The qualification that these expressions hold for all $(x, y) \in \mathcal{X} \times \mathcal{Y}$, a product space, is very important\(^1\) and sometimes forgotten.

Example 2.1. Suppose that $X, W$ are independent Exponential($\lambda$) 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 $p(x \mid y)$ such that

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

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

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

\(^1\)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, and let the joint density be $p(x,y,z)$. We say that $X$ and $Y$ are **conditionally independent** given $Z$ if

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

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 $X \perp Y$.

**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, $X_k \perp X_1, \ldots, X_{k-2} \mid X_{k-1}$. 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) $p(x \mid y,z) = p(x \mid z)$ for all $x,y,z$ such that $p(y,z) > 0$;

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

(iii) $p(x,y,z) = p(y,z) \cdot p(x \mid z)$ for all $x,y,z$;

(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 $p(y \mid z)$, and (ii) $\implies$ (iii) by multiplying by $p(z)$. (iii) $\implies$ (i) directly.

The equivalence of (iii) and (iv) is also clear (note that if $p(z) = 0$ then both sides of (iii) are 0), 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 $p(x \mid y,z)$, and it does not depend upon $y$, so we obtain (iii). \qed
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.

<table>
<thead>
<tr>
<th>Victim</th>
<th>White</th>
<th>Black</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Defendant</td>
<td>White</td>
<td>Black</td>
</tr>
<tr>
<td>Yes</td>
<td>53</td>
<td>11</td>
</tr>
<tr>
<td>No</td>
<td>414</td>
<td>37</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Victim</th>
<th>Black</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td>Defendant</td>
<td>White</td>
</tr>
<tr>
<td>Yes</td>
<td>0</td>
</tr>
<tr>
<td>No</td>
<td>16</td>
</tr>
</tbody>
</table>

The marginal table has

<table>
<thead>
<tr>
<th>Defendant</th>
<th>White</th>
<th>Black</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Yes</td>
<td>53</td>
<td>15</td>
</tr>
<tr>
<td>No</td>
<td>430</td>
<td>176</td>
</tr>
</tbody>
</table>

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 sentenced to death while no white defendants were. (In fact we will see in Chapter 3.4 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 $X \perp \perp Y \mid W, Z \implies X \perp \perp Y, W \mid Z$. 

8
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

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

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’:

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

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.

**Remark 2.8.** Since the events $\{Y = y\}$ and $\{Y = y, h(Y) = h(y)\}$ are equal for any (measurable) function $h$, it follows that

$$p(x \mid y, z) = p(x \mid y, h(y), z).$$

This can be used to prove that

$$X \perp \perp Y \mid Z \quad \implies \quad X \perp \perp h(Y) \mid Z \quad \text{and} \quad X \perp \perp Y \mid h(Y), Z,$$

both of which are very useful facts.

### 2.3 Statistical Inference

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

**Example 2.9.** Suppose that $X \sim f_\theta$ for some parameter $\theta \in \Theta$. We say that $T \equiv t(X)$ is a **sufficient statistic** for $\theta$ 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 $\theta$, this is equivalent to saying that $X \perp \perp \theta \mid T$. 

9
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

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

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

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

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

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

$$\pi(\eta, \zeta \mid x, y, z) \propto \pi(\eta) \cdot \pi(\zeta) \cdot \prod_i g_\eta(x_i, y_i) \cdot h_\zeta(y_i, z_i)$$

$$= \left\{ \pi(\eta) \prod_i g_\eta(x_i, y_i) \right\} \cdot \left\{ \pi(\zeta) \prod_i h_\zeta(y_i, z_i) \right\}$$

$$= \pi(\eta \mid x, y) \cdot \pi(\zeta \mid y, z).$$

Applying Theorem 2.4(ii) we see that $\eta \perp \zeta \mid X, Y, Z$, 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 Exponential Families and Contingency Tables

For much of the rest of the course we will be dealing with collections of random variables $X_V \equiv \{X_v : v \in V\}$, indexed by a set $V = \{1, \ldots, p\}$. Each $X_v$ takes values in the set $X_v$. For a subset of the variables $A \subseteq V$, we write $X_A$ to denote $\{X_v : v \in A\}$.

3.1 Exponential Families

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

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

If $\Theta$ is a non-empty open set then the family is said to be regular. The functions $\phi_i$ are the sufficient statistics, and the components $\theta_i$ are called the canonical parameters (or natural parameters). We can replace the sum with an inner product of vectors $\theta = (\theta_i)$ and $\phi = (\phi_i(x))$:

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

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

$$A(\theta) = \log \int \exp\{ \langle \theta, \phi(x) \rangle - C(x) \} \, dx.$$

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

Lemma 3.1. We have

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

Consequently $A(\theta)$ (and $-\log p(x; \theta)$) are convex in $\theta$. In addition, the map $\mu(\theta) : \theta \mapsto \nabla A(\theta)$ is bijective, and called the mean function.

Proof. For the first part,

$$e^{A(\theta)} \frac{\partial}{\partial \theta_i} A(\theta) = \frac{\partial}{\partial \theta_i} e^{A(\theta)} = \frac{\partial}{\partial \theta_i} \int \exp\{ \langle \theta, \phi(x) \rangle - C(x) \} \, dx$$

$$= \int \frac{\partial}{\partial \theta_i} \exp\{ \langle \theta, \phi(x) \rangle - C(x) \} \, dx$$

$$= \int \phi_i(x) \exp\{ \langle \theta, \phi(x) \rangle - C(x) \} \, dx$$

$$= e^{A(\theta)} \int \phi_i(x) \exp\{ \langle \theta, \phi(x) \rangle - A(\theta) - C(x) \} \, dx$$

$$= e^{A(\theta)} \mathbb{E}_\theta \phi_i(X).$$

The result for the Hessian follows similarly. The convexity of $-\log p(x; \theta) = A(\theta) - \langle \theta, \phi(x) \rangle$ is now immediate from the fact that its Hessian is a positive definite matrix. That $\mu(\theta)$ is bijective requires strict convexity; i.e. that the Hessian is positive definite. This follows from a slight extension to the above (see the book by Wainwright and Jordan, Proposition 3.1).
The property of convexity plays an important role in the computational advantages of exponential families. Convex functions are easy to work with for the purposes of optimization: in particular, they do not contain multiple local minima.

