UNIVERSITY OF COPENHAGEN

DEPARTMENT OF MATHEMATICAL SCIENCES

LECTURE NOTES

LECTURES ON GRAPHICAL MODELS

3rd edition

STEFFEN L. LAURITZEN

DEPARTMENT OF MATHEMATICAL SCIENCES
UNIVERSITY OF COPENHAGEN
UNIVERSITETSPARKEN 5
2100 COPENHAGEN, DENMARK

2nd printing 2020.

ISBN 978-87-70787-53-6

PREFACE

These lecture notes have as their purpose to give a rigorous introduction to the notion of conditional distributions and expectations, as well as properties of conditional independence as this concept underpins the Markov theory of graphical models. A major part of these lecture notes are based on *Conditioning and Markov properties* by Anders Rønn–Nielsen and Ernst Hansen (Third edition, 2016), in particular the treatment of conditional distributions using Markov kernels in Chapter 1, but also the measure-theoretic version of conditional independence and a large number of exercises. These were again heavily using previous lecture notes of Martin Jacobsen, Søren Tolver Jensen, and Søren Feodor Nielsen. I am indebted to this tradition of openness in the Statistics group at the Department of Mathematical Sciences and in particular to Anders and Ernst for permission to use their material of which parts have been copied and pasted directly into this document. Clearly, I take full responsibility for any error that may have crept into the notes in one way or another during this process. For basic results in measure theory, the reader is referred to Hansen (2009). Much of the remaining material builds heavily on Lauritzen (1996).

Copenhagen S.L.L.
27 September 2017.

PREFACE TO THE SECOND EDITION

In this second edition the theory of conditional independence has been streamlined using conditional expectations and the construction of Bayesian networks using combination of Markov kernels has been developed. The sections on Markov properties of directed acyclic graphs has been extended to cover the case where the measures do not have a density w.r.t. a product measure, and the section of Markov equivalence correspondingly extended.

Copenhagen S.L.L.
5 September 2018

PREFACE TO THE THIRD EDITION

In this third edition, minor errors have been corrected, a few sections have been modified, and more exercises have been added.

Copenhagen S.L.L.
14 October 2019

CONTENTS

1

CONDITIONAL DISTRIBUTIONS

Let $(\mathcal{X}, \mathbb{E})$ and $(\mathcal{Y}, \mathbb{K})$ be two measurable spaces. In this chapter we shall discuss the relation between measures on the product space $(\mathcal{X} \times \mathcal{Y}, \mathbb{E} \otimes \mathbb{K})$ and measures on the two marginal spaces $(\mathcal{X}, \mathbb{E})$ and $(\mathcal{Y}, \mathbb{K})$. Following the notation in Hansen (2009) we say that the map $f : (\mathcal{X}, \mathbb{E}) \to (\mathcal{Y}, \mathbb{K})$ is $\mathbb{E} - \mathbb{K}$–measurable, if

$$f^{-1}(K) \in \mathbb{E}$$

for all $K \in \mathbb{K}$. For ease of notation, we will say that f is $\mathbb{E}$-measurable instead of $\mathbb{E} - \mathbb{B}$–measurable, when f has values in $(\mathbb{R}, \mathbb{B})$, where $\mathbb{B}$ is the Borel σ–algebra. Similarly, if X defined on $(\Omega, \mathbb{F}, P)$ is a random variable with values in $(\mathcal{X}, \mathbb{E})$, and $\mathbb{D}$ is a sub σ–algebra of $\mathbb{F}$, we will say that X is $\mathbb{D} - \mathbb{E}$–measurable, if

$$X^{-1}(E) = (X \in E) \in \mathbb{D}$$

for all $E \in \mathbb{E}$. If $(\mathcal{X}, \mathbb{E}) = (\mathbb{R}, \mathbb{B})$ this will be abbreviated as X being $\mathbb{D}$–measurable.

1.1 Markov kernels

Definition 1.1 A $(\mathcal{X}, \mathbb{E})$–*Markov Kernel* on $(\mathcal{Y}, \mathbb{K})$ is a family of probability measures $(P_x)_{x \in \mathcal{X}}$ on $(\mathcal{Y}, \mathbb{K})$ indexed by points in $\mathcal{X}$ such that the map

$$x \mapsto P_x(B)$$

is $\mathbb{E}$–measurable for every fixed $B \in \mathbb{K}$.

Theorem 1.2 *Let $(\mathcal{X}, \mathbb{E})$ and $(\mathcal{Y}, \mathbb{K})$ be measurable spaces, let ν be a σ–finite measure on $(\mathcal{Y}, \mathbb{K})$, and assume that $f : \mathcal{X} \times \mathcal{Y} \to [0, \infty]$ is $\mathbb{E} \otimes \mathbb{K}$–measurable and has the property that*

$$0 < \int f(x, y) \, \mathrm{d}\nu(y) < \infty \qquad \textit{for all } x \in \mathcal{X} \, .$$

Then $(P_x)_{x \in \mathcal{X}}$ given by

$$P_x(B) = \frac{\int_B f(x, y) \, \mathrm{d}\nu(y)}{\int f(x, y) \, \mathrm{d}\nu(y)} \qquad \textit{for all } B \in \mathbb{K}, \ x \in \mathcal{X}$$

is an $(\mathcal{X}, \mathbb{E})$–Markov Kernel on $(\mathcal{Y}, \mathbb{K})$.

Proof For each fixed set $B \in \mathbb{K}$ we need to argue that

$$x \mapsto \frac{\int 1_{\mathcal{X}\times B}(x,y) f(x,y)\, \mathrm{d}\nu(y)}{\int f(x,y)\, \mathrm{d}\nu(y)}$$

is an $\mathbb{E}$–measurable function. But this follows from Theorem 8.7 in Hansen (2009) since the ratio of measurable functions is itself measurable. □

Lemma 1.3 *If $(\mathcal{Y}, \mathbb{K})$ has an intersection–stable generating system $\mathbb{D}$, then $(P_x)_{x\in\mathcal{X}}$ is a $(\mathcal{X}, \mathbb{E})$–Markov Kernel on $(\mathcal{Y}, \mathbb{K})$ if only*

$$x \mapsto P_x(D)$$

is $\mathbb{E}$-measurable for all fixed $D \in \mathbb{D}$.

Proof Define $\mathbb{H} = \{F \in \mathbb{K} \,:\, x \mapsto P_x(F) \text{ is } \mathbb{E}\text{–measurable}\}$ and verify that $\mathbb{H}$ is a Dynkin Class. Since $\mathbb{D} \subseteq \mathbb{H}$, we have $\mathbb{H} = \mathbb{K}$. □

Next we define the inclusion map $i_x : \mathcal{Y} \to \mathcal{X} \times \mathcal{Y}$ by

$$i_x(y) = (x, y) \qquad \text{for } y \in \mathcal{Y}\,.$$

Then i_x is $\mathbb{K} - \mathbb{E} \otimes \mathbb{K}$–measurable for each fixed $x \in \mathcal{X}$. For $G \in \mathbb{E} \otimes \mathbb{K}$ define

$$G^x = \{y \in \mathcal{Y} \,:\, (x, y) \in G\} = i_x^{-1}(G)$$

Note that G^x is $\mathbb{K}$–measurable due to the measurability of i_x. Then we have

Lemma 1.4 *Let $(P_x)_{x\in\mathcal{X}}$ be a $(\mathcal{X}, \mathbb{E})$–Markov kernel on $(\mathcal{Y}, \mathbb{K})$. For each $G \in \mathbb{E} \otimes \mathbb{K}$ the map*

$$x \mapsto P_x(G^x)$$

is $\mathbb{E}$–measurable.

Proof Let $\mathbb{H} = \{G \in \mathbb{E} \otimes \mathbb{K} \,:\, x \mapsto P_x(G^x) \text{ is } \mathbb{E}\text{–measurable}\}$ and consider a product set $A \times B \in \mathbb{E} \otimes \mathbb{K}$. Then

$$(A \times B)^x = \begin{cases} \emptyset & \text{if } x \notin A \\ B & \text{if } x \in A \end{cases}$$

so that

$$P_x((A \times B)^x) = \begin{cases} 0 & \text{if } x \notin A \\ P_x(B) & \text{if } x \in A \end{cases} = 1_A(x) \cdot P_x(B)$$

This is a product of two $\mathbb{E}$-measurable functions and hence itself $\mathbb{E}$-measurable. So we conclude that $\mathbb{H}$ contains all product sets; since this is an intersection stable generating system for $\mathbb{E} \otimes \mathbb{K}$, we have $\mathbb{H} = \mathbb{E} \otimes \mathbb{K}$, if we can show that $\mathbb{H}$ is a Dynkin class:

We already have that $\mathcal{X} \times \mathcal{Y} \in \mathbb{H}$ since it is a product set. Assume that $G_1 \subseteq G_2$ are two sets in $\mathbb{H}$. Then obviously also $G_1^x \subseteq G_2^x$ for all $x \in \mathcal{X}$, and

$$(G_2 \setminus G_1)^x = G_2^x \setminus G_1^x\,.$$

Then

$$P_x((G_2 \setminus G_1)^x) = P_x(G_2^x) - P_x(G_1^x)$$

which is a difference between two measurable functions. Hence $G_2 \setminus G_1 \in \mathbb{H}$.

Finally, assume that $G_1 \subseteq G_2 \subseteq \cdots$ is an increasing sequence of $\mathbb{H}$–sets. Similarly to above we have $G_1^x \subseteq G_2^x \subseteq \cdots$ and

$$\left(\bigcup_{n=1}^{\infty} G_n\right)^x = \bigcup_{n=1}^{\infty} G_n^x \,.$$

Then

$$P_x\left(\left(\bigcup_{n=1}^{\infty} G_n\right)^x\right) = P_x\left(\bigcup_{n=1}^{\infty} G_n^x\right) = \lim_{n\to\infty} P_x(G_n^x)$$

This limit is $\mathbb{E}$–measurable, since each of the functions $x \mapsto P_x(G_n^x)$ are measurable. Then $\bigcup_{n=1}^{\infty} G_n \in \mathbb{H}$. □

1.2 Integration of Markov kernels

Theorem 1.5 *Let μ be a probability measure on $(\mathcal{X}, \mathbb{E})$ and let $(P_x)_{x\in\mathcal{X}}$ be an $(\mathcal{X}, \mathbb{E})$–Markov kernel on $(\mathcal{Y}, \mathbb{K})$. There exists a uniquely determined probability measure λ on $(\mathcal{X} \times \mathcal{Y}, \mathbb{E} \otimes \mathbb{K})$ satisfying*

$$\lambda(A \times B) = \int_A P_x(B)\, \mathrm{d}\mu(x)$$

for all $A \in \mathbb{E}$ and $B \in \mathbb{K}$.

The probability measure λ constructed in Theorem 1.5 is called *the integration* of $(P_x)_{x\in\mathcal{X}}$ with respect to μ. The interpretation is that λ describes an experiment on $\mathcal{X} \times \mathcal{Y}$ that is performed in two steps: The first step picks a random element $x \in \mathcal{X}$ with distribution μ. The second step picks a random point $y \in \mathcal{Y}$ according to the probability measure P_x determined by the outcome x.

Proof The uniqueness follows, since λ is determined on all product sets and these form an intersection stable generating system for $\mathbb{E} \otimes \mathbb{K}$.

In order to prove the existence, we define

$$\lambda(G) = \int P_x(G^x)\, \mathrm{d}\mu(x)$$

For each $G \in \mathbb{E} \otimes \mathbb{K}$ the integrand is measurable according to Lemma 1.4. It is furthermore non–negative, such that $\lambda(G)$ is well–defined with values in $[0, \infty]$.

Now let $G_1, G_2, \ldots$ be a sequence of disjoint sets in $\mathbb{E} \otimes \mathbb{K}$. Then for each $x \in \mathcal{X}$ the sets $G_1^x, G_2^x, \ldots$ are disjoint as well. Hence

$$\lambda\Big(\bigcup_{n=1}^{\infty} G_n\Big) = \int P_x\Big(\Big(\bigcup_{n=1}^{\infty} G_n\Big)^x\Big)\, \mathrm{d}\mu(x) = \int \sum_{n=1}^{\infty} P_x(G_n^x)\, \mathrm{d}\mu(x) = \sum_{n=1}^{\infty} \lambda(G_n)\,.$$

In the second equality we have used that each P_x is a measure, and in the third equality we have used monotone convergence to interchange integration and summation. From this we have that λ is a measure. And since

$$\lambda(\mathcal{X} \times \mathcal{Y}) = \int P_x((\mathcal{X} \times \mathcal{Y})^x) \, \mathrm{d}\mu(x) = \int P_x(\mathcal{Y}) \, \mathrm{d}\mu(x) = \int 1 \, \mathrm{d}\mu(x) = 1$$

we obtain that λ is actually a probability measure. Finally, it follows that

$$\lambda(A \times B) = \int P_x((A \times B)^x) \, \mathrm{d}\mu(x) = \int 1_A(x) P_x(B) \, \mathrm{d}\mu(x) = \int_A P_x(B) \, \mathrm{d}\mu(x)$$

for all $A \in \mathbb{E}$ and $B \in \mathbb{K}$. □

Corollary 1.6 *Let μ be a probability measure on $(\mathcal{X}, \mathbb{E})$ and let $(P_x)_{x \in \mathcal{X}}$ be an $(\mathcal{X}, \mathbb{E})$–Markov kernel on $(\mathcal{Y}, \mathbb{K})$. Let λ be the integration of $(P_x)_{x \in \mathcal{X}}$ with respect to μ. Then λ satisfies*

$$\lambda(A \times \mathcal{Y}) = \mu(A) \qquad \text{for all } A \in \mathbb{E}$$

$$\lambda(\mathcal{X} \times B) = \int P_x(B) \, \mathrm{d}\mu(x) \qquad \text{for all } B \in \mathbb{K}$$

Proof The second statement is obvious. For the first result just note that $P_x(\mathcal{Y}) = 1$ for all $x \in \mathcal{X}$. □

The probability measure on $(\mathcal{Y}, \mathbb{K})$ defined by $\lambda(\mathcal{X} \times B)$ is called *the mixture* of the Markov kernel with respect to μ.

Example 1.7 Let μ be a probability measure on $(\mathcal{X}, \mathbb{E})$ and let ν be a probability measure on $(\mathcal{Y}, \mathbb{K})$. Define $P_x = \nu$ for all $x \in \mathcal{X}$. Then, trivially, $(P_x)_{x \in \mathcal{X}}$ is an $\mathcal{X}$–Markov kernel on $\mathcal{Y}$. Let λ be the integration of this kernel with respect to μ. Then for all $A \in \mathbb{E}$ and $B \in \mathbb{K}$

$$\lambda(A \times B) = \int_A \nu(B) \, \mathrm{d}\mu(x) = \mu(A) \cdot \nu(B) \, .$$

The only measure satisfying this property is the product measure $\mu \otimes \nu$, so $\lambda = \mu \otimes \nu$. Hence a product measure is a particularly simple example of a measure constructed by integrating a Markov kernel. □

Example 1.8 Let μ be the Poisson distribution with parameter λ. For each $x \in \mathbb{N}_0$ we define P_x to be the binomial distribution with parameters (x, p). Then $(P_x)_{x \in \mathbb{N}_0}$ is an $\mathbb{N}_0$–Markov kernel on $\mathbb{N}_0$.

Let ξ be the mixture of $(P_x)_{x \in \mathbb{N}_0}$ with respect to μ. This must be a probability measure on $\mathbb{N}_0$ and is thus determined by its probability mass function q. For $n \in \mathbb{N}_0$ we obtain

$$\begin{aligned} q(n) &= \sum_{k=n}^{\infty} \binom{k}{n} p^n (1-p)^{k-n} \frac{\lambda^k}{k!} e^{-\lambda} \\ &= \frac{(\lambda p)^n}{n!} e^{-\lambda} \sum_{k=n}^{\infty} \frac{\big((1-p)\lambda\big)^{k-n}}{(k-n)!} \\ &= \frac{(\lambda p)^n}{n!} e^{-\lambda} e^{(1-p)\lambda} = \frac{(\lambda p)^n}{n!} e^{-\lambda p} \, . \end{aligned}$$

Hence the mixture ξ is the Poisson distribution with parameter λp. □

Theorem 1.9. (Uniqueness of integration) *Suppose that $(\mathcal{Y}, \mathbb{K})$ has a countable generating system that is intersection stable. Let μ and $\tilde{\mu}$ be two probability measures on $(\mathcal{X}, \mathbb{E})$ and assume that $(P_x)_{x\in\mathcal{X}}$ and $(\tilde{P}_x)_{x\in\mathcal{X}}$ are two $(\mathcal{X}, \mathbb{E})$–Markov kernels on $(\mathcal{Y}, \mathbb{K})$. Let λ be the integration of $(P_x)_{x\in\mathcal{X}}$ with respect to μ, and let $\tilde{\lambda}$ be the integration of $(\tilde{P}_x)_{x\in\mathcal{X}}$ with respect to $\tilde{\mu}$. Define*

$$E_0 = \{x \in \mathcal{X} \ : \ P_x = \tilde{P}_x\}$$

Then $\lambda = \tilde{\lambda}$ if and only if $\mu = \tilde{\mu}$ and $\mu(E_0) = 1$.

Proof Let $(B_n)_{n\in\mathbb{N}}$ be a countable generating system for $(\mathcal{Y}, \mathbb{K})$. Then

$$E_0 = \bigcap_{n=1}^{\infty} \{x \in \mathcal{X} \ : \ P_x(B_n) = \tilde{P}_x(B_n)\}$$

from which we can conclude that $E_0 \in \mathbb{E}$.

Assume that $\mu = \tilde{\mu}$ and $\mu(E_0) = 1$. Then for all $A \in \mathbb{E}$ and $B \in \mathbb{K}$ we have

$$\lambda(A \times B) = \int_{A\cap E_0} P_x(B) \, \mathrm{d}\mu(x) = \int_{A\cap E_0} \tilde{P}_x(B) \, \mathrm{d}\tilde{\mu}(x) = \tilde{\lambda}(A \times B)$$

and thereby $\lambda = \tilde{\lambda}$.

Assume conversely that $\lambda = \tilde{\lambda}$. According to Corollary 1.6 we have for all $A \in \mathbb{E}$

$$\mu(A) = \lambda(A \times \mathcal{Y}) = \tilde{\lambda}(A \times \mathcal{Y}) = \tilde{\mu}(A)$$

such that $\mu = \tilde{\mu}$. The proof will be complete, if we can show that

$$\mu(\{x \in \mathcal{X} \ : \ P_x(B_n) \neq \tilde{P}_x(B_n)\}) = 0$$

for all $n \in \mathbb{N}$. For this purpose we consider the set

$$E_n^+ = \{x \in \mathcal{X} \ : \ P_x(B_n) > \tilde{P}_x(B_n)\} .$$

Using this definition gives

$$\int_{E_n^+} (P_x(B_n) - \tilde{P}_x(B_n)) \, \mathrm{d}\mu(x) = \lambda(E_n^+ \times B_n) - \tilde{\lambda}(E_n^+ \times B_n) = 0$$

and since the integrand is strictly positive on E_n^+, we can conclude that $\mu(E_n^+) = 0$. It is shown similarly that $\mu(E_n^-) = 0$, where

$$E_n^- = \{x \in \mathcal{X} \ : \ P_x(B_n) < \tilde{P}_x(B_n)\} .$$

This concludes the proof. □

1.3 Properties of the integration measure

In this section we will consider integration with respect to λ, where λ is the integrated measure of a Markov kernel $(P_x)_{x\in\mathcal{X}}$ with respect to some probability measure μ. We shall see, that such λ–integrals can be calculated by successive integration similar to what is known for product measures.

Lemma 1.10 *Let $(P_x)_{x\in\mathcal{X}}$ be a $(\mathcal{X},\mathbb{E})$–Markov kernel on $(\mathcal{Y},\mathbb{K})$ and assume that $f:\mathcal{X}\times\mathcal{Y}\to[0,\infty]$ is $\mathbb{E}\otimes\mathbb{K}$–measurable. Then the map*

$$x\mapsto\int f(x,y)\,\mathrm{d}P_x(y) \tag{1.1}$$

is $\mathbb{E}$–measurable.

Proof Firstly note that for fixed x then $f(x,y)=f\circ i_x(y)$ which is a $\mathbb{K}$–measurable function. Hence the integral in (1.1) is well–defined. Now assume that f is a simple function

$$f=\sum_{k=1}^{n}c_k1_{G_k} \tag{1.2}$$

where $c_1,\ldots,c_n\in(0,\infty)$ and $G_1,\ldots,G_n$ are disjoint sets in $\mathbb{E}\otimes\mathbb{K}$. Since

$$1_{G_k}(x,y)=1_{G_k^x}(y)$$

for all x and y, we obtain

$$\begin{aligned}\int f(x,y)\,\mathrm{d}P_x(y)&=\sum_{k=1}^{n}\int c_k1_{G_k}(x,y)\,\mathrm{d}P_x(y)\\&=\sum_{k=1}^{n}c_k\int 1_{G_k^x}(y)\,\mathrm{d}P_x(y)\\&=\sum_{k=1}^{n}c_kP_x(G_k^x)\end{aligned}$$

According to Lemma 1.4 this is a linear combination of $\mathbb{E}$–measurable functions. Hence it is $\mathbb{E}$–measurable.

Now assume that f is a general function in $\mathcal{M}^+(\mathcal{X}\times\mathcal{Y},\mathbb{E}\otimes\mathbb{K})$. Then there exists a sequence $(f_n)_{n\in\mathbb{N}}$ of non–negative simple functions with $f_n(x,y)\uparrow f(x,y)$ for all $x\in\mathcal{X}$ and $y\in\mathcal{Y}$. For fixed x we have from monotone convergence, that

$$\int f_n(x,y)\,\mathrm{d}P_x(y)\uparrow\int f(x,y)\,\mathrm{d}P_x(y)\,.$$

Hence the right hand side is the point–wise limit of $\mathbb{E}$–measurable functions. Thereby it is $\mathbb{E}$–measurable. □

Theorem 1.11. (Extended Tonelli) *Let μ be a probability measure on $(\mathcal{X}, \mathbb{E})$, and assume that $(P_x)_{x\in\mathcal{X}}$ is a $(\mathcal{X}, \mathbb{E})$–Markov kernel on $(\mathcal{Y}, \mathbb{K})$. Let λ be the integration of $(P_x)_{x\in\mathcal{X}}$ with respect to μ. For every $\mathbb{E}\otimes\mathbb{K}$–measurable function $f : \mathcal{X}\times\mathcal{Y} \to [0,\infty]$ it holds that*

$$\int f(x,y)\,\mathrm{d}\lambda(x,y) = \iint f(x,y)\,\mathrm{d}P_x(y)\,\mathrm{d}\mu(x)\,.$$

Proof The inner integral on the right hand side is $\mathbb{E}$–measurable with values in $[0,\infty]$ according to Lemma 1.10. Hence both the left-hand side and the right-hand side are well-defined.

Now assume that f is a simple function on the form (1.2). Then

$$\begin{aligned}
\int f\,\mathrm{d}\lambda &= \sum_{k=1}^{n} c_k\lambda(G_k)\\
&= \sum_{k=1}^{n} c_k \int P_x(G_k^x)\,\mathrm{d}\mu(x)\\
&= \sum_{k=1}^{n} c_k \iint 1_{G_k^x}(y)\,\mathrm{d}P_x(y)\,\mathrm{d}\mu(x)\\
&= \sum_{k=1}^{n} c_k \iint 1_{G_k}(x,y)\,\mathrm{d}P_x(y)\,\mathrm{d}\mu(x)\\
&= \iint \sum_{k=1}^{n} c_k 1_{G_k}(x,y)\,\mathrm{d}P_x(y)\,\mathrm{d}\mu(x)\\
&= \iint f(x,y)\,\mathrm{d}P_x(y)\,\mathrm{d}\mu(x)
\end{aligned}$$

which shows the result when f is a simple function.

Now let f be a general function in $\mathcal{M}^+(\mathcal{X}\times\mathcal{Y}, \mathbb{E}\otimes\mathbb{K})$. Then there exists a sequence $(f_n)_{n\in\mathbb{N}}$ of non–negative simple functions with $f_n \uparrow f$. From monotone convergence we get

$$\int f\,d\lambda = \lim_{n\to\infty}\int f_n\,d\lambda = \lim_{n\to\infty}\iint f_n(x,y)\,\mathrm{d}P_x(y)\,\mathrm{d}\mu(x)$$

But monotone convergence also yields

$$\int f_n(x,y)\,\mathrm{d}P_x(y) \uparrow \int f(x,y)\,\mathrm{d}P_x(y)$$

and applying monotone convergence once more then gives

$$\iint f_n(x,y)\,\mathrm{d}P_x(y)\,\mathrm{d}\mu(x) \uparrow \iint f(x,y)\,\mathrm{d}P_x(y)\,\mathrm{d}\mu(x)\,,$$

and this shows the theorem. □

Theorem 1.12. (Extended Fubini) *Let μ be a probability measure on $(\mathcal{X}, \mathbb{E})$ and assume that $(P_x)_{x\in\mathcal{X}}$ is a $(\mathcal{X}, \mathbb{E})$–Markov kernel on $(\mathcal{Y}, \mathbb{K})$. Let λ be the integration of $(P_x)_{x\in\mathcal{X}}$ with respect to μ. For every $\mathbb{E}\otimes\mathbb{K}$–measurable and λ–integrable function $f : \mathcal{X}\times\mathcal{Y} \to \mathbb{R}$ it holds that*

$$A_0 = \{x \in \mathcal{X} \,:\, \int |f(x,y)| \,\mathrm{d}P_x(y) < \infty\}$$

is $\mathbb{E}$–measurable with $\mu(A_0) = 1$. Furthermore, the function

$$x \mapsto g(x) := \begin{cases} \int f(x,y)\,\mathrm{d}P_x(y) & x \in A_0 \\ 0 & x \notin A_0 \end{cases}$$

is $\mathbb{E}$–measurable and μ–integrable, and

$$\int f(x,y)\,\mathrm{d}\lambda(x,y) = \int_{A_0}\int f(x,y)\,\mathrm{d}P_x(y)\,\mathrm{d}\mu(x)\,.$$

Note: The extended Tonelli's Theorem can be applied to determine whether f is λ–integrable – that is whether $\int |f|\, d\lambda < \infty$.

Proof It follows from Lemma 1.10 that $A_0 \in \mathbb{E}$. The extended Tonelli's Theorem gives

$$\iint |f(x,y)|\,\mathrm{d}P_x(y)\,\mathrm{d}\mu(x) = \int |f|\, d\lambda < \infty\,.$$

Hence the integral $\int |f(x,y)|\,\mathrm{d}P_x(y)$ must be finite for μ almost all $x \in \mathcal{X}$ such that $\mu(A_0) = 1$. For each $x \in A_0$ we have

$$\int f(x,y)\,\mathrm{d}P_x(y) = \int f^+(x,y)\,\mathrm{d}P_x(y) - \int f^-(x,y)\,\mathrm{d}P_x(y)$$

From this we see that the function g defined in the theorem is measurable according to Lemma 1.10. Furthermore we obtain from the extended Tonelli's Theorem that

$$\begin{aligned}\int |g(x)|\mathrm{d}\mu(x) &= \int_{A_0}\left|\int f(x,y)\,\mathrm{d}P_x(y)\right|\,\mathrm{d}\mu(x)\\ &\le \iint 1_{A_0\times\mathcal{Y}}(x,y)|f(x,y)|\,\mathrm{d}P_x(y)\,\mathrm{d}\mu(x)\\ &< \infty\,,\end{aligned}$$

showing that g is μ–integrable. Finally, we have from the extended Tonelli that

$$\begin{aligned}&\int_{A_0}\int f(x,y)\,\mathrm{d}P_x(y)\,\mathrm{d}\mu(x)\\ &\quad= \int_{A_0}\int f(x,y)^+\,\mathrm{d}P_x(y)\,\mathrm{d}\mu(x) - \int_{A_0}\int f(x,y)^-\,\mathrm{d}P_x(y)\,\mathrm{d}\mu(x)\\ &\quad= \int 1_{A_0}(x) f^+(x,y)\,\mathrm{d}\lambda(x,y) - \int 1_{A_0}(x) f^-(x,y)\,\mathrm{d}\lambda(x,y)\\ &\quad= \int 1_{A_0}(x) f(x,y)\,\mathrm{d}\lambda(x,y) = \int_{A_0\times\mathcal{Y}} f(x,y)\,\mathrm{d}\lambda(x,y)\,.\end{aligned}$$

But from Corollary 1.6 we have $\lambda(A_0 \times \mathcal{Y}) = \mu(A_0)$ so

$$\int_{A_0 \times \mathcal{Y}} f(x,y)\ \mathrm{d}\lambda(x,y) = \int f(x,y)\ \mathrm{d}\lambda(x,y)$$

and the proof is complete. □

1.4 Conditional distributions

In an experiment where two random variables X and Y are observed, it is often convenient to consider the probabilistic model in two steps: X is observed first, afterwards Y is observed. Here it is natural to believe that the mechanism that decides the value of Y depends on the value of X drawn. This two–step model can be constructed by considering the joint distribution of X and Y as the integration of the conditional distribution of Y given X with respect to the distribution of X.

Definition 1.13 Let X and Y be random variables defined on the probability space $(\Omega, \mathbb{F}, P)$ with values in $(\mathcal{X}, \mathbb{E})$ and $(\mathcal{Y}, \mathbb{K})$ respectively. Let $(P_x)_{x\in\mathcal{X}}$ be a $(\mathcal{X}, \mathbb{E})$–Markov kernel on $(\mathcal{Y}, \mathbb{K})$. We say that $(P_x)_{x\in\mathcal{X}}$ is the conditional distribution of Y given X, if the joint distribution $(X,Y)(P)$ is the integration of $(P_x)_{x\in\mathcal{X}}$ with respect to $X(P)$. That is, if

$$P(X \in A, Y \in B) = (X,Y)(P)(A \times B) = \int_A P_x(B)\ \mathrm{d}X(P)(x)$$

for all $A \in \mathbb{E}$ and $B \in \mathbb{K}$.

Note that we say *the* conditional distribution although according to Theorem 1.9 the Markov kernel can be changed on nullsets with respect to $X(P)$. The strictly correct term would be *a* conditional distribution.
When specifying the conditional distribution, it is not necessary to give the entire Markov kernel $(P_x)_{x\in\mathcal{X}}$. Since the Markov kernel is integrated with respect to $X(P)$ it will be enough to give $(P_x)_{x\in A_0}$, where $A_0 \in \mathbb{E}$ is any set with $P(X \in A_0) = 1$.

Conversely, a conditional distribution given by $(P_x)_{x\in A_0}$, where $P(X \in A_0) = 1$, can be extended to a "true" Markov kernel $(\tilde{P}_x)_{x\in\mathcal{X}}$ by the definition

$$\tilde{P}_x = \begin{cases} P_x & x \in A_0 \\ P_0 & x \notin A_0 \end{cases}$$

where P_0 is some probability measure on $(\mathcal{Y}, \mathbb{K})$. Note that $x \mapsto \tilde{P}_x(B)$ is measurable, since A_0 is a measurable set.

The interpretation of the conditional distribution of Y given X is that P_x describes the distribution of Y if we know that $X = x$. This interpretation is very useful although it should not be taken too seriously, since it may be difficult to give a strict mathematical description when the event $X = x$ is a nullset. This interpretation leads to the

following alternative notation for a Markov kernel $(P_x)_{x\in\mathcal{X}}$ that is a conditional distribution of Y given X:

$$P(Y \in B \,|\, X = x) = P_x(B) \qquad \text{for } B \in \mathbb{K}\,.$$

A more relaxed but useful notation will be simply talking about 'the distribution of $Y \,|\, X = x$' instead of the longer 'the distribution P_x , when $(P_x)_{x\in\mathcal{X}}$ is the conditional distribution of Y given X'. We will also from time to time write expressions like $Y \,|\, X = x \sim \nu$.

For completely arbitrary random variables, a conditional distribution may not always be well-defined. However, if the image space of Y is a *Borel space* (i.e. isomorphic to a Borel subset of the unit interval), this is the case:

Theorem 1.14 *Let X and Y be random variables defined on the probability space $(\Omega, \mathbb{F}, P)$ with values in $(\mathcal{X}, \mathbb{E})$ and $(\mathcal{Y}, \mathbb{K})$ respectively, such that $\mathcal{Y}$ is a Borel space. Then there exists a conditional distribution of Y given X.*

This result is particularly important, since *spaces of the form $\mathbb{R}$, $\mathbb{R}^n$, and $\mathbb{R}^\infty$ are all Borel spaces.* We refrain from proving Theorem 1.14 as well as these facts and refer the reader to other textbooks on probability.

Although the conditional distributions exist, it is in general difficult and not clear how the corresponding Markov kernels should be constructed. However, direct construction of the Markov kernels is possible in specific situations.

Theorem 1.15 *Assume that X and Y are random variables on $(\mathcal{X}, \mathbb{E})$ and $(\mathcal{Y}, \mathbb{K})$ such that $(P_x)_{x\in\mathcal{X}}$ is the conditional distribution of Y given X. Then X and Y are independent if and only if P_x does not depend on x, i.e. the Markov kernel can be chosen so that*

$$P_x = P_0$$

for all $x \in \mathcal{X}$. In the case of independence, then $P_x = P_0 = Y(P)$ for all $x \in \mathcal{X}$.

Proof Suppose that X and Y are independent. Then for $A \in \mathbb{E}$ and $B \in \mathbb{K}$

$$P(X \in A, Y \in B) = X(P)(A) \cdot Y(P)(B) = \int_A Y(P)(B)\, \mathrm{d}X(P)(x)$$

which shows that the constant Markov kernel $(Y(P))_{x\in\mathcal{X}}$ is the conditional distribution of Y given X.

Conversely, assume that $P_x = P_0$ for all $x \in \mathcal{X}$, where P_0 is some probability measure on $(\mathcal{Y}, \mathbb{K})$. Then for $B \in \mathbb{K}$ we have

$$\begin{aligned} P(Y \in B) = P(X \in \mathcal{X}, Y \in B) &= \int P_x(B)\, \mathrm{d}X(P)(x) \\ &= \int P_0(B)\, \mathrm{d}X(P)(x) = P_0(B) \end{aligned}$$

which shows that $Y(P) = P_0$. Furthermore for $A \in \mathbb{E}$ and $B \in \mathbb{K}$ we obtain

$$P(X \in A, Y \in B) = \int_A P_x(B)\, \mathrm{d}X(P)(x) = \int_A P_0(B)\, \mathrm{d}X(P)(x)$$
$$= X(P)(A)P_0(B) = P(X \in A)P(Y \in B)$$

leading to the conclusion that X and Y are independent. □

Hence *independence between two variables X and Y is equivalent to the conditional distribution of Y given X being constant.* In contrast, if the conditional distribution consists of very different probability measures there is a strong dependence between X and Y.

In the following theorem it is seen that if X is a discrete random variable, the conditional distribution is just given by elementary conditional probabilities.

Theorem 1.16 *Let X and Y be random variables defined on $(\Omega, \mathbb{F}, P)$ with values in $(\mathcal{X}, \mathbb{E})$ and $(\mathcal{Y}, \mathbb{K})$. Assume that $\mathcal{X}$ is finite or countable and that $\mathbb{E}$ is the paving that consists of all subsets of $\mathcal{X}$. Then the conditional distribution of Y given X is determined by*

$$P_x(B) = \frac{P(X = x, Y \in B)}{P(X = x)} \qquad \text{for } B \in \mathbb{K}\,, \tag{1.3}$$

for all $x \in \mathcal{X}$ with $P(X = x) > 0$.

Note that $P_x(B)$ is simply defined as the conditional probability of $(Y \in B)$ given the set $(X = x)$:

$$P_x(B) = \frac{P(X = x, Y \in B)}{P(X = x)} = P(Y \in B \,|\, X = x)$$

Proof Let $A_0 = \{x \in \mathcal{X} \,:\, P(X = x) > 0\}$ and note that $X(P)(A_0) = 1$ such that (1.3) defines an $(\mathcal{X}, \mathbb{E})$–Markov kernel on $(\mathcal{Y}, \mathbb{K})$ – the measurability is not a problem, since all functions on $\mathcal{X}$ are $\mathbb{E}$–measurable. For $A \subseteq \mathcal{X}$ and $B \in \mathbb{K}$ we have

$$\begin{aligned}
\int_A P_x(B)\, \mathrm{d}X(P)(x) &= \int_{A \cap A_0} P_x(B)\, \mathrm{d}X(P)(x) \\
&= \sum_{x \in A \cap A_0} \frac{P(X = x, Y \in B)}{P(X = x)} P(X = x) \\
&= \sum_{x \in A \cap A_0} P(X = x, Y \in B) \\
&= P(X \in A \cap A_0, Y \in B) \\
&= P(X \in A, Y \in B)
\end{aligned}$$

such that $(P_x)_{x \in \mathcal{X}}$ actually is the conditional distribution of Y given X. □

Example 1.17 Let X_1 and X_2 be independent random variables that are Poisson distributed with parameters λ_1 and λ_2. Then the distribution of $X = X_1 + X_2$ is a Poisson distribution with parameter $\lambda = \lambda_1 + \lambda_2$. We will find the conditional

distribution of X_1 given X by indicating $P_x(\{n\})$ for all $x, n \in \mathbb{N}_0$. This must be sufficient, since all P_x are concentrated on $\mathbb{N}_0$. Using Theorem 1.16 yields for $x \in \mathbb{N}_0$ and $n = 0, 1, \ldots, x$ that

$$\begin{aligned} P_x(\{n\}) &= \frac{P(X_1 = n, X_2 = x - n)}{P(X = x)} \\ &= \frac{P(X_1 = n)P(X_2 = x - n)}{P(X = x)} \\ &= \frac{\frac{\lambda_1^n}{n!} e^{-\lambda_1} \frac{\lambda_2^{x-n}}{(x-n)!} e^{-\lambda_2}}{\frac{\lambda^x}{x!} e^{-\lambda}} \\ &= \binom{x}{n} \left(\frac{\lambda_1}{\lambda_1 + \lambda_2}\right)^n \left(\frac{\lambda_2}{\lambda_1 + \lambda_2}\right)^{x-n} \end{aligned}$$

Hence the conditional distribution of X_1 given $X = x$ is a binomial distribution with parameters $(x, \frac{\lambda_1}{\lambda_2 + \lambda_2})$. □

Theorem 1.18 *Assume that X and Y are random variables defined on $(\Omega, \mathbb{F}, P)$ with values in $(\mathcal{X}, \mathbb{E})$ and $(\mathcal{Y}, \mathbb{K})$ respectively. Assume that $(P_x)_{x \in \mathcal{X}}$ is the conditional distribution of Y given X. Furthermore let μ and ν be σ–finite measures on $(\mathcal{X}, \mathbb{E})$ and $(\mathcal{Y}, \mathbb{K})$ respectively and assume that $X(P) = f \cdot \mu$. Finally assume that $(P_x)_{x \in \mathcal{X}}$ is a $(\mathcal{X}, \mathbb{E})$–Markov kernel on $(\mathcal{Y}, \mathbb{K})$ of the type constructed in Theorem 1.2: Assume that $P_x = g_x \cdot \nu$, where the function $(x, y) \mapsto g_x(y)$ is $\mathbb{E} \otimes \mathbb{K}$–measurable.*

Then the joint distribution of X and Y is given by $(X, Y)(P) = h \cdot \mu \otimes \nu$, where

$$h(x, y) = f(x)\, g_x(y) \qquad \text{for all } x \in \mathcal{X},\ y \in \mathcal{Y}$$

Proof Let $A \in \mathbb{E}$ and $B \in \mathbb{K}$. Then

$$\begin{aligned} (X, Y)(P)(A \times B) &= P(X \in A, Y \in B) \\ &= \int 1_A(x)\, P_x(B)\, \mathrm{d}X(P)(x) \\ &= \int 1_A(x) \Big(\int 1_B(y) g_x(y)\, \mathrm{d}\nu(y) \Big)\, f(x)\, \mathrm{d}\mu(x) \\ &= \int \int 1_{A \times B}(x, y)\, f(x) g_x(y) \mathrm{d}\nu(y) \mathrm{d}\mu(x) \\ &= \int 1_{A \times B}(x, y)\, h(x, y) \mathrm{d}(\mu \otimes \nu)(x, y) \end{aligned}$$

where the last equality is follows from Tonelli's theorem. We see that $(X, Y)(P)$ and $h \cdot \mu \otimes \nu$ coincide on all product sets, and therefore must be equal. □

The theorem states that the joint density is the product of the marginal density and the conditional densities. The next theorem gives the converse result: The densities for the conditional distribution is the ratio of the joint density and the marginal density.

Theorem 1.19 *Assume that X and Y are random variables defined on $(\Omega, \mathbb{F}, P)$ with values in $(\mathcal{X}, \mathbb{E})$ and $(\mathcal{Y}, \mathbb{K})$. Furthermore let μ and ν be σ–finite measures on $(\mathcal{X}, \mathbb{E})$ and $(\mathcal{Y}, \mathbb{K})$ and assume that $(X, Y)(P) = h \cdot \mu \otimes \nu$. Then the conditional distribution of Y given X exists. The marginal distribution of X has density with respect to μ given by*

$$f(x) = \int h(x, y) \, \mathrm{d}\nu(y)$$

Let $A_0 = \{x \in \mathcal{X} : 0 < f(x) < \infty\}$. Then $X(P)(A_0) = 1$ and the conditional distribution $(P_x)_{x \in \mathcal{X}}$ of Y given X has density with respect to ν given by

$$g_x(y) = \frac{h(x, y)}{f(x)}$$

for all $x \in A_0$.

Proof Finding the marginal density for $X(P)$ is a well–known calculation. For $A \in \mathbb{E}$ we have

$$\begin{aligned} X(P)(A) &= (X, Y)(P)(A \times \mathcal{Y}) \\ &= \int_{A \times \mathcal{Y}} h(x, y) \, \mathrm{d}(\mu \otimes \nu)(x, y) \\ &= \int_A \int h(x, y) \, \mathrm{d}\nu(y) \, \mathrm{d}\mu(x) \\ &= \int_A f(x) \, \mathrm{d}\mu(x) \end{aligned}$$

according to Tonelli. Thus $X(P)$ has density f with respect to μ.

Now define the sets

$$A_1 = \{x \in \mathcal{X} : f(x) = 0\} \quad \text{and} \quad A_2 = \{x \in \mathcal{X} : f(x) = \infty\}.$$

Since $X(P)(\mathcal{X}) = 1$ we have

$$1 \geq X(P)(A_2) = \int_{A_2} f(x) \, \mathrm{d}\mu(x) = \infty \cdot \mu(A_2)$$

so $\mu(A_2) = 0$. Clearly we have that $X(P)(A_1) = 0$, such that $X(P)(A_0) = 1$.

From Tonelli we have that $x \mapsto \int h(x, y) \mathrm{d}\nu(y) = f(x)$ is $\mathbb{E}$–measurable. Then also

$$(x, y) \mapsto 1_{A_0}(x) \frac{h(x, y)}{f(x)} = 1_{A_0}(x) g_x(y)$$

is $\mathbb{E} \otimes \mathbb{K} - \mathbb{B}$–measurable, and we have from Theorem 1.2 that $(P_x)_{x \in A_0}$ is a Markov kernel, when $P_x = g_x \cdot \nu$. Finally we have for $A \in \mathbb{E}$ and $B \in \mathbb{K}$ that

$$\begin{aligned}\int_A P_x(B)\,\mathrm{d}X(P)(x) &= \int_{A\cap A_0}\left(\int_B g_x(y)\,\mathrm{d}\nu(y)\right) f(x)\,\mathrm{d}\mu(x)\\ &= \int_{A\cap A_0}\left(\int_B \frac{h(x,y)}{f(x)}\,\mathrm{d}\nu(y)\right) f(x)\,\mathrm{d}\mu(x)\\ &= \int_A\left(\int_B h(x,y)\,\mathrm{d}\nu(y)\right)\,\mathrm{d}\mu(x)\\ &= \int_{A\times B} h(x,y)\mathrm{d}(\mu\otimes\nu)(x,y)\\ &= (X,Y)(P)(A\times B)\\ &= P(X\in A, Y\in B)\,,\end{aligned}$$

which shows, that $(P_x)_{x\in A_0}$ is the conditional distribution for Y given X. □

Example 1.20 Suppose $V=\mathbb{R}^d$ and assume the random vector X partitioned into components X_1 and X_2, where $X_1\in\mathbb{R}^r$ and $X_2\in\mathbb{R}^s$ with $r+s=d$. Its mean vector and covariance matrix can then be partitioned accordingly into blocks as

$$\xi=\begin{pmatrix}\xi_1\\ \xi_2\end{pmatrix}\quad\text{and}\quad \Sigma=\begin{pmatrix}\Sigma_{11}&\Sigma_{12}\\ \Sigma_{21}&\Sigma_{22}\end{pmatrix}$$

such that Σ_{11} has dimensions $r\times r$ and so on. Let X be distributed as $\mathcal{N}_d(\xi,\Sigma)$, where X, ξ and Σ are partitioned as above and Σ is regular. Then the conditional distribution of X_1 given $X_2=x_2$ is $\mathcal{N}_r(\xi_{1|2},\Sigma_{1|2})$, where

$$\xi_{1|2}=\xi_1+\Sigma_{12}(\Sigma_{22})^{-1}(x_2-\xi_2)\quad\text{and}\quad \Sigma_{1|2}=\Sigma_{11}-\Sigma_{12}(\Sigma_{22})^{-1}\Sigma_{21}. \tag{1.4}$$

This is seen as follows: Since Σ is positive definite we can let $K=\Sigma^{-1}$ denote the concentration matrix and assume this to be partitioned in the same fashion as Σ. By Theorem 1.19, the conditional density is proportional to the joint density of X_1 and X_2. Hence, exploiting that x_2 is fixed, we find by direct calculation that

$$\begin{aligned}&f(x_1\,|\,x_2)\propto f_{\xi,\Sigma}(x)\\ &\propto\exp\left\{-(x_1-\xi_1)^\top K_{11}(x_1-\xi_1)/2-(x_1-\xi_1)^\top K_{12}(x_2-\xi_2)\right\}.\end{aligned}$$

The linear term involving x_1 has coefficient equal to

$$K_{11}\xi_1-K_{12}(x_2-\xi_2)=K_{11}\left\{\xi_1-(K_{11})^{-1}K_{12}(x_2-\xi_2)\right\}.$$

Using (C.2) we find that

$$(K_{11})^{-1}=\Sigma_{11}-\Sigma_{12}(\Sigma_{22})^{-1}\Sigma_{21} \tag{1.5}$$

and further that

$$(K_{11})^{-1}K_{12}=-\Sigma_{12}(\Sigma_{22})^{-1}, \tag{1.6}$$

which then gives

$$f(x_1 \mid x_2) \propto \exp\left\{-(x_1 - \xi_{1|2})^\top K_{11}(x_1 - \xi_{1|2})/2\right\}$$

and the result follows. Note that the proportionality constant may in principle depend on the parameters as well as on x_2. But as the distribution is normal, this turns out not to be the case. It follows from (C.1) that we have

$$\det \Sigma = \det \Sigma_{1|2} \det \Sigma_{22} = \frac{\det \Sigma_{22}}{\det K_{11}}. \tag{1.7}$$

Note also the identities (1.5) and (1.6), which are quite useful in their own right. The first expresses that the concentration matrix of the conditional distribution is obtained from the concentration matrix of the joint distribution by deleting rows and columns corresponding to the variables conditioned upon. We thus obtain an alternative formula for the parameters of the conditional distribution

$$\xi_{1|2} = \xi_1 - (K_{11})^{-1}K_{12}(x_2 - \xi_2) \quad \text{and} \quad K_{1|2} = K_{11} \tag{1.8}$$

which may be simpler to use in certain contexts. □

We can reformulate Theorem 1.19 to obtain what is known as *Bayes' formula*.

Corollary 1.21. (Bayes' formula) *If $\pi = X(P)$ and P_x has density g_x w.r.t. ν, then the conditional distribution of X given Y exists and is determined by the Markov kernel $(\pi_y^*)_{y\in\mathcal{Y}}$ with density L_y w.r.t. π where*

$$L_y(x) = g_x(y)/c(y)$$

where

$$c(y) = \int g_x(y)\,\mathrm{d}\pi(x).$$

Proof We have by integration that

$$\begin{aligned} P(X \in A, Y \in B) &= \int_A \left\{ \int_B g_x(y)\,\mathrm{d}\nu(y) \right\} \mathrm{d}\pi(x) \\ &= \int_{A\times B} g_x(y)\,\mathrm{d}(\pi \otimes \nu)(x,y). \end{aligned}$$

Hence, $h(x,y) = g_x(y)$ is the joint density of (X,Y) w.r.t. $\pi \otimes \nu$. Using now Theorem 1.19 with x and y interchanged we get that the conditional distribution of X given $Y = y$ exists and has density

$$L_y(x) = g_x(y)/c(y)$$

w.r.t. π. □

Bayes' theorem is of course particularly important for Bayesian inference. We use the term *Fisherian model* for a parametrized family $\mathcal{P} = \{P_\theta, \theta \in \Theta\}$ of probability measures on a measurable space $(\mathcal{Y}, \mathbb{F})$. If the parameter space Θ is equipped with a σ-algebra $\mathbb{T}$ and the map $\theta \mapsto P_\theta(A)$ is measurable for all $A \in \mathbb{F}$, this family can be seen as a Markov kernel. We then define a corresponding Bayesian model as follows:

Definition 1.22 If π is a prior distribution on $(\Theta, \mathbb{T})$ and $\mathcal{P} = \{P_\theta, \theta \in \Theta\}$ is a Fisherian model on $\mathcal{Y}$ with Θ as parameter space and densities $g_\theta(\cdot)$ w.r.t. a σ-finite measure ν on $\mathcal{Y}$, the *corresponding Bayesian model with prior distribution* π for $\Theta \times \mathcal{Y}$ is the integration of $(P_\theta)_{\theta\in\Theta}$ w.r.t. π.

We can then reformulate Corollary 1.21 as

Corollary 1.23 *If π is a prior distribution on $(\Theta, \mathbb{T})$ and $\mathcal{P} = \{P_\theta, \theta \in \Theta\}$ is a Fisherian model on $\mathcal{Y}$ with Θ as parameter space and densities g_θ w.r.t. a σ-finite measure ν on $\mathcal{Y}$, then the posterior distribution of θ given Y in the corresponding Bayesian model exists and is determined by the Markov kernel $(\pi^*_y)_{y\in\mathcal{Y}}$ with density L_y w.r.t. π where*

$$L_y(\theta) \propto g_\theta(y).$$

In other words, the corollary says that the *likelihood* function $L_y(\theta) \propto g_\theta(y)$ is the density of the *posterior distribution* π^*_y w.r.t. the *prior distribution* π:

$$\pi^*_y \propto L_y \cdot \pi, \quad \text{or } \textit{posterior} \propto \textit{likelihood} \times \textit{prior}.$$

1.5 Transformations of conditional distributions

In this section we shall present a series of transformation results for conditional distributions. They have a somewhat similar content: In a framework with three or more random variables, where we know some of the conditional distributions, various other conditional distributions can be simply expressed.