**Example 3.2.** Let $X \sim \text{Poisson}(\lambda)$. We have

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

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.

### 3.2 Empirical Moment Matching

To find the maximum likelihood estimate in an exponential family, we maximize the log-likelihood (ignoring $C$, since it is constant in $\theta$)

$$l(\theta; X^{(1)}, \ldots, X^{(n)}) = \sum_{i=1}^n \phi(X^{(i)}), \theta) - nA(\theta)$$

$$n^{-1}l(\theta; X^{(1)}, \ldots, X^{(n)}) = \overline{\phi(X)} - A(\theta)$$

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 $\theta$ so that $E_{\theta} \phi(X) = \overline{\phi(X)}$: the mean of the sufficient statistics matches the empirical mean from the data.

Note also that if we differentiate just with respect to $\theta_i$, we obtain the same result for each sufficient statistic separately; hence if we update the parameters to match the moment $\overline{\phi_i(X)} = E_{\theta} \phi_i(X)$, 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 (strictly) concave and differentiable function.

### 3.3 Multivariate Gaussian Distribution

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 $\mu$ and $\Sigma$ if the joint density is

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

This is also called the multivariate normal distribution. The **concentration matrix** is $K \equiv \Sigma^{-1}$. 

12
We can rewrite this as
\[
f(x_V) = \frac{1}{(2\pi)^{p/2}} \exp \left\{ -\frac{1}{2} x_V^T K x_V + \mu^T K \mu + \frac{1}{2} \log |K| \right\}, \quad x_V \in \mathbb{R}^p.
\]
Noting that \( x_V^T K x_V = \sum_{i,j} k_{ij} x_i x_j \) we see that this is an exponential family with canonical parameters \( K \) and \( \eta \equiv -K\mu \), and sufficient statistics \( \phi(x_V) = (x_V, \frac{1}{2} x_V x_V^T) \).

We then obtain that
\[
2A(\theta) = 2A(K, \eta) = \eta^T K^{-1} \eta + \log |K|,
\]
which by differentiating gives
\[
\nabla_\eta A(\theta) = K^{-1} \eta = \mu \quad \text{and} \quad 2\nabla_K A(\theta) = K^{-T} \eta \eta^T K^{-1} + K^{-1} = \Sigma + \mu \mu^T.
\]

Proposition 3.3. Let \( X_V \) have a multivariate Gaussian distribution with concentration matrix \( K = \Sigma^{-1} \). Then \( X_i \perp \perp X_j \mid X_V \setminus \{i,j\} \) if and only if \( k_{ij} = 0 \), where \( k_{ij} \) is the corresponding entry in the concentration matrix.

Proof. The log-density is
\[
\log f(x_V) = -\frac{1}{2} (x_V - \mu)^T K (x_V - \mu) + \text{const}
\]
where the constant term does not depend upon \( x_V \). It is clear that the only term involving both \( x_i \) and \( x_j \) is \(-k_{ij}(x_i - \mu_i)(x_j - \mu_j)\). Hence, \( k_{ij} = 0 \) if and only if the log-density has separate terms for each of \( x_i \) and \( x_j \).

We will return to the multivariate Gaussian distribution in Chapter 5.

3.4 Contingency Tables

In this section we will assume that our variables \( X_v \) are discrete with a finite set of levels \( X_v \equiv \{1, \ldots, d_v\} \). 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:
\[
n(x_V) = \sum_{i=1}^n \mathbb{1}\{X_1^{(i)} = x_1, \ldots, X_p^{(i)} = x_p\},
\]
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
\[
l(p; n) = \sum_{x_V} n(x_V) \log p(x_V), \quad p(x_V) \geq 0, \quad \sum_{x_V} p(x_V) = 1
\]

\[\text{Here I use matrix calculus, see for example, “The Matrix Cookbook”, available here: https://www.math.uwaterloo.ca/~hwolkowi/matrixcookbook.pdf}\]
Letting $0_V$ mean the vector of zeros, we can rewrite this as

$$l(p; n) = \sum_{x \neq 0_V} n(x_V) \log p(x_V)/p(0_V) + n \log p(0_V),$$

We immediately obtain that the multinomial distribution is an exponential family with sufficient statistics given by the counts $n(x_V)$, and canonical parameters given by the ratios of log-probabilities. The cumulant function is $-\log p(0_V)$, but it should be written as a function of the canonical parameters; you can check that this gives

$$-\log p(0_V) = \log \left(1 + \sum_{x \neq 0_V} e^{\theta(x_V)}\right),$$

for $\theta(x_V) = \log p(x_V)/p(0_V)$, which is convex. Note that canonical parameters are only unique up to linear transformations; in particular, we could have used a different reference value. We will use an alternative parameterization below.

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 3.4.** Consider the death penalty data again:

<table>
<thead>
<tr>
<th>Victim</th>
<th>White</th>
<th>Black</th>
</tr>
</thead>
<tbody>
<tr>
<td>Defender</td>
<td>White</td>
<td>Black</td>
</tr>
<tr>
<td>Yes</td>
<td>53</td>
<td>11</td>
</tr>
<tr>
<td>No</td>
<td>414</td>
<td>37</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Victim</th>
<th>Black</th>
</tr>
</thead>
<tbody>
<tr>
<td>Defender</td>
<td>White</td>
</tr>
<tr>
<td>Yes</td>
<td>0</td>
</tr>
<tr>
<td>No</td>
<td>16</td>
</tr>
</tbody>
</table>

The marginal table has

<table>
<thead>
<tr>
<th>Defendant</th>
<th>White</th>
<th>Black</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yes</td>
<td>53</td>
<td>15</td>
</tr>
<tr>
<td>No</td>
<td>430</td>
<td>176</td>
</tr>
</tbody>
</table>

### 3.5 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 $k$ binary variables, then the contingency table will have $2^k$ cells. Even for $k = 10$ we will have over a thousand possibilities, and for $k = 50$ there are too many to cells to store in a computer’s memory.

Conditional independence can help, however; suppose that $X_A \independent X_B \mid X_S$ 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 $2^s + 2^{s+s} + 2^{b+s} = 2^s(1 + 2^s + 2^b)$ cells instead of $2^{s+a+b}$. This is a considerable saving if
s is small. With respect to calculations, if we want to find \( P(X_v = 1) \) and \( v \in A \), 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 other conditional independences then one might imagine that further computational savings are possible: indeed this is correct, and is one of the main ideas behind graphical models.

### 3.6 Log-linear models

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

\[
\log p(x_V) = \sum_{A \subseteq V} \lambda_A(x_A) = \lambda_\emptyset + \lambda_1(x_1) + \cdots + \lambda_V(x_V),
\]

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, \( d_v = 2 \), \( X_v = \{1, 2\} \)), 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 3.5.** Consider a \( 2 \times 2 \) table with probabilities \( \pi_{ij} = P(X = i, Y = j) \). The log-linear parametrization has

\[
\begin{align*}
\log \pi_{11} &= \lambda_\emptyset \\
\log \pi_{12} &= \lambda_\emptyset + \lambda_Y \\
\log \pi_{21} &= \lambda_\emptyset + \lambda_X \\
\log \pi_{22} &= \lambda_\emptyset + \lambda_X + \lambda_Y + \lambda_{XY}.
\end{align*}
\]

From this we can deduce that

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

The quantity \( \exp \lambda_{XY} \) is called the *odds ratio* between \( X \) and \( Y \), and is a fundamental quantity in statistical inference.

Multinomial models can be fitted as Poisson GLMs using the following fact:

**Proposition 3.6.** Let \( X_i \sim \text{Poisson}(\mu_i) \) independently, and let \( N = \sum_{i=1}^{k} X_i \). Then,

\[
N \sim \text{Poisson}\left(\sum_{i} \mu_i\right)
\]

\( (X_1, \ldots, X_k) \mid N = n \sim \text{Multinom}(n, (\pi_1, \ldots, \pi_k)) \),

where \( \pi_i = \mu_i / \sum_i \mu_i \).

### 3.7 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 3.7. Let \( p > 0 \) be a discrete distribution on \( X_V \) with associated log-linear parameters \( \lambda_C, C \subseteq V \). The conditional independence \( X_a \perp X_b \mid X_V \setminus \{a,b\} \) 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 \( X_A \perp X_B \mid X_C \), 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

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

where \( \lambda_{BC} \) By equating terms we can see that

\[
\lambda_W(x_W) = \lambda_W^C(x_W) \quad \text{for any } W \subseteq A \cup C \text{ with } W \cap A \neq \emptyset \\
\lambda_W(x_W) = \lambda_W^B(x_W) \quad \text{for any } W \subseteq B \cup C \text{ with } W \cap B \neq \emptyset \\
\lambda_W(x_W) = \lambda_W^C(x_W) + \lambda_W^B(x_W) - \lambda_W^C(x_W) \quad \text{for any } W \subseteq C.
\]