It is complicated to understand how conditional distributions are specified in situations with three or more random variables. The reader is encouraged to spend much time understanding the content of the results, rather than the proofs. The stated results are not very surprising if the content is understood. And the proofs are rather mechanical: Firstly, it is argued that some expression is a Markov kernel, and then it is shown that this Markov kernel is the right conditional distribution.

Assume in this section, that X, Y, X_1, X_2, Y_1 and Y_2 are random variables defined on $(\Omega, \mathbb{F}, P)$ with values in $(\mathcal{X}, \mathbb{E})$, $(\mathcal{Y}, \mathbb{K})$, $(\mathcal{X}_1, \mathbb{E}_1)$, $(\mathcal{X}_2, \mathbb{E}_2)$, $(\mathcal{Y}_1, \mathbb{K}_1)$ and $(\mathcal{Y}_2, \mathbb{K}_2)$ respectively.

Theorem 1.24. (Substitution Theorem) *Assume that $(P_x)_{x\in\mathcal{X}}$ is the conditional distribution of Y given X. Let $(\mathcal{Z}, \mathbb{H})$ be a measurable space, and let $\phi : \mathcal{X} \times \mathcal{Y} \to \mathcal{Z}$ be a measurable map. Define $Z = \phi(X, Y)$. Then the conditional distribution of Z given X exists and is determined by $(\tilde{P}_x)_{x\in\mathcal{X}}$, where*

$$\tilde{P}_x = (\phi \circ i_x)(P_x)$$

Note that this is not at all surprising: If we know that $X = x$, then we have $Z = \phi(x, Y) = (\phi \circ i_x)(Y)$, and apparently we are allowed to plug the conditional distribution into this formula.

Proof For a fixed $C \in \mathbb{H}$ we have

$$\tilde{P}_x(C) = P_x((\phi \circ i_x)^{-1}(C)) = P_x((\phi^{-1}(C))^x)\,,$$

which is a measurable function of x, since $(P_x)_{x\in\mathcal{X}}$ is a Markov kernel. Hence $(\tilde{P}_x)_{x\in\mathcal{X}}$ is a Markov kernel (each $\tilde{P}_x$ is a probability measure since it is the image measure by the function $\phi \circ i_x$).

Now let $A \in \mathbb{E}$ and $C \in \mathbb{H}$. Then

$$P(X \in A, Z \in C) = (X,Y)(P)((A \times \mathcal{Y}) \cap \phi^{-1}(C))\,.$$

It is seen that if $x \notin A$ then

$$((A \times \mathcal{Y}) \cap \phi^{-1}(C))^x = \emptyset$$

and if $x \in A$ we have

$$((A \times \mathcal{Y}) \cap \phi^{-1}(C))^x = (\phi^{-1}(C))^x = (\phi \circ i_x)^{-1}(C)\,.$$

Hence

$$P(X \in A, Z \in C) = \int_A P_x((\phi \circ i_x)^{-1}(C))\,\mathrm{d}X(P)(x) = \int_A \tilde{P}_x(C)\,\mathrm{d}X(P)(x)\,,$$

which is what we wanted to prove. □

Corollary 1.25 *Assume that X and Y are independent random variables, let $(\mathcal{Z}, \mathbb{H})$ be a measurable space, and let $\phi : \mathcal{X} \times \mathcal{Y} \to \mathcal{Z}$ be a measurable map. Define $Z = \phi(X,Y)$. Then the conditional distribution of Z given X exists and is determined by $(\tilde{P}_x)_{x\in\mathcal{X}}$, where each $\tilde{P}_x$ is given as the distribution of $\phi(x,Y)$.*

Proof This follows directly from Theorem 1.24 since the conditional distribution of Y given X is the constant Markov kernel $(Y(P))_{x\in\mathcal{X}}$. □

Example 1.26 In this example we revisit the multivariate Gaussian distribution and derive the same conditional distribution as in Example 1.20; however, this time we shall not assume that its covariance matrix Σ is positive definite. We assume the covariance matrix is partitioned as in Example 1.20. We recall from Proposition D.5 that in the normal distribution, X_1 and X_2 are independent if and only if $\Sigma_{12} = 0$.

The aim will be to find the conditional distribution of X_1 given X_2. For this define $Z = X_1 - \Sigma_{12}\Sigma_{22}^{-}X_2$ where Σ_{22}^{-} is any g-inverse to Σ_{22}; i.e. any symmetric matrix satisfying $\Sigma_{22}\Sigma_{22}^{-}\Sigma_{22} = \Sigma_{22}$. Note that if Σ_{22} is positive definite, we have $\Sigma_{22}^{-} = (\Sigma_{22})^{-1}$.
We first show that if v satisfies $v^\top\Sigma_{22} = 0$ we must also have $v^\top\Sigma_{21} = 0$; for if this were not the case, we can find u such that $\psi = v^\top\Sigma_{21}u < 0$ and hence if we let $v_\lambda^\top = (u^\top : \lambda v^\top)$ we have

$$v_\lambda^\top \Sigma v_\lambda = u^\top\Sigma_{11}u + 2\lambda\psi$$

which becomes negative when λ is large, contradicting that Σ is positive semidefinite. More generally, if $H\Sigma_{22} = 0$, then also $H\Sigma_{21} = 0$ for any $q \times s$-matrix H. We then have for Z and X_2 that (with e.g. I_r the r–dimensional identity matrix)

$$\begin{pmatrix} Z \\ X_2 \end{pmatrix} = \begin{pmatrix} I_r & -\Sigma_{12}\Sigma_{22}^- \\ 0 & I_s \end{pmatrix} \begin{pmatrix} X_1 \\ X_2 \end{pmatrix} .$$

To find the joint distribution of Z and X_2 we use Proposition D.3 and calculate

$$\begin{aligned} &\begin{pmatrix} I_r & -\Sigma_{12}\Sigma_{22}^- \\ 0 & I_s \end{pmatrix} \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix} \begin{pmatrix} I_r & 0 \\ -\Sigma_{22}^-\Sigma_{21} & I_s \end{pmatrix} \\ =&\begin{pmatrix} I_r & -\Sigma_{12}\Sigma_{22}^- \\ 0 & I_s \end{pmatrix} \begin{pmatrix} \Sigma_{11} - \Sigma_{12}\Sigma_{22}^-\Sigma_{21} & \Sigma_{12} \\ 0 & \Sigma_{22} \end{pmatrix} \\ =&\begin{pmatrix} \Sigma_{11} - \Sigma_{12}\Sigma_{22}^-\Sigma_{21} & 0 \\ 0 & \Sigma_{22} \end{pmatrix} . \end{aligned}$$

Here we have used that

$$\Sigma_{21} - \Sigma_{22}\Sigma_{22}^-\Sigma_{21} = (I_r - \Sigma_{22}\Sigma_{22}^-)\Sigma_{21} = 0$$

since also

$$(I_r - \Sigma_{22}\Sigma_{22}^-)\Sigma_{22} = \Sigma_{22} - \Sigma_{22}\Sigma_{22}^-\Sigma_{22} = 0.$$

Hence we get that

$$\begin{pmatrix} Z \\ X_2 \end{pmatrix} \sim \mathcal{N}_{r+s} \left(\begin{pmatrix} \xi_1 - \Sigma_{12}\Sigma_{22}^-\xi_2 \\ \xi_2 \end{pmatrix}, \begin{pmatrix} \Sigma_{11} - \Sigma_{12}\Sigma_{22}^-\Sigma_{21} & 0 \\ 0 & \Sigma_{22} \end{pmatrix} \right) .$$

From this we see that Z and X_2 are independent and that

$$Z \sim \mathcal{N}_r(\xi_1 - \Sigma_{12}\Sigma_{22}^-\xi_2 \, , \, \Sigma_{11} - \Sigma_{12}\Sigma_{22}^-\Sigma_{21}) \, .$$

Hence this normal distribution is also the conditional distribution of Z given X_2 (Theorem 1.15). Then using the substitution $X_1 = Z + \Sigma_{12}\Sigma_{22}^-X_2$ gives according to Corollary 1.25 and Proposition D.3 that

$$(X_1 \,|\, X_2 = x) \sim \mathcal{N}_r(\xi_1 + \Sigma_{12}\Sigma_{22}^-(x - \xi_2) \, , \, \Sigma_{11} - \Sigma_{12}\Sigma_{22}^-\Sigma_{21})$$

for any g-inverse Σ_{22}^-. □

Example 1.27 Assume that X and Y are real valued variables such that the joint distribution of (X, Y) is a Dirichlet distribution with parameters $(\lambda_1, \lambda_2, \lambda)$. Then the distribution of (X, Y) has density

$$f(x, y) = \frac{\Gamma(\lambda + \lambda_1 + \lambda_2)}{\Gamma(\lambda)\Gamma(\lambda_1)\Gamma(\lambda_2)} \, x^{\lambda_1 - 1} y^{\lambda_2 - 1} (1 - x - y)^{\lambda - 1}$$

on the set $\{(x, y) \in \mathbb{R}^2 \, : \, 0 < x, \, 0 < y, \, x + y < 1\}$. It can be shown that the marginal distribution of X is a B–distribution with parameters $(\lambda_1, \lambda_2 + \lambda)$. Hence it has density

$$g(x) = \frac{\Gamma(\lambda + \lambda_1 + \lambda_2)}{\Gamma(\lambda_1)\Gamma(\lambda_2 + \lambda)} \, x^{\lambda_1 - 1}(1-x)^{\lambda_2 + \lambda - 1}$$

for $x \in (0,1)$. The conditional distribution P_x of Y given $X = x$ for $x \in (0,1)$ must be concentrated on the interval $(0, 1-x)$ and have density

$$f_x(y) = \frac{f(x,y)}{g(x)} = \frac{\Gamma(\lambda_2 + \lambda)}{\Gamma(\lambda)\Gamma(\lambda_2)} \left(\frac{y}{1-x}\right)^{\lambda_2 - 1} \left(1 - \frac{y}{1-x}\right)^{\lambda - 1} \frac{1}{1-x} \, .$$

If P_x is transformed by the map $y \to y/(1-x)$ then a B–distribution with parameters (λ_2, λ) is obtained. According to Theorem 1.24 the constant family consisting of B–distributions with parameters (λ_2, λ) indexed by $x \in (0,1)$ must be the conditional distribution of $Y/(1-X)$ given X. It follows from Theorem 1.15 that $Y/(1-X)$ and X are independent and that $Y/(1-X)$ is B–distributed with parameters (λ_2, λ). □

The following rather deep results is a type of converse to Corollary 1.25 and shows that the functional construction in some sense is a universal representation of a Markov kernel.

Theorem 1.28 *Let X and Y be random variables with values in $(\mathcal{X}, \mathbb{E})$ and $(\mathcal{Y}, \mathbb{K})$. There exists a map $\phi : \mathcal{X} \times (0,1) \to \mathcal{Y}$, which is $\mathbb{E} \otimes \mathbb{B}_{(0,1)} - \mathbb{K}$ measurable, with the following property: if X' is a random variable with the same distribution as X, U is a real valued random variable, independent of X' and uniformly distributed on $(0,1)$, and if we let*

$$Y' = \phi(X', U)$$

then (X', Y') has the same distribution as (X, Y).

Proof Due to the underlying assumption that the spaces involved are Borel spaces, we may assume that $(\mathcal{Y}, \mathbb{K}) = (\mathbb{R}, \mathbb{B})$. Let $(P_x)_{x \in \mathcal{X}}$ be the conditional distribution of Y given X. We know that the conditional distribution of U given X' is degenerate:

$$Q_x = \nu \qquad \text{for all } x \in \mathcal{X} \, ,$$

where ν is the uniform distribution on $(0,1)$. By the substitution theorem, the conditional distribution of Y' given X' is

$$R_x = \phi \circ i_x(Q_x) = \phi \circ i_x(\nu) \, .$$

The proof is complete, once we show how to choose ϕ such that $R_x = P_x$ for every x, as the joint distribution is uniquely determined from one marginal distribution and the conditional distribution of the remaining marginal given the first.
The deep claim is not so much that it is possible to choose ϕ is such a way that

$$\phi \circ i_x(\nu) = P_x \qquad \text{for all } x \in \mathcal{X} \, . \tag{1.9}$$

For if we let F_x be the distribution function corresponding to P_x, and if we let q_x be a quantile function for F_x, it is well known that $q_x(\nu) = P_x$. So we may let

$$\phi(x,u) = q_x(u) \, ,$$

and (1.9) will be satisfied *bona fide.*

What *is* a deep claim is that the construction can be carried out in a way that guarantees ϕ to be measurable. There is a choice involved, in the sense that quantile functions are not unique, and even though the individual quantile functions are increasing, and thus necessarily measurable, the various choices may destroy joint measurability.
The key is to get rid of the choices, and find an operationally defined quantile function. A nice one is

$$q_x(p) = \inf\{y \in \mathbb{R} \,|\, F_x(y) > p\} \qquad \text{for all } x \in \mathcal{X}, p \in (0,1)\,.$$

The idea is to single out the largest possible p-quantile whenever there is a choice. Let us prove that this is in fact a quantile function: For fixed x and p, we have that

$$\{y \in \mathbb{R} \,|\, F_x(y) > p\} = \begin{cases} (y_0, \infty) \\ [y_0, \infty) \end{cases},$$

for some $y_0 \in \mathbb{R}$. Whether we have the open or the halfclosed interval, depends on the specifics of the situation, but in both cases we see that $q_x(p) = y_0$. For each n we have that $y_0 + \frac{1}{n} > y_0$, and thus

$$F_x\left(y_0 + \frac{1}{n}\right) > p\,.$$

Using right continuity of F_x, we can conclude that

$$F_x(y_0) \geq p\,.$$

Similarly, $y_0 - \frac{1}{n} < y_0$, and so

$$F_x\left(y_0 - \frac{1}{n}\right) \leq p\,.$$

Using monotonicity of F_x, we can conclude that

$$F_x(y_0-) \leq p\,.$$

Together these inequalities show that y_0 is a p-quantile for F_x. As for measurability, an elementary argument shows that

$$\{(x,p) \,|\, q_x(p) < z\} = \bigcup_{w<z, w\in\mathbb{Q}} \{(x,p) \,|\, F_x(w) > p\}\,. \tag{1.10}$$

For any fixed w, the map

$$x \mapsto F_x(w) = P_x\Big((-\infty, w]\Big)$$

is measurable, as $(P_x)_{x\in\mathcal{X}}$ is a Markov kernel. Hence $(x,p) \mapsto (F_x(w), p)$ is measurable, and thus

$$\{(x,p) \,|\, F_x(w) > p\} = \{(x,p) \,|\, F_x(w) - p > 0\}$$

is a measurable set. The fact that the right hand side of (1.10) is a countable union, shows that the left hand side is a measurable set. □

The point of Theorem 1.28 is that we may think of as any pair of variables as generated in a two-step procedure, where the generation of the second variable can be accomplished by mixing the first variable with random noise. It is the way that the mixing is carried out, that determines the joint distribution.
The *update function* ϕ is not at all unique. There are literally uncountably many ways to choose it. In certain cases it matters which one we use, in most cases it is irrelevant. However, in typical applications there is a specific update function that almost forces itself upon us.

We conclude the section by identifying some cases where the structure of conditional distribution simplifies.

Theorem 1.29 *Assume that $(P_x)_{x\in\mathcal{X}}$ is the conditional distribution of Y given X. Let $(\mathcal{Z},\mathbb{H})$ be a measurable space and let $t:\mathcal{X}\to\mathcal{Z}$ be an $\mathbb{E}-\mathbb{H}$–measurable map. Define $Z=t(X)$. Then the conditional distribution $(Q_{x,z})_{(x,z)\in\mathcal{X}\times\mathcal{Z}}$ of Y given (X,Z) is given by*

$$Q_{x,z}=P_x \qquad \textit{for all } x\in\mathcal{X}, z\in\mathcal{Z} \tag{1.11}$$

Note: This is a situation where it is quite clear that conditional distributions are not uniquely determined. The variable (X,Z) does not have values in the entire product space $\mathcal{X}\times\mathcal{Z}$ but only on the *graph of* t, meaning the set of points

$$\{(x,z)\in\mathcal{X}\times\mathcal{Z}\,:\,z=t(x)\}$$

Then $Q_{x,z}$ could be defined as any probability measure outside the graph, if only some measurability conditions are fulfilled. Hence the Markov kernel defined in (1.11) is not the only possible conditional distribution of Y given (X,Z) – it is simply a convenient choice.

Proof It is easily argued that a $(\mathcal{X}\times\mathcal{Z},\mathbb{E}\otimes\mathbb{H})$–Markov kernel $(Q_{x,z})_{(x,z)\in\mathcal{X}\times\mathcal{Z}}$ on $(\mathcal{Y},\mathbb{K})$ is defined by (1.11). For $A\in\mathbb{E}$, $B\in\mathbb{K}$ and $C\in\mathbb{H}$ we have

$$\begin{aligned}\int_{A\times C} Q_{x,z}(B)\,\mathrm{d}(X,Z)(P)(x,z) &= \int 1_{A\times C}(x,z)Q_{x,z}(B)\,\mathrm{d}((\mathrm{id},t)\circ X)(P)(x,z)\\ &= \int 1_{A\times C}\circ(\mathrm{id},t)(x)\,Q_{(\mathrm{id},t)(x)}(B)\,\mathrm{d}X(P)(x)\\ &= \int 1_{A\cap t^{-1}(C)}(x)P_x(B)\,\mathrm{d}X(P)(x)\,.\end{aligned}$$

Since $(P_x)_{x\in\mathcal{X}}$ is the conditional distribution of Y given X, the last integral can be identified as

$$\begin{aligned}P(X\in A\cap t^{-1}(C),Y\in B) &= P(X\in A,Z\in C,Y\in B)\\ &= P((X,Z)\in A\times C,Y\in B)\,.\end{aligned}$$

By fixing B and letting $A\times C$ vary it is obtained (by uniqueness of measures) that

$$\int_G Q_{x,z}(B)\,\mathrm{d}(X,Z)(P)(x,z) = P((X,Z)\in G, Y\in B)$$

for all $G \in \mathbb{E}\otimes\mathbb{H}$ and all $B\in\mathbb{K}$. Hence it is concluded that $(Q_{x,z})_{(x,z)\in\mathcal{X}\times\mathcal{Z}}$ is the conditional distribution of Y given (X,Z). □

Theorem 1.30 *Let $(P_x)_{x\in\mathcal{X}}$ be the conditional distribution of Y given X. Let $(\mathcal{Z},\mathbb{H})$ be a measurable space and let $t:\mathcal{X}\to\mathcal{Z}$ be an $\mathbb{E}-\mathbb{H}$–measurable map. Define $Z=t(X)$. If an $(\mathcal{Z},\mathbb{H})$–Markov kernel $(Q_z)_{z\in\mathcal{Z}}$ on $(\mathcal{Y},\mathbb{K})$ exists such that*

$$P_x = Q_{t(x)} \qquad \textit{for all } x\in\mathcal{X}\,,$$

then $(Q_z)_{z\in\mathcal{Z}}$ is the conditional distribution of Y given Z

A more relaxed formulation of this is that if the conditional distribution of Y given X only depends on X through $t(X)$, then this is also the conditional distribution of Y given $t(X)$.

Proof Let $C\in\mathbb{H}$ and $B\in\mathbb{K}$. According to the change–of–variable theorem we have

$$\begin{aligned}
P(Z\in C, Y\in B) &= P(X\in t^{-1}(C), Y\in B)\\
&= \int 1_{t^{-1}(C)}(x) P_x(B)\,\mathrm{d}X(P)(x)\\
&= \int 1_C\circ t(x) Q_{t(x)}(B)\,\mathrm{d}X(P)(x)\\
&= \int 1_C(z) Q_z(B)\,\mathrm{d}(t\circ X)(P)(z)\\
&= \int_C Q_z(B)\,\mathrm{d}Z(P)(z)\,.
\end{aligned}$$

Hence $(Q_z)_{z\in\mathcal{Z}}$ is the conditional distribution of Y given Z. □

1.6 Conditional moments

Clearly we can consider moments of random variables w.r.t. a conditional distribution. For the expectation we have the following formal definition of a pointwise conditional expectation.

Definition 1.31 Assume that X and Y are random variables defined on $(\Omega,\mathbb{F},P)$ with values in $(\mathcal{X},\mathbb{E})$ and $(\mathbb{R},\mathbb{B})$. Let $(P_x)_{x\in\mathcal{X}}$ be the conditional distribution of Y given X. Assume that for some $x\in\mathcal{X}$ it holds that

$$\int |y|\,\mathrm{d}P_x(y) < \infty$$

then we define *the conditional expectation of Y given $X=x$* as

$$E(Y\mid X=x) = \int y\,\mathrm{d}P_x(y)$$

A related and useful concept is that of a conditional expectation (operator) given a σ-algebra. In contrast to the conditional expectation defined pointwise for a specific x in Definition 1.31, this is a global definition and demands that the unconditional expectation exists.

Definition 1.32 Let X be a real random variable defined on $(\Omega, \mathbb{F}, P)$ with $E|X| < \infty$ and $\mathbb{D}$ be a sub σ–algebra of $\mathbb{F}$. A *conditional expectation of X given $\mathbb{D}$* is any real random variable denoted $E(X \mid \mathbb{D})$ that satisfies

1) $E(X \mid \mathbb{D})$ is $\mathbb{D}$–measurable
2) $E|E(X \mid \mathbb{D})| < \infty$
3) For all $D \in \mathbb{D}$ it holds that

$$\int_D E(X \mid \mathbb{D}) \, \mathrm{d}P = \int_D X \, \mathrm{d}P \,. \tag{1.12}$$

Conditional expectations are almost unique:

Theorem 1.33 *If U and $\tilde{U}$ are both conditional expectations of X given $\mathbb{D}$, then $U = \tilde{U}$ a.s. Further, If U is a conditional expectation of X given $\mathbb{D}$ and $\tilde{U}$ is $\mathbb{D}$-measurable with $\tilde{U} = U$ a.s., then $\tilde{U}$ is also a conditional expectation of X given $\mathbb{D}$.*

Proof If U and $\tilde{U}$ are both conditional expectations, we let $A_\epsilon = \{U - \tilde{U} > \epsilon\}$ which is clearly $\mathbb{D}$-measurable and thus we have from (1.12) that for all $\epsilon > 0$ it holds that

$$0 = \int_{A_\epsilon} (U - \tilde{U}) \, \mathrm{d}P \geq \epsilon P(A_\epsilon) \geq 0$$

and hence we have $P(A_\epsilon) = 0$; similarly we conclude that $P(B_\epsilon) = 0$ where $B_\epsilon = \{\tilde{U} - U > \epsilon\}$ and hence $U = \tilde{U}$ almost surely.
Further, if U is a conditional expectation and $U = \tilde{U}$ a.s. with $\tilde{U}$ $\mathbb{D}$-measurable, we have $E|\tilde{U}| < \infty$ and get for $D \in \mathbb{D}$ that

$$\int_D \tilde{U} \, \mathrm{d}P = \int_D U \, \mathrm{d}P = \int_D X \, \mathrm{d}P$$

and hence $\tilde{U}$ is also a conditional expectation. □

Finally, such a conditional expectation always exists, even in general (non-Borel) probability spaces. However, as this is a non-elementary result, we shall refrain from proving this here. See for example Chapter 23 of Schilling (2005) for further details.

Theorem 1.34 *If X is a real random variable with $E|X| < \infty$, then there exists a conditional expectation of X given $\mathbb{D}$.*

Furthermore we have a series of nice properties.

Theorem 1.35 *Let X, X_n and Y be real random variables, all of which are integrable. It then holds that*

(a) *If $X = c$ a.s., where $c \in \mathbb{R}$ is a constant, then $E(X|\mathbb{D}) = c$ a.s.*

(b) *For $\alpha, \beta \in \mathbb{R}$ it holds that*

$$E(\alpha X + \beta Y|\mathbb{D}) = \alpha E(X|\mathbb{D}) + \beta E(Y|\mathbb{D}) \ a.s.$$

(c) *If $X \geq 0$ $a.s.$ then $E(X|\mathbb{D}) \geq 0$ $a.s.$ If $Y \geq X$ $a.s.$ then $E(Y|\mathbb{D}) \geq E(X|\mathbb{D})$ $a.s.$*

(d) *If $\mathbb{D} \subseteq \mathbb{E}$ are sub σ-algebras of $\mathbb{F}$ then*

$$E(X|\mathbb{D}) = E[E(X|\mathbb{E})|\mathbb{D}] = E[E(X|\mathbb{D})|\mathbb{E}] \ a.s.$$

(e) *If $\sigma(X)$ and $\mathbb{D}$ are independent then*

$$E(X|\mathbb{D}) = EX \ \ a.s.$$

(f) *If X is $\mathbb{D}$-measurable then*

$$E(X|\mathbb{D}) = X \ \ a.s.$$

(g) *If it holds for all $n \in \mathbb{N}$ that $X_n \geq 0$ a.s. and $X_{n+1} \geq X_n$ a.s. with $\lim X_n = X$ a.s., then*

$$\lim_{n\to\infty} E(X_n|\mathbb{D}) = E(X|\mathbb{D}) \ \ a.s.$$

(h) *If X is $\mathbb{D}$-measurable and $E|XY| < \infty$, then*

$$E(XY|\mathbb{D}) = X\, E(Y|\mathbb{D}) \ \ a.s.$$

(i) *If $f : \mathbb{R} \to \mathbb{R}$ is a measurable function that is convex on an interval I, such that $P(X \in I) = 1$ and $E|f(X)| < \infty$, then it holds that*

$$f\big(E(X|\mathbb{D})\big) \leq E\big(f(X)|\mathbb{D}\big) \ \ a.s.$$

Proof The proof is left as Exercise 1.16. □

Now assume that X is a random variable with values in $(\mathcal{X}, \mathbb{E})$ and that Y is a real random variable with $E|Y| < \infty$. We shall then write the conditional expectation of Y given $\mathbb{D} = \sigma(X)$ in short as $E(Y \mid X)$ rather than $E(Y \mid \sigma(X))$. The resulting random variable is referred to as the conditional expectation of Y given X — as opposed to the conditional expectation of Y given $X = x$.

The fact that $E(Y \mid X)$ is $\sigma(X)$–measurable is equivalent to the existence of a measurable map $\phi : (\mathcal{X}, \mathbb{E}) \to (\mathbb{R}, \mathbb{B})$ such that

$$E(Y \mid X) = \phi(X) \quad P \text{ almost surely}$$

and the next theorem gives that ϕ and $x \mapsto E(Y \mid X = x)$ are almost identical if Y has finite expectation. In words, the theorem says that any conditional expectation is almost surely equal to the expectation in the conditional distribution.

Theorem 1.36 *Assume that X and Y are random variables defined on $(\Omega, \mathbb{F}, P)$ and with values in $(\mathcal{X}, \mathbb{E})$ and $(\mathbb{R}, \mathbb{B})$ respectively. Let $(P_x)_{x\in\mathcal{X}}$ be the conditional distribution of Y given X.*

If $E|Y| < \infty$, then $X(P)(A_0) = 1$, where

$$A_0 = \{x \in \mathcal{X} \,:\, \int |y| \,\mathrm{d}P_x(y) < \infty\}\,.$$

Define $\phi : \mathcal{X} \to \mathbb{R}$ by $\phi(x) = 1_{A_0}(x)E(Y \mid X = x)$. Then $\phi(X)$ is a version of the conditional expectation of Y given X.

Proof Consider the function $f : \mathcal{X} \times \mathbb{R} \to \mathbb{R}$ given by $f(x, y) = y$. Since

$$\int |f(x,y)| \,\mathrm{d}(X,Y)(P)(x,y) = \int |f(X,Y)| \,dP = E|Y| < \infty\,,$$

it follows from the extended Fubini, that

$$\int |y| \,\mathrm{d}P_x(y) = \int |f(x,y)| \,\mathrm{d}P_x(y) < \infty$$

for $X(P)$ almost all $x \in \mathcal{X}$, such that $X(P)(A_0) = 1$.

Since $\phi : \mathcal{X} \to \mathbb{R}$ is defined by

$$\phi(x) = \begin{cases} \int y \,\mathrm{d}P_x(y)\,, & x \in A_0 \\ 0\,, & x \notin A_0 \end{cases}$$

then it is $\mathbb{E} - \mathbb{B}$–measurable according to the extended Fubini. We will argue, that $\phi(X)$ is a conditional expectation of Y given X by verifying the conditions 1)–3) in Definition 1.32. Since ϕ is measurable we have that $\phi(X)$ is $\sigma(X)$–measurable. Furthermore — using the change-of-variable theorem, Theorem A.8 — we get

$$\begin{aligned} E|\phi(X)| &= \int |\phi(X)| \,dP \\ &= \int |\phi(x)| \,\mathrm{d}X(P)(x) \\ &= \int_{A_0} \left| \int y \,\mathrm{d}P_x(y) \right| \mathrm{d}X(P)(x) \\ &\leq \int \left(\int |y| \mathrm{d}P_x(y) \right) \mathrm{d}X(P)(x) \\ &= \int |y| \,\mathrm{d}(X,Y)(P)(x,y) < \infty \end{aligned}$$

In the last equality we have used extended Tonelli. This shows, that 2) is satisfied for $\phi(X)$. Finally we have for $A \in \mathbb{E}$

$$\begin{aligned}\int_{(X\in A)} \phi(X)\, dP &= \int_A \phi(x)\, \mathrm{d}X(P)(x) \\ &= \int_{A\cap A_0} \int y\, \mathrm{d}P_x(y)\, \mathrm{d}X(P)(x) \\ &= \iint 1_{A\cap A_0}(x) y\, \mathrm{d}P_x(y)\, \mathrm{d}X(P)(x) \\ &= \int 1_{A\cap A_0}(x) y\, \mathrm{d}(X,Y)(P)(x,y) \\ &= \int 1_{A\cap A_0}(X) Y\, dP \\ &= \int_{(X\in A\cap A_0)} Y\, dP \\ &= \int_{(X\in A)} Y\, dP\end{aligned}$$

In the fourth equality we have used the extended Fubini's theorem. This shows that also 3) is fulfilled, such that $\phi(X)$ is a conditional expectation of Y given X. □

The following result can be shown using the proof of Theorem 1.36:

Theorem 1.37 *Assume that X and Y are random variables defined on $(\Omega, \mathbb{F}, P)$ with values in $(\mathcal{X}, \mathbb{E})$ and $(\mathbb{R}, \mathbb{B})$ respectively. If $E|Y| < \infty$, then $E|E(Y \mid X)| < \infty$ and*

$$E(E(Y \mid X)) = EY$$

Proof In the proof of Theorem 1.36 we saw that $\phi(X) = E(Y \mid X)$ is integrable, and that

$$\int_{(X\in A)} \phi(X)\, \mathrm{d}P = \int_{(X\in A)} Y\, \mathrm{d}P$$

So for $A = \mathcal{X}$ we get

$$E(E(Y \mid X)) = \int \phi(X)\, \mathrm{d}P = \int Y\, \mathrm{d}P = EY$$

which completes the proof. □

Theorem 1.38 *Assume that X and Y are random variables defined $(\Omega, \mathbb{F}, P)$ with values in $(\mathcal{X}, \mathbb{E})$ and $(\mathcal{Y}, \mathbb{K})$ respectively. Suppose that the conditional distribution $(P_x)_{x\in\mathcal{X}}$ of Y given X exists. Let $\phi : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ be a measurable function and define $Z = \phi(X, Y)$. Assume that $E|Z| < \infty$. Then*

$$E(Z \mid X = x) = \int \phi(x, y)\, \mathrm{d}P_x(y)$$

for $X(P)$ almost all $x \in \mathcal{X}$.

Proof According to Theorem 1.24 we have that the conditional distribution of Z given X is given by the Markov kernel $(\tilde{P}_x)_{x\in\mathcal{X}}$, where

$$\tilde{P}_x = (\phi \circ i_x)(P_x)$$

Then according to Theorem 1.36 we have for $X(P)$ almost all $x \in \mathcal{X}$

$$E(Z \mid X = x) = \int z \, \mathrm{d}\tilde{P}_x(z) = \int (\phi \circ i_x)(y) \, \mathrm{d}P_x(y) = \int \phi(x, y) \, \mathrm{d}P_x(y)$$

which was to be shown. □

Corollary 1.39 *Assume that X and Y are independent random variables defined $(\Omega, \mathbb{F}, P)$ with values in $(\mathcal{X}, \mathbb{E})$ and $(\mathcal{Y}, \mathbb{K})$ respectively. Let $\phi : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ be a measurable function and define $Z = \phi(X, Y)$. Assume that $E|\phi(X, Y)| < \infty$. Then*

$$E(\phi(X, Y) \mid X = x) = \int \phi(x, y) \, \mathrm{d}Y(P)(y)$$

for $X(P)$ almost all $x \in \mathcal{X}$.

Proof $(Y(P))_{x\in\mathcal{X}}$ is the conditional distribution of Y given X. □

We can also use Theorem 1.38 to show the following.

Corollary 1.40 *Assume that Y and Z are real valued random variables with $E|Y| < \infty$ and $E|Z| < \infty$. Let X be a random variable with values in $(\mathcal{X}, \mathbb{E})$. Then*

$$E(Y + Z \mid X = x) = E(Y \mid X = x) + E(Z \mid X = x)$$

for $X(P)$ almost all $x \in \mathcal{X}$.

Proof Let $(P_x)_{x\in\mathcal{X}}$ be the conditional distribution of (Y, Z) given X. Then

$$E(Y + Z \mid X = x) = \int (y + z) \, \mathrm{d}P_x(y, z)$$

for $X(P)$ almost all $x \in \mathcal{X}$. And

$$E(Y \mid X = x) = \int y \, \mathrm{d}P_x(y, z) \qquad E(Z \mid X = x) = \int z \, \mathrm{d}P_x(y, z)$$

for $X(P)$ almost all $x \in \mathcal{X}$. □

If Y is real valued with $EY^2 < \infty$, and $(P_x)_{x\in\mathcal{X}}$ is the conditional distribution of Y given X, then we can define the conditional variance of Y given $X = x$ by

$$V(Y \mid X = x) = \int y^2 \, \mathrm{d}P_x(y) - \left(\int y \, \mathrm{d}P_x(y) \right)^2$$

which will be well–defined for $X(P)$ almost all $x \in \mathcal{X}$. Letting $V(Y \mid X)$ be the composition of X and $x \mapsto V(Y \mid X = x)$ gives

$$V(Y \mid X) = E(Y^2 \mid X) - E(Y \mid X)^2$$

Theorem 1.41 *Let X and Y be random variables defined on $(\Omega, \mathbb{F}, P)$ with values in $(\mathcal{X}, \mathbb{E})$ and $(\mathbb{R}, \mathbb{B})$ respectively. If $EY^2 < \infty$, then*

$$VY = E\big(V(Y \mid X)\big) + V\big(E(Y \mid X)\big)\,.$$

Proof We use the calculation

$$\begin{aligned} E\big(V(Y \mid X)\big) + V\big(E(Y \mid X)\big) &= E\big(E(Y^2 \mid X) - E(Y \mid X)^2\big) \\ &\quad + E\big(E(Y \mid X)^2\big) - \big(E\big(E(Y \mid X)\big)\big)^2 \\ &= E\big(E(Y^2 \mid X)\big) - \big(E\big(E(Y \mid X)\big)\big)^2 \end{aligned}$$

as required. □

Example 1.42 In example 1.26 we studied the situation where

$$\begin{pmatrix} X_1 \\ X_2 \end{pmatrix} \sim \mathcal{N}_{r+s}\left(\begin{pmatrix} \xi_1 \\ \xi_2 \end{pmatrix}, \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}\right),$$

and we found that

$$X_1 \mid X_2 = x \sim \mathcal{N}_r(\xi_1 + \Sigma_{12}\Sigma_{22}^{-1}(x - \xi_2)\,,\, \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21})$$

If we assume that X_1 is one–dimensional, we have defined $E(X_1 \mid X_2 = x)$ and $V(X_1 \mid X_2 = x)$. Since conditional expectations and conditional variances are calculated as expectations and variances in the conditional distributions, we have

$$\begin{aligned} E(X_1 \mid X_2 = x) &= \xi_1 + \Sigma_{12}\Sigma_{22}^{-1}(x - \xi_2) \\ V(X_1 \mid X_2 = x) &= \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21} \end{aligned}$$

Note: the conditional variance does not depend on x but is different from $V(X_1)$. □

Confusing conditional variances and ordinary variances is a quite common mistake – and that may lead to substantial problems.

1.7 Exercises

Exercise 1.1 Assume that X_1 and X_2 are independent random variables that are both binomially distributed with parameters (n, p). Define the random variable $X = X_1 + X_2$. Find the conditional distribution of X_1 given X.

Exercise 1.2 Let X and Y be random variables defined on $(\Omega, \mathbb{F}, P)$. Assume that

(a) X has the binomial distribution with parameters (n, p_1)

(b) The conditional distribution of Y given $X = x$ is binomial with parameters (n, p_2)

Find the marginal distribution of Y and try to give an intuitive explanation of the result.

Exercise 1.3 Assume that X and Y are two random variables defined on $(\Omega, \mathbb{F}, P)$ with values in $(\mathcal{X}, \mathbb{E})$ and $(\mathcal{Y}, \mathbb{K})$ respectively, and let $(P_x)_{x\in\mathcal{X}}$ be the conditional distribution of Y given X. Let $f : \mathcal{X} \times \mathcal{Y} \to [0, \infty)$ be an $\mathbb{E} \otimes \mathbb{K}$–measurable function.

(a) Show that

$$E[f(X, Y)] = \iint f(x, y)\, \mathrm{d}P_x(y)\, \mathrm{d}X(P)(x).$$

(b) Assume that X is uniformly distributed on $(0, 1)$. Assume that the conditional distribution $(P_x)_{x\in(0,1)}$ of Y given X is the exponential distribution with mean value x. Find EY.

Exercise 1.4 Let $\mathcal{Y}$ be a finite or countable set, and let $\mathbb{K}$ consist of all subsets of $\mathcal{Y}$. Assume that Y is a random variable defined on $(\Omega, \mathbb{F}, P)$ with values in $(\mathcal{Y}, \mathbb{K})$. Let $p(y)$ denote the probability function for Y. Let $\mathcal{X}$ be another finite or countable set, and assume that $t : \mathcal{Y} \to \mathcal{X}$ is some map. Define $X = t(Y)$.

(a) Show that X has probability function

$$r(x) = \sum_{y\in t^{-1}(x)} p(y)$$

(b) Show that $(P_x)_{x\in\mathcal{X}}$ is the conditional distribution of Y given X, where each P_x has probability function

$$q_x(y) = \frac{p(y)1_{\{x\}}(t(y))}{r(x)}$$

Exercise 1.5 Let $Y_1, \ldots, Y_n$ be independent and identically distributed random variables defined on $(\Omega, \mathbb{F}, P)$ with values in $\{0, 1\}$. Assume that

$$P(Y_1 = 0) = 1 - p\,, \qquad P(Y_1 = 1) = p$$

for some $0 < p < 1$. Define $t : \{0, 1\}^n \to \{0, 1, \ldots, n\}$ by

$$t(y_1, \ldots, y_n) = y_1 + \cdots + y_n$$

Define $X = t(Y_1, \ldots, Y_n)$.

(a) Realise that X has the binomial distribution with parameters (n, p) and argue that $P(X = x) > 0$ for all $x = 0, 1, \ldots, n$.

(b) Show that $(P_x)_{x=0,\ldots,n}$ is the conditional distribution of $Y = (Y_1, \ldots, Y_n)$ given X, where P_x is the uniform distribution on $\{(y_1, \ldots, y_n) \in \{0,1\}^n : y_1 + \cdots + y_n = x\}$.

Exercise 1.6 Let X_1, X_2 and X_3 be independent random variables, where each X_i has a Poisson distribution with parameter λ_i. Define $X = X_1 + X_2 + X_3$ and show that the conditional distribution of (X_1, X_2, X_3) given X is given by $(P_x)_{x\in\mathbb{N}_0}$, where each P_x (for $x > 0$) is a multinomial distribution with parameters x and $(\lambda_1/\lambda, \lambda_2/\lambda, \lambda_3/\lambda)$, where $\lambda = \lambda_1 + \lambda_2 + \lambda_3$: For all (x_1, x_2, x_3) with $x_1 + x_2 + x_3 = x$ it holds that

$$P_x\{(x_1, x_2, x_3)\} = \frac{x!}{x_1!x_2!x_3!} \left(\frac{\lambda_1}{\lambda}\right)^{x_1} \left(\frac{\lambda_2}{\lambda}\right)^{x_2} \left(\frac{\lambda_2}{\lambda}\right)^{x_2}$$

It may be useful to note that X is Poisson distributed with parameter λ.

Exercise 1.7 Let X and Y be random variables defined on $(\Omega, \mathbb{F}, P)$. Assume that

(a) X has the binomial distribution with parameters (n, p_1).

(b) The conditional distribution of Y given $X = x$ is binomial with parameters (x, p_2).

Find the marginal distribution of Y and try to give an intuitive explanation of the result.

Exercise 1.8 Let X and Y be random variables with values in $(\mathcal{X}, \mathbb{E})$ and $(\mathcal{Y}, \mathbb{K})$ respectively, such that $(P_x)_{x\in\mathcal{X}}$ is the conditional distribution of Y given X. Assume that μ and ν are σ–finite measures on $(\mathcal{X}, \mathbb{E})$ and $(\mathcal{Y}, \mathbb{K})$. Assume furthermore that $X(P)$ has density f with respect to μ, and that for each $x \in \mathcal{X}$ the probability P_x has density g_x with respect to ν, such that $(x, y) \mapsto g_x(y)$ is $\mathbb{E} \otimes \mathbb{K} - \mathbb{B}$–measurable.

(a) Show that

$$\ell(y) = \int g_x(y) f(x) \, \mathrm{d}\mu(x)$$

is the density for the marginal distribution of Y with respect to ν.

(b) Show that $Y(P)(B_0) = 1$, where $B_0 = \{y \in \mathcal{Y} : 0 < \ell(y) < \infty\}$.

(c) Show that the conditional distribution of X given Y exists and is given by $(Q_y)_{y\in\mathcal{Y}}$, where Q_y has density with respect to μ given by

$$k_y(x) = \frac{g_x(y) f(x)}{\ell(y)}$$

for $y \in B_0$.

Exercise 1.9 Assume that X is Gamma–distributed with parameters (λ, β) and that the conditional distribution of Y given X is given by $(P_x)_{x\in\mathbb{R}}$, where P_x is the Poisson distribution with parameter x.

(a) Show that the marginal distribution of Y is a negative binomial distribution and find the parameters.

(b) Show that the conditional distributions of X given Y are Γ–distributions.

Exercise 1.10 Let X and Y be real valued random variables defined on $(\Omega, \mathbb{F}, P)$. Let $C \in \mathbb{B}$ be a fixed subset of $\mathbb{R}$. Consider the following game: We are told the value of X, and are based on this information supposed to guess whether $Y \in C$ or not.

It seems natural to expect that we in two different games, where the same value of X is observed, give the same guess of whether $Y \in C$ or not – we know the same in the two situations. Hence giving a rule for guessing must be the same as indicating a set A: If we observe $X \in A$ then we guess that $Y \in C$, and if we observe $X \notin A$, then we guess that $Y \notin C$.

Obviously, different choices of A may lead to more or less successful guessing rules (we define a guessing rule to be successful, if it often leads to the right guess...). Let $(P_x)_{x\in\mathbb{R}}$ be the conditional distribution of Y given X.

(a) Show that for a given guessing rule, then

$$P(\text{right guess}) = \int_A P_x(C)\,\mathrm{d}X(P)(x) + \int_{A^c} P_x(C^c)\,\mathrm{d}X(P)(x)$$

(b) Show that the optimal guessing rule corresponds to the set

$$A_0 = \{x \in \mathbb{R} \,:\, P_x(C) \geq \frac{1}{2}\}\,.$$

(c) How is the optimal guessing rule, if X and Y are independent?

(d) How is the optimal guessing rule, if $X = Y$?

Exercise 1.11 Let X be a random variable with values in $(\mathcal{X}, \mathbb{E})$ that is defined on a probability space $(\Omega, \mathbb{F}, P)$. Let furthermore $F \in \mathbb{F}$ and consider the random variable 1_F.

(a) Find the Markov kernel $(P_z)_{z\in\{0,1\}}$ that is the conditional distribution of X given 1_F.

(b) Find the Markov kernel $(Q_x)_{x\in\mathcal{X}}$ that is the conditional distribution of 1_F given X.

Exercise 1.12 Assume that X is uniformly distributed on $(0, 1)$ and that the conditional distribution of Y given $X = x$ is a binomial distribution with parameters (n, x) We could say that Y has a binomial distribution with fixed length n and random probability parameter.

(a) What are the possible values of Y? Argue that $E|Y| < \infty$.

(b) Find $E(Y \mid X = x)$ and $E(Y \mid X)$.

(c) Find EY.

(d) Find $P(Y = k)$ for all k being a possible value of Y. What is the marginal distribution of Y?

Exercise 1.13 Let X and Y be random variables with values in $(\mathcal{X}, \mathbb{E})$ and $(\mathcal{Y}, \mathbb{K})$ respectively. Assume that (P_x) is the conditional distribution of Y given X. Let

$$A_0 = \{x \in \mathcal{X} \mid \int |y|\,\mathrm{d}P_x(y) < \infty\}$$

and assume that $X(P)(A_0) = 1$. Define

$$\phi(x) = 1_{A_0}(x) \int |y| \, \mathrm{d}P_x(y) \, .$$

Show that $E\phi(X) = E|Y|$ and conclude that

$$E\phi(X) < \infty \quad \text{if and only if} \quad E|Y| < \infty.$$

Exercise 1.14 Assume that X has the exponential distribution with mean 1, and assume that the conditional distribution of Y given $X = x$ is a Poisson distribution with parameter x. We could say that Y is Poisson distributed with random parameter.

(a) Use Exercise 1.13 to argue that $E|Y| < \infty$.

(b) Find $E(Y \mid X = x)$ and $E(Y \mid X)$.

(c) Find EY.

(d) Find $P(Y = k)$ for all k being a possible value of Y. What is the marginal distribution of Y?

Exercise 1.15 Let X and Y be independent random variables that both have the uniform distribution on $(0, 1)$. Define $Z = XY$.

(a) Find the conditional distribution of Z given X.

(b) What are the possible values of Z? Argue that $E|Z| < \infty$.

(c) Find $E(Z \mid X)$ and use this to find EZ.

(d) Find EZ without using conditional distributions.

Exercise 1.16 Prove Theorem 1.35.

Exercise 1.17 Let $A \in O_d^+$ be a matrix of a rotation in $\mathbb{R}^d$, i.e. $AA^\top = I_d$ and $\det(A) = 1$. A random variable U with values in $\mathbb{R}^d$ is said to have a *uniform distribution on the sphere* rS_{d-1} with radius r if for any such A, $V = AU$ has the same distribution as U, and $P(U/r \in S_{d-1}) = 1$, where

$$S_{d-1} = \{x \in \mathbb{R}^d : ||x|| = 1\}$$

is the unit sphere in $\mathbb{R}^d$.

(a) Let $X \sim \mathcal{N}_d(0, \sigma^2 I_d)$ and define $U = X/||X||$. Show that U is uniform on S_{d-1};

(b) Use this fact to show that the conditional distribution of X given $||X|| = r$ is uniform on rS_{d-1}.

Exercise 1.18 The spherical coordinates of a point $x = (x_1, x_2, x_3) \in \mathbb{R}^3 \setminus \{0\}$ are (ρ, ϕ, θ) determined as

$$x_1 = r \sin\phi \cos\theta, \quad x_2 = r \sin\phi \sin\theta, \quad x_3 = r \cos\phi,$$

where $r > 0$, $\phi \in [0, \pi)$, and $\theta \in [0, 2\pi)$. Now let (R, F, T) denote the spherical coordinates of X, where $X \sim \mathcal{N}_3(0, I_d)$.

(a) Find the joint density of (R, F, T) w.r.t. Lebesgue measure on $\mathbb{R}^+ \times [0, \pi) \times [0, 2\pi)$;

(b) Find the conditional distribution of (F, T) given R and show that these are independent;

(c) Deduce that if (F, T) has density $g(\phi, \theta) = \sin\phi/(4\pi)$ w.r.t. Lebesgue measure on $[0, \pi) \times [0, 2\pi)$, then

$$Y = (\sin F \cos T, \sin F \sin T, \cos T)$$

is uniform on S_2.

(d) Find the conditional distribution of T given $F = \pi/2$;

(e) Find the conditional distribution of F given $T = 0$;

(f) Comment on the result, which is closely related to the Borel–Kolmogorov paradox.

Exercise 1.19 Assume that $Y_1, Y_2, \ldots$ is a sequence of independent and identically distributed random variables such that $E|Y_1| < \infty$. Assume that N is a random variables with values in $\mathbb{N}$ such that $EN < \infty$. Assume that N and $(Y_1, Y_2, \ldots)$ are independent (we consider $(Y_1, Y_2, \ldots)$ as a random variable with values in $(\mathbb{R}^\infty, \mathbb{B}^\infty)$). Define the random variable Y by

$$Y = \sum_{k=1}^{N} Y_k$$

(a) Show that the conditional distribution $(P_n)_{n\in\mathbb{N}}$ of Y given N is determined such that P_n is the distribution of $\sum_{k=1}^{n} Y_k$. Argue similarly that the conditional distribution $(Q_n)_{n\in\mathbb{N}}$ of $\sum_{k=1}^{N} |Y_k|$ given N is determined such that Q_n is the distribution of $\sum_{k=1}^{n} |Y_k|$.

(b) Show that

$$\int |y| \, \mathrm{d}Q_n(y) = nE|Y_1|$$

for all $n \in \mathbb{N}$.

(c) Use (2) and Exercise 1.13 to obtain that

$$E\left(\sum_{k=1}^{N} |Y_k|\right) = ENE|Y_1| < \infty$$

(d) Show that $E(Y \mid N = n) = nEY_1$ and that $EY = ENEY_1$.

Exercise 1.20 Let X and Z be independent and exponentially distributed with expectation $\mathbf{E}X = \mathbf{E}Z = 1$ and define

$$Y = \phi(X, Z) = \min(X, Z).$$

a) Show that also Y is exponentially distributed and find $\mathbf{E}Y$;

b) Show that the conditional distribution of Y given $X = x$ is given by the Markov kernel $P_x, x > 0$ where

$$P_x(Y \le y) = \begin{cases} 1 - e^{-y} & \text{if } y < x \\ 1 & \text{if } y \ge x \end{cases}.$$

c) Identify the joint distribution of (X, Y), for example by specifying its joint distribution function.

d) Find the conditional distribution of X given $Y = y$.

Exercise 1.21 Let f and g be densities for distributions on $[0, \infty)$. Assume that there exists a constant $c > 0$ such that

$$f(x) \leq cg(x) \qquad \text{for all } x \in [0, \infty)$$

Think of a situation where we want to simulate random variables with a distribution that has density f, but where f is so complicated that this is not straightforward to do directly. Suppose on the other hand that g is a simple well–known density that we actually *can* simulate from. An algorithm to produce a random variable X with density f is the *acceptance–rejection algorithm*:

(i) Generate Y with density g and U uniform on $(0, 1)$ such that $Y \perp\!\!\!\perp U$

(ii) If $U \leq f(Y)/(cg(Y))$, let $X = Y$. Otherwise return to (i)

The idea of this exercise is to show that X generated in the algorithm above actually has density f.

So let Y have density g and let U be uniform on $(0, 1)$. Assume that Y and U are independent. Define the random variable

$$Z = \begin{cases} 1\,, U \leq \frac{f(Y)}{cg(Y)} \\ 0\,, U > \frac{f(Y)}{cg(Y)} \end{cases}$$

(a) Show that $P(Z = 1) = \frac{1}{c}$.

(b) Show that $P(Y \in B \,|\, Z = 1) = \int_B f(x)\,dx$ for all $B \in \mathbb{B}$.

(c) Conclude that the algorithm produces a variable X with density f, and discuss which value of c we should choose.

Exercise 1.22 Think of a situation where we want to estimate the value z that is given by

$$z = EZ$$

for some real valued random variable Z with $EZ^2 < \infty$. Let $Z_1, Z_2, \ldots, Z_n$ be independent replications of Z. Then

$$\hat{z}_n^1 = \frac{1}{n} \sum_{k=1}^{n} Z_k$$

is an estimator for z.

(a) Show that $\hat{z}_n^1$ is unbiased

$$E\hat{z}_n^1 = z$$

and find the variance $V\hat{z}_n^1$.

A method to improve the estimator could be finding some random variable X and consider the new variable $E(Z \,|\, X)$. Now let $(Z_1, X_1), \ldots, (Z_n, X_n)$ be independent replications of (Z, X), and define the estimator

$$\hat{z}_n^2 = \frac{1}{n} \sum_{k=1}^{n} E(Z_k \,|\, X_k)$$

(b) Show that $\hat{z}_n^2$ is unbiased and that

$$V\hat{z}_n^2 \leq V\hat{z}_n^1$$

Apparently this method will improve the estimator no matter which variable X we choose. But of course some choices may be more clever than others.

(c) What happens, if we use $X = 1$ (or some other constant), and why is this a bad idea anyway?

We will consider two specific examples of variables Z. In both examples we shall just let $n = 1$, since increasing values of n simply makes both variances smaller by a factor $1/n$, and thereby does not change anything in the comparison.

In the first example we shall find estimators for the very well–known value π (although we already know π much more accurately than we will ever be able to estimate, the example serves as a very good illustration of what is going on). Let

$$Z = 4 \cdot 1_{(U_1^2+U_2^2\leq 1)}\,,$$

where U_1 and U_2 are independent and both uniform on $(0, 1)$. Define the first estimator $\hat{z}_1 = Z$.

(d) Show that $E\hat{z}_1 = \pi$.

Define the estimator $\hat{z}_2$ by

$$\hat{z}_2 = E(Z \,|\, U_1)$$

(e) Show that $\hat{z}_2 = 4\sqrt{1 - U_1^2}$.

(f) Try to simulate 10000 replications of both $\hat{z}_1$ and $\hat{z}_2$. Compare the variances – and also compare with the theoretical variance of $\hat{z}_1$.

In the next example, the estimation has some real practical use. Assume that X_1 and X_2 are independent and has a distribution ν. Assume that $X_1, X_2 \geq 0$ and that ν has density f with respect to the Lebesgue measure. Furthermore think of a situation, where the distribution of $S = X_1 + X_2$ is complicated to calculate. We are interested in estimating

$$z(s) = P(S > x)$$

especially for large values of x.

The simple estimator will in this framework be

$$\hat{z}_1(x) = 1_{(X_1+X_2>x)}$$

The problem is, that if x is very large, then it is very rare that this estimator is non–zero. Even if we make many replications. Instead we shall try to construct an estimator using conditional expectations.

Firstly, we try something similar to above. Define

$$\hat{z}_2(x) = P(S > x \,|\, X_1)$$

(g) Show that

$$\hat{z}_2(x) = \bar{F}(x - X_1)\,,$$

where $\bar{F}$ is the *survival function* for ν:

$$\bar{F}(x) = \nu((x, \infty))\,.$$

Let

$$X_{(1)} = \min\{X_1, X_2\} \qquad \text{and} \qquad X_{(2)} = \max\{X_1, X_2\}$$

(h) Show that the conditional distribution of $X_{(2)}$ given $X_{(1)}$ is determined by the Markov kernel $(P_y)_{y\geq 0}$, where

$$P_y(B) = \frac{\nu(B \cap (y, \infty))}{\nu((y, \infty))}.$$

We now define a conditional estimator by

$$\hat{z}_3(x) = P(S > x \,|\, X_{(1)})$$

(i) Show that

$$\hat{z}_3(x) = \frac{\bar{F}\big(\max\{x - X_{(1)}, X_{(1)}\}\big)}{\bar{F}(X_{(1)})},$$

Now assume that ν is the Weibull distribution with shape parameter 0.5. Then the density f is given by

$$0.5x^{-0.5}e^{-x^{0.5}}$$

for $x > 0$. And $\bar{F}$ is

$$\bar{F}(x) = e^{-x^{0.5}}$$

(j) Simulate 10000 replications of the three estimators (with e.g. $x = 20$ and $x = 50$) and compare the variances.

Exercise 1.23 Let X be a real valued random variable with $E|X| < \infty$.

(a) Show that the conditional distribution of X given X is given by the Markov kernel $(\delta_x)_{x\in\mathcal{X}}$, where δ_x is the Dirac Measure in x:

$$\delta_x(B) = \begin{cases} 1\,, & x \in B \\ 0\,, & x \notin B \end{cases}$$

(b) Show that $E(X \,|\, X = x) = x$ and $E(X \,|\, X) = X$.

(c) Assume that Y is another real valued random variable with $E|Y| < \infty$ and $E|XY| < \infty$. Show that $E(XY \,|\, X = x) = xE(Y \,|\, X = x)$ and $E(XY \,|\, X) = XE(Y \,|\, X)$.

Exercise 1.24 Let W be the set $(0, 1)^2$. Assume that we generate N points in W in the following way: Let N be Poisson distributed with parameter λ and $(U_1^1, U_1^2), (U_2^1, U_2^2), \ldots, (U_N^1, U_N^2)$ be independent and identically distributed such that U_k^1 and U_k^2 are independent and uniformly distributed on $(0, 1)$. This makes each (U_k^1, U_k^2) uniformly distributed on W.
In this exercise we will show that the collection of points $(U_1^1, U_1^2), \ldots, (U_N^1, U_N^2)$ in W is a *Poisson process* on W: Define for a subset $A \subseteq W$ the random variable $N(A)$ to be the number of points in A:

$$N(A) = \sum_{k=1}^{N} 1_{(U_1^k, U_2^k)\in A)}\,.$$

Then

(i) $N(A)$ is Poisson distributed with parameter $\lambda m_2(A)$, where $m_2(A)$ is the area (2–dimensional Lebesgue measure) of A.

(ii) For disjoint sets $A_1, \ldots, A_m$ the variables $N(A_1), \ldots, N(A_m)$ are independent.

The result will follow by finding conditional distributions given N

(a) Show that for U_1 and U_2 independent and uniformly distributed on $(0, 1)$ and A some subset of W, then

$$P((U_1, U_2) \in A) = m_2(A)$$

(b) Let $A_1, \ldots, A_m$ be disjoint subsets of W such that $\bigcup_{j=1}^m A_j = W$. Argue that the conditional distribution of $(N(A_1), \ldots, N(A_m))$ given $N = n$ is a polynomial distribution with length n and probability parameters $(m_2(A_1), \ldots, m_2(A_m))$.

(c) Show that $N(A_1), \ldots, N(A_m)$ are independent and that each $N(A_j)$ is Poisson distributed with parameter $\lambda m_2(A_j)$.

Now assume that $k : W \to [0, 1]$ is a measurable function that is bounded by 1. Define for each subset A of W the number

$$K(A) = \int_A k(x, y)\, m_2(dx, dy)$$

(d) Give a suggestion for how to obtain a collection of points $(V_1^1, V_1^2), \ldots, (V_M^1, V_M^2)$ in W, such that for each subset A of W we have that the number of points in A

$$M(A) = \sum_{k=1}^{M} 1_{((V_j^1, V_j^2) \in A)}$$

is Poisson distributed with parameter $\lambda K(A)$.

Exercise 1.25 Assume that (X, Y) is a real valued random vector, such that $E|Y| < \infty$. Assume that the random vector $(X, \tilde{Y})$ has the same distribution as (X, Y), where $\tilde{Y}$ is another real valued random variable.

(a) Show that $E(Y \mid X) = E(\tilde{Y} \mid X)$ a.s.

Now assume that $X_1, \ldots, X_n$ are independent and identically distributed with $E|X_1| < \infty$. Define $S_n = X_1 + \cdots + X_n$.

(b) Argue that (X_1, S_n) has the same distribution as (X_k, S_n) for all $k = 1, \ldots, n$.

(c) Show that $E(X_1 \mid S_n) = S_n/n$.

2

CONDITIONAL INDEPENDENCE

In this chapter we will work on a general probability space $(\Omega, \mathbb{F}, P)$. All events occurring will silently be assumed to be $\mathbb{F}$-measurable, all σ-algebras occurring will silently be assumed to be subalgebras of $\mathbb{F}$, and all random variables $X : (\Omega, \mathbb{F}) \to (\mathcal{X}, \mathbb{E})$ will silently be assumed to be $\mathbb{F} - \mathbb{E}$ measurable. The general convention is that random variables with names like X or X_i or variations thereof have values in a generic space $(\mathcal{X}, \mathbb{E})$, unless it is explicitly stated that they are real valued (or integer valued or whatever). Similarly, variables with names like Y or Z will have values in $(\mathcal{Y}, \mathbb{K})$ and $(\mathcal{Z}, \mathbb{G})$ respectively.
Recall that $(\mathcal{X}, \mathbb{E})$ is a *Borel space* if it is in bijective, bimeasurable correspondence with $(\mathbb{R}, \mathbb{B})$ or a subspace of this. Such a correspondence enables us to replace $\mathcal{X}$ with $\mathbb{R}$, whenever there is an advantage in that. It turns out that every sensible space has this property, unless it its very, very huge (non-separable metric spaces, with the σ-algebra generated by the open sets, say). The above generic $\mathcal{X}$, $\mathcal{Y}$ and $\mathcal{Z}$-spaces are always assumed to be Borel spaces.

2.1 Conditional probabilities given a σ–algebra

We recall that the conditional expectation $E(Y \,|\, \mathbb{H})$ of a real valued random variable Y with $E|Y| < \infty$ given a σ-algebra $\mathbb{H}$ is any $\mathbb{H}$-measurable and integrable random variable satisfying

$$\int_H E(Y \,|\, \mathbb{H})\, dP = \int_H Y\, dP \qquad \text{for all } H \in \mathbb{H}\,. \tag{2.1}$$

as defined in Definition 1.32.
As we have considered conditional expectations given a σ-algebra, we shall also be concerned with the conditional probability given a σ-algebra. This is defined by taking conditional expectation of the indicator function 1_A, that is

$$P(A \;|\, \mathbb{H}) = E(1_A \,|\, \mathbb{H})\,.$$

The integrability condition (2.1) will in this case take the form

$$\int_H P(A \,|\, \mathbb{H})\, dP = P(A \cap H) \qquad \text{for all } H \in \mathbb{H}\,. \tag{2.2}$$

We will make frequent use of the monotonicity property of conditional expectations that ensure

$$0 \le P(A \,|\, \mathbb{H}) \le 1 \qquad \text{a.s.}$$

and even that

$$A \subseteq B \quad \Rightarrow \quad P(A \,|\, \mathbb{H}) \leq P(B \,|\, \mathbb{H}) \quad \text{a.s}$$

Furthermore, the double conditioning theorem (Theorem 1.35) says in this context that

$$E\Big(P(A \,|\, \mathbb{H}) \,|\, \mathbb{G}\Big) = P(A \,|\, \mathbb{G}) \qquad \text{a.s}$$

whenever the two σ-algebras $\mathbb{G}$ and $\mathbb{H}$ satisfies that $\mathbb{G} \subseteq \mathbb{H}$.

2.2 Conditionally independent events

Definition 2.1 Two events A and B are are *conditionally independent* given a σ-algebra $\mathbb{H}$, if

$$P(A \cap B \,|\, \mathbb{H}) = P(A \,|\, \mathbb{H})\, P(B \,|\, \mathbb{H}) \qquad \text{a.s.} \tag{2.3}$$

Symbolically, we will write $A \perp\!\!\!\perp B \,|\, \mathbb{H}$ if (2.3) is satisfied.

Speaking colloquially, we will frequently say that A and B are independent given $\mathbb{H}$ if (2.3) is satisfied - repeated use of the word *conditionally* makes the sentences sound tedious.
Please note that conditional independence represents an intricate relation between the two events and the σ-algebra. The σ-algebra $\mathbb{H}$ is really an integral part of the definition. Whether A and B are conditionally independent or not, depends crucially on which σ-algebra we use for conditioning.
If $\mathbb{H} \subseteq \mathbb{G}$ are two σ-algebras, it is completely possible that two events A and B are independent given $\mathbb{H}$, while they are not independent given the finer σ-algebra $\mathbb{G}$. But it is equally possible that A and B are independent given $\mathbb{G}$, while they are not independent given the coarser σ-algebra $\mathbb{H}$, see Example 2.3 below. Changing the σ-algebra on which we are conditioning is usually a very challenging task.

Example 2.2 Recall that a σ-algebra $\mathbb{H}$ is a *trivial* if every event in $\mathbb{H}$ has probability 0 or 1. The most obvious trivial σ-algebra is

$$\mathbb{H} = \{\emptyset, \Omega\}\,,$$

but there are plenty of other trivial algebras arising all over probability theory - tail algebras, symmetric algebras, invariant σ-algebras in ergodic theory and what not. If $\mathbb{H}$ is trivial, we observe that

$$P(A \,|\, \mathbb{H}) = P(A) \qquad \text{a.s.}$$

for any event A, since the relation

$$\int_H P(A)\, dP = P(A \cap H)\,,$$

is satisfied for all $\mathbb{H}$-sets H, both those of probability 0 (where there is nothing to prove) and those of probability 1 (where there is also nothing to prove). Hence (2.3) translates to

$$P(A \cap B) = P(A)\, P(B)\,. \tag{2.4}$$

A priori the formula has an a.s.-qualifier, but as it is a relation between deterministic numbers, it is either true or false, with no probability involved.

Hence we see that conditional independence of two events given a trivial σ-algebra is simply classical independence of the events. □

Example 2.3 If C is yet another event, and if $\mathbb{H}$ is the σ-algebra generated by that event,

$$\mathbb{H} = \{\emptyset, C, C^c, \Omega\},$$

then it is readily checked that

$$P(A \,|\, \mathbb{H}) = \begin{cases} \dfrac{P(A \cap C)}{P(C)} & \text{on } C \\ \dfrac{P(A \cap C^c)}{P(C^c)} & \text{on } C^c \end{cases} \qquad \text{a.s}$$

for any event A. If we suppose that $\mathbb{H}$ is non-trivial, meaning that $P(C) \in (0, 1)$, we see that (2.3) translates to the two conditions

$$\frac{P(A \cap B \cap C)}{P(C)} = \frac{P(A \cap C)}{P(C)} \frac{P(B \cap C)}{P(C)},$$

$$\frac{P(A \cap B \cap C^c)}{P(C^c)} = \frac{P(A \cap C^c)}{P(C^c)} \frac{P(B \cap C^c)}{P(C^c)}.$$

These two conditions cannot be deduced from each other, and they are not related to (2.4). For instance, the probability table

C			C^c		
	B	B^c		B	B^c
A	$\frac{2}{18}$	$\frac{1}{18}$	A	$\frac{2}{18}$	$\frac{4}{18}$
A^c	$\frac{4}{18}$	$\frac{2}{18}$	A^c	$\frac{1}{18}$	$\frac{2}{18}$

corresponds to a situation where $A \perp\!\!\!\perp B \,|\, \mathbb{H}$ but where A and B are dependent, as can readily be checked.
On the other hand, the probability table

C			C^c		
	B	B^c		B	B^c
A	$\frac{1}{12}$	$\frac{2}{12}$	A	$\frac{2}{12}$	$\frac{1}{12}$
A^c	$\frac{2}{12}$	$\frac{1}{12}$	A^c	$\frac{1}{12}$	$\frac{2}{12}$

corresponds to a situation where A and B are independent, but where they are *not* independent given $\mathbb{H}$. □

Example 2.4 If we have a finite partition $\mathbb{D}$ of Ω,

$$\mathbb{D} = \{D_1, \ldots, D_n\}$$

where the *atoms* of $\mathbb{D}$ (the D_i-sets) are pairwise disjoint and unite to the whole of Ω, the σ-algebra generated by $\mathbb{D}$ is the family of all unions,

$$\mathbb{H} = \left\{ \bigcup_{i \in I} D_i \,|\, I \subseteq \{1, \ldots, n\} \right\} .$$

If we let

$$\mathbb{D}^* = \{D \in \mathbb{D} \,|\, P(D) > 0\} ,$$

it is easily checked that

$$P(A \,|\, \mathbb{H}) = \sum_{D \in \mathbb{D}^*} \frac{P(A \cap D)}{P(D)} \, 1_D \qquad \text{a.s.}$$

for any event A. In this setting, condition (2.3) translates into

$$\frac{P(A \cap B \cap D)}{P(D)} = \frac{P(A \cap D)}{P(D)} \, \frac{P(B \cap D)}{P(D)} \qquad \text{for all } D \in \mathbb{D}^* .$$

Again, whether this holds or not is very sensitive to the specific atoms. If an atom is divided into two, there is no telling if A and B are independent on each of the two subatoms, just because we know if they are independent on the original atom. And similarly, if two atoms are coalesced, we may loose or create conditional independence, as the case may be. □

2.3 Conditionally independent σ-algebras

Definition 2.5 Two classes of events, $\mathcal{A}$ and $\mathcal{B}$, are conditionally independent given a σ-algebra $\mathbb{H}$ if

$$A \perp\!\!\!\perp B \,|\, \mathbb{H} \qquad \text{for all } A \in \mathcal{A}, B \in \mathcal{B} . \tag{2.5}$$

Symbolically, we will write $\mathcal{A} \perp\!\!\!\perp \mathcal{B} \,|\, \mathbb{H}$ if (2.5) is satisfied.

We will almost exclusively use this concept in situations where the two classes of events are σ-algebras, but it is nice to be allowed to formulate things in a slightly broader fashion. We may for instance see that it typically is enough to check (2.5) on two generators of the σ-algebras under consideration:

Lemma 2.6 *Let $\mathcal{A}$ and $\mathcal{B}$ be two classes of events, both stable under formation of intersections. Then*

$$\mathcal{A} \perp\!\!\!\perp \mathcal{B} \,|\, \mathbb{H} \qquad \Rightarrow \qquad \sigma(\mathcal{A}) \perp\!\!\!\perp \sigma(\mathcal{B}) \,|\, \mathbb{H} .$$

Proof A prototypical application of Dynkin's lemma. For each set $F \in \mathbb{F}$ we consider the class

$$\mathcal{C}_F = \{E \in \mathbb{F} \,|\, F \perp\!\!\!\perp E \,|\, \mathbb{H}\}\,,$$

and we observe that this is a Dynkin class. If we take $A \in \mathcal{A}$, we know that $\mathcal{B} \subseteq \mathcal{C}_A$. Using Dynkins lemma, we see that $\sigma(\mathcal{B}) \subseteq \mathcal{C}_A$. On the other hand, conditional independence of two events is a property that is symmetric in the two events, so we can reformulate this fact as $\mathcal{A} \subseteq \mathcal{C}_B$ for any set $B \in \sigma(\mathcal{B})$. Using Dynkin's lemma again establishes that $\sigma(\mathcal{A}) \subseteq \mathcal{C}_B$ for any set $B \in \sigma(\mathcal{B})$. And though this may look awkward, it is in fact the property we are after. □

Conditional independence of classes of events is of course just as sensitive to the exact choice of the σ-algebra on which we are conditioning, as conditional independence of events were. In fact, if

$$\mathbb{A} = \{\emptyset, A, A^c, \Omega\}\,, \qquad \mathbb{B} = \{\emptyset, B, B^c, \Omega\}\,,$$

then $\mathbb{A} \perp\!\!\!\perp \mathbb{B} \,|\, \mathbb{H}$ if and only if $A \perp\!\!\!\perp B \,|\, \mathbb{H}$, as is readily seen from Lemma 2.6. So the counterexamples to any kind of simple behaviour under change of the conditioning algebra given in section 2.2 also apply in this setting.

Example 2.7 Assume that $\mathbb{H}$ is a trivial σ–algebra. Then we saw in Example 2.2 that two sets A and B are conditionally independent given $\mathbb{H}$, if and only if they are truly independent. This translates directly into conditional independence of classes of events: If $\mathbb{H}$ is trivial, then any two classes $\mathcal{A}$ and $\mathcal{B}$ satisfies

$$\mathcal{A} \perp\!\!\!\perp \mathcal{B} \,|\, \mathbb{H} \qquad \Leftrightarrow \qquad \mathcal{A} \perp\!\!\!\perp \mathcal{B}$$

Assume conversely that $\mathbb{H} = \mathbb{F}$. Then for all $F \in \mathbb{F}$ we have

$$P(F \,|\, \mathbb{F}) = 1_F \qquad \text{a.s.}\,,$$

since 1_F is $\mathbb{F}$–measurable. Hence it is seen that for any choice of $\mathcal{A}$ and $\mathcal{B}$ we have with $A \in \mathcal{A}$ and $B \in \mathcal{B}$ that

$$P(A \,|\, \mathbb{F})P(B \,|\, \mathbb{F}) = 1_A \cdot 1_B = 1_{A\cap B} = P(A \cap B \,|\, \mathbb{F}) \quad \text{a.s.}\,,$$

so we conclude that $\mathcal{A}$ and $\mathcal{B}$ are always conditionally independent given $\mathbb{F}$. □

Example 2.8 Assume that $\mathcal{A}$, $\mathcal{B}$ and $\mathbb{H}$ are independent. Then with $A \in \mathcal{A}$ we observe

$$P(A \,|\, \mathbb{H}) = P(A) \qquad \text{a.s.}$$

since for $H \in \mathbb{H}$ the relation

$$\int_H P(A)\, dP = P(A)P(H) = P(A \cap H)$$

is satisfied. Then – using the independence between $\mathcal{A}$ and $\mathcal{B}$ – we obtain

$$P(A \,|\, \mathbb{H})P(B \,|\, \mathbb{H}) = P(A) \cdot P(B) = P(A \cap B) = P(A \cap B \,|\, \mathbb{H}) \quad \text{a.s.}\,.$$

In the last equality we used that $A \cap B \perp\!\!\!\perp \mathbb{H}$ since both A and B are independent of $\mathbb{H}$. We conclude that $\mathcal{A}$ and $\mathcal{B}$ are independent given $\mathbb{H}$ as well. □

Theorem 2.9 *Let $\mathbb{A}, \mathbb{B}$ and $\mathbb{H}$ be three σ-algebras. Suppose that $\mathbb{A} \perp\!\!\!\perp \mathbb{B} \mid \mathbb{H}$. If X is an $\mathbb{A}$-measurable real valued random variable, and if Y is a $\mathbb{B}$-measurable real valued random variable, such that $E|X| < \infty$, $E|Y| < \infty$ and $E|XY| < \infty$, then it holds that*

$$E(XY \mid \mathbb{H}) = E(X \mid \mathbb{H})\, E(Y \mid \mathbb{H}) \qquad a.s.$$

Proof A prototypical extension result. We know the theorem to be true for indicator variables. Hence it is true for simple variables. The monotone convergence theorem for conditional expectations will show it is true for non-negative variables, and a final handwaving will dismiss the problems of positive and negative parts. □

Conditional independence is by its very definition symmetric in the two events, or more general, in the two classes of events. Rather surprisingly, it turns out that the most fruitful way of working with the concept is through an asymmetric formulation:

Theorem 2.10 *Let $\mathbb{A}$, $\mathbb{B}$ and $\mathbb{H}$ be σ-algebras. It holds that $\mathbb{A} \perp\!\!\!\perp \mathbb{B} \mid \mathbb{H}$ if and only if*

$$P(A \mid \mathbb{B} \vee \mathbb{H}) = P(A \mid \mathbb{H}) \qquad a.s \tag{2.6}$$

for every event $A \in \mathbb{A}$.

In the theorem $\mathbb{B} \vee \mathbb{H}$ denotes the smallest σ–algebra that contains both $\mathbb{B}$ and $\mathbb{H}$. This σ–algebra must be generated by the $\cap$–stable generating system given by

$$\{B \cap H \,:\, B \in \mathbb{B}, H \in \mathbb{H}\}$$

Proof Notice that for any three events $A \in \mathbb{A}$, $B \in \mathbb{B}$ and $H \in \mathbb{H}$ we have that

$$\begin{aligned}\int_{B\cap H} P(A \mid \mathbb{H})\, dP &= \int_H 1_B\, P(A \mid \mathbb{H})\, dP = \int_H E\Big(1_B\, P(A \mid \mathbb{H}) \mid \mathbb{H}\Big)\, dP \\ &= \int_H P(A \mid \mathbb{H})\, P(B \mid \mathbb{H})\, dP. \end{aligned} \tag{2.7}$$

In the second equality we have used the integration property from the definition of conditional expectations. In the third equality we have used that if X is $\mathbb{H}$–measurable, then $E(XY \mid \mathbb{H}) = XE(Y \mid \mathbb{H})$ (we also exploit that for indicator functions we trivially have $E|X| < \infty$, $E|Y| < \infty$ and $E|XY| < \infty$ so the conditional expectations are well defined).
Suppose that $\mathbb{A}$ and $\mathbb{B}$ are conditionally independent given $\mathbb{H}$. Then we can work the above line of equations one step further to see that

$$\int_{B\cap H} P(A \mid \mathbb{H})\, dP = \int_H P(A \cap B \mid \mathbb{H})\, dP = P(A \cap B \cap H)\,.$$

Since the events of the form $B \cap H$ is a generator for the σ-algebra $\mathbb{B} \vee \mathbb{H}$ that is stable under formation of intersections, and as $P(A \mid \mathbb{H})$ is $\mathbb{H}$-measurable, and thereby in

particular $\mathbb{B} \vee \mathbb{H}$–measurable, we conclude that $P(A \,|\, \mathbb{H})$ indeed does satisfy all conditions for being the conditional probability of A given $\mathbb{B} \vee \mathbb{H}$. And hence (2.6) holds.
For the opposite implication, we may utilise (2.6) on the starting end of (2.7), and obtain that

$$\int_H P(A \,|\, \mathbb{H})\, P(B \,|\, \mathbb{H})\, dP = \int_{H \cap B} P(A \,|\, \mathbb{B} \vee \mathbb{H})\, dP = P(A \cap B \cap H)\,.$$

As $P(A \,|\, \mathbb{H})\, P(B \,|\, \mathbb{H})$ is indeed $\mathbb{H}$-measurable, we see that it satisfies all conditions for being the conditional probability of $A \cap B$ given $\mathbb{H}$. And hence A and B are conditionally independent given $\mathbb{H}$. □

The asymmetric condition (2.6) is usually paraphrased by saying that there is no extra information in $\mathbb{B}$ for making predictions on the occurrence of an $\mathbb{A}$-set, when we already have access to the information in $\mathbb{H}$. All the information in $\mathbb{B}$, useful for that prediction, is already contained in $\mathbb{H}$. The symmetry between $\mathbb{A}$ and $\mathbb{B}$ is not clearly visible here, but somehow it is still there.

In many cases, it is an advantage to use a slightly simplified version of Theorem 2.10, saying that $\mathbb{A} \perp\!\!\!\perp \mathbb{B} \,|\, \mathbb{H}$ if and only if $P(A \,|\, \mathbb{B} \vee \mathbb{H})$ is $\mathbb{H}$-measurable. More formally, we have the following corollary:

Corollary 2.11 *Let $(\Omega, \mathbb{F}, P)$ be a probability space and $\mathbb{A}$, $\mathbb{B}$, and $\mathbb{H}$ sub-σ-algebras of $\mathbb{F}$. Then $\mathbb{A} \perp\!\!\!\perp \mathbb{B} \,|\, \mathbb{H}$ if and only if for any $A \in \mathbb{A}$ there is a $\mathbb{H}$-measurable random variable Z_A so that*

$$P(A \,|\, \mathbb{B} \vee \mathbb{H}) = Z_A \text{ } P\text{-almost surely}. \tag{2.8}$$

Proof If $\mathbb{A} \perp\!\!\!\perp \mathbb{B} \,|\, \mathbb{H}$, we have from Theorem 2.10 that

$$P(A \,|\, \mathbb{B} \vee \mathbb{H}) = P(A \,|\, \mathbb{H})$$

almost surely and hence we can let $Z_A = P(A \,|\, \mathbb{H})$. Conversely, if (2.8) holds, we have for any $H \in \mathbb{H} \subseteq (\mathbb{B} \vee \mathbb{H})$ that

$$P(A \cap H) = \mathbf{E}(1_H 1_A) = \mathbf{E}\{1_H P(A \,|\, \mathbb{B} \vee \mathbb{H})\} = \mathbf{E}(1_H Z_A)$$

so Z_A is a version of $P(A \,|\, \mathbb{H})$. Theorem 2.10 now yields that $\mathbb{A} \perp\!\!\!\perp \mathbb{B} \,|\, \mathbb{H}$. □

Further, the statement in Theorem 2.10 can of course be extended to random variables:

Corollary 2.12 *Let $\mathbb{A}$, $\mathbb{B}$ and $\mathbb{H}$ be σ-algebras. If $\mathbb{A} \perp\!\!\!\perp \mathbb{B} \,|\, \mathbb{H}$ then it holds for any $\mathbb{A}$-measurable real random variable X such that $E|X| < \infty$ that*

$$E(X \,|\, \mathbb{B} \vee \mathbb{H}) = E(X \,|\, \mathbb{H}) \qquad a.s \tag{2.9}$$

Proof Follows from Theorem 2.10 by the same extension technique, that was used to prove Theorem 2.9. □

Example 2.13 Assume that $\mathbb{A}$ and $\mathbb{H}$ are σ–algebras. We clearly have $\mathbb{H} \vee \mathbb{H} = \mathbb{H}$, such that

$$P(A \,|\, \mathbb{H} \vee \mathbb{H}) = P(A \,|\, \mathbb{H})$$

Then it follows that $\mathbb{A} \perp\!\!\!\perp \mathbb{H} \,|\, \mathbb{H}$. The same argument applies to deduce that $\mathbb{A} \perp\!\!\!\perp \mathbb{B} \,|\, \mathbb{H}$ whenever $\mathbb{B} \subseteq \mathbb{H}$. □

In many cases we have σ-algebras generated by random variables. We will make no distinction between the random variable X and the σ-algebra $\sigma(X)$ generated by X, and we will write things like

$$X \perp\!\!\!\perp Y \,|\, Z \qquad \text{instead of} \quad \sigma(X) \perp\!\!\!\perp \sigma(Y) \,|\, \sigma(Z)$$

without notification.

2.4 Combination of Markov kernels

Let $(\mathcal{X}, \mathbb{E})$, $(\mathcal{Y}, \mathbb{K})$, $(\mathcal{Z}, \mathbb{G})$ be measurable spaces and consider Markov kernels P and Q, where P is a Markov kernel from $\mathcal{X}$ to $\mathcal{Y}$, and Q a Markov kernel from $\mathcal{Y}$ to $\mathcal{Z}$.

Definition 2.14 The *combination* $P \circledast Q$ of the Markov kernels P and Q is a Markov kernel from $\mathcal{X}$ to $\mathcal{Y} \times \mathcal{Z}$ determined as

$$(P \circledast Q)_x(B \times C) = \int_B Q_y(C) \, \mathrm{d}P_x(y) \tag{2.10}$$

for $A \in \mathbb{E}, B \in \mathbb{K}, C \in \mathbb{G}$.

In other words, $(P \circledast Q)_x$ is the integration of Q w.r.t. P_x. If μ is a probability measure on $\mathcal{X}$, we may think of μ as a Markov kernel from $\{0\}$ to $\mathcal{X}$ and write the integration λ of P w.r.t. μ as a combination in this sense

$$\lambda(A \times B) = \int_A P_x(B) \, \mathrm{d}\mu(x) = (\mu \circledast P)(A \times B).$$

If $\mathcal{Y} = \mathcal{Y}_1 \times \mathcal{Y}_2$ and Q is a Markov kernel from $\mathcal{Y}_2$ to $\mathcal{Z}$, Q can always be extended to a Markov kernel $\tilde{Q}$ from $\mathcal{Y}$ to $\mathcal{Z}$ as

$$\tilde{Q}_y = \tilde{Q}_{(y_1, y_2)} = Q_{y_2}.$$

We shall in the following not distinguish between Q and its extension $\tilde{Q}$.

Consider now Markov kernels P and Q as above, and further, a Markov kernel R from $\mathcal{Z}$ to $\mathcal{W}$. We then have

Proposition 2.15 *Combination of Markov kernels is associative*

$$(P \circledast Q) \circledast R = P \circledast (Q \circledast R)$$

where Markov kernels are extended whenever appropriate.

Proof This is a direct consequence of the extended Tonelli's Theorem 1.11. From (2.10) we get

$$\begin{aligned}\{(P \circledast Q) \circledast R\}_x(B \times C \times D) &= \int_{B\times C} R_z(D)\,\mathrm{d}(P \circledast Q)_x(y,z) \\ &= \int_B \int_C R_z(D)\,\mathrm{d}Q_y(z)\,\mathrm{d}P_x(y) \\ &= \int_B (Q \circledast R)_y(C \times D)\,\mathrm{d}P_x(y) \\ &= \{P \circledast (Q \circledast R)\}_x(B \times C \times D)\end{aligned}$$

as desired. □

Corollary 2.16 *If in Proposition 2.15 $\mathcal{X} = \{0\}$ and $P = \mu$ is a probability measure on $\mathcal{Y}$, we have*

$$\lambda = (\mu \circledast Q) \circledast R = \mu \circledast (Q \circledast R)$$

and further, if λ represents the joint distribution of random variables (Y, Z, W) it holds that $W \perp\!\!\!\perp_\lambda Y \mid Z$.

Proof The conditional independence statement follows from Theorem 1.30 since then $\tilde{R}_{(y,z)} = R_z$ — which represents the conditional distribution of W given (Y, Z) — depends on Z only. □

Suppose now that P is a Markov kernel from $\mathcal{X}$ to $\mathcal{Y}$ and Q a Markov kernel from $\mathcal{X}$ to $\mathcal{Z}$. We can then extend P and Q to be Markov kernels $\tilde{P}$ and $\tilde{Q}$ from $\mathcal{X} \times \mathcal{Y}$ and $\mathcal{X} \times \mathcal{Z}$ and hence combine them in both directions as $P \circledast \tilde{Q}$ or $Q \circledast \tilde{P}$. And, in fact, these two combinations are identical:

Proposition 2.17 *If P is a Markov kernel from $\mathcal{X}$ to $\mathcal{Y}$ and Q a Markov kernel from $\mathcal{X}$ to $\mathcal{Z}$ the combination of these Markov kernels is commutative and equal to the product Markov kernel:*

$$(P \circledast Q)_x = (Q \circledast P)_x = P_x \otimes Q_x.$$

Further, for any probability measure μ on $\mathcal{X}$ we then have $Y \perp\!\!\!\perp_\lambda X \mid Z$ where $\lambda = \mu \circledast P \circledast Q = \mu \circledast Q \circledast P$ represents the joint distribution of (X, Y, Z).

Proof We have

$$\begin{aligned}(P \circledast Q)_x(B \times C) &= \int_B \tilde{Q}_{(x,y)}(C)\,\mathrm{d}P_x(y) \\ &= Q_x(C) \int_B \mathrm{d}P_x(y) = P_x(B)Q_x(C)\end{aligned}$$

which establishes that $(P \circledast Q)_x = (Q \circledast P)_x = P_x \otimes Q_x$. The final statement follows from Corollary 2.16. □

2.5 Conditional independence revisited

It is convenient to collect the most fundamental properties of conditional independence in a single theorem.

Theorem 2.18 *Let $(\Omega, \mathbb{F}, P)$ be a probability space and $\mathbb{A}$, $\mathbb{B}$, $\mathbb{C}$, and $\mathbb{D}$ sub-σ-algebras of $\mathbb{F}$. Then the following properties hold*

(A1) $\mathbb{A} \perp\!\!\!\perp \mathbb{B} \,|\, \mathbb{C} \implies \mathbb{B} \perp\!\!\!\perp \mathbb{A} \,|\, \mathbb{C}$ *(symmetry);*
(A2) $(\mathbb{A} \perp\!\!\!\perp \mathbb{B} \,|\, \mathbb{C}) \wedge \mathbb{D} \subseteq \mathbb{B} \implies \mathbb{A} \perp\!\!\!\perp \mathbb{D} \,|\, \mathbb{C}$ *(reduction);*
(A3) $\mathbb{A} \perp\!\!\!\perp (\mathbb{B} \vee \mathbb{C}) \,|\, \mathbb{D} \implies \mathbb{A} \perp\!\!\!\perp \mathbb{B} \,|\, (\mathbb{C} \vee \mathbb{D})$ *(weak union);*
(A4) $(\mathbb{A} \perp\!\!\!\perp \mathbb{C} \,|\, \mathbb{B}) \wedge (\mathbb{A} \perp\!\!\!\perp \mathbb{D} \,|\, \mathbb{B} \vee \mathbb{C}) \implies \mathbb{A} \perp\!\!\!\perp (\mathbb{C} \vee \mathbb{D}) \,|\, \mathbb{B}$ *(contraction);*
(A5) $(\mathbb{A} \perp\!\!\!\perp \mathbb{B} \,|\, \mathbb{C}) \wedge (\mathbb{A} \perp\!\!\!\perp \mathbb{C} \,|\, \mathbb{B}) \implies \mathbb{A} \perp\!\!\!\perp (\mathbb{B} \vee \mathbb{C}) \,|\, \overline{\mathbb{B}} \cap \overline{\mathbb{C}}$ *(intersection);*

Proof The *symmetry* in (A1) is trivial from Definition 2.5.
Reduction (A2) is also immediate as $\mathbb{D}$ represents fewer events than $\mathbb{B}$.

To establish *weak union*, note that if $\mathbb{A} \perp\!\!\!\perp (\mathbb{B} \vee \mathbb{C}) \,|\, \mathbb{D}$ we have from Theorem 2.10 that

$$P(A \,|\, \mathbb{B} \vee \mathbb{C} \vee \mathbb{D}) = P(A \,|\, \mathbb{D}),$$

where we have also used that $\vee$ is associative. Since $P(A \,|\, \mathbb{D})$ is $\mathbb{D}$-measurable it is also $\mathbb{C} \vee \mathbb{D}$-measurable and thus

$$P(A \,|\, \mathbb{B} \vee \mathbb{C} \vee \mathbb{D}) = P(A \,|\, \mathbb{C} \vee \mathbb{D}).$$

Now use Theorem 2.10 again to conclude that $\mathbb{A} \perp\!\!\!\perp \mathbb{B} \,|\, (\mathbb{C} \vee \mathbb{D})$.
For *contraction* (A4) we get

$$P(A \,|\, \mathbb{B} \vee \mathbb{C} \vee \mathbb{D}) = P(A \,|\, \mathbb{B} \vee \mathbb{C}) = P(A \,|\, \mathbb{B})$$

where we have used Theorem 2.10 in one direction twice. Using it in the other direction gives

$$A \perp\!\!\!\perp (\mathbb{C} \vee \mathbb{D}) \,|\, \mathbb{B}.$$

For the *intersection* (A5) we get

$$P(A \,|\, \mathbb{B} \vee \mathbb{C}) = P(A \,|\, \mathbb{B}) = P(A \,|\, \mathbb{C})$$

where all equalities hold almost surely. Thus $P(A \,|\, \mathbb{B} \vee \mathbb{C})$ is almost surely equal to an $\mathbb{A}$ measurable function and almost surely equal to a $\mathbb{B}$ measurable function. Hence, by Lemma A.11, $P(A \,|\, \mathbb{B} \vee \mathbb{C})$ is almost surely equal to an $\overline{\mathbb{B}} \cap \overline{\mathbb{C}}$-measurable function and therefore

$$P(A \,|\, \mathbb{B} \vee \mathbb{C}) = P(A \,|\, \overline{\mathbb{B}} \cap \overline{\mathbb{C}})$$

whereby $\mathbb{A} \perp\!\!\!\perp (\mathbb{B} \vee \mathbb{C}) \,|\, \overline{\mathbb{B}} \cap \overline{\mathbb{C}}$. Using Corollary 2.11 completes the proof. □

Note that if we do not supplement $\mathbb{A}$ and $\mathbb{B}$ with the σ-ideal of null sets $\mathcal{I}_P$, the corresponding variant of (A5) is not true in general. This is illustrated in the following example.

Example 2.19 Let $\Omega = \{0,1\}^3$ and $\mathbb{F}$ be the σ-algebra of all subsets of Ω. Define P as

$$p_{111} = p_{000} = 1/2$$

so that all six other atoms in $\mathbb{F}$ have probability zero.
Let $\mathbb{A}_i = \sigma(X_i), i = 1,2,3$, i.e. the σ-algebras generated by coordinate projections. It then holds that

$$\mathbb{A}_1 \perp\!\!\!\perp \mathbb{A}_2 \,|\, \mathbb{A}_3 \text{ and } \mathbb{A}_1 \perp\!\!\!\perp \mathbb{A}_3 \,|\, \mathbb{A}_2.$$

But $\mathbb{A}_2 \cap \mathbb{A}_3 = \{\emptyset, \Omega\}$ so *it is not true that* $\mathbb{A}_1 \perp\!\!\!\perp (\mathbb{A}_2 \vee \mathbb{A}_3) \,|\, \mathbb{A}_2 \cap \mathbb{A}_3$. However, from (A5) we get

$$\mathbb{A}_1 \perp\!\!\!\perp \mathbb{A}_2 \vee \mathbb{A}_3 \,|\, \overline{\mathbb{A}}_2 \cap \overline{\mathbb{A}}_3.$$

Indeed in this example we have $\overline{\mathbb{A}}_2 \cap \overline{\mathbb{A}}_3 = \mathbb{F}$ since $\mathcal{I}$ consists of all subsets of the six atoms with probability zero. Hence the last conditional independence is really an uninteresting tautology. □

Translating Theorem 2.18 to random variables, we get the following:

Corollary 2.20 *Let $(\Omega, \mathbb{F}, P)$ be a probability space and X, Y, Z, W random variables on Ω. Then the following properties hold.*

(C1) $X \perp\!\!\!\perp Y \,|\, Z \implies Y \perp\!\!\!\perp X \,|\, Z$ *(symmetry);*
(C2) $(X \perp\!\!\!\perp Y \,|\, Z) \wedge (W = \phi(Y)) \implies X \perp\!\!\!\perp W \,|\, Z$ *(reduction);*
(C3) $X \perp\!\!\!\perp (Y,Z) \,|\, W \implies X \perp\!\!\!\perp Y \,|\, (Z,W)$ *(weak union);*
(C4) $(X \perp\!\!\!\perp Z \,|\, Y) \wedge (X \perp\!\!\!\perp W \,|\, (Y,Z)) \implies X \perp\!\!\!\perp (Z,W) \,|\, Y$ *(contraction);*

Proof This follows directly from the definition and Theorem 2.18 by realizing that, for example, $W = \phi(Y) \implies \sigma(W) \subseteq \sigma(Y)$. □

When X, Y, and Z are discrete random variables the condition for $X \perp\!\!\!\perp Y \,|\, Z$ simplifies as

$$P(X = x, Y = y \,|\, Z = z) = P(X = x \,|\, Z = z) P(Y = y \,|\, Z = z),$$

where the equation holds for all z with $P(Z = z) > 0$. When the three variables admit a joint density with respect to a product measure μ, we have

Proposition 2.21 *Assume that the joint distribution of (X,Y,Z) has density w.r.t. a σ-finite product measure μ on $\mathcal{X} \times \mathcal{Y} \times \mathcal{Z}$. Then we have:*

$$X \perp\!\!\!\perp Y \,|\, Z \iff f(x,y \,|\, z) = f(x \,|\, z) f(y \,|\, z), \tag{2.11}$$
$$X \perp\!\!\!\perp Y \,|\, Z \iff f(x,y,z) f(z) = f(x,z) f(y,z), \tag{2.12}$$
$$X \perp\!\!\!\perp Y \,|\, Z \iff f(x,y,z) = f(x,z) f(y,z) / f(z) \tag{2.13}$$
$$X \perp\!\!\!\perp Y \,|\, Z \iff f(x \,|\, y,z) = f(x \,|\, z) \tag{2.14}$$
$$X \perp\!\!\!\perp Y \,|\, Z \iff f(x,z \,|\, y) = f(x \,|\, z) f(z \,|\, y) \tag{2.15}$$
$$X \perp\!\!\!\perp Y \,|\, Z \iff f(x,y,z) = h(x,z) k(y,z) \text{ for some } h,k \tag{2.16}$$
$$X \perp\!\!\!\perp Y \,|\, Z \iff f(x,y,z) = f(x \,|\, z) f(y,z). \tag{2.17}$$

where these equations hold almost surely with respect to P and we have used f as a generic symbol for the densities involved.

Proof The proof is Exercise 2.5. □

Another property of the conditional independence relation is often used:
(C5) if $X \perp\!\!\!\perp Y \mid Z$ and $X \perp\!\!\!\perp Z \mid Y$ then $X \perp\!\!\!\perp (Y, Z)$.

However (C5) does not hold universally, but only under additional conditions — essentially that there be no non-trivial logical relationship between Y and Z or, formally, $\overline{\sigma(Y)} \cap \overline{\sigma(Z)}$ is trivial.

Proposition 2.22 *If the joint density of all variables with respect to a product measure is positive then the statement* (C5) *will hold true.*

Proof Assume that the variables have density $f(x, y, z) > 0$ and that $X \perp\!\!\!\perp Y \mid Z$ as well as $X \perp\!\!\!\perp Z \mid Y$. Then (2.16) gives for almost all values of (x, y, z) that

$$f(x, y, z) = k(x, z)l(y, z) = g(x, y)h(y, z) \tag{2.18}$$

for suitable strictly positive functions g, h, k, l. Thus we have that for almost all (x, y, z) that

$$g(x, y) = \frac{k(x, z)l(y, z)}{h(y, z)}. \tag{2.19}$$

Choosing a fixed $z = z_0$ we have $g(x, y) = \pi(x)\rho(y)$ where $\pi(x) = k(x, z_0)$ and $\rho(y) = l(y, z_0)/h(y, z_0)$. Thus $f(x, y, z) = \pi(x)\rho(y)h(y, z)$ and hence $X \perp\!\!\!\perp (Y, Z)$ as desired. □

Generally, the following modification of (C5) holds
(C5*) if $X \perp\!\!\!\perp Y \mid Z$ and $X \perp\!\!\!\perp Z \mid Y$ then $X \perp\!\!\!\perp (Y, Z) \mid H$,
where $H = h(Y, Z)$ represents the information which is common to (Y, Z) so that $\sigma(H) = \overline{\sigma(X)} \cap \overline{\sigma(Y)}$, see Theorem 2.18.

It is illuminating to consider the special case when all state spaces are discrete. Define the bipartite graph $\mathcal{G}^+$ with vertex set $V = \mathcal{Y} \cup \mathcal{Z}$ by letting

$$y \sim^+ z \iff f(y, z) > 0.$$

The common information H above is simply indicating the connectivity component of Y (or Z) in this graph. Hence we have the following necessary and sufficient condition for (C5):

Proposition 2.23 *Assume that all state spaces are discrete and let $\mathcal{G}^+$ be defined as above. Then* (C5) *holds for all discrete random variables X if and only if $\mathcal{G}^+$ is connected.*

Proof If $\mathcal{G}^+$ is not connected, we can let $X = H$ denote the connectivity component of Y (or Z); then, conditionally on Y or Z, X has a degenerate distribution and hence $X \perp\!\!\!\perp Y \mid Z$ and $X \perp\!\!\!\perp Z \mid Y$; but $X = H$ is certainly not independent of (Y, Z).

If $\mathcal{G}^+$ is connected — and H therefore is trivial — we first get from (2.18) that the marginal distribution of Y and Z satisfies

$$f(y,z) = \bar{k}(z)l(y,z) = \bar{g}(y)h(y,z)$$

where $\bar{k}(z) = \sum_x k(x,z)$ and $\bar{g}(y) = \sum_x g(x,y)$. Hence, if $f(y,z) > 0$ we have $\bar{k}(z) > 0$, $\bar{g}(y) > 0$, and $l(y,z)/h(y,z) = \bar{g}(y)/\bar{k}(z)$. Thus from (2.19) we have

$$g(x,y) = k(x,z)\bar{g}(y)/\bar{k}(z). \tag{2.20}$$

Next, choose a fixed $y^* \in \mathcal{Y}$ and let $(y^* = y_1, z_1, y_2 \ldots, z_n, y_n = y)$ be a path in $\mathcal{G}^+$ from y^* to an arbitrary $y \in \mathcal{Y}$; such a path exists because $\mathcal{G}^+$ is connected. Then, since $f(y_1, z_1) > 0$ and $f(y_2, z_1) > 0$, we get from (2.20)

$$g(x,y_1) = k(x,z_1)\bar{g}(y_1)/\bar{k}(z_1) \text{ and } g(x,y_2) = k(x,z_1)\bar{g}(y_2)/\bar{k}(z_1);$$

hence we find

$$g(x,y_2) = g(x,y_1)\bar{g}(y_2)/\bar{g}(y_1).$$

Proceeding in the same way with y_2, z_2, and y_3 yields

$$g(x,y_3) = g(x,y_2)\bar{g}(y_3)/\bar{g}(y_2) = g(x,y_1)\bar{g}(y_3)/\bar{g}(y_1)$$

and if we continue this along the path we get for an arbitrary y that

$$g(x,y) = g(x,y^*)\bar{g}(y)/\bar{g}(y^*).$$

We can now let $\pi(x) = g(x,y^*)$ and $\rho(y) = \bar{g}(y)/\bar{g}(y^*)$ and proceed as in the proof of Proposition 2.22. □

Note in particular that the proof of Proposition 2.23 effectively establishes (C5*) in the discrete case.