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

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

<table>
<thead>
<tr>
<th>DeathPen</th>
<th>Defendant</th>
<th>Victim</th>
<th>freq</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yes</td>
<td>White</td>
<td>White</td>
<td>53</td>
</tr>
<tr>
<td>No</td>
<td>White</td>
<td>White</td>
<td>414</td>
</tr>
<tr>
<td>Yes</td>
<td>Black</td>
<td>White</td>
<td>11</td>
</tr>
<tr>
<td>No</td>
<td>Black</td>
<td>White</td>
<td>37</td>
</tr>
<tr>
<td>Yes</td>
<td>White</td>
<td>Black</td>
<td>0</td>
</tr>
<tr>
<td>No</td>
<td>White</td>
<td>Black</td>
<td>16</td>
</tr>
<tr>
<td>Yes</td>
<td>Black</td>
<td>Black</td>
<td>4</td>
</tr>
<tr>
<td>No</td>
<td>Black</td>
<td>Black</td>
<td>139</td>
</tr>
</tbody>
</table>
```

We can fit log-linear models using the `glm()` command with a Poisson response. This gives the model \( \text{DeathPen} \perp \perp \text{Defendant} \mid \text{Victim} \).
\begin{verbatim}
> 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

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>58.035</td>
<td>408.965</td>
<td>5.965</td>
<td>42.035</td>
<td>0.403</td>
<td>15.597</td>
<td>3.597</td>
<td>139.403</td>
</tr>
</tbody>
</table>

Compare these to the actual counts: a goodness of fit test can be performed by using Pearson’s \( \chi^2 \) test or (almost equivalently) by looking at the residual deviance of the model.
\end{verbatim}
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 \( G \) 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 \( \text{bd}_G(i) \).

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 \( G \) (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 \( G \) to be the graph with vertices \( W \), and all edges from \( G \) 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
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 $G$ 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 $G$ if

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

In other words, whenever an edge is missing in $G$ 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 $G$ if for any disjoint sets $A, B, S$

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

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

19
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 \( j \). \qed

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

\[
p(x_V) = \prod_{C \in \mathcal{C}(\mathcal{G})} \psi_C(x_C)
\]

for some functions \( \psi_C \). The functions \( \psi_C \) 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 \( \mathcal{G} \).

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

\[
\prod_{C \in \mathcal{C}} \psi_C(x_C) = \prod_{C \in \mathcal{C}_A} \psi_C(x_C) \cdot \prod_{C \in \mathcal{C}_B} \psi_C(x_C) = f(x_{\tilde{A}},xs) \cdot f(x_{\tilde{B}},xs).
\]
and hence $X_A \perp X_B \mid X_S$. Then applying property 2 of Theorem \ref{th:property2} gives $X_A \perp X_B \mid X_S$.

**Theorem 4.11** (Hammersley-Clifford Theorem). If $p(x) > 0$ 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 $P(X_3 = 1) = \frac{1}{2}$, and $P(X_1 = X_2 = X_4) = 1$. Then $X_4 \perp X_1 \mid X_2, X_3$ and $X_4 \perp X_2 \mid X_1, X_3$, but $X_4 \not\perp \mid X_1, X_2 \mid X_3$.

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 $C$ 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.

**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 \( G_{\{1,2,3\}} \) and \( G_{\{3,4,5\}} \) are complete (and therefore decomposable by definition).

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

Definition 4.16. Let \( C \) be a collection of subsets of \( V \). We say that the sets \( C \) satisfy the running intersection property if there is an ordering \( C_1, \ldots, C_k \), such that for every \( j = 2, \ldots, k \) 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 \( C_1, \ldots, C_k \) satisfy the running intersection property, then there is a graph whose cliques are precisely (the inclusion maximal elements of) \( C = \{C_1, \ldots, C_k\} \).

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

Definition 4.19. Let \( G \) be an undirected graph. A cycle is a sequence of vertices \( \langle v_1, \ldots, v_k \rangle \) 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 \( \geq 4 \), 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.

Theorem 4.20. Let \( G \) be an undirected graph. The following are equivalent:

(i) \( G \) is decomposable;
Figure 4: Two undirected graphs: (a) is chordal, (b) is not.

(ii) \( G \) is triangulated;

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

(iv) the cliques of \( G \) satisfy the running intersection property.

Proof. (i) \( \implies \) (ii). We proceed by induction on \( p \), the number of vertices in the graph. Let \( G \) 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 \( G_{A\cup S} \) and \( G_{B\cup S} \) 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 \( \pi_1 \) from \( a \) to \( b \) via \( s_1 \in S \), and another path \( \pi_2 \) 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 \( \geq 4 \).

(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 \( \hat{A} \) be the connected component of \( a \) in \( G_{V\setminus S} \), and \( \hat{B} \) the rest. Then apply the result by induction to the strictly smaller graphs \( G_{\hat{A}\cup S} \) and \( G_{\hat{B}\cup S} \). 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 \( H_{k-1} = C_1 \cup \cdots \cup C_{k-1}, S_k = C_k \cap H_{k-1}, \) and \( R_k = C_k \setminus S_k \); we claim that \((H_{k-1} \setminus S_k, S_k, R_k)\) is a proper decomposition, and that the graph \( G_{H_{k-1}} \) has \( k - 1 \) cliques that also satisfy the running intersection property.

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

Proof. If \( G \) is triangulated then so are any induced subgraphs of \( G \).

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 ??). This turns out to be extremely useful for computational reasons.

4.5 Separator Sets

Let \( G \) be a decomposable graph, and let \( C_1, \ldots, C_k \) be an ordering of the cliques which satisfies running intersection. Define the \( j \)th 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 \( G \) be a graph with decomposition \((A, S, B)\), and let \( p \) be a distribution; then \( p \) factorizes with respect to \( G \) if and only if its marginals \( p(x_{A \cup S}) \) and \( p(x_{B \cup S}) \) factorize according to \( G_{A \cup S} \) and \( G_{B \cup S} \) respectively, and

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

Proof. Note that, as observed in the proof of Theorem 4.10, every clique in \( G_{A \cup S} \) is a (subset of a) clique in \( G \). Hence if \( \ref{2} \) and the factorizations with respect to those subgraphs hold, then we can see that \( p \) factorizes with respect to \( G \).

Now suppose that \( p \) factorizes with respect to \( G \), and note that this implies that \( p \) obeys the global Markov property with respect to \( G \). From the decomposition, we have \( A \perp_S B \mid S \) in \( G \), and so by the global Markov property applied to \( G \) we obtain the independence \( X_A \perp X_B \mid X_S[p] \); this gives us the equation \( \ref{2} \) by Theorem 2.4. Since this is a decomposition, all cliques of \( G \) are contained either within \( A \cup S \) or \( B \cup S \) (or both). Let \( A \) be the cliques contained in \( A \cup S \), and \( B \) the rest.

Then \( p(x_V) = \prod_{C \in A} \psi_C(x_C) \cdot \prod_{C \in B} \psi_C(x_C) = h(x_A, x_S) \cdot k(x_B, x_S) \). Substituting \( p(x_V) \) into \( \ref{2} \) and integrating both sides with respect to \( x_A \) gives

\[
p(x_S) \cdot k(x_B, x_S) \int h(x_A, x_S) \, dx_A = p(x_S) \cdot p(x_B, x_S)
\]

\[
p(x_S) \cdot k(x_B, x_S) \cdot \hat{h}(x_S) = p(x_S) \cdot p(x_B, x_S),
\]

which shows that \( p(x_B, x_S) = \psi_S'(x_S) \prod_{C \in B} \psi_C(x_C) \) as required. \( \square \)

Theorem 4.24. Let \( G \) be a decomposable graph with cliques \( C_1, \ldots, C_k \). Then \( p \) factorizes with respect to \( G \) 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, the quantities \( p(x_{C_i \setminus S_i} \mid x_{S_i}) \) are variation independent (i.e. they may jointly take any set of values that would be valid individually), 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. \hfill \Box

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

$$l(p; n) = \sum_{x_{V}} n(x_{V}) \log p(x_{V})$$

$$= \sum_{x_{V}} n(x_{V}) \sum_{i=1}^{k} \log p(x_{C_i \setminus S_i} \mid x_{S_i})$$

$$= \sum_{i=1}^{k} \sum_{x_{C_i}} n(x_{C_i}) \log p(x_{C_i \setminus S_i} \mid x_{S_i}),$$

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

$$\hat{p}(x_{C_i \setminus S_i} \mid x_{S_i}) = \frac{n(x_{C_i})}{n(x_{S_i})},$$

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, from our discussion about exponential families 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

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

25
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 $G$ (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:

$$
p^{(t+1)}(x_V) = p^{(t)}(x_V) \cdot \frac{p(x_C)}{p^{(t)}(x_C)}
= p^{(t)}(x_{V \setminus C} \mid x_C) \cdot p(x_C).
$$

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

**Algorithm 1** Iterative Proportional Fitting (IPF) algorithm.