2.5.1 *Independence models*

It is illuminating to think of the properties (C1)–(C5) as purely formal expressions, with a meaning that is not necessarily tied to probability. If we interpret the symbols used for random variables as abstract symbols for pieces of knowledge obtained from, say, reading books, and further interpret the symbolic expression $X \perp\!\!\!\perp Y \mid Z$ as:

Knowing Z, reading Y is irrelevant for reading X,

the properties (C1)–(C4) translate to the following:

(I1) if, knowing Z, reading Y is irrelevant for reading X, then so is reading X for reading Y;

(I2) if, knowing Z, reading Y is irrelevant for reading the book X, then reading any chapter U of Y is irrelevant for reading X;

(I3) if, knowing Z, reading both of Y and W is irrelevant for reading the book X, reading Y remains irrelevant after having also read W;

(I4) if, knowing Y, reading the book Z is irrelevant for reading X and even after having also read Z, reading W is irrelevant for reading X, then reading both of Z and W is irrelevant for reading X.

Thus one can view the relations (A1)–(A4) as pure formal properties of the notion of irrelevance. The property (A5) is slightly more subtle. In a certain sense, also the symmetry (A1) is a somewhat special property of probabilistic conditional independence, rather than general irrelevance.
It is thus tempting to use the relations such as these as formal axioms for conditional independence or irrelevance. Indeed, it was conjectured (Pearl, 1988) that the properties (C1)–(C4) were sound and complete axioms for probabilistic conditional independence. However, the completeness fails. In fact, finite axiomatization of probabilistic conditional independence is not possible (Studený, 1992).

Let V be a finite set. An *independence model* $\perp_\sigma$ over V is a ternary relation over subsets of a finite set V. The independence model is a *semi-graphoid* if it holds for all mutually disjoint subsets A, B, C, D:

(S1) if $A \perp_\sigma B \,|\, C$ then $B \perp_\sigma A \,|\, C$ (symmetry);
(S2) if $A \perp_\sigma (B \cup D) \,|\, C$ then $A \perp_\sigma B \,|\, C$ and $A \perp_\sigma D \,|\, C$ (decomposition);
(S3) if $A \perp_\sigma (B \cup D) \,|\, C$ then $A \perp_\sigma B \,|\, (C \cup D)$ (weak union);
(S4) if $A \perp_\sigma B \,|\, C$ and $A \perp_\sigma D \,|\, (B \cup C)$, then $A \perp_\sigma (B \cup D) \,|\, C$ (contraction);

The independence model is a *graphoid* if it also satisfies

(S5) if $A \perp_\sigma B \,|\, (C \cup D)$ and $A \perp_\sigma C \,|\, (B \cup D)$ then $A \perp_\sigma (B \cup C) \,|\, D$ (intersection).

Finally, the graphoid is *compositional* if also

(S6) if $A \perp_\sigma B \,|\, C$ and $A \perp_\sigma D \,|\, C$ then $A \perp_\sigma (B \cup D) \,|\, C$ (composition).

The composition property ensures that pairwise conditional independence implies setwise conditional independence, i.e. that

$$A \perp_\sigma B \,|\, C \iff \alpha \perp_\sigma \beta \,|\, C, \quad \forall \alpha \in A, \beta \in B.$$

Example 2.24. (Probabilistic independence models) An important class of independence models are generated by probability distributions. For a system V of labeled random variables $X_v, v \in V$ with distribution P we can define an independence model $\perp\!\!\!\perp_P$ by

$$A \perp\!\!\!\perp_P B \,|\, C \iff X_A \perp\!\!\!\perp_P X_B \,|\, X_C,$$

where $X_A = (X_v, v \in A)$ denotes the variables with labels in A. The general properties (C1)–(C4) of probabilistic conditional independence imply that *probabilistic independence models are semi-graphoids,* but they are not generally graphoids, nor do they generally satisfy composition. From Proposition 2.22 it follows that *if P has strictly positive density w.r.t. a product measure, $\perp\!\!\!\perp_P$ is a graphoid,* but this condition is not neccessary, as indicated in Proposition 2.23. We shall later see that *if P is a regular multivariate Gaussian distribution, $\perp\!\!\!\perp_P$ is a compositional graphoid,*

but in general this is not so for an arbitrary P, reflecting that pairwise independence does not generally imply joint independence. □

Fundamentally, graphical models exploit that graph separation is an independence model, and indeed it is a full compositional graphoid.

Example 2.25. (Separation in an undirected graph) A very important example of a model for the irrelevance axioms above is that of separation in undirected graphs. Let A, B, and C be subsets of the vertex set V of a finite undirected graph $\mathcal{G} = (V, E)$. Define

$$A \perp_{\mathcal{G}} B \,|\, C \iff C \text{ separates } A \text{ from } B \text{ in } \mathcal{G}.$$

Then it is not difficult to see that *undirected graph separation* $\perp_{\mathcal{G}}$ *is a compositional graphoid.* □

Example 2.26. (Second order independence) Sets of random variables A and B are *partially uncorrelated* for fixed C if their residuals after linear regression on X_C are uncorrelated:

$$\mathrm{Cov}\{X_A - \mathbf{E}^*(X_A \,|\, X_C), X_B - \mathbf{E}^*(X_B \,|\, X_C)\} = 0,$$

where $\mathbf{E}^*(X_A \,|\, X_C)$ is the linear regression of X_A on X_C; this is defined as the affine function $\mathbf{E}^*(X_A \,|\, X_C) = \alpha + \beta^\top X_C$ that minimizes the second moment of the residual $\mathbf{E}||X_A - (\alpha + \beta^\top X_C)||^2$. In other words we write $A \perp_2 B \,|\, C$ if all partial correlations $\rho_{AB \cdot C}$ are equal to zero. The relation $\perp_2$ satisfies the semigraphoid axioms (S1)–(S4), composition (S6), and (S5) if there is no non-trivial linear relation between the variables in V. □

Example 2.27. (Geometric orthogonality) As another example, consider geometric orthogonality in Euclidean vector spaces or in Hilbert spaces. Let L, M, and N be linear subspaces of a Euclidean space V and define

$$L \perp M \,|\, N \iff (L \ominus N) \perp (M \ominus N), \tag{2.21}$$

where $L \ominus N = L \cap N^\perp$. Note that this is not the same as standard orthogonality as this usually is defined to imply $L \cap M = \{0\}$ for orthogonal L and M. If (2.21) is satisfied, then L and M are said to *meet orthogonally in* N. Again, it is not hard to see that the orthogonal meet has the following properties:

(O1) if $L \perp M \,|\, N$ then $M \perp L \,|\, N$;
(O2) if $L \perp M \,|\, N$ and U is a linear subspace of L, then $U \perp M \,|\, N$;
(O3) if $L \perp (M + U) \,|\, N$, then $L \perp M \,|\, (N + U)$;
(O4) if $L \perp M \,|\, N$ and $L \perp U \,|\, (M + N)$, then $L \perp (M + U) \,|\, N$.
(O5) if $L \perp M \,|\, N$ and $L \perp N \,|\, M$, then $L \perp (M + N) \,|\, (M \cap N)$.
(O6) if $L \perp M \,|\, N$ and $L \perp U \,|\, N$ then $L \perp (M + U) \,|\, N$.

The direct analogue of (C5) does not hold in general; for example if $M = N$ we may have

$$L \perp M \mid N \text{ and } L \perp N \mid M,$$

but if M and N are not orthogonal then it is false that $L \perp (M + N)$. □

Example 2.28. (Separation in a bidirected graph) There are other notions of graph separation that determine compositional graphoid independence models. If A, B, and C are subsets of the vertex set V of a finite *bidirected* graph $\mathcal{B} = (V, E)$, we say that C *b-separates* A from B in $\mathcal{B}$ and write $A \perp_{\mathcal{B}} B \,|\, C$ if all paths from A to B intersect $V \setminus (A \cup B \cup C)$. This notion is dual to separation in undirected graphs and *bidirected graph separation* $\perp_{\mathcal{B}}$ *determines a compositional graphoid;* see also Theorem 2.31 below. □

Example 2.29. (Variation independence) Let $\mathcal{U} \subseteq \mathcal{X} = \times_{v \in V} \mathcal{X}_v$ and define for $S \subseteq V$ and $u_S^* \in \mathcal{X}_S \cap \mathcal{U}$ the *S-section* $\mathcal{U}^{u_S^*}$ of $\mathcal{U}$ as

$$\mathcal{U}^{u_S^*} = \{u_{V \setminus S} : u_S = u_S^*, u \in \mathcal{U}\}.$$

Define further the conditional independence relation $\ddagger_{\mathcal{U}}$ as

$$A \ \ddagger_{\mathcal{U}} \ B \,|\, S \iff \forall u_S^* : \mathcal{U}^{u_S^*} = \{\mathcal{U}^{u_S^*}\}_A \times \{\mathcal{U}^{u_S^*}\}_B$$

i.e. if and only if the S-sections all have the form of a product space. *The relation* $\ddagger_{\mathcal{U}}$ *satisfies the semigraphoid axioms.* Note in particular that $A \ddagger_{\mathcal{U}} B \,|\, S$ holds if $\mathcal{U}$ is the support of a probability measure satisfying $A \perp\!\!\!\perp B \,|\, S$. □

2.5.2 *Graphical independence models*

A particularly important class of independence models are given by various forms of separation in graphs. We have already seen such examples, the standard separation in undirected graphs in Example 2.25, and its dual, b-separation for bidirected graphs, in Example 2.28. These are both special instances of a more general concept of graph separation for graphs with three types of edge, to be discussed below.

2.5.3 *General graph separation*

We say that a walk ω from α to β in a general graph $\mathcal{G}$ is *active* relative to S, if all collider sections on ω intersect S, and all non-collider sections are disjoint from S. A walk that is not active relative to S is said to be *blocked* by S. If ω is an active walk from $\alpha \in A$ to $\beta \in B$, we also say that ω *connects* A and B.

Definition 2.30 Two subsets A and B of the vertex set V of a graph $\mathcal{G} = (V, E)$ are said to be *g-separated* by S if all walks from A to B are blocked by S and we then write $A \perp_{\mathcal{G}} B \,|\, S$.

We note that if $\mathcal{G}$ is an undirected graph, there are no collider sections, so an active walk is a walk that is disjoint from S; hence, this notion of separation specializes to standard separation for undirected graphs. Similarly, for bidirected graphs all non-endpoints of a walk are collider singletons, so an active path is a path that runs entirely within S. Hence the g-separation in Definition 2.30 also specializes to b-separation in bidirected graphs. For that reason we shall also just say that A and B are separated by S in $\mathcal{G}$ when the conditions in Definition 2.30 are fulfilled and the context will identify the precise meaning of this.

In the following we shall consider other interesting special cases, but first we wish to establish that the independence model $\perp_{\mathcal{G}}$ so defined is indeed a compositional graphoid:

Theorem 2.31 *Let $\mathcal{G} = (V, E)$ be a general graph. Then the independence model $A \perp_{\mathcal{G}} B \,|\, C$ determined by g-separation forms a compositional graphoid.*

Proof Let $\mathcal{G} = (V, E)$ and consider disjoint subsets A, B, C, and D of V. We verify each of the six properties separately.

Symmetry: If $A \perp_{\mathcal{G}} B \,|\, C$ then $B \perp_{\mathcal{G}} A \,|\, C$. This is obvious from the definition.

Decomposition: If $A \perp_{\mathcal{G}} (B \cup D) \,|\, C$ then $A \perp_{\mathcal{G}} B \,|\, C$ and $A \perp_{\mathcal{G}} D \,|\, C$. Also immediate from the definition.

Weak union: If $A \perp_{\mathcal{G}} (B \cup D) \,|\, C$ then $A \perp_{\mathcal{G}} B \,|\, (C \cup D)$. Using decomposition yields $A \perp_{\mathcal{G}} D \,|\, C$ and $A \perp_{\mathcal{G}} B \,|\, C$. Now suppose, for contradiction, that there exist a connecting walk ω from A to B relative to $C \cup D$. Then ω must have at least one collider section, or it would also be connecting between A and B relative to C. All collider sections on ω must intersect $(C \cup D)$ for ω to be connecting. Also, ω must have a collider section intersecting D but disjoint from C for else it would also be connecting relative to C. Consider the collider section ρ which is closest to A on ω and let δ be the point nearest to A on ρ. The vertices between A and δ on ω are not in $B \cup D$ and hence the subwalk of ω consisting of these vertices connects between A and δ relative to C. This contradicts the assumption that $A \perp_{\mathcal{G}} (B \cup D) \,|\, C$.

Contraction: If $A \perp_{\mathcal{G}} B \,|\, C$ and $A \perp_{\mathcal{G}} D \,|\, (B \cup C)$ then $A \perp_{\mathcal{G}} (B \cup D) \,|\, C$. Suppose, for contradiction, that there exists a walk between A and $B \cup D$ which is connecting relative to C. Consider a shortest walk of this type and call it ω. This walk must be between A and D since $A \perp_{\mathcal{G}} B \,|\, C$. In addition, since all collider sections on ω intersect C. As $A \perp_{\mathcal{G}} D \,|\, (B \cup C)$, ω must have a non-collider section that intersects B. This contradicts the fact that ω is a shortest walk between A and $B \cup D$ which is connecting relative to C.

Intersection: If $A \perp_{\mathcal{G}} B \,|\, (C \cup D)$ and $A \perp_{\mathcal{G}} D \,|\, (C \cup B)$ then $A \perp_{\mathcal{G}} (B \cup D) \,|\, C$. Suppose, for contradiction, that there exists a walk between A and $B \cup D$ that is connecting relative to C. Consider a shortest walk of this type and call it ω. The walk ω is either between A and B or between A and D. By symmetry we may assume that ω is between A and B. Since all collider sections on ω intersect C and $A \perp_{\mathcal{G}} B \,|\, (C \cup D)$ ω must have a non-collider section that intersects D. This contradicts that ω is a shortest walk between A and $B \cup D$ that is connecting relative to C.

Composition: If $A \perp_{\mathcal{G}} B \,|\, C$ and $A \perp_{\mathcal{G}} D \,|\, C$ then $A \perp_{\mathcal{G}} (B \cup D) \,|\, C$. This is obvious from the definition.

The proof is now complete. □

Theorem 2.31 implies that we can focus on establishing conditional independence for pairs of nodes, formulated in the corollary below.

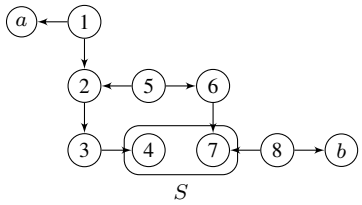


FIG. 2.1. Example of an active walk: $(a, 1, 2, 3, 4, 3, 2, 5, 6, 7, 8, b)$. The colliders on the walk are 4 and 7 which are both in S and all non-colliders are outside S; hence the walk connects a and b relative to S. The shortest walk $(a, 1, 2, 5, 6, 7, 8, b)$ from a to b is blocked by 2 which is a collider on that walk and outside S.

Corollary 2.32 *Let $\mathcal{G} = (V, E)$ and A, B, and C be disjoint subsets of V. Then $A \perp_{\mathcal{G}} B \,|\, C$ if and only if $\alpha \perp_{\mathcal{G}} \beta \,|\, C$ for every pair $\alpha \in A$ and $\beta \in B$.*

Proof This follows because $\perp_{\mathcal{G}}$ satisfies decomposition and composition. □

2.5.4 *Directed acyclic graphs*

For the case of a directed acyclic graph $\mathcal{D}$ the notion of g-separation in Definition 2.30 is equivalent (Koster, 2002) to what has been known as *d-separation* and differs from g-separation by using paths instead of walks and considering paths to be active if colliders are in $\mathrm{An}(S)$ and all non-colliders outside of S. We shall use the term d-separation also for the version with walks and denote this by $\perp_{\mathcal{D}}$.
To answer a query whether a given triple (A, B, S) satisfies $A \perp_{\mathcal{D}} B \,|\, S$ one must potentially check an infinite number of walks. However, there is an alternative method for checking d-separation in terms of standard separation in a suitable undirected graph, associated with the query.
More precisely we say that A is *m-separated* from B by S and we write $A \perp_m B \,|\, S$ if S separates A from B in $(\mathcal{D}_{\mathrm{An}(A \cup B \cup S)})^m$, i.e.

$$A \perp_m B \,|\, S \iff A \perp_{(\mathcal{D}_{\mathrm{An}(A \cup B \cup S)})^m} B \,|\, S.$$

We then have:

Proposition 2.33 *Let A, B and S be disjoint subsets of a directed, acyclic graph $\mathcal{G}$. Then $A \perp_{\mathcal{D}} B \,|\, S \iff A \perp_m B \,|\, S$.*

Proof Since d-separation is a special instance of g-separation, it follows from Corollary 2.32 that we only need to consider the case where A and B are singletons a and b. We show the result by showing that every g-connecting walk between a and b corresponds to a connecting walk in $(\mathcal{D}_{\mathrm{An}(a \cup b \cup S)})^m$ and *vice versa.*
Suppose S does not d-separate a from b. Then there is a connecting walk ω from a to b such as, for example, indicated in Fig. 2.1. This walk must lie proceed entirely within $\mathrm{An}(a \cup b \cup S)$. For if some vertex $\gamma \in \omega$ is a collider on ω we must have $\gamma \in S$ or the walk would be blocked; and if γ is neither an ancestor of S nor a collider on the walk, at least one of the subwalks away from γ would lead to a or b. If γ is a collider,

$\mathrm{pa}(\gamma) \cap \omega$ must be outside S or the walk would be blocked. If $\mathrm{pa}(\gamma) \cap \omega = \{\delta\}$ is a singleton, as in vertex 5 of Fig. 2.1, we shorten the walk omitting the subwalk (δ, γ, δ) from ω and obtain a walk from a to b not including γ; see Fig. 2.2. If

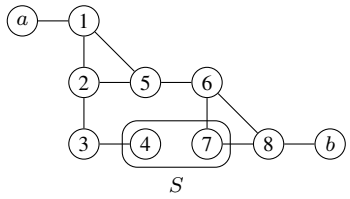


FIG. 2.2. The moral graph corresponding to the connecting walk in $\mathcal{D}$. Modifying the walk $\omega = (a, 1, 2, 3, 4, 3, 2, 5, 6, 7, 8, b)$ between a and b to become $\omega^* = (a, 1, 2, 3, 2, 5, 6, 8, b)$ which connects a and b in the moral graph.

$\mathrm{pa}(\gamma) \cap \omega = \{\delta_1, \delta_2\}$ as for vertex 7 in Fig. 2.1, the marriage in the moral graph between δ_1 and δ_2 means we can similarly replace $(\delta_1, \gamma, \delta_2)$ with (δ_1, δ_2) and still have a walk in the moral graph. Continuing in this fashion for all colliders on ω yields a walk ω^* from a to b in $(\mathcal{D}_{\mathrm{An}(a \cup b \cup S)})^m$, circumventing S.
Suppose conversely that a is not separated from b in $(\mathcal{D}_{\mathrm{An}(a \cup b \cup S)})^m$. Then there is a walk ω in this graph that circumvents S. The walk has pieces that correspond to edges in the original graph and pieces that correspond to marriages. Each marriage edge (δ_1, δ_2) connects the parents of some vertex γ. If γ is in S we can simply replace the piece (δ_1, δ_2) of the walk with $(\delta_1, \gamma, \delta_2)$ as γ does not block this part of the walk in $\mathcal{D}$. If $\gamma \notin S$ but has a descendant $\sigma \in S$, we instead modify the walk by replacing (δ_1, δ_2) with $(\delta_1, \gamma, \ldots, \sigma, \ldots, \gamma, \delta_2)$, thus extending ω with a walk from γ to its first descendant $\sigma \in S$ and back again to γ. If neither of these, a or b must be a descendant of γ since the ancestral set was smallest. In the latter case, a new walk can be created in the new graph with one collider less, using the line of descent, such as illustrated in Fig. 2.3. Continuing this substitution process eventually leads to an active walk from a to b. Thus a is not d-separated from b by S and the proof is complete. □

The use of Proposition 2.33 for deciding whether two sets A and B are separated by S is illustrated in the following example.

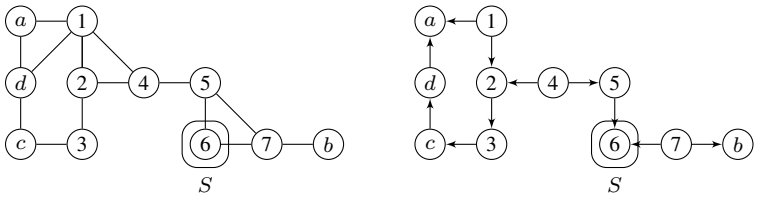


FIG. 2.3. The walk $(a, 1, 4, 5, 7, b)$ in $(\mathcal{D}_{\mathrm{An}(a \cup b \cup S)})^m$ enables the construction of a connecting walk $(a, d, c, 3, 2, 4, 5, 6, 7, b)$ in $\mathcal{D}$.

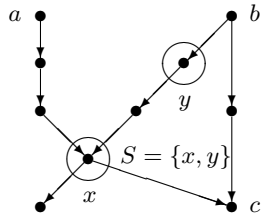


FIG. 2.4. Is $\{a\}$ d-separated from b by $S = \{x, y\}$ in this graph? By $S^* = \{x\}$? By $S^{**} = \{y\}$?

Example 2.34 Consider the directed acyclic graph in Fig. 2.4 and the problem of deciding whether $a \perp_{\mathcal{D}} b \,|\, S$.
There are two paths from a to b, one of them passing through y and one of them avoiding y; the first path is blocked by y, whereas the second path is blocked both by x and c. Thus we have $a \perp_{\mathcal{D}} b \,|\, \{x, y\}$. If we consider $S^* = \{x\}$ then the first path (via y) becomes d-connecting so S^* is not d-separating. For $S^{**} = \{y\}$, the first path is still blocked both by y and x, whereas the second path is blocked by the collider node c, and hence it also holds that $a \perp_{\mathcal{D}} b \,|\, \{y\}$. In this particular case we need only consider paths and not walks.
The moral graph of the smallest ancestral set containing all the variables involved is shown to the left in Fig. 2.5. It is immediate that S separates a from b in the graph to

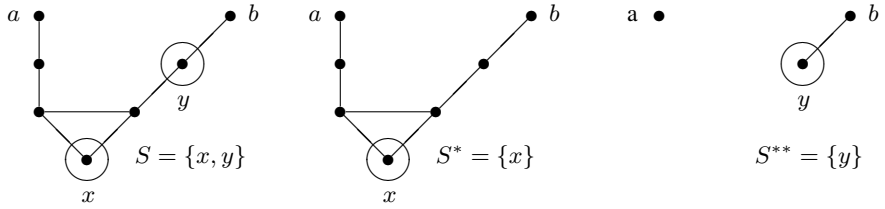


FIG. 2.5. The moral graphs of the smallest ancestral sets in the graph of Fig. 2.4 containing all variable involved are shown from left to right. S separates a from b in the graph to the left, S^{**} separates in the graph to the right, wheras S^* does not separate in the graph in the middle.

the left, implying $a \perp_{\mathcal{D}} b \,|\, S$ and similarly we have $a \perp_{\mathcal{D}} b \,|\, S^{**}$, whereas $\neg(a \perp_{\mathcal{D}} b \,|\, S^*)$. □

2.6 Markov properties

In this section we consider relationships between probabilistic independence models — as described in Example 2.24 — and graphical independence models. Such relations take the form of *Markov properties* which typically involve statements of the

form that certain graph separations imply conditional independence statements in the probabilistic independence model. More generally, $\mathcal{G}$-Markov properties are relationships between an independence model $\perp_\sigma$ and separation properties in $\mathcal{G}$.

Thus we shall consider the situation where we have a collection of random variables $(X_\alpha)_{\alpha\in V}$ taking values in Borel spaces $(\mathcal{X}_\alpha)_{\alpha\in V}$. The probability spaces are either real finite-dimensional vector spaces or finite and discrete sets. For A being a subset of V we let $\mathcal{X}_A = \times_{\alpha\in A}\mathcal{X}_\alpha$ and further $\mathcal{X} = \mathcal{X}_V$. Typical elements of $\mathcal{X}_A$ are denoted as $x_A = (x_\alpha)_{\alpha\in A}$. Similarly $X_A = (X_\alpha)_{\alpha\in A}$. We then use the short notation $A \perp\!\!\!\perp B \mid C$ for $X_A \perp\!\!\!\perp X_B \mid X_C$ and so on. If we specifically want to emphasize the dependence of the conditional independence relation on the specific probability distribution P, we write $\perp\!\!\!\perp_P$, but most of the time P will be fixed and given by the context.
If a graph $\mathcal{G} = (V, E)$ is given with vertex set equal to the labels of the random variables $(X_\alpha)_{\alpha\in V}$ we define

Definition 2.35 P is said to satisfy the *global Markov property* w.r.t. the graph $\mathcal{G}$ if for all disjoint subsets A, B, S of V

$$\text{(G)} \qquad A \perp_{\mathcal{G}} B \,|\, S \implies A \perp\!\!\!\perp_P B \,|\, S.$$

In other words, separation in the graph $\mathcal{G}$ implies conditional independence; or the map $A \to X_A$ is an independence homomorphism. We shall also say that X (or P) *is globally Markov* w.r.t. $\mathcal{G}$ if (G) holds. Further, we say:

Definition 2.36 P is said to be *faithful* to the graph $\mathcal{G}$ if for all disjoint subsets A, B, S of V

$$A \perp_{\mathcal{G}} B \,|\, S \iff A \perp\!\!\!\perp_P B \,|\, S.$$

Thus a distribution P is faithful to $\mathcal{G}$ if and only if the independence models $\perp_{\mathcal{G}}$ and $\perp\!\!\!\perp_P$ are isomorphic. For a faithful distribution, the graph therefore gives a full picture of the conditional independence relations among subsets of variables. Note that even for a faithful distribution, there might be conditional independence relations among other functions of the random variables than the coordinate projections $X \to X_A, A \subseteq V$.
Whereas the global Markov property and the notion of faithfulness relates universally to any graph with a separation property, there are a variety of Markov properties that make reference to special types of graph. We shall discuss these in the following subsections.

2.6.1 *Markov properties on undirected graphs*

Consider an undirected graph $\mathcal{G} = (V, E)$ and a collection of random variables $(X_\alpha)_{\alpha\in V}$ as above.
In addition to the global Markov property defined by separation in Definition 2.35 above, we have a range of different Markov properties. A probability measure P on $\mathcal{X}$ is said to obey

(P) *the pairwise Markov property* relative to $\mathcal{G}$, if for any pair (α, β) of vertices

$$\alpha \not\sim \beta \implies \alpha \perp\!\!\!\perp \beta \,|\, V \setminus \{\alpha, \beta\};$$

(L) *the local Markov property* relative to $\mathcal{G}$, if for any vertex $\alpha \in V$

$$\alpha \perp\!\!\!\perp V \setminus \mathrm{cl}(\alpha) \,|\, \mathrm{bd}(\alpha);$$

(G) *the global Markov property* relative to $\mathcal{G}$, if separation implies conditional independence

$$A \perp_{\mathcal{G}} B \,|\, S \implies A \perp\!\!\!\perp B \,|\, S.$$

The Markov properties are related as described in the proposition below.

Proposition 2.37 *For any undirected graph $\mathcal{G}$ and any probability distribution on $\mathcal{X}$ it holds that*

$$\text{(G)} \implies \text{(L)} \implies \text{(P)}. \tag{2.22}$$

Proof Firstly, (G) implies (L) because $\mathrm{bd}(\alpha)$ separates α from $V \setminus \mathrm{cl}(\alpha)$. Assume next that (L) holds. We have $\beta \in V \setminus \mathrm{cl}(\alpha)$ because α and β are non-adjacent. Hence

$$\mathrm{bd}(\alpha) \cup ((V \setminus \mathrm{cl}(\alpha)) \setminus \{\beta\}) = V \setminus \{\alpha, \beta\},$$

and it follows from (L) and (C3) that

$$\alpha \perp\!\!\!\perp V \setminus \mathrm{cl}(\alpha) \mid V \setminus \{\alpha, \beta\}.$$

Application of (C2) then gives $\alpha \perp\!\!\!\perp \beta \,|\, V \setminus \{\alpha, \beta\}$ which is (P). □

It is worth noting that the proof of (2.22) only exploits the properties (C1)–(C4) of conditional independence and hence also holds for any semi-graphoid independence model $\perp_\sigma$.

The various Markov properties are different in general; but if the conditional independence relation $\perp\!\!\!\perp_P$ induced by P is a graphoid, the Markov properties are all equivalent. The result is stated in the theorem below, due to Pearl and Paz (1987); see also Pearl (1988).

Theorem 2.38. (Pearl and Paz) *If a probability distribution P on $\mathcal{X}$ is such that its independence model $\perp\!\!\!\perp_P$ is a graphoid, it holds that*

$$\text{(G)} \iff \text{(L)} \iff \text{(P)}.$$

Proof We need to show that (P) implies (G), so assume that $A \perp_{\mathcal{G}} B \,|\, S$, (P) holds, and $\perp\!\!\!\perp_P$ satisfies (S1)–(S5). Without loss of generality we may assume that both A and B are non-empty. The proof is then reverse induction on the number of vertices $n = |S|$ in S. If $n = |V| - 2$ then both A and B consist of one vertex and the required conditional independence follows from (P).

So assume $|S| = n < |V| - 2$ and that separation implies conditional independence for all separating sets S with more than n elements. We first assume that

$V = A \cup B \cup S$, implying that at least one of A and B has more than one element, A, say. If $\alpha \in A$ then $A \setminus \{\alpha\} \perp_{\mathcal{G}} B \mid S \cup \{\alpha\}$ and also $\alpha \perp_{\mathcal{G}} B \mid S \cup A \setminus \{\alpha\}$; thus by the induction hypothesis

$$A \setminus \{\alpha\} \perp\!\!\!\perp_P B \mid S \cup \{\alpha\} \text{ and } \alpha \perp\!\!\!\perp_P B \mid S \cup A \setminus \{\alpha\}.$$

Now (S5) for $\perp\!\!\!\perp_P$ gives $A \perp\!\!\!\perp_P B \mid S$.
If $A \cup B \cup S \subset V$ we choose $\alpha \in V \setminus (A \cup B \cup S)$. Then $A \perp_{\mathcal{G}} B \mid S \cup \{\alpha\}$ implying $A \perp\!\!\!\perp_P B \mid S \cup \{\alpha\}$. Further, either $\alpha \perp_{\mathcal{G}} B \mid A \cup S$ separates B from $\{\alpha\}$ or $\alpha \perp_{\mathcal{G}} A \mid B \cup S$. Assuming the former gives $\alpha \perp\!\!\!\perp_P B \mid A \cup S$. Using (S5) and (S2) we derive that $A \perp\!\!\!\perp_P B \mid S$. The latter case is similar. □

Note again that the proof only exploits (S1)–(S5) and therefore applies to any graphoid independence model $\perp_\sigma$.

The global Markov property (G) is important because it gives a general criterion for deciding when two groups of variables A and B are conditionally independent given a third group of variables S.

As conditional independence is intimately related to factorization, so are the Markov properties. In the following, symbols ψ_a, ϕ_a for $a \subseteq V$, etc. denote functions that depend on x through its coordinates in a only, i.e.

$$x_a = y_a \implies \psi_a(x) = \psi_a(y).$$

Further we let $f = \mathrm{d}P/\mathrm{d}\mu$ denote the density of P w.r.t. a product measure $\mu = \otimes_{\alpha \in V} \mu_\alpha$ on $\mathcal{X}$. We then define:

Definition 2.39 A probability measure P on $\mathcal{X}$ is said to *factorize* according to $\mathcal{G}$ if for all complete subsets $a \subseteq V$ there exist non-negative functions ψ_a that depend on x through x_a only so that

$$f(x) = \prod_{a \text{ complete}} \psi_a(x). \tag{2.23}$$

The functions ψ_a are not uniquely determined. There is arbitrariness in the choice of μ, but also groups of functions ψ_a can be multiplied together or split up in different ways. In fact one can without loss of generality assume — although this is not always practical — that only cliques appear as the sets a, i.e. that

$$f(x) = \prod_{c \in \mathcal{C}} \psi_c(x), \tag{2.24}$$

where $\mathcal{C}$ is the set of cliques of $\mathcal{G}$. If P factorizes, we say that P has property (F) and the set of such probability measures is denoted by $M_F(\mathcal{G})$. We have

Proposition 2.40 *For any undirected graph $\mathcal{G}$ and any probability distribution on $\mathcal{X}$ it holds that*

$$\text{(F)} \implies \text{(G)} \implies \text{(L)} \implies \text{(P)}.$$

Proof We only have to show that (F) implies (G) as the remaining implications are given in Proposition 2.37. Let (A, B, S) be any triple of disjoint subsets such that S separates A from B. Let $\tilde{A}$ denote the connectivity components in $\mathcal{G}_{V\setminus S}$ which contain A and let $\tilde{B} = V \setminus (\tilde{A} \cup S)$. Since A and B are separated by S, their elements are in different connectivity components of $\mathcal{G}_{V\setminus S}$ and any clique of $\mathcal{G}$ is either a subset of $\tilde{A} \cup S$ or of $\tilde{B} \cup S$. If $\mathcal{C}_A$ denotes the cliques contained in $\tilde{A} \cup S$, we thus obtain from (2.24) that

$$f(x) = \prod_{c\in\mathcal{C}} \psi_c(x) = \prod_{c\in\mathcal{C}_A} \psi_c(x) \prod_{c\in\mathcal{C}\setminus\mathcal{C}_A} \psi_c(x) = h(x_{\tilde{A}\cup S})k(x_{\tilde{B}\cup S}).$$

Hence (2.16) gives that $\tilde{A} \perp\!\!\!\perp \tilde{B} \mid S$. Applying (C2) twice gives the desired independence. □

In the case where P has a positive density we can use the Möbius inversion lemma to show that (P) implies (F), and thus all Markov properties are equivalent. This result seems to have been discovered independently in various forms by a number of authors (Averintsev, 1970; Spitzer, 1971; Besag, 1972; Besag, 1974), see Clifford (1990). The result is often referred to as the Hammersley–Clifford Theorem. However, Hammersley and Clifford (1971) actually prove that (L) implies (F) under the positivity assumption, so this is slightly inaccurate. The proof given below is essentially identical to the proof given by Grimmett (1973); see also Koster (1994).

Theorem 2.41 *A probability distribution P with positive density f with respect to a product measure μ satisfies the pairwise Markov property with respect to an undirected graph $\mathcal{G}$ if and only if it factorizes according to $\mathcal{G}$.*

Proof If P factorizes, it is pairwise Markov as shown in Proposition 2.40, so we just have to show that (P) implies (F).
Since the density is positive, we may take logarithms on both sides of (2.23). Hence this equation can be rewritten as

$$\log f(x) = \sum_{a:a\subseteq V} \phi_a(x), \tag{2.25}$$

where $\phi_a(x) = \log \psi_a(x)$ and $\phi_a \equiv 0$ unless a is a complete subset of V.
Assume then that P is pairwise Markov and choose a fixed but arbitrary element $x^* \in \mathcal{X}$. Define for all $a \subseteq V$

$$H_a(x) = \log f(x_a, x^*_{a^c}),$$

where $(x_a, x^*_{a^c})$ is the element y with $y_\gamma = x_\gamma$ for $\gamma \in a$ and $y_\gamma = x^*_\gamma$ for $\gamma \notin a$. Since x^* is fixed, H_a depends on x through x_a only. Let further for all $a \subseteq V$

$$\phi_a(x) = \sum_{b:b\subseteq a} (-1)^{|a\setminus b|} H_b(x).$$

From this relation it is also clear that ϕ_a depends on x through x_a only. Next we can apply Lemma A.12 (Möbius inversion) to obtain that

$$\log f(x) = H_V(x) = \sum_{a:a\subseteq V} \phi_a(x)$$

such that we have proved the theorem if we can show that $\phi_a \equiv 0$ whenever a is not a complete subset of V. So let us assume that $\alpha, \beta \in a$ and $\alpha \not\sim \beta$. Let further $c = a \setminus \{\alpha, \beta\}$. If we write H_a as short for $H_a(x)$ we have

$$\phi_a(x) = \sum_{b:b\subseteq c} (-1)^{|c\setminus b|} \left\{ H_b - H_{b\cup\{\alpha\}} - H_{b\cup\{\beta\}} + H_{b\cup\{\alpha,\beta\}} \right\}. \tag{2.26}$$

Let $d = V \setminus \{\alpha, \beta\}$. Then, by the pairwise Markov property and (2.17), we have

$$\begin{aligned}
H_{b\cup\{\alpha,\beta\}}(x) - H_{b\cup\{\alpha\}}(x) &= \log \frac{f(x_b, x_\alpha, x_\beta, x^*_{d\setminus b})}{f(x_b, x_\alpha, x^*_\beta, x^*_{d\setminus b})} \\
&= \log \frac{f(x_\alpha \,|\, x_b, x^*_{d\setminus b}) f(x_\beta, x_b, x^*_{d\setminus b})}{f(x_\alpha \,|\, x_b, x^*_{d\setminus b}) f(x^*_\beta, x_b, x^*_{d\setminus b})} \\
&= \log \frac{f(x^*_\alpha \,|\, x_b, x^*_{d\setminus b}) f(x_\beta, x_b, x^*_{d\setminus b})}{f(x^*_\alpha \,|\, x_b, x^*_{d\setminus b}) f(x^*_\beta, x_b, x^*_{d\setminus b})} \\
&= \log \frac{f(x_b, x^*_\alpha, x_\beta, x^*_{d\setminus b})}{f(x_b, x^*_\alpha, x^*_\beta, x^*_{d\setminus b})} \\
&= H_{b\cup\{\beta\}}(x) - H_b(x).
\end{aligned}$$

Thus all terms in the curly brackets in (2.26) add to zero and henceforth the entire sum is zero. This completes the proof. □

When (A, B, S) form a decomposition of $\mathcal{G}$ the Markov properties decompose accordingly. This is expressed formally in three propositions below.

Proposition 2.42 *Assume that (A, B, S) decompose $\mathcal{G} = (V, E)$. Then a probability distribution P factorizes with respect to $\mathcal{G}$ if and only if both its marginal distributions $P_{A\cup S}$ and $P_{B\cup S}$ factorize with respect to $\mathcal{G}_{A\cup S}$ and $\mathcal{G}_{B\cup S}$ respectively and the densities f satisfy*

$$f(x) f_S(x_S) = f_{A\cup S}(x_{A\cup S}) f_{B\cup S}(x_{B\cup S}). \tag{2.27}$$

Proof Suppose that p factorizes with respect to $\mathcal{G}$ such that

$$f(x) = \prod_{c\in\mathcal{C}} \psi_c(x).$$

Since (A, B, S) decomposes $\mathcal{G}$, all cliques are either subsets of $A \cup S$ or of $B \cup S$. Let $\mathcal{A}$ denote the cliques that are subsets of $A \cup S$ and $\mathcal{B}$ those that are subsets of $B \cup S$. Then

$$f(x) = \prod_{c\in\mathcal{A}} \psi_c(x) \prod_{c\in\mathcal{B}\setminus\mathcal{A}} \psi_c(x) = h(x_{A\cup S})k(x_{B\cup S}).$$

By direct integration we find

$$f_{A\cup S}(x_{A\cup S}) = h(x_{A\cup S})\bar{k}(x_S)$$

where

$$\bar{k}(x_S) = \int k(x_{B\cup S})\mu_B(dx_B),$$

and similarly with the other marginals. This gives (2.27) as well as the factorizations of both marginal densities.
Conversely, assume that (2.27) holds and that $f_{A\cup S}$ and $f_{B\cup S}$ factorize. Then let

$$\psi_S(x_S) = \begin{cases} \frac{1}{f_S(x_S)} & \text{if } f_S(x_S) \neq 0 \\ 0 & \text{otherwise.} \end{cases}$$

Since f_S is obtained from integration of f, the latter must also be almost everywhere zero when f_S is. Hence

$$\tilde{f}(x) = f_{A\cup S}(x_{A\cup S}) f_{B\cup S}(x_{B\cup S})\psi_S(x_S)$$

is a density for P and P factorizes. □

Recursive use of Proposition 2.42 yields

Corollary 2.43 *If $\mathcal{G}$ is decomposable and P has density f with respect to a product measure μ on $\mathcal{X}$, it holds that*

$$f(x) \prod_{S\in\mathcal{S}} f_S(x_S)^{\nu(S)} = \prod_{C\in\mathcal{C}} f_C(x_C), \tag{2.28}$$

where $\mathcal{S}$ are the separators of $\mathcal{G}$ with multiplicities $\nu(S)$, $\mathcal{C}$ the set of cliques of $\mathcal{G}$, and $f_A, A \in \mathcal{C}\cup\mathcal{S}$ are the marginal densities of $X_A, A \in \mathcal{C}\cup\mathcal{S}$.

2.6.2 *Markov properties on directed acyclic graphs*

For a directed acyclic graph, there are again several relevant Markov properties other than the global Markov property (G) as defined by graph separation in Definition 2.35 or, equivalently, via m-separation, cf. Proposition 2.33.

An ordering $V = \{1, \ldots, d\}$ of the vertices of a DAG $\mathcal{D}$ is *topological* if all arrows point from lower to higher: $\alpha \to \beta \implies \alpha < \beta$. The *predecessors* of a vertex β are $\mathrm{pr}\,\beta = \{\alpha \in V : \alpha < \beta\}$. Now an independence model $\perp_\sigma$ is said to obey

(O) *the ordered Markov property* relative to $(\mathcal{D}, <)$, if any vertex α is is conditionally independent of its predecessors, given its parents:

$$\alpha \perp_\sigma \mathrm{pr}(\alpha) \setminus \mathrm{pa}(\alpha) \mid \mathrm{pa}(\alpha).$$

(L) *the local Markov property* relative to $\mathcal{D}$, if if any variable is conditionally independent of its non-descendants, given its parents:

$$\alpha \perp_\sigma \mathrm{nd}(\alpha) \setminus \mathrm{pa}(\alpha) \mid \mathrm{pa}(\alpha).$$

(G) *the global Markov property* relative to $\mathcal{D}$, if separation implies conditional independence

$$A \perp_{\mathcal{D}} B \mid S \implies A \perp_\sigma B \mid S.$$

In constrast to the undirected case, these Markov properties are all equivalent without further regularity assumptions. Indeed we have:

Theorem 2.44 *Let $\mathcal{D}$ be a directed, acyclic graph and $\perp_\sigma$ a semigraphoid independence model. on $\mathcal{X}$. Then the following conditions are equivalent*

(G) *$\perp_\sigma$ obeys the directed global Markov property, relative to $\mathcal{D}$;*
(L) *$\perp_\sigma$ obeys the directed local Markov property, relative to $\mathcal{D}$;*
(O) *$\perp_\sigma$ obeys the ordered local Markov property relative to $(\mathcal{D}, <)$ where $<$ is a topological ordering of $\mathcal{D}$.*

Proof That (G) implies (L) follows by observing that $\{\alpha\} \cup \mathrm{nd}(\alpha)$ is an ancestral set and that $\mathrm{pa}(\alpha)$ obviously separates $\{\alpha\}$ from $\mathrm{nd}(\alpha) \setminus \mathrm{pa}(\alpha)$ in $(\mathcal{G}_{\{\alpha\}\cup\mathrm{nd}(\alpha)})^m$.

To show that (L) implies (G) we use induction on the number $|V|$ of vertices in $\mathcal{D}$. For less than or equal to two vertices there is nothing to show. So assume the conclusion holds for all DAGs with less than or equal to n vertices and assume $|V| = n + 1$. We then consider three disjoint subsets (A, B, S) and may w.l.o.g. assume $\mathrm{An}(A \cup \mathcal{B} \cup S) = V$ for else the inductive assumption would yield the desired conclusion.

Hence assume that $A \perp_{\mathcal{G}} B \mid S$ in $\mathcal{G} = \mathcal{D}^m$. Let v^* be a terminal vertex in $\mathcal{D}$. The separation implies that either $v^* \in A \cup S$ or $v^* \in B \cup S$ since parents are married in $\mathcal{D}^m$, so assume the former.

Consider first the case $v^* \in A$. Since v^* has no descendants, (L) yields $v^* \perp_\sigma (V \setminus \mathrm{pa}(v^*)) \mid \mathrm{pa}(v^*)$; separation implies that also $\mathrm{pa}(v^*) \subseteq A \cup S$; now weak union (S3) yields $v^* \perp_\sigma B \mid (A \setminus \{v^*\}) \cup S$. Also, S must separate $A \setminus \{v^*\}$ from B in $(\mathcal{D}_{V\setminus v^*})^m$ since this has no more edges than $\mathcal{D}^m$. From the inductive hypothesis we now conclude that $(A \setminus \{v^*\}) \perp_\sigma B \mid S$. The case $v^* \in B$ is similar.

Next we consider the case $v^* \in S$. First, realise that if S separates A from B in $\mathcal{D}^m$, then $S \setminus \{v^*\}$ separates A from B in $(\mathcal{D}^m)_{V\setminus v^*}$ and hence also in $(\mathcal{D}^m)_{V\setminus v^*})$. The inductive hypothesis yields $A \perp_\sigma B \mid (S \setminus \{v^*\})$. If $\mathrm{pa}(v^*) \subseteq A \cup S$, (L) in combination with (S3) yields $v^* \perp_\sigma B \mid (A \cup (S \setminus \{v^*\}))$; contraction (S4) yields $(A \cup \{v^*\}) \perp_\sigma B \mid (S \setminus \{v^*\})$; finally weak union (S3) yields $A \perp_\sigma B \mid S$ as desired. The case with $\mathrm{pa}(v^*) \subseteq B \cup S$ is similar.

To show that (L) implies (O) we observe that $\mathrm{pr}(\alpha) \subseteq \mathrm{nd}(\alpha)$ so (S2)

$$\alpha \perp_\sigma \mathrm{nd}(\alpha) \setminus \mathrm{pa}(\alpha) \implies \alpha \perp_\sigma \mathrm{pr}(\alpha) \setminus \mathrm{pa}(\alpha).$$

Finally we show that (O) implies (L) by induction on the number $n = |V|$ of vertices of $\mathcal{D}$. For $n \leq 2$ there is nothing to show. Assume the statement is true for $|V| \leq n$ and

consider a DAG with $n+1$ vertices so that (O) holds for $\perp_\sigma$ and let $v^* = n+1$. Then for any $\alpha \leq n$ the inductive assumption implies

$$\alpha \perp_\sigma \mathrm{nd}(\alpha) \setminus (\mathrm{pa}(\alpha) \cup \{v^*\}) \mid \mathrm{pa}(\alpha). \tag{2.29}$$

If $v^* \in \mathrm{de}(\alpha)$ we thus have $\alpha \perp_\sigma \mathrm{nd}(\alpha) \setminus \mathrm{pa}(\alpha) \mid \mathrm{pa}(\alpha)$. Otherwise we must have $\mathrm{pa}(v^*) \subseteq \mathrm{nd}(\alpha)$ and $v^* \perp_\sigma V \setminus \{v^*\} \mid \mathrm{pa}(v^*)$, which by (S2) and (S3) implies $v^* \perp_\sigma \alpha \,|\, (\mathrm{nd}(\alpha) \setminus \{v^*\})$. Combining this fact with (2.29) and (S4) now yields $\alpha \perp_\sigma \mathrm{nd}(\alpha) \setminus \mathrm{pa}(\alpha) \mid \mathrm{pa}(\alpha)$. The case $\alpha = v^*$ is trivial. Hence this completes the proof. □

Corollary 2.45 *If $\perp_\sigma$ satisfies* (O) *w.r.t. a specific topological ordering it satisfies* (O) *w.r.t. any topological ordering.*

Proof This is immediate since (O) is equivalent to (G) and (L) and these do not refer to any ordering. □

Consider next a system $(P^v)_{v\in V}$ of Markov kernels that are *adapted* to $\mathcal{D}$ in the sense that each P^v is a Markov kernel from $\mathcal{X}_{\mathrm{pa}(v)}$ to $\mathcal{X}_v$. We then define the *recursive combination* of $(P^v)_{v\in V}$ w.r.t. a topological numbering of $\mathcal{D}$ as $P^< = \circledast_{v\in V} P^v$, where combinations are made according to $<$. We then have:

Proposition 2.46 *The recursive combination is well-defined and independent of the specific topological ordering.*

Proof This is shown by induction on the number $|V|$ of vertices in $\mathcal{D}$. For $|V| \leq 2$ this is immediate.
Next assume the statement is true for $|V| \leq n$ and consider a DAG with $n+1$ vertices. We can then let v^* be a terminal vertex in $\mathcal{D}$ and by Proposition 2.15 write

$$P = \left(\circledast_{v\in V\setminus\{v^*\}} P^v\right) \circledast P^{v^*}$$

where v^* is the last vertex in a topological ordering, so the combination is well-defined. If two topological orderings both have $v^* = n+1$, the inductive assumption also ensures that the corresponding recursive combinations are identical. If two different terminal vertices v^* and v^{**} are last in two topological orderings it follows from Proposition 2.17 that

$$\left(\circledast_{v\in V\setminus\{v^*,v^{**}\}} P^v\right) \circledast P^{v^*} \circledast P^{v^{**}} = \left(\circledast_{v\in V\setminus\{v^*,v^{**}\}} P^v\right) \circledast P^{v^{**}} \circledast P^{v^*}$$

and hence the inductive assumption ensures that the recursive combinations are identical. □

Remark: We may therefore without ambiguity omit the specific topological ordering and write $P = \circledast_{v\in V} P^v$.

Definition 2.47 The *Bayesian network* $(\mathcal{D}, P)$ generated by $\mathcal{D}$ and $(P^v)_{v\in V}$ is the recursive combination $P = \circledast_{v\in V} P^v$ of $(P^v)_{v\in V}$ w.r.t. $\mathcal{D}$.

For a DAG $\mathcal{D}$ we say that P admits a *recursive kernel factorization* (R) if there exists a a system $(P^v)_{v\in V}$ of Markov kernels that are adapted to $\mathcal{D}$ in the sense that each P^v is a Markov kernel from $\mathcal{X}_{\mathrm{pa}(v)}$ to $\mathcal{X}_v$ and $P = \circledast_{v\in V} P^v$. We first note

Proposition 2.48 *If P admits a recursive kernel factorization according to the directed, acyclic graph $\mathcal{D}$ and A is an ancestral set, then the marginal distribution P_A admits a recursive factorization according to $\mathcal{G}_A$.*

Proof For any ancestral set A there is a topological ordering of V so that $\alpha < \beta$ for all $\alpha \in A$ and $\beta \in V \setminus A$. Hence we may write

$$P = \circledast_{v\in V} P^v = (\circledast_{v\in A} P^v) \circledast \left(\circledast_{v\in V\setminus A} P^v\right) = P_A \circledast Q$$

so the marginal distribution simply factorizes as $P_A = \circledast_{v\in A} P^v$. □

We then have the following theorem:

Theorem 2.49 *Let $\mathcal{D}$ be a directed, acyclic graph and P a probability distribution on $\mathcal{X}$. Then the following conditions are equivalent*

(G) *P obeys the directed global Markov property, relative to $\mathcal{D}$;*
(L) *P obeys the directed local Markov property, relative to $\mathcal{D}$;*
(O) *P obeys the ordered local Markov property relative to $(\mathcal{D}, <)$ where $<$ is a topological ordering of $\mathcal{D}$;*
(R) *P admits a recursive kernel factorization according to $\mathcal{D}$.*

Proof That (G), (L), and (O) are equivalent follows from Theorem 2.44 since $\perp\!\!\!\perp_P$ is a semigraphoid.
That (R) is equivalent to (O) follows by induction on the number of vertices and Corollary 2.16: If (O) holds and v^* is the largest vertex in a topological ordering we have $P = P_{V\setminus\{v^*\}} \circledast Q$ where Q represents the conditional distribution of X_{v^*} given $X_{V\setminus\{v^*\}}$ the latter only depending on $x_{\mathrm{pa}(v^*)}$ by (O) and can we thus let $Q = P^{v^*}$. Hence induction yields the recursive factorization (R).
On the other hand, if (R) holds, Corollary 2.16 implies $v^* \perp\!\!\!\perp_P V \setminus \mathrm{pa}(v^*)$ and the remaining conditional independence relations follow by induction. □

Example 2.50 [Structural equation models] A specific way to construct a Bayesian network is through a system of *structural equations*. More precisely define

$$X_v \leftarrow \phi_v(X_{\mathrm{pa}(v)}, U_v), \quad v \in V \tag{2.30}$$

where the assigments are made according to a topological ordering of the vertices of a DAG $\mathcal{D}$ and U_v are independent and uniformly distributed on $(0,1)$. Then, by Corollary 1.25, this defines a system of Markov kernels that is adapted to $\mathcal{D}$ and the recursive combination of these Markov kernels represents the distribution of $X_v, v \in V$. □

In fact, by Theorem 1.28, any adapted system of Markov kernels and thus any Bayesian network can be represented by a system of structural equations as above.

Generally, a Bayesian network defined by stuctural equations may not have a density w.r.t. any product measure. But if it has, the results above can be strengthened. We say that a probability distribution P admits a *recursive density factorization* according to $\mathcal{D}$, if there exist non-negative functions, henceforth referred to as *kernels,* $k^\alpha(\cdot,\cdot), \alpha \in V$ defined on $\mathcal{X}_\alpha \times \mathcal{X}_{\mathrm{pa}(\alpha)}$, such that

$$\int k^\alpha(y_\alpha, x_{\mathrm{pa}(\alpha)})\mu_\alpha(dy_\alpha) = 1$$

and P has density f with respect to $\mu = \otimes_{\alpha\in V}\mu_\alpha$, where

$$f(x) = \prod_{\alpha\in V} k^\alpha(x_\alpha, x_{\mathrm{pa}(\alpha)}).$$

We then also say that P has property (F). It is an easy induction argument to show that

Proposition 2.51 *P admits a recursive density factorization if and only if the kernels $k^\alpha(\cdot, x_{\mathrm{pa}(\alpha)})$ are densities for the conditional distribution of X_α, given $X_{\mathrm{pa}(\alpha)} = x_{\mathrm{pa}(\alpha)}$.*

Proof Left to the reader as Exercise 2.8. □

Also it is immediate that if we form the undirected moral graph $\mathcal{D}^m$ (marrying parents and deleting directions) such as described towards the end of Section B.1, we have

Lemma 2.52 *If P admits a recursive density factorization according to the directed, acyclic graph $\mathcal{D}$, it factorizes according to the moral graph $\mathcal{D}^m$ and therefore obeys the global Markov property relative to $\mathcal{D}^m$.*

Proof The factorization follows from the fact that, by construction, the sets $\{\alpha\} \cup \mathrm{pa}(\alpha)$ are complete in $\mathcal{D}^m$ and we can therefore let $\psi_{\{\alpha\}\cup\mathrm{pa}(\alpha)} = k^\alpha$. The remaining part of the statement follows from the fact that (F) implies (G) in the undirected case; see Proposition 2.40. □

Generally, if P admits a recursive density factorization it also admits a recursive kernel factorization.
We may now extend Theorem 2.49 to

Theorem 2.53 *Let $\mathcal{D}$ be a directed, acyclic graph. For a probability distribution P on $\mathcal{X}$ which has density with respect to a product measure μ, the following conditions are equivalent:*

(F) *P admits a recursive density factorization according to $\mathcal{D}$;*
(R) *P admits a recursive kernel factorization according to $\mathcal{D}$;*
(G) *P obeys the directed global Markov property, relative to $\mathcal{D}$;*
(L) *P obeys the directed local Markov property, relative to $\mathcal{D}$;*
(O) *P obeys the ordered Markov property, relative to $\mathcal{D}$.*

Proof This follows as (R) and (F) are equivalent when the joint distribution have a density w.r.t. a product measure, cf. Theorems 1.18 and 1.19. □

2.6.3 *Markov equivalence*

Consider two graphs $\mathcal{G}_1$ and $\mathcal{G}_2$ as well as their associated independence models $\perp_{\mathcal{G}_1}$ and $\perp_{\mathcal{G}_2}$. It may well happen that even though the graphs are different, their independence models might be identical, see for example Figure 2.6 below. Here all

FIG. 2.6. Four Markov equivalent graphs. Their associated independence models are in all cases $\alpha \perp_{\mathcal{G}} \gamma \,|\, \beta$.

independence models are the same although the graphs are different. This also means that any probability distribution P which satisfies the global Markov property for any of them, automatically satisfies the global Markov property for all of them. We formally define

Definition 2.54 Two graphs $\mathcal{G}_1$ and $\mathcal{G}_2$ are *Markov equivalent* if and only if their independence models coincide, i.e. if $A \perp_{\mathcal{G}_1} B \,|\, S \iff A \perp_{\mathcal{G}_2} B \,|\, S$.

Generally, independence models based on directed acyclic graphs are different from those based on corresponding undirected graphs, but occasionally, as in Figure 2.6, they are identical. More precisely we first define:

Definition 2.55 A directed acyclic graph $\mathcal{D}$ is said to be *perfect* if all parents of common children are married, i.e. if for any triple α, β, γ with $\alpha \to \gamma \leftarrow \beta$, we have $\alpha \sim \beta$.

Perfect DAGs are Markov equivalent to their skeleton. More precisely:

Proposition 2.56 *A directed acyclic graph $\mathcal{D}$ is Markov equivalent to its skeleton $\mathcal{G} = \mathrm{ske}(\mathcal{D})$ if and only if $\mathcal{D}$ is perfect.*

Proof The proof is by induction on the number of vertices $|V|$ of $\mathcal{D}$. For $|V| \leq 3$ the only non-perfect DAG is a triple α, β, γ with $\alpha \to \gamma \leftarrow \beta$ and $\alpha \not\sim \beta$, where then $\alpha \perp_{\mathcal{D}} \beta$ is the only independence whereas, for the skeleton, the only independence is $\alpha \perp_{\mathcal{G}} \beta \,|\, \gamma$.
For $|V| > 3$, consider a triplet A, B, S and let v^* be a terminal vertex outside $A \cup B \cup S$. If this does not exist, we have $\mathrm{An}(A \cup B \cup S) = V$ and we have $A \perp_m B \,|\, S \iff A \perp_{\mathcal{G}} B \,|\, S$ where $\mathcal{G} = \mathrm{ske}(\mathcal{D})$ is the skeleton of $\mathcal{D}$. Else remove v^* and use the induction hypothesis.
For the converse, the inductive assumption implies that we only have to consider the case where a terminal vertex v^* has unmarried parents α and β. These are then separated in the skeleton by $V \setminus \{\alpha, \beta\}$ but not d-separated as the walk $\alpha \to v^* \leftarrow \beta$ is active relative to $V \setminus \{\alpha, \beta\}$. □

In general, the question of Markov equivalence of two directed acyclic graphs is slightly more subtle. Before we explore this further, we need a few lemmas.

Lemma 2.57 *Let $\mathcal{D} = (V, E)$ be a directed acyclic graph. Then for $\alpha \neq \beta$ it holds that $\alpha \not\sim \beta$ if and only if there exists $S \subseteq V \setminus \{\alpha, \beta\}$ such that $\alpha \perp_{\mathcal{D}} \beta \,|\, S$.*

Proof If $\alpha \sim \beta$, the walk (α, β) connects α and β given any S and hence we do not have $\alpha \perp_{\mathcal{D}} \beta \,|\, S$.
Conversely, if $\alpha \not\sim \beta$ we let $A = \mathrm{An}(\alpha, \beta)$ be the smallest ancestral set containing $\{\alpha, \beta\}$ and subsequently $S = A \setminus \{\alpha, \beta\}$. It is also true that $\alpha \not\sim \beta$ in the moral graph $(\mathcal{D}_A)^m$ as α and β have no common children in A: since all elements of A are ancestors of α or β a directed path from a common child to either of them would create a cycle. Hence $\alpha \perp_m \beta \,|\, S$ and thus, by Proposition 2.33 $\alpha \perp_{\mathcal{D}} \beta \,|\, S$. □

Note that the same result is trivially true for an undirected graph $\mathcal{G}$, where we can use $S = V \setminus \{\alpha, \beta\}$. Indeed it is true for any general graphical independence model $\perp_{\mathcal{G}}$ as defined earlier, but we refrain from showing the lemma in this generality here.
Further we define

Definition 2.58 An arrow $\alpha \to \beta$ is *covered* in a DAG $\mathcal{D}$ if $\mathrm{pa}(\beta) = \mathrm{pa}(\alpha) \cup \alpha$.

Note that if $\mathcal{D}^*$ is obtained from $\mathcal{D}$ by replacing the arrow $\alpha \to \beta$ with $\beta \to \alpha$ then $\beta \to \alpha$ is covered in $\mathcal{D}^*$ if and only if $\alpha \to \beta$ is covered in $\mathcal{D}$. Covered edges can be reversed without changing the independence model, as the next lemma says.

Lemma 2.59 *Let $\mathcal{D}$ be a DAG and let $\mathcal{D}^*$ be obtained from $\mathcal{D}$ by replacing the edge $\alpha \to \beta$ with $\beta \to \alpha$. Then $\mathcal{D}^*$ is a DAG which is Markov equivalent to $\mathcal{D}$ if and only if $\alpha \to \beta$ is covered.*

Proof Assume first that the arrow $\alpha \to \beta$ is covered. To show that $\mathcal{D}^*$ is a DAG we assume for contradiction that there is a directed cycle in $\mathcal{D}^*$ which then must include the arrow $\beta \to \alpha$ since $\mathcal{D}$ was a DAG. Let this cycle be

$$\omega^* = (\alpha_1 \to \cdots \to \gamma \to \beta \to \alpha \to \cdots \to \alpha_1).$$

But if $\alpha \to \beta$ is covered and $\gamma \to \beta$ then γ is also a parent of α in $\mathcal{D}$ implying that

$$\omega = (\alpha_1 \to \cdots \to \gamma \to \alpha \to \cdots \to \alpha_1)$$

is a cycle in $\mathcal{D}$, contradicting that $\mathcal{D}$ is acyclic.

To show that $\mathcal{D}^*$ and $\mathcal{D}$ are Markov equivalent we have to show that if there is a walk in $\mathcal{D}$ from u to v which is connecting relative to S there is also a connecting walk from u to v in $\mathcal{D}^*$. We may without loss of generality assume that this walk includes the edge $\alpha \to \beta$ since otherwise the statement is obvious; thus the walk has the form

$$\omega = (u \sim \cdots \sim \gamma_1 \sim \alpha \to \beta \sim \gamma_2 \cdots \sim v).$$

If $\beta \in S$, β must be a collider in $\mathcal{D}$ on this walk and we must have $\gamma_2 \in \mathrm{pa}(\beta)$, and $\alpha, \gamma_2 \notin S$ since otherwise the walk would not be connecting. Since $\alpha \to \beta$ is covered we have $\gamma_2 = \alpha$ or $\gamma_2 \in \mathrm{pa}(\alpha)$; if $\alpha = \gamma_2$ the walk

$$\omega' = (u \sim \cdots \sim \gamma_1 \sim \alpha = \gamma_2 \sim \cdots \sim v)$$

would still be connecting and if $\gamma_1 = \gamma_2$ the walk

$$\omega'' = (u \sim \cdots \sim \gamma_1 = \gamma_2 \sim \cdots \sim v)$$

connects. If $\gamma_1 \neq \gamma_2$ and $\gamma_1 \in \mathrm{pa}(\alpha)$ we also have $\gamma_1 \in \mathrm{pa}(\beta)$ and the walk

$$\omega''' = (u \sim \cdots \sim \gamma_1 \to \beta \leftarrow \gamma_2 \cdots \sim v)$$

connects. Every time the walk includes $\alpha \to \beta$ we modify the walk as above and eventually obtain a walk $\tilde{\omega}$ which connects in both $\mathcal{D}$ and $\mathcal{D}^*$.
If $\beta \notin S$ it is never a collider on ω. If now $\tilde{\gamma}_1 \in \mathrm{pa}(\beta)$ we modify ω as

$$\tilde{\omega}' = (u \sim \cdots \sim \tilde{\gamma}_1 \to \beta \to \tilde{\gamma}_2 \cdots \sim v)$$

and this walk will still be connecting. Working through all the appearances of $\alpha \to \beta$ eventually yields a walk that is connecting in both $\mathcal{D}$ and $\mathcal{D}^*$.

For the converse we consider an edge $\alpha \to \beta$ which is not covered. Then there is a $\gamma \in \mathrm{pa}(\beta) \setminus \mathrm{pa}(\alpha)$ with $\gamma \neq \alpha$ or a $\gamma \in \mathrm{pa}(\alpha) \setminus \mathrm{pa}(\alpha)$. Consider the former case: If $\alpha \to \gamma$, reversing $\alpha \to \beta$ creates a cycle $(\alpha \to \gamma \to \beta \to \alpha)$ and thus $\mathcal{D}^*$ is not acyclic. Hence we may assume that $\alpha \not\sim \gamma$ and thus by Lemma 2.57 there is an S such that $\alpha \perp_{\mathcal{D}} \gamma \,|\, S$. Then $\beta \notin S$ for else would $(\alpha \to \beta \leftarrow \gamma)$ connect. But then $(\gamma \to \beta \to \alpha)$ connects in $\mathcal{D}^*$ and thus $\mathcal{D}$ and $\mathcal{D}^*$ are not Markov equivalent.
If $\gamma \in \mathrm{pa}(\alpha) \setminus \mathrm{pa}(\beta)$ we argue similarly, just reversing the role of $\mathcal{D}$ and $\mathcal{D}^*$. This completes the proof. □

We define an *unshielded collider tripath* to be an induced subgraph of the form $\alpha \to \gamma \leftarrow \beta$, i.e. arrows meet head-to-head at γ but $\alpha \not\sim \beta$. Such structures are preserved under the operation of reversing a covered edge:

Lemma 2.60 *Let $\mathcal{D}$ be a DAG and let $\mathcal{D}^*$ be obtained from $\mathcal{D}$ by replacing a covered edge $\alpha \to \beta$ with $\beta \to \alpha$. Then $\mathcal{D}$ and $\mathcal{D}^*$ have the same skeleton and the same unshielded collider tripaths.*

Proof The skeleton is clearly unchanged by any edge reversal. Also, none of the edges involved in an unshielded collider tripath $\alpha \to \gamma \leftarrow \beta$ are covered as then α is a parent of γ but not a parent of β. □

Further we have

Lemma 2.61 *If two directed acyclic graphs $\mathcal{D}_1 = (V, E_1)$ and $\mathcal{D}_2 = (V, E_2)$ with $\mathcal{D}_1 \neq \mathcal{D}_2$ have the same skeleton $\mathrm{ske}(\mathcal{D}_1) = \mathrm{ske}(\mathcal{D}_2)$ and the same unshielded collider tripaths, then there exists a covered edge $\alpha \to \beta$ in $E_1 \setminus E_2$.*

Proof The proof is constructive and gives a simple algorithm for identifying such an edge. First we consider a topological ordering of $\mathcal{D}_1$ and let β be the minimal element of this ordering satisfying $\mathrm{pa}_1(\beta) \setminus \mathrm{pa}_2(\beta) \neq \emptyset$. Further, let α be the maximal element of $\mathrm{pa}_1(\beta) \setminus \mathrm{pa}_2(\beta)$. We shall argue by contradiction that $\alpha \to \beta$ is indeed covered.

So assume for contradiction that there exists a vertex $\gamma \in \mathrm{pa}_1(\beta) \setminus \mathrm{pa}_1(\alpha)$ with $\gamma \neq \alpha$. Then we must have $\alpha \to \gamma$ in E_1 for else we would have an unshielded collider tripath contradicting that $\alpha \to \beta \in E_1 \setminus E_2$. But then either $\alpha \to \gamma$ is in $E_1 \setminus E_2$ or $\gamma \to \beta$ must be in $E_1 \setminus E_2$ for else $\mathcal{D}_2$ would contain a directed cycle. But the former contradicts the minimality of β and the latter the maximality of α.

If there were a vertex $\gamma \in \mathrm{pa}_1(\alpha) \setminus \mathrm{pa}_1(\beta)$ we would have $\gamma \in \mathrm{pa}_2(\alpha)$ since β was chosen to be minimal. Since $\alpha \to \beta$ cannot be part of an unshielded collider tripath, there must be an edge $\beta \to \gamma$ in E_1 which would create a cycle.
Hence we conclude that $\alpha \to \beta$ is covered. □

We are then ready to show the main result about Markov equivalence of directed acyclic graphs.

Theorem 2.62 *Two directed acyclic graphs $\mathcal{D}_1 = (V, E_1)$ and $\mathcal{D}_2 = (V, E_2)$ are Markov equivalent if and only if they have the same skeleton $\mathrm{ske}(\mathcal{D}_1) = \mathrm{ske}(\mathcal{D}_2)$ and the same unshielded collider tripaths.*

Proof From Lemma 2.57 it follows that two Markov equivalent DAGs must have the same skeleton. If we have two DAGs $\mathcal{D}_1$ and $\mathcal{D}_2$ with the same skeleton and the same unshielded colliders, Lemma 2.61 yields that we can find a covered edge in $E_1 \setminus E_2$. Reversing this edge yields a new DAG $\mathcal{D}_1'$ with $E_1' \setminus E_2$ having one element less. Lemma 2.60 ensures that also $\mathcal{D}_1'$ and $\mathcal{D}_1$ have the same skeleton and same unshielded collider tripaths and Lemma 2.59 ensures $\mathcal{D}_1'$ is Markov equivalent to $\mathcal{D}_1$. Continuing in this matter eventually transforms $\mathcal{D}_1$ to $\mathcal{D}_2$ and every step preserves Markov equivalence, hence $\mathcal{D}_1$ is Markov equivalent to $\mathcal{D}_2$.

For the converse we simply realise that if $\alpha \to \gamma \leftarrow \beta$ is an unshielded collider tripath in $\mathcal{D}_1$ which is not unshielded in $\mathcal{D}_2$, any set S which satisfies $\alpha \perp_{\mathcal{D}_1} \beta \,|\, S$ cannot include γ where as it must include γ to ensure $\alpha \perp_{\mathcal{D}_2} \beta \,|\, S$. □

Note that it follows from the proof that *if $\mathcal{D}_1$ and $\mathcal{D}_2$ are Markov equivalent, there is a sequence of covered arrow reversals transforming $\mathcal{D}_1$ into $\mathcal{D}_2$.*

2.7 Exercises

Exercise 2.1 Let Y and Z be real valued random variables such that $EY^2 < \infty$ and $EZ^2 < \infty$. Let X be a random variable with values in the measurable space $(\mathcal{X}, \mathbb{E})$. Define the *conditional covariance* between Y and Z given X by

$$\mathrm{Cov}(Y, Z \mid X) = \mathbf{E}(YZ \mid X) - \mathbf{E}(Y \mid X)\mathbf{E}(Z \mid X)$$

(a) Show that

$$\mathrm{Cov}(Y, Z) = \mathbf{E}(\mathrm{Cov}(Y, Z \mid X)) + \mathrm{Cov}(\mathbf{E}(Y \mid X), \mathbf{E}(Z \mid X))$$

(b) Assume that $Y \perp\!\!\!\perp Z \mid X$. Show that $\mathrm{Cov}(Y, Z \mid X) = 0$ a.s.

Now assume that X is a real valued random variable with $\mathbf{E}X^2 < \infty$, and assume that Y_1 and Y_2 are two other random variables with the same conditional distribution $(P_x)_{x\in\mathbb{R}}$ given X, where

$$P_x = \mathcal{N}(x, 1)$$

Assume that $Y_1 \perp\!\!\!\perp Y_2 \mid X$.

(c) Show that $\mathbf{E}Y_1^2 = \mathbf{E}Y_2^2 < \infty$.

(d) Show that $\mathrm{Cov}(Y_1, Y_2) = \mathbf{V}(X)$.

Exercise 2.2 Assume that X_1 and X_2 are real valued random variables. Let $(P_x)_{x\in\mathbb{R}}$ be the conditional distribution of X_1 given $X_1 + X_2$.

Define for each $x \in \mathbb{R}$ and $B \in \mathbb{B}$ the set

$$x - B = \{x - y \,:\, y \in B\}$$

and define the collection of measures $(Q_x)_{x\in\mathbb{R}}$ by

$$Q_x(B) = P_x(x - B)$$

(a) Show that $(Q_x)_{x\in\mathbb{R}}$ is the conditional distribution of X_2 given $X_1 + X_2$.

Define for each $x \in \mathbb{R}$ the measure R_x on $(\mathbb{R}^2, \mathbb{B}^2)$ by

$$R_x(A \times B) = P_x(A \cap (x - B))$$

for $A, B \in \mathbb{B}$.

(b) Show that $(R_x)_{x\in\mathbb{R}}$ is the conditional distribution of (X_1, X_2) given $X_1 + X_2$.

(c) Assume that $X_1 \perp\!\!\!\perp X_2 \mid X_1 + X_2$. Show that for all $x \in \mathbb{R}$ it holds that $P_x(A) \in \{0, 1\}$ for all $A \in \mathbb{B}$. Conclude that $P_x = \delta_{\phi(x)}$, where $\phi(x)$ is some real number depending on x.

(d) Show that the function ϕ from (3) is measurable.

(e) Show that if $X_1 \perp\!\!\!\perp X_2 \mid X_1 + X_2$, then there exists measurable functions ϕ_1 and ϕ_2, such that

$$X_1 = \phi_1(X_1 + X_2) \quad \text{a.s.,} \quad \text{and} \quad X_2 = \phi_2(X_1 + X_2) \quad \text{a.s.}$$

.

(f) Give an example of real random variables X_1, X_2 and X_3, where

$$X_1 \perp\!\!\!\perp X_2 \,|\, X_1 + X_2 + X_3 \,.$$

.

Exercise 2.3 Assume that X is a real random variable and that $(P_x)_{x\in\mathbb{R}}$ is the conditional distribution of Y given X, where P_x is the exponential distribution with mean $|x|$.

(a) Find a measurable function

$$\phi : \mathbb{R} \times (0,1) \to \mathbb{R}$$

such that if X' has the same distribution as X and if U is uniform on $(0,1)$ and independent of X', then (X', Y') has the same distribution as (X, Y), where $Y' = \phi(X', U)$.

(b) Assume that $X \sim \mathcal{N}(0,1)$. Simulate 10000 independent replications of (X, Y), and plot the points $(X_n, Y_n)_{n=1,\ldots,10000}$.

(c) Find $E(Y \,|\, X = x)$ and $V(Y \,|\, X = x)$ theoretically and use this to explain the plot.

Exercise 2.4 Assume that U_1, U_2 and U_3 are independent and identically distributed according to the uniform distribution on $(0,1)$. Define the real random variables X, Y and Z as follows:

$$\begin{aligned} X &= U_1^2 \\ Y &= -X\log(1-U_2) \\ Z &= -2X\log(1-U_3) \end{aligned}$$

(a) Find the conditional distribution of Y given X and the conditional distribution of Z given X.

(b) Show that $U_2 \perp\!\!\!\perp U_3 \,|\, U_1$

(c) Show that $(U_1, U_2) \perp\!\!\!\perp (U_1, U_3) \,|\, U_1$

(d) Show that $Y \perp\!\!\!\perp Z \,|\, U_1$

(e) Show that $Y \perp\!\!\!\perp Z \,|\, X$

(f) Find the conditional distribution of $(Y, Z) \,|\, X$.

(g) Let U_4 be another uniformly distributed random variable independent of (U_1, U_2, U_3), and define $Z_2 = -2U_4^2\log(1-U_3)$. Argue that $Y \perp\!\!\!\perp Z_2$. Simulate 10000 replications of (Y, Z) and (Y, Z_2), plot these replications in two plots and explain the difference.

Exercise 2.5 Prove Proposition 2.21.

Exercise 2.6 Show that if the distribution of $X \,|\, Z$ is degenerate so that X in effect is a deterministic function of Z, then $X \perp\!\!\!\perp Y \,|\, Z$ for all possible random variables Y.

Exercise 2.7 Let (X, Y, Z) be random variables with a discrete and finite state space.

(a) Show that if (X, Y, Z) are all binary, it holds that

$$X \perp\!\!\!\perp Y \text{ and } X \perp\!\!\!\perp Y \,|\, Z \implies (X, Z) \perp\!\!\!\perp Y \text{ or } X \perp\!\!\!\perp (Y, Z);$$

(b) Find a counterexample to the analogous result when state spaces are discrete but may have more than two states.

Exercise 2.8 Prove Proposition 2.51.

Exercise 2.9 Let $X = (X_1, X_2, X_3)^\top \sim \mathcal{N}_3(0, \Sigma)$ be a multivariate Gaussian distribution.

(a) Show that $X_1 \perp\!\!\!\perp X_3 \,|\, X_2$ if and only if $\sigma_{13}\sigma_{22} = \sigma_{12}\sigma_{23}$;

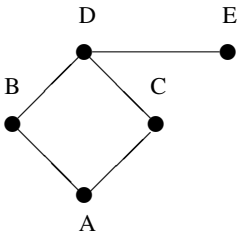
(b) Use this to show that for multivariate Gaussian variables it holds that

$$X_1 \perp\!\!\!\perp X_3 \text{ and } X_1 \perp\!\!\!\perp X_3 \,|\, X_2 \implies (X_1, X_2) \perp\!\!\!\perp X_3 \text{ or } X_1 \perp\!\!\!\perp (X_2, X_3).$$

Exercise 2.10 Show that graph separation $\perp_{\mathcal{B}}$ in a bidirected graph $\mathcal{G}$ is a compositional graphoid without using Proposition 2.31.

Exercise 2.11 Show that graph separation $\perp_{\mathcal{G}}$ in an undirected graph $\mathcal{G}$ is a compositional graphoid without using Proposition 2.31.

Exercise 2.12 Consider the graph below:

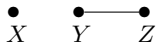


(a) Write down all conditional independence statements for this graph corresponding to the pairwise Markov property;

(b) Write down all conditional independence statements for this graph corresponding to the local Markov property;

(c) Write down some of the conditional independence statements for this graph which follow from the global Markov property and which are not listed above.

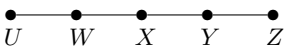
Exercise 2.13 Let $X = Y = Z$ with $P\{X = 1\} = P\{X = 0\} = 1/2$. Show that this distribution satisfies (P) but not (L) with respect to the graph below.



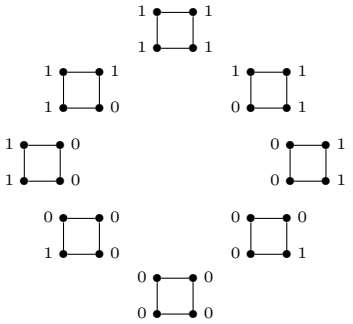
Exercise 2.14 Let U and Z be independent with

$$P(U = 1) = P(Z = 1) = P(U = 0) = P(Z = 0) = 1/2,$$

$W = U$, $Y = Z$, and $X = WY$. Show that this distribution satisfies (L) but not (G) w.r.t. the graph below.



Exercise 2.15 Consider the distribution over four binary variables which gives probability 1/8 to all of the 8 configurations displayed in the figure below:



Note that there are only four variables. The only reason that the graph (four-cycle) is repeated is to see the obvious pattern in the configuration.
Show that this distribution satisfies (G) with respect to the four cycle, but the distribution does not factorize with respect to this graph, i.e., it does not satisfy (F).
Hint: Assume that it factorizes and show that if it is positive on these configurations, it must be positive on all 16 possible configurations of the four binary variables.

Exercise 2.16 Let the joint distribution of (X, Y) have density $f(x, y)$ w.r.t. a product measure $\nu \otimes \mu$. The *conditional entropy* $H(X \mid Y)$ is defined as the average entropy in the conditional distribution

$$\begin{aligned} H(X \mid Y) &= \mathbf{E}\Big[\mathbf{E}\left\{-\log f(X \mid Y) \mid Y\right\}\Big] \\ &= \int_y \left\{ \int_x -f(x \mid y) \log f(x \mid y) \nu(dx) \right\} f(y)\, \mu(dy). \end{aligned}$$

(a) Use Jensen's inequality to show that

$$H(X \mid Y) \leq H(X),$$

i.e. the *entropy is always reduced by conditoning*.

(b) Show that

$$H(X, Y) = H(X \mid Y) + H(Y).$$

(c) For three random variables, show that

$$H(X, Y, Z) + H(Z) \leq H(X, Z) + H(Y, Z).$$

(d) Show further that

$$X \perp\!\!\!\perp Y \mid Z \iff H(X, Y, Z) + H(Z) = H(X, Z) + H(Y, Z).$$

Exercise 2.17 Let $\perp_\sigma$ be an independence model on a finite set V and let $M \subseteq V$. The *marginal independence model* $\perp_{\sigma M}$ is simply the restriction of $\perp_\sigma$ to triples (A, B, S) with

$(A \cup B \cup S) \subseteq M$. The *conditional independence model* $\perp_\sigma^M$ is defined for triples (A, B, S) with $(A \cup B \cup S) \subseteq V \setminus M$ as

$$A \perp_\sigma^M B \,|\, S \iff A \perp_\sigma B \,|\, (S \cup M).$$

(a) Show that $\perp_{\sigma M}$ and $\perp_\sigma^M$ inherit the properties of $\perp_\sigma$, such that, e.g., if $\perp_\sigma$ is a compositional graphoid, so are its marginal and conditional;

(b) Show that for a probability distribution P with associated independence model $\perp\!\!\!\perp_P$, the independence model of the marginal distribution P_M is indeed the marginal of $\perp\!\!\!\perp_P$;

(c) Show that for a probability distribution P with associated independence model $\perp\!\!\!\perp_P$, the independence model of the conditional distributions (P_{x_M}) of $X_{V\setminus M}$ given X_M is indeed the conditional model of $\perp\!\!\!\perp_P$;

(d) Show that if $\perp\!\!\!\perp_P$ is globally Markov w.r.t. an undirected graph $\mathcal{G}$, then the conditional distributions (P_{x_M}) of $X_{V\setminus M}$ given X_M are globally Markov w.r.t. the induced subgraph $\mathcal{G}_{V\setminus M}$;

(e) Show that in general the M-marginal of $\perp\!\!\!\perp_P$ is not globally Markov w.r.t. $\mathcal{G}_M$.

Exercise 2.18 Consider a DAG $\mathcal{D}$ with arrows
$1 \to 2, 2 \to 5, 2 \to 3, 5 \to 6, 4 \to 5, 4 \to 7, 5 \to 7$.

(a) Draw the DAG;

(b) List all conditional independence relations corresponding to the local, directed Markov property;

(c) Find the ancestral sets generated by the following subsets:

  (a) $\{5\}$;
  (b) $\{2, 7\}$;
  (c) $\{4, 6\}$;

(d) Which of the following separation statements are true? For those that are not true, identify an active walk.

  (a) $2 \perp_\mathcal{D} 4 \,|\, 5$;
  (b) $2 \perp_\mathcal{D} 7 \,|\, 5$,
  (c) $1 \perp_\mathcal{D} 7 \,|\, 5, 6$;
  (d) $1 \perp_\mathcal{D} 4 \,|\, 6$;

Exercise 2.19 Consider the following directed acyclic graphs, and in each case, list all DAGs in their Markov equivalence class and verify in every single case whether they are Markov equivalent to an undirected graph.

(a) $1 \to 2, 3 \to 2, 2 \to 4, 4 \to 5, 2 \to 5$;

(b) $1 \to 2, 2 \to 3, 2 \to 4, 4 \to 5, 6 \to 5$;

(c) $1 \to 2, 2 \to 3, 2 \to 4, 4 \to 5, 5 \to 6$;

Exercise 2.20 Consider a directed acyclic graph $\mathcal{D}$ with arrows
$A \to B, B \to D, B \to E, C \to E, D \to E, D \to F, E \to F$.

(a) Form the moral graph $\mathcal{D}^m$ of $\mathcal{D}$.

(b) Assume P satisfies the local directed Markov property with respect to $\mathcal{D}$. Which of the following statements can be concluded? Explain your reasoning.

$$C \perp\!\!\!\perp D \,|\, B, \quad A \perp\!\!\!\perp C \,|\, E, \quad B \perp\!\!\!\perp F \,|\, \{E, A\}.$$

(c) Consider the following directed acyclic graphs obtained from $\mathcal{D}$ by reversing arrows:

(i) $\mathcal{D}_1$ has reversed the arrow from $A \to B$, i.e. it has arrows $B \to A, B \to D, B \to E, C \to E, D \to E, D \to F, E \to F$;
(ii) $\mathcal{D}_2$ has reversed the arrow from $D \to E$, i.e. it has arrows $A \to B, B \to D, B \to E, C \to E, E \to D, D \to F, E \to F$.
(iii) $\mathcal{D}_3$ has reversed the arrow from $C \to E$, i.e. it has arrows $A \to B, B \to D, B \to E, E \to C, D \to E, D \to F, E \to F$.

Which of these directed acyclic graphs are Markov equivalent to $\mathcal{D}$?

(d) Which of the directed acyclic graphs above are Markov equivalent to an undirected graph?

3

LOCAL COMPUTATION

3.1 Local computation

Local computation algorithms have been developed with a variety of purposes. For example:

- Kalman filter and smoother (Thiele, 1880; Kalman and Bucy, 1961);
- Solving sparse linear equations; (Parter, 1961);
- Decoding digital signals; (Viterbi, 1967; Bahl, Cocke, Jelinek and Raviv, 1974);
- Estimation in hidden Markov models; (Baum, 1972);
- Peeling in pedigrees; (Elston and Stewart, 1971; Cannings, Thompson and Skolnick, 1976);
- Belief function evaluation; (Kong, 1986; Shenoy and Shafer, 1986);
- Probability propagation. (Pearl, 1986; Lauritzen and Spiegelhalter, 1988; Jensen, Lauritzen and Olesen, 1990);

Also dynamic programming, linear programming, optimizing decisions, calculating Nash equilibria in cooperative games, and many others. This list is far from exhaustive. An abstract framework has been discussed by Shenoy and Shafer (1990) and Lauritzen and Jensen (1997).

All algorithms are using, explicitly or implicitly, a graph decomposition and a junction tree or similar structure. The basic idea is to arrange computations to be performed locally, i.e. in cliques of a decomposable graph and thus effectively in a smaller state space than that associated with all the variables in V.

3.2 Probability propagation

3.2.1 *Basic problem*

We consider a factorizing density on $\mathcal{X} = \times_{v \in V} \mathcal{X}_v$ with V and $\mathcal{X}_v$ finite:

$$p(x) = \prod_{C \in \mathcal{C}} \phi_C(x).$$

The *potentials* $\phi_C(x)$ depend on $x_C = (x_v, v \in C)$ only. The basic task is to calculate the *marginal* probability

$$p(x_E^*) = \sum_{y_{V \setminus E}} p(x_E^*, y_{V \setminus E})$$

for $E \subseteq V$ and fixed x_E^*, but this sum has too many terms if V is large as then $\mathcal{X}$ is huge and has cardinality at least $2^{|V|}$. A second purpose of the computation is to get the *prediction* $p(x_v \mid x_E^*) = p(x_v, x_E^*)/p(x_E^*)$ for $v \in V$. We first consider a simple example:

Example 3.1 Assume that the density factorizes as

$$p(x, y, z, w) = \phi(x, y)\psi(y, z)\eta(z, w)$$

and assume each of X, Y, Z, and W has, say, 100 states. The joint state space has thus 10^8 states, and to calculate $p(x)$ directly from $p(x, y, z, w)$ by brute force involves 10^6 terms in the sum for every x, hence 10^8 arithmetic operations are needed. This is possible, but time consuming, and in networks with many variables, direct calculation becomes impossible.
Instead, we may use the factorization $p(x, y, z, w) = \phi(x, y)\psi(y, z)\eta(z, w)$ as follows:

1. Calculate $\eta^*(z) = \sum_w \eta(z, w)$, with 10000 additions;
2. Calculate $\psi^*(y, z) = \psi(y, z)\eta^*(z)$ with 10000 multiplications
3. Calculate $\psi^*(y) = \sum_z \psi^*(y, z)$, with 10000 additions;
4. Calculate $\phi^*(x, y) = \phi(x, y)\psi^*(y)$ with 10000 multiplications;
5. Calculate $\phi^*(x) = \sum_y \phi^*(x, y)$, with 10000 additions.

Now the marginal $p^*(x)$ is equal to $\phi^*(x)$ so we calculated our quantity of interest with only 50000 operations. Note in particular that we never explicitly formed the product $p(x, y, z, w) = \phi(x, y)\psi(y, z)\eta(z, w)$. The product only appears conceptually, guiding the specific computations. □

3.2.2 *Setting up the structure*

The typical application of a local computation algorithm involves first specifying a Bayesian network. Starting from a DAG, the local computational structure is set up in several steps:

1. *Moralisation:* Constructing $\mathcal{D}^m$ and exploiting that if P factorizes over $\mathcal{D}$, it factorizes over $\mathcal{D}^m$ (Lemma 2.52);
2. *Triangulation:* Adding edges to find a *chordal cover* $\tilde{\mathcal{G}}$ of $\mathcal{G}$, i.e. a chordal graph $\tilde{\mathcal{G}}$ with $\mathcal{G} \subseteq \tilde{\mathcal{G}}$. This step is non-trivial (NP-complete) to optimize;
3. *Constructing a junction tree:* Using maximum cardinality search, the cliques of $\tilde{\mathcal{G}}$ are found and arranged in a junction tree.
4. *Initialization:* Assigning appropriate potential functions ϕ_C to cliques and separators; see below.

The complete process above is often referred to as *compilation*. Computation is then performed by *message passing* after observations have been incorporated, to be explained in the following.

3.2.3 *The basic invariant*

3.2.3.1 *Initialization* From a Bayesian network over a directed acyclic graph $\mathcal{D}$ with conditional densities $k_v(x_v \mid x_{\mathrm{pa}(v)})$ we first assign every v to a clique C_v of $\tilde{\mathcal{G}}$ with the property that $\{v\} \cup \mathrm{pa}(v) \subseteq C_v$. Such a clique will always exist after moralization. There may be more than one possible choice for C_v, but we then choose arbitrarily among these. We then assign potential functions

$$\phi_C(x_C) = \prod_{v:C_v=C} k_v(x_v \,|\, x_{\mathrm{pa}(v)});$$

with the usual convention that the product over the empty set is equal to 1 so we get $\phi_C \equiv 1$ for cliques C that have no node assigned. We also assign potentials to separators, initially $\phi_S \equiv 1$ for all $S \in \mathcal{S}$, where $\mathcal{S}$ is the set of separators in the junction tree.
We now define the *joint potential* of the junction tree as

$$\kappa(x) = f(x) = \prod_{v\in V} k_v(x_v \,|\, x_{\mathrm{pa}(v)}) = \frac{\prod_{C\in\mathcal{C}} \phi_C(x_C)}{\prod_{S\in\mathcal{S}} \phi_S(x_S)}. \tag{3.1}$$

and emphasize that only the clique potentials are actually computed whereas the product above is a conceptual quantity.
From Bayes' formula in Corollary 1.21 we also note that the conditional density of $X_{V\setminus E}$ given $X_E = x_E^*$ is determined as

$$f(x \,|\, x_E^*) = \kappa(x)/p(x_E^*).$$

Formally, we shall *incorporate evidence* $X_E = x_E^*$ by multiplying the clique potentials with appropriate indicator functions, i.e.

$$\phi_{C_v}(x) \leftarrow \phi_{C_v}(x) 1_{\{x_v^*\}}(x_v), \quad v \in E.$$

We shall see below that the expression on the right-hand side of (3.1) will remain invariant under the message passing process.
For simplicity we shall in the following assume that all state spaces $\mathcal{X}_v$ are discrete and finite and densities are always expressed w.r.t. counting measure.

3.2.3.2 *Marginalization* We define the *A-marginal* of a potential ϕ_B for $A \subseteq V$ as

$$\phi_B^{\downarrow A}(x) = \phi_B^{\downarrow A}(x_A) = \sum_{y_B : y_{A\cap B} = x_{A\cap B}} \phi_B(y)$$

Since ϕ_B depends on x through x_B only, it is true that if $B \subseteq V$ is 'small', its A-marginal can be computed easily. Note also that the marginal $\phi^{\downarrow A}$ depends on x_A only and that marginalization satisfies the following properties:

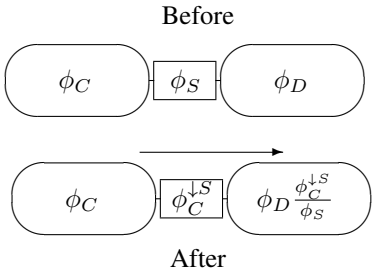
Consonance: For subsets A and B it holds that $\phi^{\downarrow(A\cap B)} = \left(\phi^{\downarrow B}\right)^{\downarrow A}$;
Distributivity: If ϕ_C depends on x_C only and $C \subseteq B$, it holds that $(\phi\phi_C)^{\downarrow B} = \left(\phi^{\downarrow B}\right)\phi_C$.

The distributivity ensures that we can move factors in a sum outside of the summation sign.

3.2.4 *Message passing*

The computation now proceeds by neighbouring cliques communicating by appropriate messages, exploiting the separators as transmitters. The basic operation of sending a message is described below.

3.2.4.1 *Messages* When C sends a message to D, the following happens:



In words, ϕ_D receives the message from C by multiplying its potential by the ratio $\phi_C^{\downarrow S}/\phi_S$; the separator potential is subsequently replaced by the S-marginal $\phi_C^{\downarrow S}$ of the C-potential. The computations are completely local, only involving variables associated with the communicating cliques. We note in particular that the expression for the joint potential

$$\kappa(x) = \frac{\prod_{C\in\mathcal{C}} \phi_C(x_C)}{\prod_{S\in\mathcal{S}} \phi_S(x_S)}$$

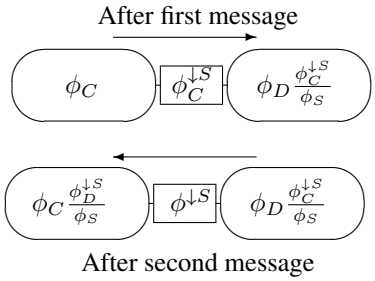
is *invariant under the message passing* since $\phi_C\phi_D/\phi_S$ is:

$$\frac{\phi_C\, \phi_D \frac{\phi_C^{\downarrow S}}{\phi_S}}{\phi_C^{\downarrow S}} = \frac{\phi_C\phi_D}{\phi_S}.$$

Also, after the message has been sent, *D contains the D-marginal of $\phi_C\phi_D/\phi_S$*. To see this we calculate

$$\left(\frac{\phi_C\phi_D}{\phi_S}\right)^{\downarrow D} = \frac{\phi_D}{\phi_S}\phi_C^{\downarrow D} = \frac{\phi_D}{\phi_S}\phi_C^{\downarrow S}.$$

3.2.4.2 *Second message* Suppose that after the first message, D returns a message to C, i.e. the following happens:



After two messages all sets contain the relevant marginal of $\phi = \phi_C\phi_D/\phi_S$. To see this, we argue as follows: The total marginal is

$$\phi^{\downarrow S} = \left(\frac{\phi_C\phi_D}{\phi_S}\right)^{\downarrow S} = (\phi^{\downarrow D})^{\downarrow S} = \left(\phi_D\frac{\phi_C^{\downarrow S}}{\phi_S}\right)^{\downarrow S} = \frac{\phi_C^{\downarrow S}\,\phi_D^{\downarrow S}}{\phi_S}.$$

The clique C contains

$$\phi_C \frac{\phi^{\downarrow S}}{\phi_C^{\downarrow S}} = \frac{\phi_C}{\phi_S}\phi_D^{\downarrow S} = \phi^{\downarrow C}$$

since, as before,

$$\left(\frac{\phi_C \phi_D}{\phi_S}\right)^{\downarrow C} = \frac{\phi_C}{\phi_S}\phi_D^{\downarrow C} = \frac{\phi_C}{\phi_S}\phi_D^{\downarrow S}.$$

Note that any further messages between C and D are neutral! Nothing will change if a message is repeated.

3.2.5 *Message scheduling*

We now schedule the message passing for the entire junction tree in two phases

(i) COLLECTEVIDENCE: messages are sent from leaves towards arbitrarily chosen root R. *After* COLLECTEVIDENCE*, the root potential* ϕ_R *satisfies*

$$\phi_R(x_R) = \kappa^{\downarrow R}(x_R) = p(x_R, x_E^*).$$

(ii) DISTRIBUTEEVIDENCE: messages are sent from root R towards leaves. *After* COLLECTEVIDENCE *and subsequent* DISTRIBUTEEVIDENCE*, it holds for all* $B \in \mathcal{C} \cup \mathcal{S}$ *that*

$$\phi_B(x_B) = \kappa^{\downarrow B}(x_B) = p(x_B, x_E^*).$$

Hence $p(x_E^*) = \sum_{x_S} \phi_S(x_S)$ for any $S \in \mathcal{S}$ and $p(x_v \mid x_E^*)$ can readily be computed from any ϕ_S with $v \in S$.

3.2.5.1 *Correctness of algorithm* The correctness of the algorithm is easily established by induction: The previous considerations in fact establish correctness for a junction tree with only two cliques since we have found that after two messages all messages are neutral and every clique or separator contains the marginal of the joint potential.

Now consider a leaf clique L of the junction tree and let $V^* = \cup_{C:C\in\mathcal{C}\setminus\{L\}} C$. We can then think of L and V^* forming a junction tree of two cliques with separator $S^* = L \cap C^*$ where C^* is the neighbour of L in the junction tree.

After a message has been sent from L to V^* in the COLLECTEVIDENCE phase, ϕ_{V^*} is equal to the V^*-marginal of κ.

By induction, when all messages have been sent except the one from the neighbour clique C^* to L, all cliques other than L contain the relevant marginal of κ, and

$$\phi_{V^*} = \frac{\prod_{C:C\in\mathcal{C}\setminus\{L\}} \phi_C}{\prod_{S:S\in\mathcal{S}\setminus\{S^*\}} \phi_S}.$$

Now let V^* send its message back to L. To do this, it needs to calculate $\phi_{V^*}^{\downarrow S^*}$. But since $S^* \subseteq C^*$, and $\phi_{C^*} = \phi_{V^*}^{\downarrow C^*}$ we have

$$\phi_{V^*}^{\downarrow S^*} = \phi_{C^*}^{\downarrow S^*}$$

and sending a message from V^* to L is thus equivalent to sending a message from C^* to L. Thus, after this message has been sent, $\phi_L = \kappa^{\downarrow L}$ as desired.

3.2.6 *Alternative scheduling of messages*

There are many other valid ways of scheduling the messages. One alternative is known as local control:

3.2.6.1 *Local control:* Allow clique to send message if and only if it has already received message from all other neighbours. Such messages are *live.*
Using this protocol, there will be one clique who first receives messages from all its neighbours. This is effectively the root R in COLLECTEVIDENCE and DISTRIBUTEEVIDENCE.
Additional messages never do any harm (ignoring efficiency issues) as κ is invariant under message passing.
Exactly two live messages along every branch is needed.

3.2.7 *Alternative computations*

There are many other variants of the message passing procedure, one important being that of *maximization of a function.* To do this by the schemes above, we simply replace sum-marginal with *A–maxmarginal:*

$$\phi_B^{\downarrow A}(x) = \max_{y_B : y_{A\cap B} = x_{A\cap B}} \phi_B(y)$$

This also satisfies *consonance and distributivity* and COLLECTEVIDENCE *yields maximal value* f. Further, DISTRIBUTEEVIDENCE *yields configuration with maximum probability.*
Since (3.1) remains invariant under both kinds of message passing, *one can switch freely between max- and sum-propagation.*

3.2.8 *An example*

To illustrate the previous developments, we consider in detail a simple example. More precisely, consider the directed acyclic graph in Fig. 3.1. and consider the Bayesian

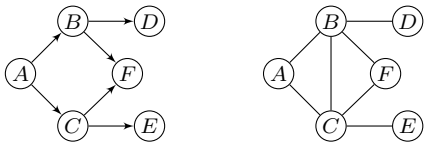


FIG. 3.1. A directed acyclic graph to be used as base for a Bayesian network and its moral graph.

network with all variables taking values in $\{0, 1\}$ and conditional probabilities

$$
\begin{aligned}
P(A=1) &= 3/4;\\
P(B=1\,|\,A=1) &= 3/4;\ P(B=1\,|\,A=0) = 1/4;\\
P(C=1\,|\,A=1) &= 2/3;\ P(C=1\,|\,A=0) = 1/2;\\
P(D=1\,|\,B=1) &= 3/4;\ P(D=1\,|\,B=0) = 1/2;\\
P(E=1\,|\,C=1) &= 3/5;\ P(E=1\,|\,C=0) = 1/2;\\
P(F=1\,|\,B,C) &= (B+C)/2.
\end{aligned}
$$

We now wish to calculate, for example, $P(B=1\,|\,E=1,F=1)$ by probability propagation and proceed through each of the steps previously described.
Moralization yields the undirected graph to the right in Fig. 3.1. As this graph is already chordal, there is no need for *triangulation*.
Next, the cliques of the graph are arranged in a junction tree, for example as displayed in Fig. 3.2. For the *initialization* we assign nodes A, B, C to ABC, D to BD, E to

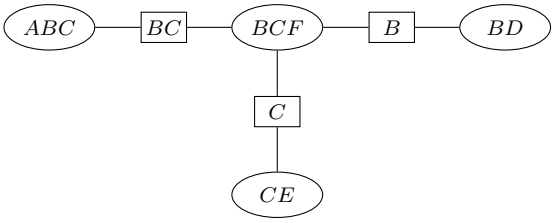


FIG. 3.2. Junction tree for the moral graph of the Bayesian network in Fig. 3.1.