```python
function IPF(collection of consistent margins $q(x_{C_i})$ for sets $C_1, \ldots, C_k$)
    set $p(x_V)$ to uniform distribution;
    while $\max_i \max_{x_{C_i}} |p(x_{C_i}) - q(x_{C_i})| > \text{tol}$ do
        for $i$ in 1, $\ldots, k$ do
            update $p(x_V)$ to $p(x_{V \setminus C_i} \mid x_{C_i}) \cdot q(x_{C_i})$;
        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 (3) 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 $G$: this can be seen easily by considering the factorization. Performing each step increases the likelihood, and since the log-likelihood is strictly concave, this sort of co-ordinate based iterative updating scheme will converge to the global maximum.

**Example 4.26.** Consider the 4-cycle in Figure 5(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’).

![Figure 5: A non-decomposable graph, and a possible triangulation of it.](image)
The marginals over the cliques are:

\[
\begin{array}{c|c|c}
X_1 & X_2 & X_3 & X_4 & \text{count} \\
0 & 0 & 0 & 0 & 5 \\
1 & 0 & 0 & 0 & 10 \\
0 & 1 & 0 & 0 & 20 \\
1 & 1 & 0 & 0 & 1 \\
0 & 0 & 1 & 0 & 0 \\
1 & 0 & 1 & 0 & 3 \\
0 & 1 & 1 & 0 & 4 \\
1 & 1 & 1 & 0 & 0 \\
0 & 0 & 0 & 1 & 24 \\
1 & 0 & 0 & 1 & 0 \\
0 & 1 & 0 & 1 & 9 \\
1 & 1 & 0 & 1 & 3 \\
0 & 0 & 1 & 1 & 1 \\
1 & 0 & 1 & 1 & 2 \\
0 & 1 & 1 & 1 & 4 \\
1 & 1 & 1 & 1 & 10 \\
\end{array}
\]

<table>
<thead>
<tr>
<th>step 0</th>
<th>step 1</th>
<th>step 2</th>
<th>step 3</th>
<th>step 4</th>
<th>( \hat{n} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>6</td>
<td>7.5</td>
<td>13</td>
<td>13</td>
<td>12.59</td>
<td>12.6</td>
</tr>
<tr>
<td>6</td>
<td>3.75</td>
<td>6.5</td>
<td>6.5</td>
<td>6.97</td>
<td>6.95</td>
</tr>
<tr>
<td>6</td>
<td>9.25</td>
<td>11.97</td>
<td>11.97</td>
<td>11.59</td>
<td>11.58</td>
</tr>
<tr>
<td>6</td>
<td>3.5</td>
<td>4.53</td>
<td>4.53</td>
<td>4.86</td>
<td>4.87</td>
</tr>
<tr>
<td>6</td>
<td>7.5</td>
<td>2</td>
<td>1.17</td>
<td>1.13</td>
<td>1.13</td>
</tr>
<tr>
<td>6</td>
<td>3.75</td>
<td>1</td>
<td>0.58</td>
<td>0.63</td>
<td>0.63</td>
</tr>
<tr>
<td>6</td>
<td>9.25</td>
<td>6.53</td>
<td>3.81</td>
<td>3.69</td>
<td>3.69</td>
</tr>
<tr>
<td>6</td>
<td>3.5</td>
<td>2.47</td>
<td>1.44</td>
<td>1.55</td>
<td>1.55</td>
</tr>
<tr>
<td>6</td>
<td>7.5</td>
<td>13</td>
<td>13</td>
<td>13.33</td>
<td>13.35</td>
</tr>
<tr>
<td>6</td>
<td>3.75</td>
<td>6.5</td>
<td>6.5</td>
<td>6.11</td>
<td>6.1</td>
</tr>
<tr>
<td>6</td>
<td>9.25</td>
<td>11.97</td>
<td>11.97</td>
<td>12.28</td>
<td>12.27</td>
</tr>
<tr>
<td>6</td>
<td>3.5</td>
<td>4.53</td>
<td>4.53</td>
<td>4.26</td>
<td>4.28</td>
</tr>
<tr>
<td>6</td>
<td>7.5</td>
<td>2</td>
<td>2.83</td>
<td>2.91</td>
<td>2.91</td>
</tr>
<tr>
<td>6</td>
<td>3.75</td>
<td>1</td>
<td>1.42</td>
<td>1.33</td>
<td>1.33</td>
</tr>
<tr>
<td>6</td>
<td>3.5</td>
<td>2.47</td>
<td>3.5</td>
<td>3.29</td>
<td>3.3</td>
</tr>
</tbody>
</table>