EC, and F to BCF which yields the clique potentials in Table 3.1, whereas the separator potentials all are initialized with values 1.
We have now initialized the system so we can calculate any conditional probability of interest.
To obtain $P(B=1\,|\,E=1,F=1)$, we first incorporate the information that $E=F=1$ into the tables and obtain the updated tables in Table 3.2. This is simply done by replacing entries corresponding to values of $E=0$ or $F=0$ by zero.
We are now ready for message passing, but need to choose a root clique; if we only want to get $P(B=1\,|\,E=1,F=1)$, we can make things easy for ourselves by choosing a root clique containing B, for example BD. Then we can avoid DISTRIBUTEEVIDENCE as COLLECTEVIDENCE to BD alone yields the correct BD marginal. DISTRIBUTEEVIDENCE is only needed if we wish to calculate more probabilities.

TABLE 3.1. Initial clique potentials for the junction tree in Fig. 3.2.

Clique	State			Potential
ABC	1	1	1	$3/4 \times 3/4 \times 2/3 = 3/8$
	1	1	0	$3/4 \times 3/4 \times 1/3 = 3/16$
	1	0	1	$3/4 \times 1/4 \times 2/3 = 1/8$
	1	0	0	$3/4 \times 1/4 \times 1/3 = 1/16$
	0	1	1	$1/4 \times 1/4 \times 1/2 = 1/32$
	0	1	0	$1/4 \times 1/4 \times 1/2 = 1/32$
	0	0	1	$1/4 \times 3/4 \times 1/2 = 3/32$
	0	0	0	$1/4 \times 3/4 \times 1/2 = 3/32$
BCF	1	1	1	1
	1	1	0	0
	1	0	1	$1/2$
	1	0	0	$1/2$
	0	1	1	$1/2$
	0	1	0	$1/2$
	0	0	1	0
	0	0	0	1
BD	1	1		$3/4$
	1	0		$1/4$
	0	1		$1/2$
	0	0		$1/2$
CE	1	1		$3/5$
	1	0		$2/5$
	0	1		$1/2$
	0	0		$1/2$

TABLE 3.2. Revised clique potentials for the junction tree in Fig. 3.2 after incorporating information that $E = F = 1$.

Clique	State			Potential
ABC	1	1	1	$3/8$
	1	1	0	$3/16$
	1	0	1	$1/8$
	1	0	0	$1/16$
	0	1	1	$1/32$
	0	1	0	$1/32$
	0	0	1	$3/32$
	0	0	0	$3/32$
BCF	1	1	1	1
	1	1	0	0
	1	0	1	$1/2$
	1	0	0	0
	0	1	1	$1/2$
	0	1	0	0
	0	0	1	0
	0	0	0	0
BD	1	1		$3/4$
	1	0		$1/4$
	0	1		$1/2$
	0	0		$1/2$
CE	1	1		$3/5$
	1	0		0
	0	1		$1/2$
	0	0		0

For COLLECTEVIDENCE we now first send messages from ABC to BCF and from CE to BCF. We find the separator potential on BC by marginalizing the ABC potential over A and similarly for the separator C. We obtain the separator potentials in Table 3.3.

TABLE 3.3. Separator potentials after sending messages from ABC to BCF and from CE to BCF.

Sep	State		Pot	Sep	State	Pot	Sep	State	Pot
BC	1	1	$3/8 + 1/32 = 13/32$	C	1	3/5	B	1	1
	1	0	$3/16 + 1/32 = 7/32$		0	1/2		0	1
	0	1	$1/8 + 3/32 = 7/32$						
	0	0	$1/16 + 3/32 = 5/32$						

The clique potentials in clique BCF then changes after message passing from ABC to BCF and from CE to BCF as in Table 3.4.

TABLE 3.4. Clique potential for BCF after incorporating $E = F = 1$ and sending messages from ABC to BCF and from CE to BCF.

Clique	State			Potential
BCF	1	1	1	$1 \times 13/32 \times 3/5 = 39/160$
	1	1	0	0
	1	0	1	$1/2 \times 7/32 \times 1/2 = 7/128$
	1	0	0	0
	0	1	1	$1/2 \times 7/32 \times 3/5 = 21/320$
	0	1	0	0
	0	0	1	0
	0	0	0	0

Finally, COLLECTEVIDENCE is completed after sending the last message from BCF to BD, yielding the separator potentials in Table 3.5 and the clique potentials in Table 3.6.

TABLE 3.5. Separator potentials after completion of COLLECTEVIDENCE to the root BD.

Sep	State		Pot	Sep	State	Pot	Sep	State	Pot
BC	1	1	13/32	C	1	3/5	B	1	39/160+7/128=191/640
	1	0	7/32		0	1/2		0	21/320
	0	1	7/32						
	0	0	5/32						

TABLE 3.6. Clique potentials for the junction tree in Fig. 3.2 after incorporating information that $E = F = 1$ and COLLECTEVIDENCE to the root BD.

Clique	State	Potential
ABC	1 1 1	$3/8$
	1 1 0	$3/16$
	1 0 1	$1/8$
	1 0 0	$1/16$
	0 1 1	$1/32$
	0 1 0	$1/32$
	0 0 1	$3/32$
	0 0 0	$3/32$
BCF	1 1 1	$39/160$
	1 1 0	0
	1 0 1	$7/128$
	1 0 0	0
	0 1 1	$21/320$
	0 1 0	0
	0 0 1	0
	0 0 0	0
BD	1 1	$3/4 \times 191/640 = 573/2560$
	1 0	$1/4 \times 191/640 = 191/2560$
	0 1	$1/2 \times 21/320 = 21/640$
	0 0	$1/2 \times 21/320 = 21/640$
CE	1 1	$3/5$
	1 0	0
	0 1	$1/2$
	0 0	0

The root clique BD now contains the correct (unnormalized) marginal potential and the normalizing constant can be obtained by adding clique potentials in clique BD to yield

$$P(E = 1, F = 1) = 573/2560{+}191/2560{+}21/640{+}21/640 = 932/2560 = 233/640.$$

The conditional probability we were looking for is then obtained by normalizing the clique potential of BD and adding appropriate entries

$$P(B = 1 \,|\, E = 1, F = 1) = \frac{640}{233}(573/2560 + 191/2560) = 191/233 \approx 0.81974$$

After COLLECTEVIDENCE we can find the remaining conditional probabilities by DISTRIBUTEEVIDENCE, sending messages away from the root, should we so wish. The first message is sent to BCF, changing the separator potential in B to be the B-marginal of BD, i.e.

$$\phi_B(1) = 573/2560 + 191/2560 = 764/2560 = 191/640,$$
$$\phi_B(0) = 21/640 + 21/640 = 21/320,$$

and the BCF clique is then updated by the ratio of the new potential to the old, to yield Table 3.7; note that indeed this does not change the BCF potential as the original BD potential was not modified, so the BD clique has nothing new to report to BCF.

TABLE 3.7. Potential for BCF after incorporating $E = F = 1$, COLLECTEVIDENCE to BD, and sending messages from BD to BCF.

Clique	State			Potential
BCF	1	1	1	$39/160 \times 191/640 \times 640/191 = 39/160$
	1	1	0	0
	1	0	1	$7/128 \times 191/640 \times 640/191 = 7/128$
	1	0	0	0
	0	1	1	$21/320 \times 21/320 \times 320/21 = 21/320$
	0	1	0	0
	0	0	1	0
	0	0	0	0

We next calculate separator potentials associated with messages from BCF to ABC and CE and display these in Table 3.8.

TABLE 3.8. Final separator potentials after incorporating evidence $E = F = 1$, COLLECTEVIDENCE, and subsequent DISTRIBUTEEVIDENCE.

Sep	State		Pot	Sep	State	Pot	Sep	State	Pot
BC	1	1	$39/160$	C	1	39/160+21/320=99/320	B	1	191/640
	1	0	$7/128$		0	7/128		0	21/320
	0	1	$21/320$						
	0	0	0						

Note that all separator potentials add up to the general normalizing constant $233/640$.

Finally, updating the clique potentials in ABC and CE yields the potentials in Table 3.9.

TABLE 3.9. Final clique potentials after incorporating information that $E = F = 1$, COLLECTEVIDENCE, and subsequent DISTRIBUTEEVIDENCE.

Clique	State			Potential
ABC	1	1	1	$3/8 \times 39/160 \times 32/13 = 9/40$
	1	1	0	$3/16 \times 7/128 \times 32/7 = 3/64$
	1	0	1	$1/8 \times 21/320 \times 32/7 = 3/80$
	1	0	0	$1/16 \times 0 = 0$
	0	1	1	$1/32 \times 39/160 \times 32/13 = 3/160$
	0	1	0	$1/32 \times 7/128 \times 32/7 = 1/128$
	0	0	1	$3/32 \times 21/320 \times 32/7 = 9/320$
	0	0	0	0
BCF	1	1	1	$39/160$
	1	1	0	0
	1	0	1	$7/128$
	1	0	0	0
	0	1	1	$21/320$
	0	1	0	0
	0	0	1	0
	0	0	0	0
BD	1	1		$573/2560$
	1	0		$191/2560$
	0	1		$21/640$
	0	0		$21/640$
CE	1	1		$3/5 \times 99/320 \times 5/3 = 99/320$
	1	0		0
	0	1		$1/2 \times 7/128 \times 2/1 = 7/128$
	0	0		0

Again we note that all clique potentials add up to $233/640$ and if we wish to calculate, say $P(A = 1 \mid E = 1, F = 1)$, we simply add up and normalize

$$P(A = 1 \mid E = 1, F = 1) = \frac{640}{233}(9/40 + 3/64 + 3/80) = 198/233 \approx 0.84979.$$

Suppose we instead wish to identify the most probable configuration of the variables A, B, C, D given $E = 1$ and $F = 1$. This can be found by using maximum in the marginalizations instead of sums. We do not need to start afresh, as the current

potentials in the junction tree have all information needed. For a change, we might now choose the clique BCF as root and the first step is then to send messages from all the other cliques to BCF yielding new separator potentials as displayed in Table 3.10.

TABLE 3.10. Separator potentials after incorporating evidence $E = F = 1$, standard COLLECTEVIDENCE and DISTRIBUTEEVIDENCE, and subsequent max-messages to new root clique BCF.

Sep	State		Pot	Sep	State	Pot	Sep	State	Pot
BC	1	1	$9/40$	C	1	99/320	B	1	573/2560
	1	0	$3/64$		0	7/128		0	21/640
	0	1	$3/80$						
	0	0	0						

The updated clique potential for BCF is displayed in Table 3.11.

TABLE 3.11. Potential for BCF after incorporating $E = F = 1$, COLLECTEVIDENCE and DISTRIBUTEEVIDENCE, and sending max-messages to BCF.

Clique	State			Potential
BCF	1	1	1	$39/160 \times 9/40 \times 160/39 \times 99/320 \times 320/99 \times 573/2560 \times 640/191 = 27/160$
	1	1	0	0
	1	0	1	$7/128 \times 3/64 \times 128/7 \times 7/128 \times 128/7 \times 573/2560 \times 640/191 = 9/256$
	1	0	0	0
	0	1	1	$21/320 \times 3/80 \times 320/21 \times 99/320 \times 320/99 \times 21/640 \times 320/21 = 3/160$
	0	1	0	0
	0	0	1	0
	0	0	0	0

Next, for DISTRIBUTEEVIDENCE, we send max-messages away from BCF to obtain the final separator potentials, displayed in Table 3.12.

TABLE 3.12. Final separator potentials after incorporating $E = F = 1$, standard COLLECTEVIDENCE and DISTRIBUTEEVIDENCE, and subsequent full max-propagation.

Sep	State		Pot	Sep	State	Pot	Sep	State	Pot
BC	1	1	$27/160$	C	1	27/160	B	1	27/160
	1	0	$9/256$		0	9/256		0	3/160
	0	1	$9/320$						
	0	0	0						

The final step is incorporating the messages in the other cliques, the results being displayed in Table 3.13

TABLE 3.13. Final clique potentials after incorporating information that $E = F = 1$, and a full max-propagation.

Clique	State			Potential
ABC	1	1	1	$9/40 \times 27/160 \times 40/9 = 27/160$
	1	1	0	$3/64 \times 9/256 \times 64/3 = 9/256$
	1	0	1	$3/80 \times 9/320 \times 80/3 = 9/320$
	1	0	0	0
	0	1	1	$3/160 \times 27/160 \times 40/9 = 9/640$
	0	1	0	$1/128 \times 9/256 \times 64/3 = 3/512$
	0	0	1	$9/320 \times 9/320 \times 320/9 = 9/320$
	0	0	0	0
BCF	1	1	1	$27/160$
	1	1	0	0
	1	0	1	$9/256$
	1	0	0	0
	0	1	1	$3/160$
	0	1	0	0
	0	0	1	0
	0	0	0	0
BD	1	1		$573/2560 \times 27/160 \times 2560/573 = 27/160$
	1	0		$191/2560 \times 27/160 \times 2560/573 = 9/160$
	0	1		$21/640 \times 3/160 \times 640/21 = 3/160$
	0	0		$21/640 \times 3/160 \times 640/21 = 3/160$
CE	1	1		$99/320 \times 27/160 \times 320/99 = 27/160$
	1	0		0
	0	1		$7/128 \times 9/256 \times 128/7 = 9/256$
	0	0		0

Note again that all clique and separating potentials have the same maximal value which is the probability of the most likely configuration jointly with the evidence

$$P(A = B = C = D = E = F = 1) = 27/160 \approx .16875$$

and hence

$$P(A = B = C = D = 1 \,|\, E = F = 1) = 27/160 \times 640/233 = 108/233 \approx .54936.$$

3.3 Exercises

Exercise 3.1 Consider the DAG $\mathcal{D}$ with arrows
$A \to C, B \to C, B \to D, C \to E, D \to F, E \to G, E \to H, F \to G, G \to J, I \to J$.

(a) Find the moral graph $\mathcal{D}^m$ of $\mathcal{D}$;

(b) Find a *minimal chordal cover* $\mathcal{G}$ of $\mathcal{D}^m$, i.e. a chordal graph $\mathcal{G} \supset \mathcal{D}^m$ with the property that removal of any edge in $\mathcal{G}$ which is not an edge in $\mathcal{D}^m$ will not be chordal;

(c) Arrange the cliques of $\mathcal{G}$ in a junction tree;

(d) For a specification of all conditional distributions $p_{v \mid \mathrm{pa}(v)}, v \in V$, allocate appropriate potentials to the junction tree to prepare for probability propagation.

Exercise 3.2 Consider random variables $X_1, \ldots, X_6$ taking values in $\{-1, 1\}$ and having distribution P with joint probability mass function determined as

$$p(x) \propto \exp\{\theta(x_1x_2 + x_2x_3 + x_2x_5 + x_3x_4 + x_3x_5 + x_3x_6)\},$$

where $\theta \neq 0$.

(a) Find the dependence graph of P determined as the smallest graph $\mathcal{G}$ so that P is Markov (pairwise, local, and globally) w.r.t. $\mathcal{G}$ and identify its cliques;

(b) Verify that $\mathcal{G}$ is chordal;

(c) Set up an appropriate junction tree for probability propagation;

(d) Allocate potentials to cliques;

(e) Calculate $P(X_6 = 1 \mid X_1 = 1, X_4 = 1)$ by probability propagation.

4

MULTIVARIATE NORMAL MODELS

4.1 Basic facts and concepts

4.1.1 *Notation*

The graphical models in this chapter have a particularly simple interpretation and a rather detailed statistical theory. The models assume that the variables observed follow a regular multivariate normal distribution. Conditional independence restrictions in the multivariate normal distribution can be expressed in a simple fashion through zero restrictions on the inverse covariance matrix.
We will need some special notation for vectors in $\mathbb{R}^{|\Gamma|}$ and matrices with entries indexed by Γ. An arbitrary element of $\mathbb{R}^{|\Gamma|}$ is denoted as any of

$$y = y_\Gamma = (y_\gamma)_{\gamma\in\Gamma},$$

and for an arbitrary subset $d \subseteq \Gamma$ we let

$$y_d = (y_\gamma)_{\gamma\in d}$$

denote a $|d|$-dimensional subvector of y.
A $|\Gamma| \times |\Gamma|$ matrix with entries indexed by Γ is written as any of

$$A = A_\Gamma = A_{\Gamma\Gamma} = \{a_{\gamma\mu}\}_{\gamma,\mu\in\Gamma},$$

whereas for two arbitrary subsets d and e of Γ we let

$$A_{de} = \{a_{\gamma\mu}\}_{\gamma\in d,\mu\in e}$$

denote a $|d| \times |e|$ submatrix of A. For a partitioning $\Gamma = d \cup e$ with $d \cap e = \emptyset$ we can then use any of the block matrix notations

$$A = \begin{pmatrix} A_d & A_{de} \\ A_{ed} & A_e \end{pmatrix} = \begin{pmatrix} A_{dd} & A_{de} \\ A_{ed} & A_{ee} \end{pmatrix}.$$

For a $|d| \times |e|$ matrix $A = \{a_{\gamma\mu}\}_{\gamma\in d,\mu\in e}$ we let $[A]^\Gamma$ denote the matrix obtained from A by filling up with zero entries to obtain full dimension $|\Gamma| \times |\Gamma|$, i.e.

$$\left([A]^\Gamma\right)_{\gamma\mu} = \begin{cases} a_{\gamma\mu} \text{ if } \gamma \in d, \mu \in e \\ 0 \quad \text{otherwise.} \end{cases}$$

When matrix operations are combined with forming submatrices, we use the convention that the matrix operation is performed first, i.e.

$$A_d^{-1} = \left(A^{-1}\right)_d.$$

4.1.2 *The saturated model*

The model where no conditional independence restrictions are assumed to hold is called *the saturated model* as in the previous chapter. This model is concerned with a sample $(y^1, \ldots, y^n)$ of independent random vectors from a multivariate normal distribution $\mathcal{N}_{|\Gamma|}(0, \Sigma)$, where Σ is unknown and arbitrary apart from the restriction that Σ is assumed to be positive definite. The case where also the mean ξ is unknown is not much different, but the notation is considerably more cumbersome.

4.1.2.1 *Exact results* Using (D.1), we get the likelihood function, expressed in the parameter K as

$$\begin{aligned} L(K) &= (2\pi)^{-n|\Gamma|/2}(\det K)^{n/2} \prod_{\nu=1}^{n} \exp\{-\langle y^\nu, Ky^\nu\rangle/2\} \\ &\propto (\det K)^{n/2} \exp\left\{-\sum_{\nu=1}^{n}(y^\nu)^\top K(y^\nu)/2\right\} \\ &= (\det K)^{n/2} \exp\big\{-\operatorname{tr}(Ky^\top y/2)\big\} \end{aligned} \tag{4.1}$$

We have let y be the $n \times |\Gamma|$ matrix with $(y^\nu)^\top$ as rows. For later use we introduce the matrix of *sum of squares and products* W as

$$w = \sum_{\nu=1}^{n} y^\nu (y^\nu)^\top = y^\top y.$$

.

To maximize the likelihood function, we choose to take advantage of the theory of exponential models. Although it is unnecessary in this particular case, it turns out to be convenient when we later discuss graphical models with conditional independence restrictions.

The expression (4.1) identifies the statistical model determined by the family of multivariate normal distributions with unknown concentration matrix K as an exponential model. To see this, we first recall from (C.5) that

$$\langle A, B\rangle = \operatorname{tr}(A^\top B)$$

defines an inner product on the vector space of matrices of any fixed dimension, in particular also on the subspace $\mathcal{S}_{|\Gamma|}$ of symmetric $|\Gamma| \times |\Gamma|$ matrices.

We define the canonical parameter as $\theta = K$, the base measure μ as Lebesgue measure on $\mathbb{R}^{n\times|\Gamma|}$, and the canonical statistic as $t(y) = -w/2$. Then the exponent in (4.1) can be written as

$$-\operatorname{tr}(Kw/2) = \langle \theta, t(y)\rangle.$$

The cumulant function is found as

$$\psi(K) = \log\{(2\pi)^{n|\Gamma|/2}(\det K)^{-n/2}\} = (n|\Gamma|/2)\log(2\pi) - (n/2)\log\det K. \tag{4.2}$$

Since the integral

$$\int_{\mathbb{R}^{|\Gamma|}} e^{-y^\top Ky/2}\, dy$$

is finite if and only if K is positive definite, it follows that the model is a regular exponential model.
The basic estimation result for the saturated model is given in the theorem below.

Theorem 4.1 *In the saturated multivariate normal model, the maximum likelihood estimate of the unknown covariance matrix exists if and only if*

$$w = y^\top y$$

is positive definite. This happens with probability one if $n \geq |\Gamma|$ *and never when* $n < |\Gamma|$. *When the estimate exists it is given as*

$$\hat{\Sigma} = w/n = y^\top y/n.$$

Proof The rank of W is at most n by construction, so if $n < |\Gamma|$ the maximum likelihood estimate does not exist. The matrix W is positive semidefinite if and only if one of its principal minors (subdeterminants along the diagonal) is equal to zero. The set of y's such that this happens is thus the intersection of a set of polynomial equations, also known as an *algebraic variety*. Such a set is either everything (which happens when $n < |\Gamma|$) or a Lebesque null-set. If we choose $y^\nu, \nu = 1, \ldots, n$ to contain at least n linearly independent vector, $W = w(y)$ is positive definite, so if $n \geq |\Gamma|$ it must be a Lebesgue null-set.
The main result about estimation in exponential models asserts that if the maximum likelihood estimate exists it is determined by the equation

$$\mathbf{E}(-Y^\top Y/2) = -n\Sigma/2 = -y^\top y/2$$

which completes the proof. □

4.1.3 *Conditional independence*

Before the graphical models are described in detail, it seems appropriate to clarify the connection between conditional independence and the multivariate normal distribution. Let $Y = (Y_\gamma)_{\gamma\in\Gamma}$ be a random vector in $\mathbb{R}^{|\Gamma|}$ following a multivariate normal distribution with mean 0 and covariance matrix Σ. Assume the covariance to be regular such that the concentration matrix $K = \Sigma^{-1}$ is well defined. Conditional independence in the multivariate normal distribution is simply reflected in the concentration matrix of the distribution through zero entries. This fact is formalized below.

Proposition 4.2 *Assume that* $Y \sim \mathcal{N}_{|\Gamma|}(0, \Sigma)$, *where* Σ *is regular. Then it holds for* $\gamma, \mu \in \Gamma$ *with* $\gamma \neq \mu$ *that*

$$Y_\gamma \perp\!\!\!\perp Y_\mu \mid Y_{\Gamma\setminus\{\gamma,\mu\}} \iff k_{\gamma\mu} = 0,$$

where $K = \{k_{\alpha\beta}\}_{\alpha,\beta\in\Gamma} = \Sigma^{-1}$ *is the concentration matrix of the distribution.*

Proof This is a direct consequence (1.8) which identifies the matrix

$$K_{\{\gamma,\mu\}} = \begin{pmatrix} k_{\gamma\gamma} & k_{\gamma\mu} \\ k_{\mu\gamma} & k_{\mu\mu} \end{pmatrix} \tag{4.3}$$

as the concentration matrix of the conditional distribution of $Y_{\{\gamma,\mu\}}$ given $Y_{\Gamma\setminus\{\gamma,\mu\}}$. The covariance matrix of this conditional distribution is therefore equal to

$$\Sigma_{\gamma,\mu\,|\,\Gamma\setminus\{\gamma,\mu\}} = \frac{1}{\det K_{\{\gamma,\mu\}}} \begin{pmatrix} k_{\mu\mu} & -k_{\gamma\mu} \\ -k_{\mu\gamma} & k_{\gamma\gamma} \end{pmatrix}. \tag{4.4}$$

The desired independence now follows from Corollary D.5. □

This fundamental relation forms the basis for all models treated in this chapter. Corresponding to the different Markov properties studied in Chapter 2, we have multivariate normal models defined through restricting particular elements in suitable concentration matrices to be equal to zero.
The entries in the concentration matrix K have a simple interpretation. It follows from (1.4) and (1.5) that the diagonal elements $k_{\gamma\gamma}$ are reciprocals of the conditional variances, given the remaining variables, i.e.

$$k_{\gamma\gamma} = \mathbf{V}(Y_\gamma \,|\, Y_{\Gamma\setminus\{\gamma\}})^{-1}$$

for all $\gamma \in \Gamma$. Let further $C = \{c_{\alpha\beta}\}_{\alpha,\beta\in\Gamma}$ be the matrix obtained by scaling K to have all diagonal elements equal to one,

$$c_{\gamma\mu} = \frac{k_{\gamma\mu}}{\sqrt{k_{\gamma\gamma}k_{\mu\mu}}}.$$

Then $c_{\gamma\mu}$, the off-diagonal elements in C, are equal to the negative *partial correlation coefficients*

$$\rho_{\gamma\mu\,|\,\Gamma\setminus\{\gamma,\mu\}} = \frac{\mathrm{Cov}(Y_\gamma, Y_\mu \,|\, Y_{\Gamma\setminus\{\gamma,\mu\}})}{\mathbf{V}(Y_\gamma \,|\, Y_{\Gamma\setminus\{\gamma,\mu\}})^{1/2}\mathbf{V}(Y_\mu \,|\, Y_{\Gamma\setminus\{\gamma,\mu\}})^{1/2}} = -c_{\gamma\mu}.$$

This follows from (4.4) since

$$\mathbf{V}(Y_\gamma \,|\, Y_{\Gamma\setminus\{\gamma,\mu\}}) = \frac{k_{\mu\mu}}{k_{\gamma\gamma}k_{\mu\mu} - (k_{\gamma\mu})^2}$$

and

$$\mathrm{Cov}(Y_\gamma, Y_\mu \,|\, Y_{\Gamma\setminus\{\gamma,\mu\}}) = \frac{-k_{\gamma\mu}}{k_{\gamma\gamma}k_{\mu\mu} - (k_{\gamma\mu})^2}.$$

Note that it also holds that

$$\left(\rho_{\gamma\mu\,|\,\Gamma\setminus\{\gamma,\mu\}}\right)^2 = (c_{\gamma\mu})^2 = 1 - \frac{\det\Sigma \,\det\Sigma_{\Gamma\setminus\{\gamma,\mu\}}}{\det\Sigma_{\Gamma\setminus\{\gamma\}} \,\det\Sigma_{\Gamma\setminus\{\mu\}}}. \tag{4.5}$$

This follows from the relations

$$\det \Sigma = \det \Sigma_{\Gamma\setminus\{\gamma\}} \mathbf{V}(Y_\gamma \,|\, Y_{\Gamma\setminus\{\gamma\}}) = \det \Sigma_{\Gamma\setminus\{\gamma\}} / k_{\gamma\gamma}$$

and

$$\det \Sigma_{\Gamma\setminus\{\mu\}} = \det \Sigma_{\Gamma\setminus\{\gamma,\mu\}} \mathbf{V}(Y_\gamma \,|\, Y_{\Gamma\setminus\{\gamma,\mu\}}) = \det \Sigma_{\Gamma\setminus\{\gamma,\mu\}} \frac{k_{\mu\mu}}{k_{\gamma\gamma}k_{\mu\mu} - (k_{\gamma\mu})^2},$$

which both are easy consequences of (1.7). From Example 1.20 we have that the conditional distribution of Y_γ given $Y_{\Gamma\setminus\{\gamma\}} = y_{\Gamma\setminus\{\gamma\}}$ is univariate normal. Writing the conditional expectation as

$$\xi_\gamma + \sum_{\mu\in\Gamma\setminus\{\gamma\}} \beta_{\gamma\mu\,|\,\Gamma\setminus\{\gamma\}}(y_\mu - \xi_\mu)$$

and using (1.6) we find the *partial regression coefficient* as

$$\beta_{\gamma\mu\,|\,\Gamma\setminus\{\gamma\}} = -k_{\gamma\mu}/k_{\gamma\gamma}.$$

4.1.4 *Interaction*

It is illuminating to investigate the additive terms in the logarithm of the normal density, thereby highlighting the analogy to interaction expansions of discrete models. Using the expression (D.1) for the multivariate normal density we get

$$\log f(y) = c - \langle y, K(y)\rangle/2 = c - \frac{1}{2}\sum_{\gamma\in\Gamma} k_{\gamma\gamma}y_\gamma^2 - \sum_{\{\gamma,\mu\}} k_{\gamma\mu}y_\gamma y_\mu \qquad (4.6)$$

where $\{\gamma,\mu\}$ in the sum above represent all unordered pairs of elements of Γ, and c is a constant.

The expansion shows that the logarithm of the density is additively composed of *quadratic main effects* with coefficients $-k_{\gamma\gamma}/2$and *quadratic interactions* with coefficients $-k_{\gamma\mu}$. We will sometimes use the terms interactions and main effects referring directly to the coefficients and also omit the negative signs and the division by two. So, for example, we will refer to $k_{\gamma\gamma}$ as the quadratic main effect of the variable γ, although this, strictly speaking, should refer to $-k_{\gamma\gamma}y_\gamma^2/2$.

We emphasize that the interaction terms of highest order in (4.6) involve pairs of variables, and there are no terms involving groups of variables with three or more elements. This is in contrast to the discrete case and it follows in particular that within the normal distribution there are no hierarchical interaction models which are not conformal.

4.2 Covariance selection models

The interaction models for the multivariate normal distribution are called covariance selection models. They are determined by assuming conditional independence of selected pairs of variables, given the remaining ones.

Thus, if $\mathcal{G} = (\Gamma, E)$ is an undirected graph and $Y = Y_\Gamma$ is a random variable taking values in $\mathbb{R}^{|\Gamma|}$, the *Gaussian graphical model* or *covariance selection model* for Y with graph $\mathcal{G}$ is given by assuming that Y follows a multivariate normal distribution which obeys the undirected pairwise Markov property with respect to $\mathcal{G}$. Since the density is positive and continuous, this implies the global and local Markov properties and the density factorizes.
It follows from Proposition 4.2 that this is equivalent to assuming the quadratic interactions $k_{\gamma\mu}$ to be equal to zero for all pairs γ, μ which are not adjacent in $\mathcal{G}$. The expression (4.6) for the normal density then reduces to

$$\log f(y) = c - \frac{1}{2}\sum_{\gamma\in\Gamma} k_{\gamma\gamma} y_\gamma^2 - \sum_{\gamma\mu\in E} k_{\gamma\mu} y_\gamma y_\mu.$$

Let $\mathcal{S}(\mathcal{G})$ denote the set of symmetric matrices A satisfying for all $\gamma, \mu \in \Gamma$ that

$$\gamma \not\sim \mu \implies a_{\gamma\mu} = 0$$

and $\mathcal{S}^+(\mathcal{G})$ those elements of $\mathcal{S}(\mathcal{G})$ that are positive definite. Then the covariance selection model for Y can be compactly described as

$$Y \sim \mathcal{N}_{|\Gamma|}(0, \Sigma), \quad \Sigma^{-1} \in \mathcal{S}^+(\mathcal{G}).$$

4.2.1 *Maximum likelihood estimation*

4.2.1.1 *The likelihood equations* Consider a sample $(y^1, \ldots, y^n)$ from a covariance selection model. The likelihood function is obtained from (4.1):

$$L(K) \propto (\det K)^{n/2} \exp\left\{-\operatorname{tr}(K y^\top y/2)\right\}.$$

For an arbitrary matrix A, we let $A(\mathcal{G})$ denote the matrix with entries

$$A(\mathcal{G})_{\gamma\mu} = \begin{cases} 0 & \text{if } \gamma \not\sim \mu \\ a_{\gamma\mu} & \text{otherwise.} \end{cases}$$

Exploiting that $K \in \mathcal{S}(\mathcal{G})$ we find that

$$\operatorname{tr}(K y^\top y) = \operatorname{tr}\{K w(\mathcal{G})\},$$

where we have let $w = y^\top y$. Thus the likelihood function reduces to

$$L(K) \propto (\det K)^{n/2} \exp\left[-\operatorname{tr}\{K w(\mathcal{G})/2\}\right]. \tag{4.7}$$

The restriction which is imposed on the distribution of Y by the model is linear in the canonical parameter K. Hence the hypothesis $K \in \mathcal{S}^+(\mathcal{G})$ is an affine hypothesis and it follows that a covariance selection model is itself a regular exponential model with canonical statistic equal to $-w(\mathcal{G})/2$. The following result about maximum likelihood estimation then follows directly.

Theorem 4.3 *In the covariance selection model, the maximum likelihood estimate of the unknown covariance matrix exists if*

$$w = y^\top y$$

is positive definite. If $n \geq |\Gamma|$ this happens with probability one. When the estimate exists it is determined as the unique solution to the system of equations

$$n\hat{\sigma}_{\rho\rho} = w_{\rho\rho}, \quad n\hat{\sigma}_{\gamma\mu} = w_{\gamma\mu}, \quad \rho \in \Gamma, \{\gamma, \mu\} \in E,$$

which also satisfies the model restriction $\Sigma^{-1} \in \mathcal{S}^+(\mathcal{G})$.

Proof The maximum likelihood estimates are obtained directly as for the saturated model:

$$\mathbf{E}\{-SS(\mathcal{G})/2\} = -n\Sigma(\mathcal{G})/2 = -w(\mathcal{G})/2$$

whence the result follows. □

Note that the condition $n \geq |\Gamma|$ for existence of the maximum likelihood estimate in this case is only sufficient, not necessary. From the likelihood equations it follows that a necessary condition is that $n \geq \max_{C \in \mathcal{C}} |C|$, as otherwise w_C would not all be positive definite. However, this condition is not sufficient for the existence. The problem has been studied in some detail by Buhl (1993), Uhler (2012), and Gross and Sullivant (2018). We shall return to a discussion of this issue later in Section 4.3.2.
An alternative way of writing the estimating equations is

$$n\hat{\Sigma}_{cc} = w_{cc} \quad \text{for all } c \in \mathcal{C}, \tag{4.8}$$

where $\mathcal{C}$ is the set of cliques of $\mathcal{G}$.
In a general covariance selection model no exact distributional results concerning the estimate of the covariance matrix are available.
The asymptotic distribution of the maximum likelihood estimate is multivariate normal from standard exponential family theory. The asymptotic covariance in the case of a general covariance selection model is less straightforward.

4.2.1.2 *Iterative proportional scaling* Theorem 4.3 identifies which equations to solve in order to maximize the likelihood function, but it gives no advice on doing so. In general the equations concerning the estimates for the covariance matrix have to be solved by iterative methods. Below we describe one of these, which is based upon the method of iterative partial maximization described in Section A.4. It consists of iteratively and successively adjusting the covariance matrices for the clique marginals appearing in (4.8). The algorithm was discussed in detail by Speed and Kiiveri (1986) along with other algorithms.
Let w from a sample $(y^1, \ldots, y^n)$ be given, and consider a covariance selection model with graph $\mathcal{G}$. For $K \in \mathcal{S}^+(\mathcal{G})$ and $c \in \mathcal{C}$, define the operation of 'adjusting the c-marginal' by

$$(T_c K)_{cc} = K_{cc} + n(w_{cc})^{-1} - (K_{cc}^{-1})^{-1} = K_{cc} + n(w_{cc})^{-1} - (\Sigma_{cc})^{-1}, \tag{4.9}$$

leaving all other entries of K unchanged. This operation is clearly well defined if w_{cc} is positive definite. If we let $a = \Gamma \setminus c$ and exploit (C.2) we find

$$(K^{-1})_{cc} = \Sigma_{cc} = \left\{K_{cc} - K_{ca}(K_{aa})^{-1}K_{ac}\right\}^{-1}, \tag{4.10}$$

giving the alternative expression

$$T_c K = \begin{pmatrix} n(w_{cc})^{-1} + K_{ca}(K_{aa})^{-1}K_{ac} & K_{ca} \\ K_{ac} & K_{aa} \end{pmatrix}. \tag{4.11}$$

If a is large and c is small, the expression (4.11) is computationally heavy as a large matrix K_{aa} needs to be inverted, so it may be easier to calculate the update using (4.9). The latter demands that Σ_{cc} is available. After updating K, Σ can be updated simply using (C.3) with $\Delta = n(w_{cc})^{-1} - (\Sigma_{cc})^{-1}$ and $C^\top = (I_c : 0_a)$ so that $C\Delta C^\top = [\Delta]^\Gamma$. We then get

$$T_c\Sigma = \Sigma - \Sigma[H]^\Gamma\Sigma$$

where $H = (\Delta^{-1} + (\Sigma_{cc})^{-1})^{-1}$.
Using the expression (4.11) for T_cK, we find the covariance $\tilde{\Sigma}_{cc}$ corresponding to the adjusted concentration matrix

$$\begin{aligned} \tilde{\Sigma}_{cc} &= (T_cK)^{-1}_{cc} \\ &= \left\{n(w_{cc})^{-1} + K_{ca}(K_{aa})^{-1}K_{ac} - K_{ca}(K_{aa})^{-1}K_{ac}\right\}^{-1} \\ &= w_{cc}/n, \end{aligned} \tag{4.12}$$

hence T_cK does indeed adjust the marginals. From (4.9) it is seen that the pattern of zeros in K is preserved, i.e. T_cK is in $\mathcal{S}(\mathcal{G})$ if K is, and applying Lemma C.1 to (4.11) shows that it stays positive definite. Hence the adjusted concentration matrix T_cK is in $\mathcal{S}^+(\mathcal{G})$ if K is.
In fact, it is not difficult to see that the operation T_c also scales proportionally in the sense that

$$f\{y \,|\, (T_cK)^{-1}\} = f(y \,|\, \Sigma)\frac{f(y_c \,|\, w_{cc}/n)}{f(y_c \,|\, \Sigma_{cc})}.$$

This clearly demonstrates the analogy to the procedure used for hierarchical log–affine models.
Next we choose any ordering $(c_1, \ldots, c_k)$ of the cliques in $\mathcal{G}$. Choose further an arbitrary starting value $K_0 \in \mathcal{S}^+(\mathcal{G})$ and define recursively for $r = 0, 1, \ldots$

$$K_{r+1} = (T_{c_1} \cdots T_{c_k})K_r. \tag{4.13}$$

Then we have

Theorem 4.4 *Consider a sample from a covariance selection model with graph $\mathcal{G}$ and assume that w is such that the maximum likelihood estimate $\hat{K}$ of K exists. Then*

$$\hat{K} = \lim_{r\to\infty} K_r.$$

Proof We must realize that this is a special instance of iterative partial maximization, discussed in Section A.4. To do this, we let

$$\Theta_0 = \left\{K \in \mathcal{S}^+(\mathcal{G}) \mid L(K) \geq L(K_0)\right\},$$

where $K_0 \in \mathcal{S}^+(\mathcal{G})$ is chosen arbitrarily, for example as $K_0 = I$. Since a covariance selection model is a regular exponential model and the maximum likelihood estimate is assumed to exist, Θ_0 is compact.
It is obvious that the transformation T_c is continuous for all $c \in \mathcal{C}$ and, as mentioned, also that it maps $\mathcal{S}^+(\mathcal{G})$ into itself.
Next we establish that T_cK maximizes the likelihood function over the section

$$\Theta_c = \Theta_c(K) = \left\{A \in \mathcal{S}^+(\mathcal{G}) \mid A_{aa} = K_{aa}, A_{ac} = K_{ac}\right\},$$

where we have let $a = \Gamma \setminus c$. In Θ_c it holds that

$$\begin{aligned}&\operatorname{tr}\{Kw(\mathcal{G})/2\}\\ &= \operatorname{tr}\{K_{cc}w_{cc}/2\} + \operatorname{tr}\{A_{aa}w(\mathcal{G})_{aa}/2\} + \operatorname{tr}\{A_{ac}w(\mathcal{G})_{ca}\}.\end{aligned}$$

Using the expression (4.7) for the likelihood function, we identify the subfamily determined by the section as an exponential family with canonical statistic $-w_{cc}/2$, leading to the likelihood equations

$$-n\Sigma_{cc}/2 = -w_{cc}/2.$$

Using (4.12) identifies $T_c(K)$ as maximizer. Since we know already that the global maximum of L is unique, the theorem now follows from Proposition A.15. □

4.3 Decomposable models

Here we study the special features of covariance selection models whose interaction graphs are decomposable. Theorem B.14 implies that these models are built up from saturated models by successive direct joins. This structure makes it possible to break down the statistical analysis of a decomposable model into small analyses of saturated submodels in an elegant way.

4.3.1 *Basic factorizations*

As shown in Proposition B.10, we can number the cliques of a decomposable graph $\mathcal{G}$ to form a perfect sequence, i.e. a sequence $C_1, \ldots, C_k$ where each combination of subgraphs induced by $H_{j-1} = C_1 \cup \cdots \cup C_{j-1}$ and C_j is a decomposition. Repeated use of (2.27) gives

$$f(y) = \frac{\prod_{j=1}^k f(y_{C_j})}{\prod_{j=2}^k f(y_{S_j})} = \frac{\prod_{C\in\mathcal{C}} f(y_C)}{\prod_{S\in\mathcal{S}} f(y_S)^{\nu(S)}}, \tag{4.14}$$

where $S_j = H_{j-1} \cap C_j$ is the sequence of separators and $\nu(S)$ is the multiplicity of S, i.e. the number of times S occurs in a perfect sequence; see Proposition B.18 and Corollary B.18. We further find

$$K = \Sigma^{-1} = \sum_{C\in\mathcal{C}} [K_C]^\Gamma - \sum_{S\in\mathcal{S}} \nu(S)\,[K_S]^\Gamma$$
$$= \sum_{C\in\mathcal{C}} \left[(\Sigma_C)^{-1}\right]^\Gamma - \sum_{S\in\mathcal{S}} \nu(S)\left[(\Sigma_S)^{-1}\right]^\Gamma$$

as well as

$$\det\Sigma = \frac{\prod_{C\in\mathcal{C}} \det\Sigma_C}{\prod_{S\in\mathcal{S}} (\det\Sigma_S)^{\nu(S)}}. \tag{4.15}$$

4.3.2 *Maximum likelihood estimation*

4.3.2.1 *Exact results* Previously we derived a formula for combining maximum likelihood estimates of concentration matrices in two covariance selection models to find the estimate in the model formed by their direct join.
Combining this with the usual simple estimates in the saturated models, explicit formulae for the maximum likelihood estimate in a decomposable covariance selection model can be derived. More precisely we find that

$$\hat{K} = n\left\{\sum_{j=1}^{k}\left[\left(w_{C_j}\right)^{-1}\right]^\Gamma - \sum_{j=2}^{k}\left[\left(w_{S_j}\right)^{-1}\right]^\Gamma\right\}. \tag{4.16}$$

Using the alternative expression where the separators and cliques are not numbered and the expression for the determinant (4.15), we also get

Proposition 4.5 *In a decomposable covariance selection model with graph $\mathcal{G}$, the maximum likelihood estimate of the mean vector and concentration matrix exists with probability one if and only if $n > \max_{C\in\mathcal{C}} |C|$. It is then given as*

$$\hat{\xi} = \bar{y}, \quad \hat{K} = n\left\{\sum_{C\in\mathcal{C}}\left[(w_C)^{-1}\right]^\Gamma - \sum_{S\in\mathcal{S}} \nu(S)\left[(w_S)^{-1}\right]^\Gamma\right\}, \tag{4.17}$$

where $\mathcal{C}$ is the set of cliques of $\mathcal{G}$ and $\mathcal{S}$ the separators with multiplicities $\nu(S)$. The determinant of the estimate can be calculated as

$$\det\hat{K} = n^{|\Gamma|}\frac{\prod_{S\in\mathcal{S}} (\det w_S)^{\nu(S)}}{\prod_{C\in\mathcal{C}} \det w_C}. \tag{4.18}$$

4.4 The graphical lasso

The *graphical lasso* maximizes a penalized likelihood function

$$2\ell_{pen}(K)/n = \log\det K - \mathrm{tr}(KS) - \lambda||K||_1$$

where $S = W/n$. There are efficient methods for solving this problem, using lasso regression and techniques from convex optimization. The maximizing value $\hat{K}^\lambda$ will

typically have $\hat{k}^{\lambda}_{uv} = 0$ for several uv and will thus identify an independence graph. In that way, the graphical lasso is often used for graphical model selection.
It is important to realize that the graphical lasso is *not scale-invariant*. More precisely, if we let $Y = A^{-1}X$, where A is a diagonal matrix, then Y has covariance $\Sigma_Y = A^{-1}\Sigma_X A^{-1}$ and $S_Y = A^{-1}S_X A^{-1}$; similarly, the concentration matrix of Y is $K^Y = AK^X A$. Thus

$$\begin{aligned} 2\ell_{pen}(K^Y)/n &= 2\sum \log a_u + \log\det K^X \\ &\quad - \operatorname{tr}(AK^X AA^{-1}SA^{-1}) - \lambda||AK^X A||_1 \\ &= \text{const} + \log\det K^X - \operatorname{tr}(K^X S_X) - \lambda||AK^X A||_1 \\ &\sim 2\ell_{pen}(K^X)/n - \lambda(||AK^X A||_1 - ||K^X||_1). \end{aligned}$$

Data are therefore often first scaled by using the empirical correlation matrix R as input instead of the covariance matrix $S = W/n$, yielding a scale-invariant procedure.

4.4.1 *A constrained optimization problem*

Consider the convex optimization problem for $c > 0$:

$$\begin{array}{ll} \text{minimize} & -\log\det(K) + \operatorname{tr}(KS) \\ \text{subject to} & ||K||_1 \le c, \end{array}$$

The Lagrangian for this problem is

$$L(K, \lambda) = -\log\det(K) + \operatorname{tr}(KS) + \lambda(||K||_1 - c)$$

which, save for a constant and a sign, is the lasso-penalized likelihood.
The KKT conditions for an optimal pair (K^*, λ^*) are thus

$$\lambda^*(||K^*||_1 - c) = 0; \quad S - \Sigma^* + \lambda^*\Gamma^* = 0$$

where $\Sigma^* = (K^*)^{-1}$ and $\Gamma^* \in \operatorname{sign}(K^*)$. Thus the constrained problem is 'equivalent' to the penalized problem, through the correspondence $c \leftrightarrow \lambda^*$. Note in particular the subgradient equation for optimality:

$$S - \Sigma + \lambda^*\Gamma = 0$$

which is simply the subgradient equation for the lasso-penalized likelihood.

4.4.2 *Blocking the subgradient equation*

If we write the subgradient equation in block matrix form with the lower right corner being 1×1 we get

$$\begin{pmatrix} S_{11} & s_{12} \\ s_{12}^\top & s_{22} \end{pmatrix} - \begin{pmatrix} \Sigma_{11} & \sigma_{12} \\ \sigma_{12}^\top & \sigma_{22} \end{pmatrix} + \lambda \begin{pmatrix} \Gamma_{11} & \gamma_{12} \\ \gamma_{12}^\top & 1 \end{pmatrix} = 0.$$

Focusing on the upper right block of this equation we get

$$s_{12} - \sigma_{12} + \lambda\gamma_{12} = 0.$$

Using the identity $(\Sigma_{11})^{-1}\sigma_{12} = -k_{22}^{-1}k_{12} = \beta$ and thus $\text{sign}(k_{12}) = -\text{sign}(\beta)$ we can rewrite this equation as

$$\Sigma_{11}\beta - s_{12} + \lambda\,\text{sign}(\beta) = 0. \tag{4.19}$$

Now recall that the lasso regression problem is

$$\text{minimize} \qquad (y - Z\beta)^\top(y - Z\beta)/2n + \lambda||\beta||_1.$$

The subgradient equation for this problem becomes

$$\frac{1}{n}\left(Z^\top Z\beta - Z^\top y\right) + \lambda\,\text{sign}(\beta) = 0. \tag{4.20}$$

Comparing this to the subgradient equation (4.19) for the graphical lasso we see that they differ only by using $Z^\top Z/n$ instead of Σ_{11} since $Z^\top y/n = s_{12}$.
There is a simple iterative cyclic descent algorithm for solving the lasso equation (4.20), and the same algorithm can therefore be used to solve equation (4.19). Define the *soft threshold function* as

$$T(x,t) = \text{sign}(x)(|x| - t)^+.$$

The algorithm then becomes:

Algorithm 4.1 GRAPHICAL LASSO for maximizing the penalized Gaussian likelihood

Input: Empirical covariance matrix S; penalty parameter λ;
Output: Glasso estimate $\hat{K}^\lambda$; concentration graph $\hat{\mathcal{G}}^\lambda$.

1. **Initialize** $\Sigma \leftarrow S + \lambda I$; $\beta_{uv} \leftarrow 0, u, v \in V$.
2. **Repeat** for $v \in V$ **until** convergence
   (a) **For** $u \in V \setminus v$ **until convergence**:
   $\beta_{uv} \leftarrow T\left(s_{uv} - \sum_{w\neq v}\sigma_{uw}\beta_{wv};\ \lambda\right)/\sigma_{vv}$;
   (b) **For** $u \in V \setminus \{v\}$ **do** $\sigma_{uv} \leftarrow \sum_{w\neq v}\sigma_{uw}\beta_{wv}$;
3. **For** $v \in V$ **do**:
   (a) $k_{vv} \leftarrow 1/(\sigma_{vv} - \sum_{w\neq v}\sigma_{vw}\beta_{wv})$
   (b) **For** $u \in V \setminus v$ **do** $k_{uv} \leftarrow -\beta_{uv}k_{vv}$.
4. **Return** K and incidence graph of K.

An alternative algorithm for maximizing the penalized likelihood is just modifying IPS updates on 2×2 matrices. More precisely, we are simply iteratively maximizing the penalized likelihood over K_{cc} for $c = \{u, v\}$, keeping (K_{ca}, K_{aa}) fixed, where

Algorithm 4.2 Modified 2×2 IPS algorithm for computing $\hat{K}^\lambda$ with lasso penalty.