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 marginals.
5 Gaussian Graphical Models

Recall that $X_V$ has a multivariate Gaussian distribution with parameters $\mu$ and $\Sigma$ if the joint density is

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

**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 $X_U \sim N_q(\mu_U, \Sigma_U).$

*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 $\Sigma$ 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. \hfill $\square$

For simplicity of notation, we will assume throughout that $\mu = 0$. Note that the dependence structure is entirely determined by $\Sigma$, and $\mu$ is an orthogonal parameter to $\Sigma$.

### 5.1 Gaussian Graphical Models

We only consider cases in which $\Sigma$ 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 $\Sigma$ is Markov with respect to a graph, without ambiguity.

Recall that $X_A \perp \perp X_B$ if and only if $\Sigma_{AB} = 0$, and note that a corollary of this is that $X \perp \perp Y$ and $X \perp \perp Z$ does imply $X \perp \perp Y, Z$ for jointly Gaussian random variables.

**Theorem 5.2.** Let $X_V \sim N_p(\mu, \Sigma)$ for positive definite $\Sigma$, 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 $a \neq b$ in $\mathcal{G}$.

*Proof.* This follows immediately from Proposition 3.3. \hfill $\square$

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 $V$ (i.e. it has the same dimension as $M$) whose $A,A$-entries are $M$ and with zeroes elsewhere.
For example, if \(|V| = 3\) then

\[
M = \begin{pmatrix} a & b \\ b & c \end{pmatrix}, \quad \{M\}_{12,12} = \begin{pmatrix} a & b & 0 \\ b & c & 0 \\ 0 & 0 & 0 \end{pmatrix},
\]

where 12 is used as an abbreviation for \(\{1, 2\}\) in the subscript.

**Lemma 5.3.** Let \(G\) be a graph with decomposition \((A, S, B)\), and \(X_V \sim N_\mu(0, \Sigma)\). Then \(p(x_V)\) is Markov with respect to \(G\) if and only if

\[
\Sigma^{-1} = \{(\Sigma_{A\cup S, A\cup S})^{-1}\}_{A\cup S, A\cup S} + \{(\Sigma_{B\cup S, B\cup S})^{-1}\}_{B\cup S, B\cup S} - \{(\Sigma_{S,S})^{-1}\}_{S,S},
\]

and \(\Sigma_{A\cup S, A\cup S}\) and \(\Sigma_{B\cup S, B\cup S}\) are Markov with respect to \(G_{A\cup S}\) and \(G_{B\cup S}\) 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 \(G_{A\cup S}\) and \(G_{B\cup S}\) 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:

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

which is a quadratic polynomial in the variables \(x_v\). By, comparing coefficients for each term we obtain that

\[
\Sigma^{-1} = \{(\Sigma_{A\cup S, A\cup S})^{-1}\}_{A\cup S, A\cup S} + \{(\Sigma_{B\cup S, B\cup S})^{-1}\}_{B\cup S, B\cup S} - \{(\Sigma_{S,S})^{-1}\}_{S,S}.
\]

This gives the result. \(\square\)

Applying the previous result to a decomposable graph repeatedly we see that \(X_V\) is Markov with respect to \(G\) if and only if

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

5.2 **Maximum Likelihood Estimation**

Let \(X^{(1)}_V, \ldots, X^{(n)}_V\) be i.i.d. \(N_\mu(0, \Sigma)\); then from Section 3 the sufficient statistic for \(\Sigma\) is the sample covariance matrix:

\[
W \equiv \frac{1}{n} \sum_{i=1}^{n} X^{(i)}_V X^{(i)T}_V.
\]

In addition, \(\hat{\Sigma} = W\) is also the MLE for \(\Sigma\) under the unrestricted model (i.e. when all edges are present in the graph). Let \(\hat{\Sigma}^{G}\) denote the MLE for \(\Sigma\) under the restriction that the distribution satisfies the Markov property for \(G\), and \(\hat{K}^{G}\) its inverse.

Recall that if \(i \not\sim j\) then \(k_{ij} = 0\), so the sufficient statistics for a graph \(G\) reduce to the entries in \(W\) that correspond to edges in the graph. The MLE involves picking \(\hat{K}\) such that:

\[
\hat{k}_{ij} = 0 \quad \text{whenever } i \not\sim j
\]

\[
\hat{\sigma}_{ij} = W_{ij} \quad i \sim j;
\]

29
For a decomposable graph $\mathcal{G}$ with cliques $C_1, \ldots, C_k$ this means that the MLE can be written in the form
\[
\left(\Sigma^{\mathcal{G}}\right)^{-1} = \sum_{i=1}^{k} \left\{(W_{C_i,C_i})^{-1}\right\}_{C_i,C_i} - \sum_{i=2}^{k} \left\{(W_{S_i,S_i})^{-1}\right\}_{S_i,S_i}.
\]
This matches the sufficient statistics so that $\Sigma_{C_i,C_i} = W_{C_i,C_i}$ for each $i$.

## 5.3 Data Examples

**Example 5.4.** 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 $10^3$)

<table>
<thead>
<tr>
<th></th>
<th>mechanics</th>
<th>vectors</th>
<th>algebra</th>
<th>analysis</th>
<th>statistics</th>
</tr>
</thead>
<tbody>
<tr>
<td>mechanics</td>
<td>5.24</td>
<td>-2.43</td>
<td>-2.72</td>
<td>0.01</td>
<td>-0.15</td>
</tr>
<tr>
<td>vectors</td>
<td>-2.43</td>
<td>10.42</td>
<td>-4.72</td>
<td>-0.79</td>
<td>-0.16</td>
</tr>
<tr>
<td>algebra</td>
<td>-2.72</td>
<td>-4.72</td>
<td>26.94</td>
<td>-7.05</td>
<td>-4.70</td>
</tr>
<tr>
<td>analysis</td>
<td>0.01</td>
<td>-0.79</td>
<td>-7.05</td>
<td>9.88</td>
<td>-2.02</td>
</tr>
<tr>
<td>statistics</td>
<td>-0.15</td>
<td>-0.16</td>
<td>-4.70</td>
<td>-2.02</td>
<td>6.45</td>
</tr>
</tbody>
</table>

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 6 gives an excellent fit (see Examples 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.
6 Directed Graphical Models

Undirected graphs represent symmetrical relationships between random variables: the vertices in an undirected graph are not typically ordered. However, 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 are ‘prior’ to the outcome (in a temporal sense or otherwise).

A further limitation of undirected graphs is that they are only able to represent conditional independences; they can only represent marginal independences if the relevant variables are in disconnected components. In practice, marginal independences arise very naturally if we have independent inputs to a system, and an output that is a (random) function of the inputs.

An example is given in Figure 7. Suppose that within the general population academic and sporting abilities are uncorrelated, but that either may be sufficient to gain admission to the elite Yarvard University. Then—as we will see—conditional upon admission to Yarvard we would expect academic and sporting abilities to be negatively associated.

Such situations are naturally represented by a directed graph.

Definition 6.1. A directed graph $G$ 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 $v \rightarrow w$, and say that $v$ is a parent of $w$, and conversely $w$ a child of $v$. Examples are given in Figures 7 and 8(a).

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

For example, in the graph in Figure 8(a), 1 and 2 are parents of 3. There is a path $1 \rightarrow 3 \leftarrow 2 \rightarrow 5$, and there is a directed path $1 \rightarrow 3 \rightarrow 5$ from 1 to 5.

The set of parents of $w$ is $\text{pa}_G(w)$, and the set of children of $v$ is $\text{ch}_G(v)$.

Definition 6.2. A graph contains a directed cycle if there is a directed path from $v$ to $w$ together with an edge $w \rightarrow v$. 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 $i \in \text{pa}_G(j)$ 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.

We say that $a$ is an ancestor of $v$ if either $a = v$, or there is a directed path $a \rightarrow \cdots \rightarrow v$. The set of ancestors of $v$ is denoted by $\text{an}_G(v)$. The ancestors of 4 in the DAG in Figure 8(a) are $\text{an}_G(4) = \{2, 4\}$. The descendants of $v$ are defined analogously and denoted $\text{de}_G(v)$; the non-descendants of $v$ are $\text{nd}_G(v) \equiv V \setminus \text{de}_G(v)$. The non-descendants of 4 in Figure 8(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 $G$ be a directed acyclic graph with vertices $V$. We say that a probability distribution $p(x_V)$ factorizes with respect to $G$ if

$$p(x_V) = \prod_{v \in V} p(x_v \mid x_{\text{pa}_G(v)}), \quad x_V \in X_V.$$

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

$$p(x_v \mid x_{\text{pre}_<(v)}) = p(x_v \mid x_{\text{pa}_G(v)}), \quad v \in V$$

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

$$X_v \independent X_{\text{pre}_<(v) \setminus \text{pa}_G(v)} \mid X_{\text{pa}_G(v)} [p]. \quad (4)$$

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 (4) implies

$$X_v \independent X_{\text{nd}_G(v) \setminus \text{pa}_G(v)} \mid X_{\text{pa}_G(v)} [p]. \quad (5)$$

Distributions are said to obey the local Markov property with respect to $G$ if they satisfy (5) for every $v \in V$.

For example, the local Markov property applied to each vertex in Figure 8(a) would require that

$$X_1 \independent X_2, X_4 \quad X_2 \independent X_1 \quad X_3 \independent X_4 \mid X_1, X_2$$

$$X_4 \independent X_1, X_3 \mid X_2 \quad X_5 \independent X_1, X_2 \mid X_3, X_4$$

32
Figure 8: (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 8(a); however $\{1, 3\}$ is not, because $\{2\}$ is an ancestor of $\{3\}$ but it is 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 3.

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 $\text{ang}(A \cup B \cup C)$.

**Definition 6.5.** A $v$-structure is a triple $i \rightarrow k \leftarrow j$ 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 8.

**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 $\mathcal{G}^m$. 

33
Using this proposition, we see that the DAG in Figure 8(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 8(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 $G$ if whenever $A$ and $B$ are separated by $C$ in $(G_{an(A \cup B \cup C)})^m$ we have $X_A \perp \perp X_B \mid X_C \mid p$.

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 $G$. We state the result formally here, but the proof is not given in this course.

**Theorem 6.8 (Completeness of global Markov property.).** Let $G$ be a DAG. There exists a probability distribution $p$ such that $X_A \perp \perp X_B \mid X_C \mid p$ if and only if $A \perp \perp B \mid C$ in $(G_{an(A \cup B \cup C)})^m$. 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 $G$ be a DAG and $p$ a probability distribution. Then the following are equivalent:

1. $p$ factorizes according to $G$;
2. $p$ is globally Markov with respect to $G$;
3. $p$ is locally Markov with respect to $G$.