Input: Graph $\mathcal{G} = (V, E)$, sample covariance matrix S, penalty parameter λ.
Output: Glasso estimate $\hat{K}^\lambda$; concentration graph $\hat{\mathcal{G}}^\lambda$.

1. **Initialize** $S = S + \lambda I$, $K = \{\mathrm{diag}(S)\}^{-1}$, $\Sigma = \mathrm{diag}(S)$;
2. **Repeat** for each $c = \{u, v\}$ **until convergence**:

   Calculate: $\Delta = (S_{cc})^{-1} - (\Sigma_{cc})^{-1}$.
   And further

$$s^*_{uv} = \begin{cases} (\sqrt{1 + 4\delta^2_{uv} s_{uu} s_{vv}} - 1)/(2\delta_{uv}) \text{ if } \delta_{uv} \neq 0 \\ 0 \qquad\qquad\qquad\qquad \text{otherwise} \end{cases}$$

   and further

$$\tilde{s}_{uv} = \begin{cases} s_{uv} + \lambda \text{ if } s_{uv} + \lambda < s^*_{uv} \\ s_{uv} - \lambda \text{ if } s_{uv} - \lambda > s^*_{uv} \\ s^*_{uv} \qquad \text{otherwise.} \end{cases}$$

   and further

$$\tilde{S}_{cc} = \begin{pmatrix} s_{uu} & \tilde{s}_{uv} \\ \tilde{s}_{uv} & s_{vv} \end{pmatrix}, \quad \tilde{\Delta} = (\tilde{S}_{cc})^{-1} - (\Sigma_{cc})^{-1}$$

   Update $K_{cc} = K_{cc} + \tilde{\Delta}$; $H = \{\tilde{\Delta}^{-1} + (\Sigma_{cc})^{-1}\}^{-1}$; $\Sigma - \Sigma[H]^\Gamma \Sigma$
3. **Return** K, Σ, and incidence graph of K.

$a = V \setminus c$, and then cycling through all pairs c until convergence. This algorithm is again a special case of iterative partial maximization and it is described in more detail in Algorithm 4.2.
To see that this algorithm is convergent, we just need to verify that each update depends continuously on its input and use Proposition A.15, since the algorithm is a special case of Iterative Partial Maximization.

4.5 Exercises

Exercise 4.1 Show that any independence model generated by a regular Gaussian distribution is a compositional graphoid.

Exercise 4.2 A positive definite symmetric matrix K is an *M-matrix* (after Minkowski) if all off-diagonal elements are non-positive, i.e. $k_{\alpha\beta} \leq 0$ for all $\alpha \neq \beta$. Let $\Sigma = K^{-1}$.
Show that if K is a M-matrix, all off-diagonal elements of Σ are non-negative i.e. $\sigma_{\alpha\beta} \geq 0$ for all $\alpha \neq \beta$.

Exercise 4.3 Let X_1, X_2, X_3, X_4, X_5 be independent with $X_i \sim \mathcal{N}(0,1)$. Define recursively

$$Y_1 \leftarrow X_1,\ Y_2 \leftarrow X_2 + Y_1,\ Y_3 \leftarrow 2X_3 + Y_2,\ Y_4 \leftarrow X_4 + Y_3,\ Y_5 \leftarrow X_5 + 2Y_4.$$

(a) Find the covariance matrix Σ of Y;

(b) Find the concentration matrix $K = \Sigma^{-1}$ of Y.

(c) Construct the dependence graph of Y;

(d) Find the conditional distribution of Y_3 given $Y_1 = y_1, Y_2 = y_2, Y_4 = y_4, Y_5 = y_5$.

Exercise 4.4 Consider a Gaussian distribution $\mathcal{N}_5(0, \Sigma)$ with $K = \Sigma^{-1}$ satisfying the conditional independence restrictions of the graph $\mathcal{G} = (V, E)$ with $V = \{1, 2, 3, 4, 5\}$ and $E = \{\{1, 2\}, \{1, 3\}, \{1, 4\}, \{1, 5\}\}$.

(a) Show that the determinant of Σ satisfies

$$\det \Sigma = \prod_{i=1}^{5} \sigma_{ii} \prod_{j=2}^{5} (1 - \rho_{1j}^2)$$

where ρ_{ij} is the correlation between X_i and X_j;

(b) Express the covariance σ_{23} in terms of the variances $\sigma_{11}, \sigma_{22}, \sigma_{33}, \sigma_{44}, \sigma_{55}$ and the covariances $\sigma_{12}, \sigma_{13}, \sigma_{14}, \sigma_{15}$.

Exercise 4.5 Consider a Gaussian distribution $\mathcal{N}_4(0, \Sigma)$ with $K = \Sigma^{-1}$ satisfying the conditional independence restrictions of the graph $\mathcal{G} = (V, E)$ with $V = \{1, 2, 3, 4\}$ and $E = \{\{1, 2\}, \{2, 3\}\{1, 4\}\{3, 4\}\}$.

(a) Find two equations of degree 3 in σ_{13} and σ_{24} expressing these in terms of $\sigma_{11}, \sigma_{22}, \sigma_{33}, \sigma_{44}$ and the covariances $\sigma_{12}, \sigma_{13}, \sigma_{14}, \sigma_{15}$;

Hint: Express the appropriate inverse element of the covariance matrix as a cofactor;

(b) Consider the likelihood equations based on observing a Wishart matrix $W = w$ with $W \sim \mathcal{W}(n, \Sigma)$. Use the answer under (a) to establish an equation of degree 5 for the maximum likelihood estimate of σ_{13}.

(c) Assume next that $\sigma_{11} = \sigma_{22} = \sigma_{33} = \sigma_{44} = 1$ and $\sigma_{12} = \sigma_{23} = \sigma_{34} = \rho$ and $\sigma_{14} = -\rho$. Show that then $\rho^2 < 1/2$.

Exercise 4.6 Consider a Gaussian distribution $\mathcal{N}_4(0, \Sigma)$ with $K = \Sigma^{-1}$ satisfying the conditional independence restrictions of the graph $\mathcal{G} = (V, E)$ with $V = \{1, 2, 3, 4\}$ and $E = \{\{1, 2\}, \{2, 3\}, \{3, 4\}, \{1, 4\}\}$. Assume that the following Wishart matrix has been observed, with 10 degrees of freedom:

$$\begin{pmatrix} 5 & 1 & 4 & 4 \\ 1 & 10 & 2 & 5 \\ 4 & 2 & 10 & 2 \\ 4 & 5 & 2 & 8 \end{pmatrix}.$$

(a) Perform one full cycle of the IPS algorithm to find the MLE of the concentration matrix, starting with $K = I$.

(b) Assume next that $K = \Sigma^{-1}$ satisfies the conditional independence restrictions of the graph $\mathcal{G}^* = (V, E^*)$ with $V = \{1, 2, 3, 4\}$ and $E^* = \{\{1, 2\}, \{2, 3\}, \{3, 4\}\}$. Find the maximum likelihood estimate of the concentration matrix.

APPENDIX A

SOME MATHEMATICAL PREREQUISITES

A.1 Measurable spaces

In this section we recall some of the main definitions and results from measure theory that are used throughout the book. Consider a set $\mathcal{X}$ and let $\mathbb{E}$ be a collection of subsets of $\mathcal{X}$.

Definition A.1 We say that $\mathbb{E}$ is a σ–algebra on $\mathcal{X}$, if it holds that

- $\mathcal{X} \in \mathbb{E}$
- If $A \in \mathbb{E}$, then $A^c \in \mathbb{E}$
- If $A_1, A_2, \ldots \in \mathbb{E}$, then $\cup_{n=1}^{\infty} A_n \in \mathbb{E}$

If $\mathcal{X}$ is some set, and $\mathbb{E}$ is a σ–algebra on $\mathcal{X}$, then we say that the pair $(\mathcal{X}, \mathbb{E})$ is a measurable space. If $\mathbb{D}$ is a collection of subsets of $\mathcal{X}$, then we define $\sigma(\mathbb{D})$ to be the smallest σ–algebra on $\mathcal{X}$ that contains $\mathbb{D}$. For a σ–algebra $\mathbb{E}$ on $\mathcal{X}$ and a collection $\mathbb{H}$ of subsets of $\mathcal{X}$, we say that $\mathbb{H}$ is a generating system for $\mathbb{E}$, if $\mathbb{E} = \sigma(\mathbb{H})$.
If it for some collection $\mathbb{H}$ of subsets of $\mathcal{X}$ holds for all $A, B \in \mathbb{H}$ that $A \cap B \in \mathbb{H}$, then we say that $\mathbb{H}$ is stable under finite intersections.

Definition A.2 We say that $\mathbb{H}$ is a Dynkin class on $\mathcal{X}$, if it holds that

1) $\mathcal{X} \in \mathbb{H}$,
2) If $A, B \in \mathbb{H}$ with $A \subseteq B$, then $B \setminus A \in \mathbb{H}$
3) If $A_1, A_2, \ldots \in \mathbb{H}$ with $A_1 \subseteq A_2 \subseteq \ldots$, then $\cup_{n=1}^{\infty} A_n \in \mathbb{H}$

We have

Theorem A.3. (Dynkin's lemma) *Let $\mathbb{D} \subseteq \mathbb{H} \subseteq \mathbb{E}$ be collections of subsets of $\mathcal{X}$. Assume that $\mathbb{E} = \sigma(\mathbb{D})$ and that $\mathbb{D}$ is stable under finite intersections. If furthermore $\mathbb{H}$ is a Dynkin class, then $\mathbb{H} = \mathbb{E}$.*

Definition A.4 Let $(\mathcal{X}, \mathbb{E})$ be a measurable space. We say that a function $\mu : \mathbb{H} \to [0, \infty]$ is a measure (on $(\mathcal{X}, \mathbb{H})$), if

1) $\mu(\emptyset) = 0$
2) If $A_1, A_2, \ldots \in \mathbb{H}$ are pairwise disjoint sets, then $\mu(\cup_{n=1}^{\infty} A_n) = \sum_{n=1}^{\infty} \mu(A_n)$

We say that a measure μ on $(\mathcal{X}, \mathbb{E})$ is a probability measure, if $\mu(\mathcal{X}) = 1$. In the affirmative we call $(\mathcal{X}, \mathbb{E}, \mu)$ a probability space.

Theorem A.5. (Uniqueness theorem for probability measures) *Let μ and ν be two probability measures on $(\mathcal{X}, \mathbb{E})$. Let $\mathbb{H}$ be a generating system for $\mathbb{E}$ which is stable under finite intersection. If $\mu(A) = \nu(A)$ for all $A \in \mathbb{H}$, then $\mu(A) = \nu(A)$ for all $A \in \mathbb{E}$.*

Let $(\mathcal{X}, \mathbb{E})$ and $(\mathcal{Y}, \mathbb{K})$ be two measurable spaces. Then we can consider the product space $(\mathcal{X} \times \mathcal{Y}, \mathbb{E} \otimes \mathbb{K})$. Here the product σ–algebra $\mathbb{E} \otimes \mathbb{K}$ is generated by the system of all product sets

$$\mathbb{D} = \{A \times B \,:\, A \in \mathbb{E},\, B \in \mathbb{K}\}$$

Note that $\mathbb{D}$ is stable under intersections. If λ and $\tilde{\lambda}$ are two measures on $(\mathcal{X} \times \mathcal{Y}, \mathbb{E} \otimes \mathbb{K})$ that are equal on product sets

$$\lambda(A \times B) = \tilde{\lambda}(A \times B)$$

for all $A \in \mathbb{E}$ and $B \in \mathbb{K}$, then according to Theorem A.5 we have $\lambda = \tilde{\lambda}$.

Let μ be a measure on $(\mathcal{X}, \mathbb{E})$ and ν a measure on $(\mathcal{Y}, \mathbb{K})$. Then $\mu \otimes \nu$ denotes the uniquely determined measure defined by $(\mu \otimes \nu)(A \times B) = \mu(A)\nu(B)$.

Theorem A.6. (Tonelli's theorem) *Let μ be a probability measure on $(\mathcal{X}, \mathbb{E})$ and ν be probability measure on $(\mathcal{Y}, \mathbb{K})$, and assume that f is nonnegative and $\mathbb{E} \otimes \mathbb{K}$ measurable. Then*

$$\int f(x,y)\mathrm{d}(\mu \otimes \nu)(x,y) = \iint f(x,y)\mathrm{d}\nu(y)\mathrm{d}\mu(x).$$

Theorem A.7. (Fubini's theorem) *Let μ be a probability measure on $(\mathcal{X}, \mathbb{E})$ and ν be probability measure on $(\mathcal{Y}, \mathbb{K})$, and assume that f is $\mathbb{E} \otimes \mathbb{K}$ measurable and $\mu \otimes \nu$ integrable. Then $y \mapsto f(x,y)$ is integrable with respect to ν for μ-almost all x, the set where this is the case is measurable, and it holds that*

$$\int f(x,y)\mathrm{d}(\mu \otimes \nu)(x,y) = \iint f(x,y)\mathrm{d}\nu(y)\mathrm{d}\mu(x).$$

We will also need the following abstract change-of-variable theorem

Theorem A.8 *Let μ be a measure on $(\mathcal{X}, \mathbb{E})$ and let $(\mathcal{Y}, \mathbb{K})$ be some other measurable space. Let $t : \mathcal{X} \to \mathcal{Y}$ be measurable, and let $f : \mathcal{Y} \to \mathbb{R}$ be Borel measurable. Then f is $t(\mu)$-integrable if and only if $f \circ t$ is μ-integrable, and in the affirmative, it holds that $\int f \mathrm{d}t(\mu) = \int f \circ t \mathrm{d}\mu$.*

Assume that $(\Omega, \mathbb{F}, P)$ is a probability space and $(\mathcal{X}, \mathbb{E})$ is some measurable space. We say that $X : \Omega \to \mathcal{X}$ is a random variable on $(\Omega, \mathbb{F})$ with values in $(\mathcal{X}, \mathbb{E})$, if it is $\mathbb{F} - \mathbb{E}$–measurable. That is

$$X^{-1}(A) = (X \in A) \in \mathbb{F}$$

for all $A \in \mathbb{E}$. For a random variable X on $(\Omega, \mathbb{F})$ with values in $(\mathcal{X}, \mathbb{E})$ we define $\sigma(X)$ to be the smallest σ–algebra that makes X measurable. Then $\sigma(X)$ is the sub-σ–algebra of $\mathbb{F}$ given by

$$\sigma(X) = \{(X \in A) \,:\, A \in \mathbb{E}\}$$

We then have the following useful result

Theorem A.9 *Assume that X is a random variable with values in $(\mathcal{X}, \mathbb{E})$ and that Z is a real–valued random variable.Then Z is $\sigma(X)$–measurable if and only if there exists a measurable function $\phi : (\mathcal{X}, \mathbb{E}) \to (\mathbb{R}, \mathbb{B})$ such that $Z = \phi \circ X$.*

We further need the following concept. Let $(\Omega, \mathbb{F})$ be a measurable space.

Definition A.10 A subset $\mathcal{I} \subseteq \mathbb{F}$ is a *σ-ideal* in $\mathbb{F}$ if

i) $\emptyset \in \mathcal{I}$;
ii) $A_1, \ldots, A_n, \ldots \in \mathcal{I} \implies \bigcup_1^\infty A_i \in \mathcal{I}$;
iii) $A \in \mathcal{I}, F \in \mathbb{F} \implies A \cap F \in \mathcal{I}$.

Note that $\{\emptyset\}$ is always a σ-ideal, the *trivial σ-ideal*. We shall in particular be interested in the *σ-ideal $\mathcal{I}_P$ of P-null sets* where

$$\mathcal{I} = \mathcal{I}_P = \{F \in \mathbb{F} : P(F) = 0\}$$

which clearly is a σ-ideal. We shall typically write $\mathcal{I}$ for $\mathcal{I}_P$ when it is clear from the context.
The *null-extension* $\overline{\mathbb{A}}$ is the smallest σ-algebra generated by $\mathbb{A}$ and $\mathcal{I}$. We have the following Lemma.

Lemma A.11 *Assume that X is a random variable on $(\Omega, \mathbb{F}, P)$ and $\mathbb{A}$ and $\mathbb{B}$ are sub-σ-algebras of $\mathbb{F}$. If there is an $\mathbb{A}$-measurable random variable Y and a $\mathbb{B}$-measurable Z so that $X = Y = Z$ almost surely, then there is a random variable W so that $X = W$ almost surely and W is $\overline{\mathbb{A}} \cap \overline{\mathbb{B}}$-measurable.*

Proof We have $X = Y = Z$ except on the set $D = D_Y \cup D_Z$ where D_Y and D_Z are null-sets. Now define $W = X(1 - 1_D) = Y(1 - 1_D) = Z(1 - 1_D)$. Clearly, W is $\overline{\mathbb{A}} \cap \overline{\mathbb{B}}$-measurable and $W = X$ almost surely. □

A.2 Möbius inversion

An important combinatorial trick is contained in the following

Lemma A.12. (Möbius inversion) *Let Ψ and Φ be functions defined on the set of all subsets of a finite set V, taking values in an Abelian group. Then the following two statements are equivalent:*

(1) *for all $a \subseteq V$: $\Psi(a) = \sum_{b:b\subseteq a} \Phi(b)$;*
(2) *for all $a \subseteq V$: $\Phi(a) = \sum_{b:b\subseteq a} (-1)^{|a\setminus b|} \Psi(b)$.*

Proof We show (2) $\implies$ (1):

$$\begin{aligned}
\sum_{b:b\subseteq a} \Phi(b) &= \sum_{b:b\subseteq a} \sum_{c:c\subseteq b} (-1)^{|b\setminus c|} \Psi(c) \\
&= \sum_{c:c\subseteq a} \Psi(c) \left\{ \sum_{b:c\subseteq b\subseteq a} (-1)^{|b\setminus c|} \right\} \\
&= \sum_{c:c\subseteq a} \Psi(c) \left\{ \sum_{h:h\subseteq a\setminus c} (-1)^{|h|} \right\}.
\end{aligned}$$

The latter sum is equal to zero unless $a \setminus c = \emptyset$, i.e. if $c = a$, because any finite, non-empty set has the same number of subsets of even as of odd cardinality. The proof of (1) $\implies$ (2) is performed analogously. □

The Abelian group referred to in the lemma can be the real numbers, but often also just the additive group of a real vector space L, the vector space of linear maps on L or a vector space $\mathcal{S}$ of symmetric matrices, etc. More general versions of the lemma exist that relate to general lattices rather than the lattice of subsets of a set; see for example Aigner (1979).

A.3 Convexity and optimization

This section contains a brief summary of some important elements in the theory of convex optimization. We refer the reader to Boyd and Vandenberghe (2004) for further details, proofs not given here, and general algorithms for solving convex optimization problems.

A.3.1 *Convex sets and functions*

A subset $C \subseteq V$ of a finite-dimensional vector space V is said to be *convex* if it contains the line segment connecting any two points of C, i.e. if

$$c_1, c_2 \in C \implies tc_1 + (1-t)c_2 \in C \text{ for all } t \text{ with } 0 \leq t \leq 1. \tag{A.1}$$

Thus the empty set is convex and any singleton set is convex. *Convexity of sets is closed under intersection* so $\cap_{\alpha \in A} C_\alpha$ is convex if all C_α are convex.
A real-valued function $f : C \to \mathbb{R}$, defined on a convex set C is said to be *convex* if it for all $x_1, x_2 \in C$ satisfies the inequality

$$f(tx_1 + (1-t)x_2) \leq tf(x_1) + (1-t)f(x_2) \in C \text{ for all } t \text{ with } 0 < t < 1. \tag{A.2}$$

We say that f is if the inequality in (A.2) is strict unless $x_1 = x_2$. The real functions $f(x) = |x|$ and $f(x) = x^2$ are examples of strictly convex functions. For any convex function it holds that the level sets

$$C_a = \{x : f(x) \leq a\},\ a \in \mathbb{R}$$

are all convex sets. Note that the converse to this statement is false in general unless $L = \mathbb{R}$.
It can be practical to define f for all $x \in V$ by letting $f(x) = \infty$ for $x \notin C$; the inequality (A.2) is then satisfied for all $x_1, x_2 \in V$. If we do not explicitly say otherwise, we shall always consider f extended in this way. The *domain* of f is the set of points where f is finite.

$$\operatorname{dom} f = \{x \in V : f(x) < \infty\}.$$

A function f is *concave* if $-f$ is convex. So if f is concave, we have $\operatorname{dom} f = \{x \in V : f(x) > \infty\}$ and the level sets $\{x : f(x) \geq a$ are convex. If f is both concave and convex it is *affine*.

Suppose V is a Euclidean space with inner product $\langle \cdot, \cdot \rangle$ and that f is convex with $\operatorname{dom} f$ open and f is differentiable for all $x \in \operatorname{dom} f$. It then holds for all $x, y \in \operatorname{dom} f$ that

$$f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle \tag{A.3}$$

where the gradient $\nabla f(x)$ is determined as

$$\frac{\partial}{\partial t} f(x + tu)\bigg|_{t=0} = \langle \nabla f(x), u \rangle.$$

If f is not differentiable everywhere, we consider its *subgradient* $\partial f(x)$ as the set of vectors $v \in V$ such that

$$f(y) \geq f(x) + \langle v, y - x \rangle \text{ for all } y. \tag{A.4}$$

If f is differentiable at x we have $\partial f = \{\nabla f\}$. For the function $f(x) = |x|$ we have $\partial f(x) = \operatorname{sign}(x)$, where $\operatorname{sign}(x) = \{1\}$ if $x > 0$, $\operatorname{sign}(x) = \{-1\}$ if $x < 0$, and $\operatorname{sign}(0) = [-1, 1]$.
If f is convex, then (A.4) implies that x^* is a global minimum for f if and only if $0 \in \partial f(x^*)$. If f is strictly convex, this mimimum is unique. These facts are behind importance of convexity in optimization problems.
If f is twice differentiable, we consider the Hessian $Hf(x)$ of x:

$$\frac{\partial^2}{\partial s \partial t} f(x + su + tv)\bigg|_{s=t=0} = \langle u, Hf(x)v \rangle.$$

Then a twice differentiable real-valued function f with $\operatorname{dom} f = C$ is convex if and only if $\operatorname{dom} f$ is convex and $-Hf$ is positive semidefinite for all $x \in \operatorname{dom} f$; it is strictly convex if and only if $-Hf$ is positive definite.

A.3.2 *Convex optimization problems*

We shall consider a *convex optimization problem* in *standard form*:

$$\begin{array}{ll} \text{minimize} & f(x) \\ \text{subject to} & g_i(x) \leq 0,\ i = 1, \ldots, k, \\ & h_i(x) = 0,\ j = 1, \ldots, l, \end{array} \tag{A.5}$$

where f is a convex *objective function*, $g_i, i = 1, \ldots, k$ are convex *inequality constraint functions*, and $h_j, j = 1, \ldots, l$ are affine *equality constraint functions.* Either or both types of constraint function may be absent; in the latter case we have an *unconstrained* problem. Similarly, a *concave optimization problem* has standard form

$$\begin{array}{ll} \text{maximize} & f(x) \\ \text{subject to} & g_i(x) \geq 0,\ i = 1, \ldots, k, \\ & h_i(x) = 0,\ j = 1, \ldots, l, \end{array} \tag{A.6}$$

where f and g_i are concave and h_i affine. Clearly any concave problem can be modified into a convex problem by appropriate sign changes. The *domain* of the problems is in both cases the convex set $\mathcal{D}$ where

$$\mathcal{D} = \operatorname{dom} f \bigcap_{i=1}^{k} \operatorname{dom} g_i.$$

A point $x \in \mathcal{D}$ is *feasible* if it satisfies all the constraints. The set $\mathcal{F}$ of feasible points is convex. The problem is said to be feasible if $\mathcal{F}$ is non-empty. Thus, in a convex optimization problem we are minimizing the convex function f over the convex set $\mathcal{F}$. The *optimal value* of the problem is

$$f^* = \inf\{f(x) \mid x \in \mathcal{F}\}$$

which may be $-\infty$ or ∞, the latter if $\mathcal{F} = \emptyset$. A point $x^* \in \mathcal{D}$ is a *solution* to the problem or an *optimum* if x^* is feasible and if $f(x^*) = f^*$. If there is a feasible point $\tilde{x} \in \operatorname{int} \mathcal{F}$ with $0 \in \partial f(\tilde{x})$, then $\tilde{x}$ is a solution to the problem.

A.3.3 *Duality and optimality*

We shall associate a *dual problem* to any *primal* convex problem in standard form. The dual problem is sometimes better behaved that the *primal problem* and the magic of convex optimization is that in many cases the solution to the dual problem has the same optimum value as the primal. Finally, there is often a simple way of recovering the solution to the primal problem from the dual.
The fundamental function that exposes this duality is the *Lagrangian* associated with the primal problem (A.5) defined as

$$L(x, \lambda, \nu) = f(x) + \sum_{i=1}^{k} \lambda_i g_i(x) + \sum_{j=1}^{l} \nu_j h_j(x), \tag{A.7}$$

with $\operatorname{dom} L = \mathcal{D} \times (\mathbb{R})^k \times \mathbb{R}^l$. The Lagrangian incorporates information from the constraint functions into the objective function by adding linear combinations of these. The variables λ, ν are *Lagrange multipliers* and are also known as the *dual variables* of the convex problem.
The *Lagrange dual function* d is simply the minimum value of the Lagrangian over x:

$$d(\lambda, \nu) = \inf_{x \in \mathcal{D}} L(x, \lambda, \nu). \tag{A.8}$$

The *dual function is concave* and may in principle take the value $-\infty$. Also, if $\lambda_i \geq 0$ for all i we have for any feasible point $x \in \mathcal{F}$ that $L(x, \lambda, \nu) \leq f(x)$ so then

$$d(\lambda, \nu) = \inf_{x \in \mathcal{D}} L(x, \lambda, \mu) \leq \inf_{x \in \mathcal{F}} L(x, \lambda, \nu) \leq f^*$$

for all (λ, ν) with $\lambda_i \geq 0, i = 1, \ldots, k$; therefore we obtain a lower bound on the optimal value by maximizing the dual function in the following concave problem

$$\begin{array}{ll} \text{maximize} & d(\lambda, \nu) \\ \text{subject to} & \lambda_i \geq 0,\ i = 1, \ldots, k. \end{array} \tag{A.9}$$

This optimization problem is the *dual problem* to (A.5) which is then called the *primal problem*. The pair (λ, ν) is said to be *dual feasible* if $\lambda_i \geq 0$ for all i and $d(\lambda, \nu) > -\infty$. A pair (λ^*, ν^*) that is optimal for (A.9) is *dual optimal* and the the optimum value shall be denoted d^*. We thus always have

$$d^* \leq f^*.$$

For most, but not all, convex problems we also have $f^* = d^*$ and this phenomenon is referred to as *strong duality*. Strong duality is ensured by what is known as *Slater's condition* (Slater, 1950). A feasible point $x \in \mathcal{F}$ is said to be *strictly feasible* if for all non-affine $g_i, i = 1, \ldots, k$ it holds that $g_i(x) < 0$, i.e. the inequality constraints are strict.

Theorem A.13. (Slater) *If there is an $x \in \operatorname{ri} \mathcal{D}$ which is strictly feasible, then there is a dual optimum (λ^*, ν^*) with*

$$-\infty < d(\lambda^*, \nu^*) = d^* = f^*.$$

Proof See Section 5.3.2 in Boyd and Vandenberghe (2004). □

Next we need to know how we can recognize optima for the dual and primal problems. If x is primal feasible and (λ, ν) are dual feasible, then we always have

$$f(x) - f^* \leq f(x) - d(\lambda, \nu)$$

and the upper bound on the right hand side is known as the *duality gap* associated with x and (λ, ν). If this gap is equal to zero, then x is primal optimal and (λ, ν) is dual optimal. Such pairs of primal-dual optima are characterized by what is known as the *Karush–Kuhn–Tucker conditions* (KKT conditions). More precisely:

Theorem A.14. (Karush–Kuhn–Tucker) *Consider a convex optimization problem satisfying Slater's condition. Then (x^*, λ^*, ν^*) is a primal-dual optimal pair with zero duality gap if and only if x^* and (λ^*, ν^*) are feasible and satisfy:*

$$\lambda_i^* g_i(x^*) = 0 \quad \textit{for all } i = 1, \ldots, k, \tag{A.10}$$

$$0 \in \partial f(x^*) + \sum_{i=1}^{k} \lambda_i^* \partial g_i(x^*) + \sum_{j=1}^{l} \nu_j^* \partial h_j(x^*). \tag{A.11}$$

Proof Since $\lambda_i^* \geq 0$, the Lagrangian $L(x, \lambda^*, \nu^*)$ is convex in x and is therefore minimized at $x = x^*$ if and only if (A.11) holds. We then get

$$d(\lambda^*, \nu^*) = L(x^*, \lambda^*, \nu^*) = f(x^*) + \sum_{i=1}^{k} \lambda_i^* g_i(x^*) + \sum_{j=1}^{l} \nu_j^* h_j(x^*) = f(x^*)$$

where the last equality holds if and only if (A.10) holds since $h_j(x^*) = 0$, $\lambda_i^* \geq 0$, and $g_i(x^*) \leq 0$. Hence the KKT conditions are satisfied if and only if the duality gap $f(x^*) - d(\lambda^*, \nu^*)$ is equal to zero. □

The conditions (A.10) are referred to as *complementary slackness* as they express that at most one of the constraints $g_i(x) \leq 0$ and $\lambda_i > 0$ is active. The condition (A.11) is *stationarity of the Lagrangian*.

A.4 Iterative partial maximization

For computation of maximum likelihood estimates we shall rely on procedures involving iterative partial maximization in the sense that the likelihood function is maximized over different sections in the parameter space. This is then repeated cyclically.

We consider a continuous real-valued function L on a compact set Θ, and assume that the value $\hat{\theta}$ that maximizes L is uniquely determined.

We assume further that there for all $\theta^* \in \Theta$ are sections $\Theta_i(\theta^*), i = 1, \ldots, k$ in Θ in such a way that L is globally maximized at θ^* if and only if L is maximized over all of the sections.

Finally we assume that the operations of maximizing L over sections is continuous and well defined, i.e. there are continuous transformations T_i of Θ into itself such that if $\theta \in \Theta_i(\theta^*)$ for $i = 1, \ldots, k$,

$$L\{T_i(\theta^*)\} > L(\theta), \quad \text{if } \theta \neq T_i(\theta^*).$$

In other words, $T_i(\theta^*)$ is the uniquely determined point where L is maximized over the section $\Theta_i(\theta^*)$.

Now let θ_0 be arbitrary and define recursively

$$\theta_{n+1} = T_1 \cdots T_k(\theta_n), \quad n \geq 0.$$

Then we can show

Proposition A.15 *Under the assumptions given above the sequence (θ_n) converges to $\hat{\theta}$, the unique point where L attains its maximum.*

Proof Since Θ is assumed compact, the sequence (θ_n) has a convergent subsequence (θ_{n_k}) with limit θ^*, say. We need to show that $\hat{\theta} = \theta^*$. Let $S = T_1 \cdots T_k$. Since each T-operation is a partial maximization, $L(\theta_n)$ must be non-decreasing in n. Since also each operation is continuous, we have

$$L\{S(\theta^*)\} = \lim_{k\to\infty} L\{S(\theta_{n_k})\} \leq \lim_{k\to\infty} L(\theta_{n_{k+1}}) = L(\theta^*).$$

But using that T_i partially maximizes L gives

$$L\{S(\theta^*)\} \geq L\{T_2 \cdots T_k(\theta^*)\} \geq \cdots \geq L(\theta^*).$$

Thus there must everywhere be equality. Uniqueness of the partial maxima yields, when the chain of inequalities is read from right to left, that

$$\theta^* = T_k(\theta^*) = \cdots = T_1(\theta^*).$$

Finally, since the global maximum was uniquely determined by maximizing L over all sections, the proof is complete. □

The above result is the basis of a class of algorithms used to maximize likelihood functions. Sections are chosen appropriately such that the partial maximization problems are relatively simple. A starting value θ_0 is found and is iteratively changed by partial maximization over sections. In all cases the existence, uniqueness and necessary continuity properties will be established separately, but convergent algorithms necessarily appear.

APPENDIX B

SOME GRAPH THEORY

B.1 Notation and terminology

A *graph*, as we use it throughout this book, is a triple $\mathcal{G} = (V, E, \epsilon)$ consisting of a finite set V of *vertices* or *nodes*, a finite set E of *edges* and a map $\epsilon : E \to V \times V$ that with each edge associates two vertices as its *endpoints*. The endpoints may not necessarily be distinct in which case the edge is a *loop*. When nodes α and β are the endpoints of an edge, they are *adjacent* and we write $\alpha \sim \beta$ and we say the edge is *between* its two endpoints. Vertices and edges may have additional attributes. Vertices can for example be *discrete* or *continuous*. Edges can for example be *undirected*, *directed*, or *bidirected*, or have other types. We write $\alpha - \beta$, $\alpha \to \beta$, and $\alpha \leftrightarrow \beta$ to denote undirected, directed, and bidirected edges. Note that the endpoints of a directed edge is an ordered pair (α, β), whereas endpoints of undirected and bidirected edges are unordered pairs $\{\alpha, \beta\}$. In the latter case we shall often write the edge as $\alpha\beta$. A graph can be visually represented by a picture, where nodes are represented by circles, and edges by *lines*, *arrows*, and *arcs*, and we shall often refer to the edges by these names. We note that our graphs are *labelled* so that the two graphs in Fig. B.1 below are considered different.

FIG. B.1. Two different labelled graphs.

In most cases our graphs are *simple,* i.e. there are no multiple edges between endpoints and they have no loops. For simple graphs we can identify the edge set E with the set of its endpoints and represent the graph as the ordered pair $\mathcal{G} = (V, E)$.
If the graph has only undirected edges it is an *undirected* graph, if all edges are directed, the graph is said to be *directed*, and if all if all edges are bidirected, it is a *bidirected* graph.
A graph $\mathcal{G}'$ with vertex set V' and edge set E' is a *subgraph* of a graph $\mathcal{G}$ with vertices V and edges E if $V' \subseteq V$, $E' \subseteq E$, and every edge in E' has the same endpoints in $\mathcal{G}'$ as in $\mathcal{G}$. If $A \subseteq V$ is a subset of the vertex set, it *induces* a subgraph $\mathcal{G}_A = (A, E_A)$, where the edge set E_A consists of the edges in E with both endpoints in A. Similarly, a subset $F \subseteq E$ induces a subgraph $\mathcal{G}_F = (V_F, F)$, where V_F are the endpoints of edges in F.
A graph is *complete* if all vertices are adjacent. A subset is *complete* if it induces a complete subgraph. A complete subset that is maximal (with respect to inclusion $\subseteq$) is called a *clique*.

If there is an arrow from α pointing towards β, α is said to be a *parent* of β and β a *child* of α. The set of parents of β is denoted as $\mathrm{pa}(\beta)$ and the set of children of α as $\mathrm{ch}(\alpha)$. If there is a line between α and β, α and β are said to be *neighbours*, and if $\alpha \leftrightarrow \beta$ they are *spouses*. The neighbours of a vertex α is denoted as $\mathrm{ne}(\alpha)$ and the spouses of α are $\mathrm{sp}(\alpha)$. The expressions $\mathrm{pa}(A)$, $\mathrm{ch}(A)$, $\mathrm{ne}(A)$, and $\mathrm{sp}(A)$ denote the collection of parents, children, neighbours, and spouses of vertices in A that are not themselves elements of A:

$$\begin{aligned}\mathrm{pa}(A) &= \cup_{\alpha\in A}\,\mathrm{pa}(\alpha) \setminus A\\ \mathrm{ch}(A) &= \cup_{\alpha\in A}\,\mathrm{ch}(\alpha) \setminus A\\ \mathrm{ne}(A) &= \cup_{\alpha\in A}\,\mathrm{ne}(\alpha) \setminus A\\ \mathrm{sp}(A) &= \cup_{\alpha\in A}\,\mathrm{sp}(\alpha) \setminus A.\end{aligned}$$

The *boundary* $\mathrm{bd}(A)$ of a subset A of vertices is the set of vertices in $V\setminus A$ that are adjacent to vertices in A. In symbols we then have $\mathrm{bd}(A) = \mathrm{pa}(A) \cup \mathrm{ne}(A) \cup \mathrm{sp}(A)$. The *closure* of A is $\mathrm{cl}(A) = A \cup \mathrm{bd}(A)$. The *skeleton* $\mathrm{ske}(\mathcal{G})$ of a graph $\mathcal{G}$ is the undirected graph where $\alpha - \beta$ in $\mathrm{ske}(\mathcal{G})$ if and only if $\alpha \sim \beta$ in $\mathcal{G}$. See Fig. B.2 for illustration of various graphtheoretic concepts.

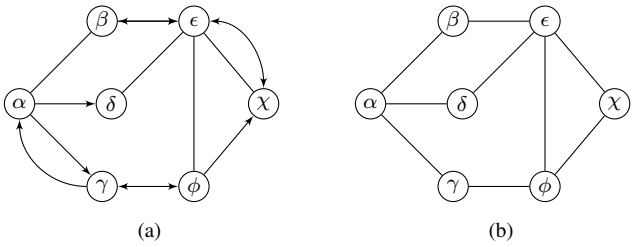


FIG. B.2. Illustration of graph theoretic concepts. In (a) we have $\alpha \to \gamma$, $\delta - \epsilon$, and $\chi \leftrightarrow \epsilon$ but $\alpha \not\to \beta, \delta \not\sim \chi$ whereas, for example, $\epsilon \sim \phi$. Also $\mathrm{pa}(\chi) = \{\phi\}$, $\mathrm{ch}(\gamma) = \{\alpha\}$, $\mathrm{sp}(\epsilon) = \{\beta, \chi\}$, $\mathrm{bd}(\epsilon) = \{\beta, \delta, \chi, \phi\}$, and $\mathrm{cl}(\{\beta, \epsilon\}) = \{\alpha, \beta, \delta, \epsilon, \chi, \phi\}$. The graph in (b) is the skeleton of the graph in (a).

A *walk* ω of length n *from* α to β or *between* α and β is a sequence $\omega = (\alpha = \alpha_0, e_1, \alpha_1, \ldots, e_n, \alpha_n = \beta)$ of vertices and edges such that for $1 \le i \le n$, the edge e_i has endpoints α_{i-1} and α_i. Note that the walk is uniquely determined by its sequence of edges so we may occasionally omit the vertices and write $\omega = (e_1, \ldots, e_n)$ or, for example, $\omega = (\alpha, e_1, \ldots, e_n, \beta)$ to emphasize the *endpoints* of the walk. If the graph is simple, a walk is also uniquely determined by its sequence of vertices and we shall then write $\omega = (\alpha_0, \alpha_1, \ldots, \alpha_n)$.
A *section* ρ of a walk ω is a maximal undirected subwalk. Thus, a walk has a unique decomposition into sections; sections may also be single nodes. A section ρ is a *collider* on a walk ω if two arrowheads meet on the walk at ρ, i.e. if either of the following situations occur $w = (\cdots, \to \rho \leftarrow, \cdots)$, $w = (\cdots, \leftrightarrow \rho \leftarrow, \cdots)$, $w = (\cdots, \to \rho \leftrightarrow, \cdots)$, or $w = (\cdots, \leftrightarrow \rho \leftrightarrow, \cdots)$.

A *path* is a walk with no repeated vertex, i.e. a path does not intersect itself. An n*-cycle* is a path of length n with the modification that it begins and ends in the same point, i.e. as $(\alpha, e_1, \ldots, e_n, \alpha)$. If $\pi_1 = (\alpha, e_1, \ldots, e_n, \beta)$ and $\pi_2 = (\beta, e_{n+1}, \ldots, e_{n+m}, \gamma)$ are paths, their *combination* $\omega_{12} = \pi_1 \circ \pi_2$ is the walk $\omega_{12} = (\alpha, e_1, \ldots, e_p, \delta, e_q, \ldots, e_{n+m}, \gamma)$, where δ is the first node of π_1 which is on both paths and an endpoint of both e_p and e_q. If $\delta = \beta$ then ω_{12} is simply the *concatenation* $(\alpha, e_1, \ldots, e_{n+m}, \gamma)$ of the two paths. In general, the concatenation of two paths will be a walk and not a path as the paths may intersect in more than one point.

Two vertices α and β are said to *connect* in $\mathcal{G}$ if there is a walk or, equivalently, a path from α to β in $\mathcal{G}$ in which case we write $\alpha \rightleftharpoons \beta$. Clearly, $\rightleftharpoons$ is an equivalence relation and the corresponding equivalence classes $[\alpha]$ where

$$\beta \in [\alpha] \iff \alpha \rightleftharpoons \beta$$

are the *connected components* of $\mathcal{G}$. If $\alpha \in A \subseteq V$, the symbol $[\alpha]_A$ denotes the connected component of α within $\mathcal{G}_A$. Note also that $\alpha \rightleftharpoons \beta$ in a graph $\mathcal{G}$ if and only if $\alpha \rightleftharpoons \beta$ in the skeleton $\mathrm{ske}(\mathcal{G})$.

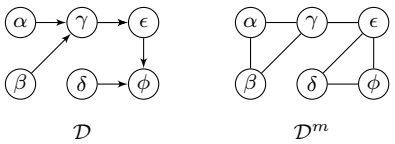
A walk $w = (\alpha = \alpha_0, e_1, \alpha_1, \ldots, e_n, \alpha_n = \beta)$ from α to β or path is *directed* if all edges e_i are directed from α_{i-1} to α_i. It is *semi-directed* if it has no arcs and all directed edges e_i point from α_{i-1} to α_i. If there is a directed walk or path from α to β we write $\alpha \mapsto \beta$. The vertices α such that $\alpha \mapsto \beta$ are the *ancestors* $\mathrm{an}(\beta)$ of β, and the *descendants* $\mathrm{de}(\alpha)$ of α are the vertices β such that $\alpha \mapsto \beta$. The *non-descendants* are $\mathrm{nd}(\alpha) = V \setminus (\mathrm{de}(\alpha) \cup \{\alpha\})$. If $\mathrm{pa}(\alpha) \subseteq A$ for all $\alpha \in A$ we say that A is an *ancestral* set. In a directed graph the set A is ancestral if and only if $\mathrm{an}(\alpha) \subseteq A$ for all $\alpha \in A$. The intersection of a collection of ancestral sets is again ancestral. Hence, for any subset A of vertices there is a smallest ancestral set containing A which is denoted by $\mathrm{An}(A)$. Note that in an undirected graph, ancestral sets are simply unions of connected components.

Certain types of graph are of special interest to us. A *tree* is a simple, connected, undirected graph without cycles. It has a unique path between any two vertices. A *forest* is an undirected graph where all connected components are trees. We have also interest in *directed acyclic graphs* which are simple, directed graphs without directed cycles. A *rooted tree* is the directed acyclic graph obtained from a tree by choosing a vertex as root and directing all edges away from this root.

For an directed acyclic graph $\mathcal{D}$ we define its *moral graph* $\mathcal{D}^m$ as the simple, undirected graph with the same vertex set but with α and β adjacent in $\mathcal{D}^m$ if and only if either $\alpha \sim \beta$ in $\mathcal{D}$ or if α and β have common child. This operation, known as *moralization* is illustrated in Fig. B.3.

If no edges have to be added to form the moral graph, the DAG is said to be *perfect*. Thus a DAG is perfect if and only if its moral graph and its skeleton coincide $\mathcal{D}^m = \mathrm{ske}(\mathcal{D})$. We warn the reader that the notion of a perfect graph in most graph theory literature refers to something quite different.

A vertex γ in a DAG $\mathcal{D}$ said to be *terminal* if none of the vertices in C have children. *A DAG has always at least one terminal chain component*.



$\mathcal{D}$ $\mathcal{D}^m$

FIG. B.3. A directed acyclic graph $\mathcal{D}$ and its moral graph.

B.2 Undirected graphs

This section is devoted to studying special issues associated with undirected graphs and we shall assume these to be simple, i.e. without loops or multiple edges.

B.2.1 *Separation and connectivity*

A subset $S \subseteq V$ is said to be an (α, β)-*separator* in an undirected graph $\mathcal{G}$ if all paths (or, equivalently, all walks) from α to β intersect S. Thus, in an undirected graph, C is an (α, β)-separator if and only if $[\alpha]_{V \setminus C} \neq [\beta]_{V \setminus C}$. The subset S is said to *separate* A from B in $\mathcal{G}$ if it is an (α, β)-separator for every $\alpha \in A, \beta \in B$; if this is the case we write $A \perp_{\mathcal{G}} B \mid S$.

B.2.2 *Decomposition*

In this subsection we study decompositions and decomposable graphs. Since the notion is fundamental, we state formally

Definition B.1 A triple (A, B, C) of disjoint subsets of the vertex set V of an undirected, graph $\mathcal{G}$ is said to form a *decomposition* of $\mathcal{G}$ if $V = A \cup B \cup C$, $A \perp_{\mathcal{G}} B \mid C$, and C is a complete subset of V.

When this is the case we say that (A, B, C) *decomposes* $\mathcal{G}$ into the *components* $\mathcal{G}_{A \cup C}$ and $\mathcal{G}_{B \cup C}$. Note that we allow any of the sets in (A, B, C) to be empty. If the sets A and B in (A, B, C) are both non-empty, we say that the decomposition is *proper.* A graph is said to be *prime* if no proper decomposition exists. Fig. B.4 shows an example of a prime graph and an example of a decomposition is shown in Fig. B.5. It holds that

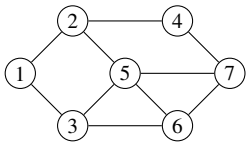


FIG. B.4. An example of a prime graph. This graph has no complete separators.

any finite undirected graph can be recursively decomposed into its uniquely defined prime components (Wagner, 1937; Tarjan, 1985; Diestel, 1987; Diestel, 1990), as illustrated in Fig. B.6.

A decomposable graph is one that can be successively decomposed into its cliques or, in other words, a graph with only cliques as its prime components. Again we choose to state this formally through a recursive definition as

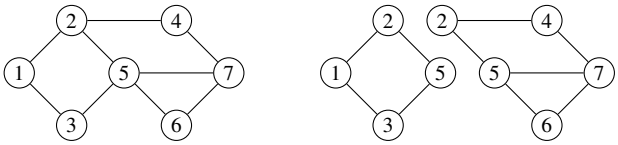


FIG. B.5. Decomposition with $A = \{1, 3\}$, $B = \{4, 6, 7\}$ and $C = \{2, 5\}$.

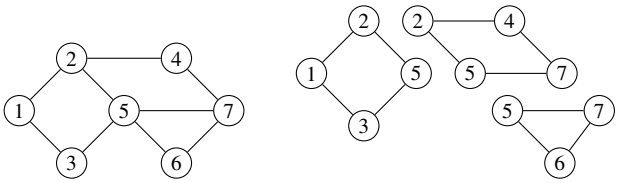


FIG. B.6. Decomposition of a graph into its unique prime components.

Definition B.2 An undirected graph is said to be *decomposable* if it is complete, or if there exists a proper decomposition (A, B, C) into decomposable subgraphs $\mathcal{G}_{A\cup C}$ and $\mathcal{G}_{B\cup C}$.

Note that the definition makes sense because the decomposition is assumed to be proper, such that both subgraphs $\mathcal{G}_{A\cup C}$ and $\mathcal{G}_{B\cup C}$ have fewer vertices than the original graph $\mathcal{G}$.

A *chordal* graph is an undirected graph with the property that every cycle of length $n \geq 4$ possesses a *chord,* i.e. two non-consecutive vertices that are neighbours. Chordal graphs are sometimes also known as *triangulated* graphs (Berge, 1973; Rose, 1970; Lauritzen, 1996) or *rigid circuit* graphs (Dirac, 1961). As this defines a chordal graph in terms of forbidden subgraphs, it follows immediately that the property must be stable under the operation of taking subgraphs, stated formally below.

Proposition B.3 *If $\mathcal{G}$ is chordal and $A \subset V$, then $\mathcal{G}_A$ is chordal.*

A classical result states that decomposable graphs are chordal and *vice versa*:

Proposition B.4 *The following conditions are equivalent for an undirected graph $\mathcal{G}$:*

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

(ii) *$\mathcal{G}$ is chordal;*

(iii) *every minimal (α, β)-separator is complete.*

Proof We show this partly by induction on the number of vertices $|V|$ of $\mathcal{G}$. The result is trivial for a graph with no more than three vertices since the three conditions then are automatically fulfilled. So assume the result holds for all graphs with $|V| \leq n$ and consider a graph $\mathcal{G}$ with $n + 1$ vertices. We then argue cyclically as
(i) $\Longrightarrow$ (ii) $\Longrightarrow$ (iii) $\Longrightarrow$ (i).
First we show (i) $\Longrightarrow$ (ii). Suppose that $\mathcal{G}$ is decomposable. If it is complete, it is obviously chordal. Otherwise it can be decomposed into decomposable subgraphs $\mathcal{G}_{A\cup C}$ and $\mathcal{G}_{B\cup C}$, both with fewer vertices. The inductive assumption implies that these are chordal. Thus the only possibility for a chordless cycle is one that intersects

both A and B. But, because C separates A from B, such a cycle must intersect C at least twice. But then it contains a chord because C is complete.
Then (ii) $\implies$ (iii). Let C be a minimal (α, β)-separator. If C has only one vertex, it is complete. If not it contains at least two, γ_1 and γ_2, say. Since C is a minimal separator, there will be paths from α to β via γ_1 and back via γ_2. The sequence

$$(\alpha, \ldots, \gamma_1, \ldots, \beta, \ldots, \gamma_2, \ldots, \alpha)$$

forms a cycle, with the modification that it can have repeated points. These, and chords other than a link between γ_1 and γ_2, can be used to shorten the cycle, still leaving at least one vertex in the component $[\alpha]_{V\setminus C}$ and one in $[\beta]_{V\setminus C}$. This produces a cycle of length at least 4, which must have a chord. Hence we get $\gamma_1 \sim \gamma_2$. Repeating the argument for all pairs of vertices in C gives that C is complete.
And finally that (iii) $\implies$ (i). Suppose that every minimal (α, β)-separator is complete. If $\mathcal{G}$ is complete there is nothing to show. Else it has at least two non-adjacent vertices α and β. Let C be a minimal (α, β)-separator and partition the vertex set into $[\alpha]_{V\setminus C}$, $[\beta]_{V\setminus C}$, C, and the set of remaining vertices D. Then, since C is complete, the triple (A, B, C), where $A = [\alpha]_{V\setminus C} \cup D$, and $B = [\beta]_{V\setminus C}$, form a decomposition of $\mathcal{G}$. But each of the subgraphs $\mathcal{G}_{A\cup C}$ and $\mathcal{G}_{B\cup C}$ must be decomposable. For if C_1 is a minimal (α_1, β_1)-separator in $\mathcal{G}_{A\cup C}$, it is contained in a minimal (α_1, β_1)-separator in $\mathcal{G}$ which is complete by assumption and C_1 is therefore itself complete. The inductive assumption implies then that $\mathcal{G}_{A\cup C}$ is decomposable, and similarly with $\mathcal{G}_{B\cup C}$. Thus we have decomposed $\mathcal{G}$ into decomposable subgraphs and the proof is complete. □

The smallest graph that is not decomposable is therefore a 4-cycle and shown in Fig. B.7.