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 $W = an_G(A \cup B \cup C)$, and suppose that there is a separation between $A$ and $B$ given $C$ in $(G_{W})^m$. The distribution $p(x_W)$ can be written as

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

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

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

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

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

as required. \qed
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

\[
l(p; n) = \sum_{x_V} n(x_V) \log p(x_V)
\]

\[
= \sum_{x_V} n(x_V) \sum_{v \in V} \log p(x_v \mid x_{pa(v)})
\]

\[
= \sum_{v \in V} \sum_{x_v \in V, x_{pa(v)}} n(x_v, x_{pa(v)}) \log p(x_v \mid x_{pa(v)})
\]

\[
= \sum_{v \in V} \sum_{x_v \in x_{pa(v)}} n(x_v, x_{pa(v)}) \log p(x_v \mid x_{pa(v)}),
\]

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

\[
\hat{p}(x_v \mid x_{pa(v)}) = \frac{n(x_v, x_{pa(v)})}{n(x_{pa(v)})};
\]

hence

\[
\hat{p}(x_V) = \prod_{v \in V} \hat{p}(x_v \mid x_{pa(v)}) = \prod_{v \in V} \frac{n(x_v, x_{pa(v)})}{n(x_{pa(v)})}.
\]

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 \( \theta_v \) for each conditional distribution \( p(x_v \mid x_{pa(v)}; \theta_v) \), then we can perform our inference on each \( \theta_v \) separately.

Formally: the MLE for \( \theta \) satisfies

\[
p(x_V; \hat{\theta}) = \prod_{v \in V} p(x_v \mid x_{pa(v)}; \hat{\theta}_v),
\]

\( x_V \in X_V \).

In addition, if we have independent priors \( \pi(\theta) = \prod_{v} \pi(\theta_v) \), then

\[
\pi(\theta \mid x_V) \propto \pi(\theta) \cdot p(x_V \mid \theta)
\]

\[
= \prod_v \pi(\theta_v) \cdot p(x_v \mid x_{pa(v)}, \theta_v),
\]

which factorizes into separate functions for each \( \theta_v \), showing that the \( \theta_v \) are independent conditional on \( X_V \). Hence

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

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

In other words, the data from \( X_v, X_{pa(v)} \) are sufficient for each \( \theta_v \). This means that if no vertex has many parents, even very large graphs represent manageable models. For 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 Examples Sheet 3).
Figure 9: (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.

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 \( a \rightarrow b \) 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 9 (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 \( G \) and \( G' \) are Markov equivalent if any \( p \) which is Markov with respect to \( G \) is also Markov with respect to \( G' \), 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 \( G = (V,D) \), define the skeleton of \( G \) as the undirected graph \( \text{skel}(G) = (V,E) \), 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 \( G \).

For example, the skeleton of the graphs in Figures 9(a)–(d) is the graph in Figure 9(e).

**Lemma 6.10.** Let \( G \) and \( G' \) be graphs with different skeletons. Then \( G \) and \( G' \) are not Markov equivalent.

**Proof.** Suppose without loss of generality that \( i \rightarrow j \) in \( G \) but that \( i \not\rightarrow j \) in \( G' \). 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 \( G \) is clearly satisfied, since each variable is independent of its non-descendants given its parents. For \( G' \), 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 $X_i \perp X_j \mid X_{C \setminus \{c\}}$. Repeating this we end up with a requirement that $X_i \perp X_j$, which does not hold by construction. Hence $p$ is not Markov with respect to $G'$, and the graphs are not Markov equivalent.

**Theorem 6.11.** Directed graphs $G$ and $G'$ 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 $G$ and $G'$ have different skeletons then the induced models are different by the previous Lemma. Otherwise, suppose that $a \rightarrow c \leftarrow b$ is a $v$-structure in $G$ but not in $G'$. 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 $G$.

In $G'$ there is no $v$-structure, so either $a \rightarrow c \rightarrow b$, $a \leftarrow c \rightarrow b$, or $a \leftarrow c \leftarrow b$. In particular, either $a$ or $b$ is a child of $c$. Now let $A = \text{ang}(\{a, b, c\})$; we claim that there is no $d \in A$ such that $a \rightarrow d \leftarrow b$. 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 $(G'_A)^m$, there is no edge between $a$ and $b$, so $a \perp b \mid A \setminus \{a,b\}$ in $(G'_A)^m$. 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 $G'$ 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 9(e) represents the same model as each of the directed graphs in Figures 9(a)–(c), so clearly in some cases this occurs.

However the graph in Figure 9(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 $G$ is a DAG, then $p$ being Markov with respect to $G$ implies that it is also Markov with respect to $G_m$.

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

The neighbours of $v$ in $G_m$ are its parents in $G$, and in the moral graph $G_m$ these are all adjacent, so there is a decomposition $(\{v\}, \text{pa}_G(v), W)$ in $G_m$, where $W = V \setminus (\{v\} \cup \text{pa}_G(v))$. 37
By Lemma 4.23, we have $X_v \perp \perp X_{\text{pa}(v)}$, and that $p(x_{V \setminus \{v\}})$ is Markov with respect to $(G^m)_{V \setminus \{v\}}$. Now, since $G$ has no v-structures, $(G^m)_{V \setminus \{v\}} = (G_{V \setminus \{v\}})^m$, so by the induction hypothesis, $p(x_{V \setminus \{v\}})$ is Markov with respect to $G_{V \setminus \{v\}}$.

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.