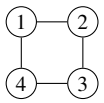


FIG. B.7. The smallest graph that is not decomposable.

B.2.3 *Simplicial subsets and perfect sequences*

Closely related to the notion of a decomposition is the notion of a *simplicial subset,* which is a subset B with complete boundary $\mathrm{bd}(B)$. Clearly, when a subset is simplicial the triple $(V \setminus \mathrm{cl}(B), B, \mathrm{bd}(B))$ is a decomposition of $\mathcal{G}$. A vertex α is said to be simplicial if the subset $\{\alpha\}$ is. The notion is illustrated in Fig. B.8. The following lemma, due to Dirac (1961), plays a central role.

Lemma B.5. (Dirac) *Let $\mathcal{G}$ be a chordal graph with at least two vertices. Then $\mathcal{G}$ has at least two simplicial vertices. If $\mathcal{G}$ is not complete these can be chosen to be non-adjacent.*

Proof Induction on $|V|$. If $|V| = 2$ the lemma is obviously true. Assume that the lemma holds for all graphs with $|V| \leq n$ and let $|V| = n + 1$. If $\mathcal{G}$ is complete the

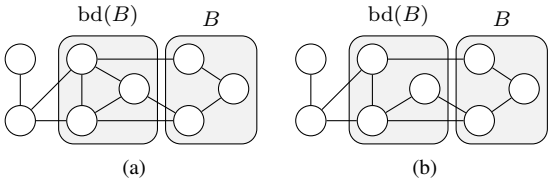


FIG. B.8. Simplicial subsets. In (a), B is simplicial. In (b), B is not simplicial because $\mathrm{bd}(B)$ is not complete.

statement is obvious. Otherwise there exists a proper decomposition (A, B, C) of $\mathcal{G}$. The induction assumption used on $\mathcal{G}_{A\cup C}$ yields a pair (α_1, α_2) of non-adjacent vertices that are simplicial in $\mathcal{G}_{A\cup C}$. At least one of these, α_1 say, must then be in A, because C is complete. By symmetry there is a vertex β in B that is simplicial in $\mathcal{G}_{B\cup C}$. Because C separates A from B, (α_1, β) must be a pair of non-adjacent vertices that are simplicial in $\mathcal{G}$. □

Let now $B_1, \cdots, B_k$ be a sequence of subsets of the vertex set V of an undirected graph $\mathcal{G}$. Let

$$H_j = B_1 \cup \cdots \cup B_j, \quad R_j = B_j \setminus H_{j-1}, \quad S_j = H_{j-1} \cap B_j.$$

The sequence is said to be *perfect* if the following conditions are fulfilled:

(i) for all $i > 1$ there is a $j < i$ such that $S_i \subseteq B_j$;
(ii) the sets S_i are complete for all i;

The condition (i) is known as the *running intersection property.* We term the sets H_j the *histories,* R_j the *residuals,* and S_j the *separators* of the sequence. The justification for the use of the term *separator* is based on Lemma B.6 below. A *perfect numbering* of the vertices V of $\mathcal{G}$ is a numbering $\alpha_1, \ldots, \alpha_k$ such that

$$B_j = \mathrm{cl}(\alpha_j) \cap \{\alpha_1, \ldots, \alpha_j\}, \quad j \geq 1$$

is a perfect sequence of sets. Note that this implies that the sets B_j are all complete. Perfect sequences and numberings play important roles in the understanding and manipulation of decomposable graphs, partly because, as we shall see in Proposition B.10, their existence is a characteristic for decomposable graphs, but also because they form the basis for recursive computational procedures. Before we show the characterization results, we need the following lemmas:

Lemma B.6 *Let $B_1, \ldots, B_k$ be a perfect sequence of sets which contains all cliques of an undirected graph $\mathcal{G}$. Then for every j, S_j separates $H_{j-1} \setminus S_j$ from R_j in $\mathcal{G}_{H_j}$ and hence (H_{j-1}, R_j, S_j) decomposes $\mathcal{G}_{H_j}$.*

Proof Let p be the highest number such that B_p is a clique. Then $H_p = V$ and hence $R_j = \emptyset$, so the separation is trivial for $j > p$. Next we must show that S_p separates $H_{p-1} \setminus S_p$ from R_p in $\mathcal{G}$. But suppose there were an edge between $\alpha \in R_p$ and $\beta \in B_j \setminus S_p$ for some $j < p$. Then $\{\alpha, \beta\}$ must be subset of some clique of $\mathcal{G}$. But this

cannot be B_p, as $\beta \notin B_p$ and not B_i for some $i < p$ as $\alpha \notin H_{p-1}$. Since all cliques are in the sequence, the edge can therefore not exist and S_p must separate $H_{p-1} \setminus S_p$ from R_p.
Now $B_1, \ldots, B_{p-1}$ is a perfect sequence of sets that contains all cliques of $\mathcal{G}_{H_{p-1}}$. For $S_p \subseteq B_i$ for some $i < p$ and hence, if $R_p = \emptyset$ then $B_i = B_p$ is a clique. If $R_p \neq \emptyset$ the subgraph $\mathcal{G}_{H_{p-1}}$ has one clique fewer. We repeat the argument and continue until the sequence is reduced to a single set. □

If a perfect sequence of sets does not contain all cliques, the sets S_j may not separate; see Fig. B.9.

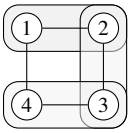


FIG. B.9. A perfect sequence of sets that does not decompose the graph.

Lemma B.7 *Let $C_1, \ldots, C_p$ be the cliques of $\mathcal{G}$ and assume that they form a perfect sequence. Next let the vertices of $\mathcal{G}$ be numbered with first those in C_1, then those in R_2, R_3 and so on. The numbering $\alpha_1, \ldots, \alpha_k$ so obtained is perfect.*

Proof This is immediate. □

The 'converse' to Lemma B.7 is false in the sense that the sequence of cliques induced by a perfect numbering of the vertices might not be perfect. The induced sequence is here formed by numbering the cliques according to their highest numbered vertex. A counterexample is provided in Fig. B.10.

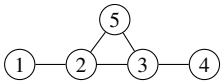


FIG. B.10. A perfect vertex numbering that does not induce a perfect clique numbering. The numbering of the vertices is perfect, but the cliques, numbered as $(\{1,2\},\{3,4\},\{2,3,5\})$, do not form a perfect sequence of sets.

Lemma B.8 *Let $C_1, \ldots, C_k$ be a perfect sequence. Assume that $C_t \subseteq C_p$ for some $t \neq p$ and that p is minimal with this property for fixed t. Then*

(i) *if $p < t$ then $C_1, \ldots, C_{t-1}, C_{t+1}, \ldots, C_k$ is a perfect sequence;*
(ii) *if $p > t$ then $C_1, \ldots, C_{t-1}, C_p, C_{t+1}, \ldots, C_{p-1}, C_{p+1}, \ldots, C_k$ is a perfect sequence.*

Proof Case (i) is immediate. In case (ii) we first argue that $S_p = C_t$. For we have

$$S_p = C_p \cap (C_1 \cup \cdots \cup C_{p-1}) \supseteq C_p \cap C_t = C_t$$

but also that $S_p \subseteq C_k$ for some $k < p$ from the running intersection property. Hence

$$C_t \subseteq S_p \subseteq C_k.$$

The minimality of p then implies $k = t$. Next

$$S^* = C_p \cap H_{t-1} \subseteq S_p = C_t,$$

whereby also $S^* = S_t$, as then

$$S^* = (C_p \cap H_{t-1}) \cap C_t = S_t.$$

Hence S^* is complete and contained in some C_k for $k < t$.
For $t < k < p$ we have

$$\begin{aligned} C_k \cap (H_{t-1} \cup C_p \cup C_{t+1} \cup \cdots \cup C_{k-1}) &= C_k \cap (H_{k-1} \cup C_p) \\ &= S_k \cup \{C_k \cap (C_p \setminus C_t)\}. \end{aligned}$$

But as $C_k \cap C_p \subseteq S_p = C_t$, then $\{C_k \cap (C_p \setminus C_t)\} = \emptyset$ and therefore

$$C_k \cap (H_{t-1} \cup C_p \cup C_{t+1} \cup \cdots \cup C_{k-1}) = S_k.$$

For $k > p$ the separators in the new sequence are trivially identical to those in the original sequence. □

Perfect sequences of vertices contain all cliques as stated below.

Lemma B.9 *Let $\alpha_1, \ldots, \alpha_k$ be a perfect numbering of the vertices of an undirected graph $\mathcal{G}$. Then the sets $B_j = \mathrm{cl}(\alpha_j) \cap \{\alpha_1, \ldots, \alpha_j\}$ form a perfect sequence that contains all cliques of $\mathcal{G}$.*

Proof The sequence $B_1, \ldots, B_k$ is perfect by definition. B_k is necessarily a clique. An induction argument now gives the result, as $\alpha_1, \ldots \alpha_{k-1}$ is a perfect numbering of the vertices of $\mathcal{G}_{V \setminus \{\alpha_k\}}$ and the cliques of $\mathcal{G}$ consist of B_k and those cliques of $\mathcal{G}_{V \setminus \{\alpha_k\}}$ that are not subsets of B_k. □

We now have a way of constructing a perfect sequence of cliques from a perfect sequence of vertices by thinning, i.e. simply by using Lemma B.8 to remove redundant sets from the sequence constructed in Lemma B.9. As a consequence we obtain a recursive characterization of decomposable graphs:

Proposition B.10 *For an undirected, graph $\mathcal{G}$, the following conditions are equivalent*

(i) *The graph $\mathcal{G}$ is decomposable.*
(ii) *For any $\alpha \in V$, the vertices of $\mathcal{G}$ admit a perfect numbering with $\alpha_1 = \alpha$;*
(iii) *The vertices of $\mathcal{G}$ admit a perfect numbering;*
(iv) *The cliques of $\mathcal{G}$ can be numbered to form a perfect sequence;*

Proof That (i) implies (ii) is seen by induction on the number of vertices as follows. If $\mathcal{G}$ is decomposable, then by Lemma B.5 $\mathcal{G}$ has a simplicial vertex other than α and

we can label this as α_k. The induction assumption gives us a perfect numbering of the remaining $k-1$ vertices with $\alpha_1 = \alpha$.
Clearly, (ii) implies (iii).
Also (iii) implies (iv) by using Lemma B.9 and the thinning procedure described in Lemma B.8. That (iv) implies (i) follows by Lemma B.6 and the definition of decomposability. □

A perfect numbering of the vertices of $\mathcal{G}$ induces a linear ordering of these and therefore a directed acyclic version $\mathcal{G}^<$ of $\mathcal{G}$ with arrows pointing from vertices with low numbers to vertices with high numbers. Since this graph is clearly perfect, $\mathcal{G}^<$ is called a *perfect directed version* of $\mathcal{G}$.
Proposition B.10 implies that *an undirected graph is chordal if and only if it has a perfect directed version.* It follows that *the skeleton* $\mathrm{ske}(\mathcal{D})$ *of a perfect directed acyclic graph* $\mathcal{D}$ *is chordal.*
The statement (ii) in Proposition B.10 can be strengthened and also perfect sequences of cliques can be arranged to begin anywhere. More precisely we have the useful lemma below.

Lemma B.11 *Let C^* be a clique in a chordal graph $\mathcal{G}$. Then the cliques of $\mathcal{G}$ can be ordered as a perfect sequence $C_1, \ldots, C_k$ with $C_1 = C^*$.*

Proof We use induction on the number of vertices $n = |V|$ of $\mathcal{G}$. For $n \leq 2$ the statement is obvious. Assume then the lemma to hold for all graphs with $n \leq p$ and let $\mathcal{G}$ have $p+1$ vertices. If $\mathcal{G}$ is complete the lemma is obviously true. Otherwise, by Dirac's Lemma B.5, $\mathcal{G}$ has at least two non-adjacent simplicial vertices, i.e. one of them, say α, is not in C^*. This vertex must be a member of exactly one clique C_α. The cliques of $\mathcal{G}'$ are the cliques of $\mathcal{G}$ except C_α, possibly with $C_\alpha \setminus \{\alpha\}$ adjoined. The inductive assumption implies that the cliques of $\mathcal{G}' = \mathcal{G}_{V\setminus\{\alpha\}}$ admit a perfect numbering $C_1, \ldots, C_{k-1}$ or $C_1, \ldots, C_{k-1}, C'_k$ with $C_1 = C^*$. Letting $C_k = C_\alpha$ we obtain a perfect numbering of the cliques of $\mathcal{G}$ with the desired property. □

B.3 Hypergraphs

B.3.1 *Basic concepts*

A *hypergraph* is a collection $\mathcal{H}$ of subsets of a finite set H, the *base set.* The elements of $\mathcal{H}$ are called *hyperedges.* In most cases of interest to us, the base set will be the union of the hyperedges, i.e. $H = \cup_{h\in\mathcal{H}} h$. This will henceforth be assumed to be the case, when not otherwise explicitly stated.
A typical hypergraph is a set of complete subsets of a graph $\mathcal{G}$, for example the set of cliques $\mathcal{C}(\mathcal{G})$ of the graph, denoted the *clique hypergraph* of $\mathcal{G}$. A hypergraph is *simple* if it has only one hyperedge. A simple hypergraph is the clique hypergraph of a complete graph.
If all hyperedges in $\mathcal{H}$ are pairwise incomparable in the sense that none is a subset of the other, we say that $\mathcal{H}$ is *reduced.* The examples above are reduced hypergraphs. The operation $\mathrm{red}\,\mathcal{H}$ produces a reduced hypergraph from $\mathcal{H}$ by removing all hyperedges that are contained in other hyperedges. If we define join and meet operations for two hypergraphs as

$$\mathcal{H}_1 \vee \mathcal{H}_2 = \mathrm{red}(\mathcal{H}_1 \cup \mathcal{H}_2)$$
$$\mathcal{H}_1 \wedge \mathcal{H}_2 = \mathrm{red}\{h_1 \cap h_2 \mid h_1 \in \mathcal{H}_1, h_2 \in \mathcal{H}_2\},$$

the class of reduced hypergraphs forms a distributive lattice with the partial order

$$\mathcal{H}_1 \preceq \mathcal{H}_2 \iff \text{for all } h_1 \in \mathcal{H}_1 \text{ there exists an } h_2 \in \mathcal{H}_2 \text{ with } h_1 \subseteq h_2.$$

Two arbitrary hypergraphs are equivalent if their reductions are equal:

$$(\mathcal{H}_1 \preceq \mathcal{H}_2 \text{ and } \mathcal{H}_1 \preceq \mathcal{H}_2) \iff \mathrm{red}(\mathcal{H}_1) = \mathrm{red}(\mathcal{H}_2).$$

The join $\mathcal{H} = \mathcal{H}_1 \vee \mathcal{H}_2$ of two hypergraphs is said to be *direct* if their meet is simple, i.e. if $\mathcal{H}_1 \wedge \mathcal{H}_2 = \{h\}$. Note that then necesssarily $h = H_1 \cap H_2$.

B.3.2 *Graphs and hypergraphs*

As mentioned above, each undirected graph $\mathcal{G}$ has an associated clique hypergraph $\mathcal{C}(\mathcal{G})$, but conversely with any hypergraph $\mathcal{H}$ we can associate its graph $\mathcal{G}(\mathcal{H}) = (V, E)$, where $V = H$ and

$$(\alpha, \beta) \in E \iff \{\alpha, \beta\} \subseteq h \text{ for some } h \in \mathcal{H}.$$

Clearly we have

$$\mathcal{G}\{\mathcal{C}(\mathcal{G})\} = \mathcal{G} \text{ and } \mathcal{H} \preceq \mathcal{C}\{\mathcal{G}(\mathcal{H})\}.$$

If it also holds that $\mathcal{H}$ contains the cliques of $\mathcal{G}(\mathcal{H})$,

$$\mathcal{C}\{\mathcal{G}(\mathcal{H})\} \preceq \mathcal{H},$$

we say that the hypergraph $\mathcal{H}$ is *conformal.* Then the reduced hypergraph $\mathrm{red}(\mathcal{H})$ consists exactly of the cliques of $\mathcal{G}(\mathcal{H})$. It obviously holds that

$$\mathcal{H}_1 \preceq \mathcal{H}_2 \implies \mathcal{G}(\mathcal{H}_1) \subseteq \mathcal{G}(\mathcal{H}_2)$$
$$\mathcal{G}_1 \subseteq \mathcal{G}_2 \implies \mathcal{C}(\mathcal{G}_1) \preceq \mathcal{C}(\mathcal{G}_2).$$

Further, one readily verifies from the definitions that

$$\mathcal{G}(\mathcal{H}_1 \vee \mathcal{H}_2) = \mathcal{G}(\mathcal{H}_1) \cup \mathcal{G}(\mathcal{H}_2) \tag{B.1}$$
$$\mathcal{G}(\mathcal{H}_1 \wedge \mathcal{H}_2) = \mathcal{G}(\mathcal{H}_1) \cap \mathcal{G}(\mathcal{H}_2) \tag{B.2}$$
$$\mathcal{C}(\mathcal{G}_1 \cap \mathcal{G}_2) = \mathcal{C}(\mathcal{G}_1) \wedge \mathcal{C}(\mathcal{G}_2), \tag{B.3}$$

whereas in general

$$\mathcal{C}(\mathcal{G}_1 \cup \mathcal{G}_2) \succeq \mathcal{C}(\mathcal{G}_1) \vee \mathcal{C}(\mathcal{G}_2). \tag{B.4}$$

In the case of direct joins and decompositions we have

Lemma B.12 *If $\mathcal{H}$ is the direct join of hypergraphs $\mathcal{H}_1$ and $\mathcal{H}_2$, then the triple $(H_1 \setminus H_2, H_2 \setminus H_1, H_1 \cap H_2)$ is a decomposition of $\mathcal{G}(\mathcal{H})$. If conversely (A, B, C) is a decomposition of the graph $\mathcal{G}$, then*

$$\mathcal{C}(\mathcal{G}) = \mathcal{C}(\mathcal{G}_{A\cup C} \cup \mathcal{G}_{B\cup C}) = \mathcal{C}(\mathcal{G}_{A\cup C}) \vee \mathcal{C}(\mathcal{G}_{B\cup C}) \tag{B.5}$$

and the join is direct.

Proof It follows from (B.1) that $H_1 \cap H_2$ separates $H_1 \setminus H_2$ from $H_2 \setminus H_1$. Since the join is direct, (B.2) gives that $H_1 \cap H_2$ is complete.
In the case where (A, B, C) forms a decomposition, (B.4) implies that it is enough to show that

$$\mathcal{C}(\mathcal{G}) \preceq \mathcal{C}\left(\mathcal{G}_{A\cup C}\right) \vee \mathcal{C}\left(\mathcal{G}_{B\cup C}\right).$$

But if $c \in \mathcal{C}(\mathcal{G})$, it must be contained in either $A \cup C$ or $B \cup C$ since C separates A from B in $\mathcal{G}$. Assume then $c \subseteq A \cup C$. Because c is a clique it is in $\mathcal{C}(\mathcal{G}_{A\cup C})$. □

An important corollary to this is

Corollary B.13 *If $\mathcal{H}$ is the direct join of conformal hypergraphs $\mathcal{H}_1$ and $\mathcal{H}_2$, then $\mathcal{H}$ is itself conformal.*

Proof We must show that $\mathcal{C}\left\{\mathcal{G}(\mathcal{H})\right\} \preceq \mathcal{H}$. We find

$$\begin{aligned}\mathcal{C}\left\{\mathcal{G}(\mathcal{H})\right\} &= \mathcal{C}\left\{\mathcal{G}(\mathcal{H}_1) \vee \mathcal{G}(\mathcal{H}_2)\right\} \\ &= \mathcal{C}\left\{\mathcal{G}(\mathcal{H}_1)\right\} \vee \mathcal{C}\left\{\mathcal{G}(\mathcal{H}_2)\right\} = \mathcal{H}_1 \vee \mathcal{H}_2 = \mathcal{H},\end{aligned}$$

where we have used (B.5) to obtain the second equality. □

A *decomposable* hypergraph $\mathcal{H}$ is a hypergraph that either is simple or can be obtained by direct joins of hypergraphs that have fewer hyperedges. We then have the following central result.

Theorem B.14 *A hypergraph $\mathcal{H}$ is decomposable if and only if it is the clique hypergraph of a decomposable graph. In particular, all decomposable hypergraphs are conformal.*

Proof Simple hypergraphs are obviously conformal with complete graphs as their graphs. Corollary B.13 ensures that this continues to hold when forming direct joins. From Lemma B.12 we have that direct joins of hypergraphs match decompositions of the associated graphs. Thus decomposable graphs must correspond to decomposable hypergraphs and vice versa. □

B.3.3 *Junction trees and forests*

An important structure associated with computational aspects of decomposable hypergraphs is a tree with a particular property. More precisely, a tree $\mathcal{T}$ with the set $\mathcal{H}$ of hyperedges as vertices of the tree is called a *junction tree* for $\mathcal{H}$ if it holds for any two hyperedges a and b in $\mathcal{H}$ and any h on the unique path in $\mathcal{T}$ between a and b that

$$a \cap b \subseteq h. \tag{B.6}$$

We refer to (B.6) as the *junction property.* It can alternatively be expressed as follows. The subset of hyperedges that contain a given subset $a \subseteq V$ forms a connected subtree $\mathcal{T}_a$ of $\mathcal{T}$ for all a.
A *junction forest* for $\mathcal{H}$ is a collection $\mathcal{F}$ of trees $\mathcal{T}_i$ that are junction trees for $\mathcal{H}_i$, with $\mathcal{H} = \vee_i \mathcal{H} i$ and

$$\mathcal{H}_i \wedge \mathcal{H}_j = \emptyset \quad \text{for } i \neq j.$$

Hence, hyperedges a and b that are in different trees of a junction forest are disjoint, and thus if $a \cap b \neq \emptyset$ there is a path in $\mathcal{F}$ between a and b.

Consider an arbitrary forest $\mathcal{F}$ with the hyperedges $\mathcal{H}$ as vertex set and two hyperedges h_1 and h_2 which are adjacent in $\mathcal{F}$. If the link between h_1 and h_2 is removed, then the tree containing these two hyperedges disconnects. Let $\mathcal{H}_1$ be the set of hyperedges that are still connected to h_1 and let $\mathcal{H}_2$ denote the set of remaining hyperedges in $\mathcal{H}$. The key to the relation between decomposability and junction trees and forests is the following lemma.

Lemma B.15 *If $\mathcal{F}$ is a junction forest for a reduced hypergraph $\mathcal{H}$ then for any neighbours h_1 and h_2 in $\mathcal{F}$, $\mathcal{H}$ is the direct join of the components $\mathcal{H}_1$ and $\mathcal{H}_2$.*

Proof Choose two neighbours h_1 and h_2 in $\mathcal{F}$ and define $\mathcal{H}_1$ and $\mathcal{H}_2$ as above. We recall that

$$\mathcal{H}_1 \wedge \mathcal{H}_2 = \mathrm{red}\{a \cap b \mid a \in \mathcal{H}_1, b \in \mathcal{H}_2\}.$$

Assume first that $\mathcal{F}$ is a junction forest for $\mathcal{H}$. For any $a \in \mathcal{H}_1$ and $b \in \mathcal{H}_2$ with $a \cap b \neq \emptyset$, both h_1 and h_2 are on the unique path between a and b in $\mathcal{F}$. Hence, by the junction property (B.6),

$$a \cap b \subseteq h_1 \cap h_2$$

and hence

$$\mathcal{H}_1 \wedge \mathcal{H}_2 = \{h_1 \cap h_2\},$$

whereby $\mathcal{H}$ is the direct join of $\mathcal{H}_1$ and $\mathcal{H}_2$. □

Proposition B.16 *A reduced hypergraph $\mathcal{H}$ is decomposable if and only if there is a junction forest $\mathcal{F}$ for $\mathcal{H}$.*

Proof The proof is by induction on the number of hyperedges in $\mathcal{H}$. The statement is trivial for a simple hypergraph. Assume then that the statement holds for any hypergraph with at most n hyperedges and let $\mathcal{H}$ have $n+1$ hyperedges.
First let $\mathcal{H}$ be decomposable. Then it is the direct join of reduced hypergraphs $\mathcal{H}_1$ and $\mathcal{H}_2$ where both of these have fewer hyperedges. As the join is direct we can choose $h_1 \in \mathcal{H}_1$ and $h_2 \in \mathcal{H}_2$ such that

$$\mathcal{H}_1 \wedge \mathcal{H}_2 = \{h_1 \cap h_2\}.$$

The inductive assumption gives two junction forests $\mathcal{F}_1$ and $\mathcal{F}_2$ for $\mathcal{H}_1$ and $\mathcal{H}_2$. Form now $\mathcal{F}$ from $\mathcal{F}_1$ and $\mathcal{F}_2$ by taking their union and adding an edge between h_1 and h_2 if $h_1 \cap h_2 \neq \emptyset$. We must show that $\mathcal{F}$ is a junction forest for $\mathcal{H}$. So let $a, b \in \mathcal{H}$. If both are in $\mathcal{H}_1$ or both in $\mathcal{H}_2$ and h is on the path between a and b, (B.6) follows from the fact that $\mathcal{F}_1$ and $\mathcal{F}_2$ were junction forests. Else we might assume that $a \in \mathcal{H}_1$ and $b \in \mathcal{H}_2$. Since $\mathcal{H}$ is the direct join of $\mathcal{H}_1$ and $\mathcal{H}_2$ we have

$$a \cap b \subseteq h_1 \cap h_2.$$

If $a \cap b \neq \emptyset$ we have also that $a \cap h_1 \neq \emptyset$ and there is therefore a path from a to h_1 in $\mathcal{F}_1$ and similarly a path from b to h_2 in forest $\mathcal{F}_2$, hence a path from a to b in $\mathcal{F}$. If h is

on the path between a and b it is either on the path from a to h_1 or from b to h_2. In the former case we find

$$a \cap b \subseteq a \cap h_1 \cap h_2 \subseteq a \cap h_1 \subseteq h,$$

where the junction property of $\mathcal{F}_1$ has been used to give the last inclusion. If h is on the path from b to h_2 we argue analogously. Hence the junction property for $\mathcal{F}$ is established.
Assume conversely that $\mathcal{F}$ is a junction forest for $\mathcal{H}$. By Lemma B.15, $\mathcal{H}$ is the direct join of hypergraphs $\mathcal{H}_1$ and $\mathcal{H}_2$ that both have fewer hyperedges. Clearly, the induced subgraphs $\mathcal{F}_1 = \mathcal{F}_{\mathcal{H}_1}$ and $\mathcal{F}_2 = \mathcal{F}_{\mathcal{H}_2}$ are junction forests. Hence $\mathcal{H}_1$ and $\mathcal{H}_2$ are decomposable by the inductive assumption. As $\mathcal{H}$ is the direct join of decomposable hypergraphs it is itself decomposable. □

Let now $\mathcal{F}$ be a junction forest for the clique hypergraph $\mathcal{C}$ of a decomposable graph $\mathcal{G}$ and let $\mathcal{S}$ denote the set of intersections of pairs of neighbours in $\mathcal{F}$

$$\mathcal{S} = \{C_i \cap C_j : C_i \sim C_j\}.$$

Further, let F_i, F_j be the base sets of the components $\mathcal{C}_i$ and $\mathcal{C}_j$ as in Lemma B.15. We then have

Corollary B.17 *Every set $S_{ij} = C_i \cap C_j$ separates $F_i \setminus S_{ij}$ from $F_j \setminus S_{ij}$ in $\mathcal{G}$ and thus $(F_i \setminus S_{ij}, F_j \setminus S_{ij}, S_{ij})$ forms a decomposition of $\mathcal{G}$.*

Proof Lemma B.15 yields that $\mathcal{C}$ is the direct join of $\mathcal{C}_i$ and $\mathcal{C}_j$; the statement now follows from Lemma B.12. □

In fact, although there in general are many possible junction forests for $\mathcal{C}$, the set $\mathcal{S} = \mathcal{S}_{\mathcal{F}}$ of *separators* in the junction tree does not depend on the particular junction forest chosen. Also, if we let $\nu_{\mathcal{F}}(S) = |\{ij \in E(\mathcal{F}) : S = C_i \cap C_j\}|$ denote the number of times S occurs in $\mathcal{S}_{\mathcal{F}}$ we have:

Proposition B.18 *If $\mathcal{F}$ and $\mathcal{F}'$ are two junction forests for $\mathcal{C}$ then $\mathcal{S}_{\mathcal{F}} = \mathcal{S}_{\mathcal{F}'}$ and $\nu_{\mathcal{F}}(S) = \nu_{\mathcal{F}'}(S)$ for all $S \subseteq V$.*

Proof The proof is induction after the cardinality of $\mathcal{C}$. For two cliques this is obviously true. Let L be a leaf in $\mathcal{F}$ with associated separator $S \in \mathcal{S}_{\mathcal{F}}$. Then $L \cap C \subseteq S$ for all $C \in \mathcal{C}_- = \mathcal{C} \setminus \{F\}$ and hence L is also a leaf in $\mathcal{F}'$ and S a separator in $S_{\mathcal{F}'}$. Using the inductive assumption on $\mathcal{C}_-$ with associated junction forests $\mathcal{F}_-$ and $\mathcal{F}'_-$ yields the result. □

Thus it makes sense to say that $\mathcal{S}$ is the set of *separators* of $\mathcal{C}$ or of $\mathcal{G}(\mathcal{C})$ and $\nu_{\mathcal{F}}(S) = \nu(S)$ are the *multiplicities* of S.

Corollary B.19 *For each S, the multiplicity $\nu(S)$ is equal to the number of times S occurs in any perfect sequence.*

Proof This again follows by induction as any leaf of a junction tree can be the last clique in some perfect sequence. □

The multiplicities $\nu(S)$ satisfy a number of combinatorial identities. More precisely, we have

Proposition B.20 *Let $\mathcal{C}$ be a decomposable hypergraph with base set V, $\mathcal{S}$ the associated separators, and $\nu(S)$ their multiplicities. We then have*

$$\sum_{C\in\mathcal{C}} |C| = |V| + \sum_{S\in\mathcal{S}} |S|\nu(S), \quad |\mathcal{C}| = |\mathcal{F}| + \sum_{S\in\mathcal{S}} \nu(S), \tag{B.7}$$

where $|\mathcal{F}|$ denotes the number of trees in any junction forest for $\mathcal{C}$.

Proof We use induction on the number $|\mathcal{C}|$ of hyperedges. For $|\mathcal{C}| = 2$ this is obviously true. Let L be a leaf in $\mathcal{F}$ with associated separator $S_L \in \mathcal{S}$. Let $\mathcal{C}^* = \mathcal{C} \setminus \{L\}$, with base set V^* and let $\mathcal{S}^*$, and ν^* denote the separators and multiplicities for $\mathcal{C}^*$. Using now the inductive assumption we have

$$\sum_{C\in\mathcal{C}} |C| = |L| + \sum_{C\in\mathcal{C}^*} |C| = |L| + |V^*| + \sum_{S\in\mathcal{S}^*} |S|\nu^*(S).$$

Now, $|L| + |V^*| = |V| + |S_L|$ and $\mathcal{S} = \mathcal{S}^* \cup \{S_L\}$. Further, $\nu(S) = \nu^*(S)$ for $S \neq S_L$; if $S_L \neq \emptyset$ we have $\nu(S_L) = \nu^*(S_L) + 1$ and $\nu(S_L) = \nu^*(S_L)$ otherwise; the first relation follows. The second relation follows similarly by noting that $|\mathcal{F}^*| = |\mathcal{F}|$ if $S_L \neq \emptyset$ and have $|\mathcal{F}| = |\mathcal{F}^*| + 1$ otherwise. □

Finally we wish to emphasize the fundamental relation between junction forests and sequences of sets satisfying the running intersection property.
Consider a sequence $B_1, \ldots, B_k$ of finite and distinct sets that satisfies the running intersection property, i.e. for all $i > 1$ there is a $j < i$ such that $S_i \subseteq B_j$ where

$$S_i = B_i \cap (B_1 \cup \cdots \cup B_{i-1}),$$

and define the hypergraph $\mathcal{H} = \{B_1, \ldots, B_k\}$. Construct then an undirected graph $\mathcal{F}$ with these sets as vertices by successively choosing j for each i such that $S_i \subseteq B_j$ and then letting $i \sim j$.

Proposition B.21 *The graph $\mathcal{F}$ is a junction forest for $\mathcal{H}$.*

Proof We use induction on the number k of sets in the sequence. The statement is trivial for $k \leq 2$. Assume the statement to hold for sequences of length at most n and assume $k = n + 1$.
Using the construction until B_{k-1} gives a graph $\mathcal{F}_{k-1}$ which is a junction forest by the inductive assumption. Adding the edge from B_k to B^*, where $S_k \subseteq B^*$, produces clearly a forest but the junction property must be checked. It is enough to consider $a \in \mathcal{F}_{k-1}$, $b = B_k$ and h on the path between them. Obviously, then h is also on the path from a to B^*. Using that $a \subseteq B_1 \cup \cdots \cup B_{k-1}$ and the junction property of $\mathcal{F}_{k-1}$ we obtain

$$a \cap B_k \subseteq a \cap (B_1 \cup \cdots \cup B_{k-1}) \cap B_k \subseteq a \cap B^* \subseteq h.$$

Hence $\mathcal{F}$ is a junction forest. □

Conversely, let $\mathcal{F}$ be a junction forest and choose roots arbitrarily in all trees, thereby directing all edges in $\mathcal{F}$. This induces a natural partial order on the vertices of $\mathcal{F}$ by having a before b if there is a directed path from a to b. Assume now that $b_1, \ldots, b_k$ is any numbering of the vertices of $\mathcal{F}$ that is compatible with this ordering.

Proposition B.22 *The sets $b_1, \ldots, b_k$ have running intersection property.*

Proof Consider b_i for $i > 1$ and assume this to be part of the tree $\mathcal{T}$ in $\mathcal{F}$ with chosen root R. All sets on the path from R to b_i must be among $b_1, \ldots b_{i-1}$ or the numbering would not be compatible. Let b^* be the hyperedge on this path which is nearest to b_i. Suppose b_l for $1 < l < i-1$ is in a different tree. Then $b_i \subseteq b_l = \emptyset$. Else b^* is on the path between b_l and b_i. The junction property thus implies $b_l \cap b_j \subseteq b^*$ and therefore

$$b_i \cap (b_1 \cup \cdots \cup b_{i-1}) \subseteq b^*,$$

which shows that we can choose $b_j = b^*$ and have the running intersection property. □

In this way there is a simple relation between all possible perfect orderings of the cliques of a decomposable graph and all junction forests for such graphs.

B.4 Algorithms

B.4.1 *Identifying chordal graphs*

There are several algorithms for identifying chordal graphs. The most straight-forward algorithm is a greedy algorithm for checking chordality based on the fact that chordal graphs are those that admit perfect numberings:

Algorithm B.1 GREEDY ALGORITHM for checking chordality of a graph and identifying a perfect numbering

Input: An undirected graph $\mathcal{G}$.
Output: If V is chordal: a perfect numbering of V; FALSE if V is not chordal.

1. Look for a vertex v^* with $\mathrm{bd}(v^*)$ complete
2. If no such vertex exists return FALSE
3. Number v^* as $v^* = |V|$ and let $\mathcal{G} = \mathcal{G}_{V \setminus v^*}$
4. If $V \neq \emptyset$ go to 1
5. Else return numbering.

The worst-case complexity of this algorithm is $O(|V|^2)$ as $|V| - k$ vertices must be queried to find the vertex to be numbered as $|VZ| - k$. The algorithm is illustrated in Fig. B.11 and Fig. B.12.
The next simple algorithm is due to Tarjan and Yannakakis (1984) and has complexity $O(|V| + |E|)$ as for every node, all neighbours of that node must be visited. It checks chordality of the graph and generates a perfect numbering if the graph is chordal. In addition, as we shall see below, the cliques of the chordal graph can be identified as the algorithm runs. It begins by initiating the first vertex in a perfect sequence, rather than the last vertex. The algorithm is illustrated in Fig. B.14 and Fig. B.13.

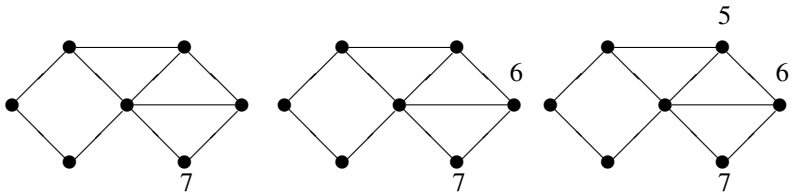


FIG. B.11. The greedy algorithm at work. This graph is *not* chordal, as there is no candidate for number 4.

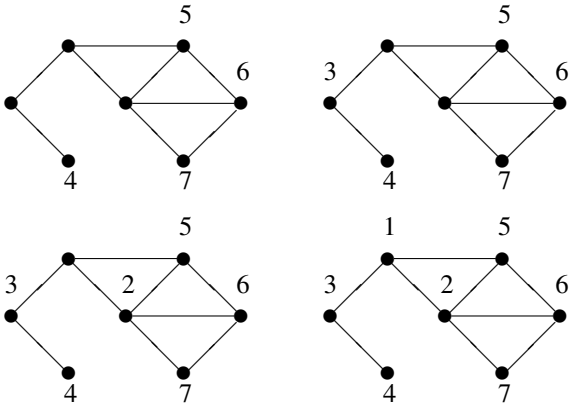


FIG. B.12. The greedy algorithm at work. Initially the algorithm proceeds as in Fig. B.11. This graph is chordal and the numbering obtained is a perfect numbering.

B.4.2 *Finding cliques and constructing a junction tree*

Finding the cliques of a general graph is an NP-complete problem. But the cliques of a chordal graph can be found in a simple fashion from a MCS numbering $V = \{1, \ldots, |V|\}$. More precisely we let

---

Algorithm B.2 MAXIMUM CARDINALITY SEARCH for checking chordality of a graph and identifying a perfect numbering

---

Input: An undirected graph $\mathcal{G}$.

Output: If V is chordal: a perfect numbering of V; FALSE if V is not chordal.

1. Choose $v_0 \in V$ arbitrarily and number v_0 as $v_0 = 1$;
2. When vertices $\{1, 2, \ldots, j\}$ have been identified, choose $v = j + 1$ among $V \setminus \{1, 2, \ldots, j\}$ with highest cardinality of its numbered neighbours
3. If $\mathrm{bd}(j + 1) \cap \{1, 2, \ldots, j\}$ is not complete, return FALSE
4. If $|V| = j + 1$ the graph is chordal and the numbering is perfect
5. Else repeat from 2.

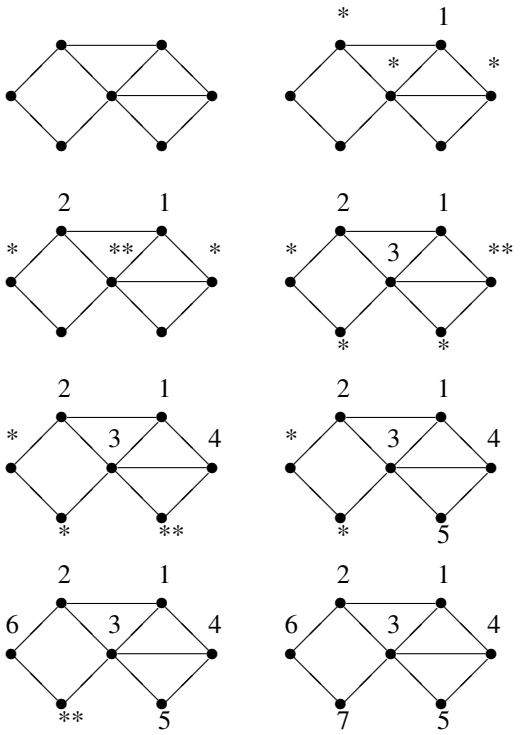


FIG. B.13. Maximum Cardinality Search at work. When a vertex is numbered, a counter for each of its unnumbered neighbours is increased with one, marked here with the symbol $*$. The counters keep track of the numbered neighbours of any vertex and are used to identify the next vertex to be numbered. This graph is *not* chordal as discovered at the last step because 7 does not have a complete boundary.

$$B_\lambda = \mathrm{bd}(\lambda) \cap \{1, \ldots, \lambda - 1\}$$

and $\pi_\lambda = |B_\lambda|$. Say that λ is a *ladder vertex* if $\lambda = |V|$ or if $\pi_{\lambda+1} < \pi_\lambda + 1$ and let Λ be the set of ladder vertices.
It then holds that the cliques of $\mathcal{G}$ are $C_\lambda = \{\lambda\} \cup B_\lambda, \lambda \in \Lambda$. For a proof of this assertion see e.g. Cowell *et al.* (1999, page 56).

Example B.25 For the MCS ordering in Fig. B.14 we find $\pi_\lambda = (0, 1, 2, 2, 2, 1, 1)$ yielding the ladder nodes $\{3, 4, 5, 6, 7\}$ and the corresponding cliques

$$\mathcal{C} = \{\{1, 2, 3\}, \{1, 3, 4\}, \{3, 4, 5\}, \{2, 6\}, \{6, 7\}\}.$$

A junction tree can be constructed directly from the MCS ordering $C_\lambda, \lambda \in \Lambda$. More precisely, since

$$B_\lambda = \mathrm{bd}(\lambda) \cap \{1, \ldots, \lambda - 1\}$$

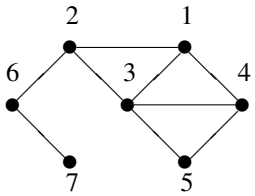


FIG. B.14. MCS numbering for a chordal graph. The algorithm runs essentially as in the non-chordal case.

is complete for all $\lambda \in \Lambda$ it holds that

$$C_\lambda \cap (\cup_{\lambda' < \lambda} C_{\lambda'}) = C_\lambda \cap C_{\lambda^*} = S_\lambda$$

for some $\lambda^* < \lambda$. A junction tree is now easily constructed by attaching C_λ to any C_{λ^*} satisfying the above. Although λ^* may not be uniquely determined, S_λ is. Indeed, the sets S_λ are the minimal complete separators and *the numbers* $\nu(S)$ *are* $\nu(S) = |\{\lambda \in \Lambda : S_\lambda = S\}|$. Junction trees can be constructed in many other ways as well (Jensen and Jensen, 1994).

B.4.3 *Junction trees of prime components*

In general, the *prime components* of any undirected graph can be identified and arranged in a junction tree in a similar way using an algorithm of Tarjan (1985), see also Leimer (1993).
Then *every pair of neighbours* (C, D) *in the junction tree represents a decomposition of* $\mathcal{G}$ into $\mathcal{G}_{\tilde{C}}$ and $\mathcal{G}_{\tilde{D}}$, where $\tilde{C}$ is the set of vertices in cliques connected to C but separated from D in the junction tree, and similarly with $\tilde{D}$.
Tarjan's algorithm is based on first numbering the vertices by a slightly more sophisticated algorithm (Rose *et al.*, 1976) known as *Lexicographic Search* (LEX) which runs in $O(|V|^2)$ time.

APPENDIX C

LINEAR ALGEBRA AND RANDOM VECTORS

C.1 Matrix results

For a block matrix of the form

$$E = \begin{pmatrix} A & B \\ C & D \end{pmatrix}$$

where A and D are $p \times p$ and $q \times q$ matrices respectively and D is regular, we define the *Schur complement of D within E*, denoted (E/D) to be the matrix

$$(E/D) = A - BD^{-1}C.$$

It then holds that

Lemma C.1 *A symmetric block matrix E is positive definite if and only if (E/D) and D are both positive definite.*

Proof Since Σ is symmetric, we have $C = B^\top$. If we then let

$$u = \begin{pmatrix} x \\ y - D^{-1}B^\top x \end{pmatrix}$$

we have a one-to-one correspondence between u and (x, y). Further we find

$$\begin{aligned} u^\top E u &= x^\top A x + 2x^\top B(y - D^{-1}B^\top x) \\ &\quad + (y - D^{-1}B^\top x)^\top D(y - D^{-1}B^\top x) \\ &= x^\top (A - BD^{-1}B^\top)x + y^\top D y = x^\top (E/D)x + y^\top D y. \end{aligned}$$

Hence the result follows. □

Similarly we have that if D is non-singular, the determinant of a partitioned matrix can be factorized as

$$\det E = \det \begin{pmatrix} A & B \\ C & D \end{pmatrix} = \det(E/D) \det D. \tag{C.1}$$

The correctness of (C.1) follows from the calculation

$$\begin{aligned} \det \begin{pmatrix} A & B \\ C & D \end{pmatrix} &= \det \begin{pmatrix} A & B \\ C & D \end{pmatrix} \det \begin{pmatrix} I_p & 0 \\ -D^{-1}C & I_q \end{pmatrix} = \\ &= \det \left\{ \begin{pmatrix} A & B \\ C & D \end{pmatrix} \begin{pmatrix} I_p & 0 \\ -D^{-1}C & I_q \end{pmatrix} \right\} \\ &= \det \begin{pmatrix} (E/D) & B \\ 0 & D \end{pmatrix} = (\det E/D)(\det D). \end{aligned}$$

Further, the inverse of a partitioned matrix is given as

$$E^{-1} = \begin{pmatrix} A & B \\ C & D \end{pmatrix}^{-1} = \begin{pmatrix} (E/D)^{-1} & -(E/D)^{-1}G \\ -F(E/D)^{-1} & D^{-1} + F(E/D)^{-1}G \end{pmatrix}. \tag{C.2}$$

The inverse on the left-hand side exists if and only if the inverses on the right-hand side exist. Here $F = D^{-1}C$, and $G = BD^{-1}$. Performing the multiplication shows the correctness. Finally, we shall make use of the following formula — known as Harville's variant of the *Woodbury matrix identity* — valid for a non-singular symmetric matrix M of order m, a symmetric matrix Δ of order d, and C being any $m \times d$ matrix:

$$(M + C\Delta C^\top)^{-1} = M^{-1} - M^{-1}C\Delta(I + C^\top M^{-1}C\Delta)^{-1}C^\top M^{-1}. \tag{C.3}$$

This formula is useful when M has previously been inverted and d is much smaller than m. To see that (C.3) is correct, we let $Q = (I + C^\top M^{-1}C\Delta)$ and multiply the right-hand side with $(M + C\Delta C^\top)$ from the left to get

$$\begin{aligned} &(M + C\Delta C^\top)(M^{-1} - M^{-1}C\Delta Q^{-1}C^\top M^{-1}) = \\ &\quad I + C\Delta C^\top M^{-1} - C\Delta Q^{-1}C^\top M^{-1} - C\Delta C^\top M^{-1}C\Delta Q^{-1}C^\top M^{-1} \\ &\quad I + C\Delta C^\top M^{-1} - C\Delta(I + C^\top M^{-1}C\Delta)Q^{-1}C^\top M^{-1} = I. \end{aligned}$$

C.2 Random vectors

It seems convenient for the purposes in this book to deal with multivariate distributions in some generality, allowing these to be discussed in general Euclidean vector spaces V rather than in the particular case of $V = \mathbb{R}^n$ with the standard inner product. In particular, when discussing the exact and asymptotic distribution of maximum likelihood estimates, it is most natural to work with random variables that, for example, take values in the vector space of symmetric matrices.

We do not intend to dwell on formal details of the theory. Therefore it will create few difficulties and indeed be very close to the more usual matrix formulation. The reader is referred to Eaton (1983) for a comprehensive and detailed exposition along similar lines.

First we have to discuss the notion of mean and covariance of a random vector X taking values in V, where V is a Euclidean space with inner product $\langle \cdot, \cdot \rangle$.

Definition C.2 An element ξ of V is said to be the *mean vector* or *expectation* of X if it holds that

$$\langle v, \xi \rangle = \mathbf{E}\langle v, X \rangle \quad \text{for all } v \in V.$$

We allow ourselves to write $\xi = \mathbf{E}X$ and have therefore that $\langle v, \mathbf{E}X \rangle = \mathbf{E}\langle v, X \rangle$.

Definition C.3 A bilinear form Σ on V is said to be the *covariance* of X if it holds that

$$\mathrm{Cov}\left(\langle u, X \rangle, \langle v, X \rangle\right) = \Sigma(u, v) \quad \text{for all } u, v \in V.$$

We write $\mathbf{V}X = \Sigma$. Note that the covariance, as well as any other bilinear form, is determined from its values on the diagonal $\Sigma(u, u)$, for the bilinearity gives

$$\Sigma(u, v) = \frac{1}{4}\left\{\Sigma(u + v, u + v) - \Sigma(u - v, u - v)\right\}.$$

To any such bilinear form there is a linear operator, which we also denote by Σ, such that

$$\Sigma(u, v) = \langle u, \Sigma v\rangle.$$

This is referred to as the *covariance operator* of X. If $(e_1, \ldots, e_p)$ is an orthonormal basis of V, we let

$$\sigma_{ij} = \Sigma(e_i, e_j) = \langle e_i, \Sigma e_j\rangle = \mathrm{Cov}\left(\langle e_i, X\rangle, \langle e_j, X\rangle\right).$$

The $p \times p$-matrix of these numbers is the *covariance matrix* of X and we also denote this by $\mathbf{V}X = \Sigma$. So the same symbol is used to refer to the covariance, the covariance operator and the covariance matrix, and the context will determine the exact meaning of the symbol.
The covariance Σ is called *regular* if $\Sigma(u, u) > 0$ for all $u \neq 0$. In this case its matrix is positive definite and the covariance determines an inner product on V which we shall denote as $\langle \cdot, \cdot \rangle_\Sigma$, i.e.

$$\langle u, v\rangle_\Sigma = \Sigma(u, v) = \langle u, \Sigma v\rangle.$$

When the covariance is regular, the inverse operator $K = \Sigma^{-1}$ is called the *concentration operator* and its matrix with respect to a chosen basis is called the *concentration matrix*. The concentration operator determines a symmetric bilinear form as usual by

$$K(u, v) = \langle u, Kv\rangle.$$

This bilinear form is called the *concentration* of the distribution.
The concentration operator K is equivalently defined through the relation

$$\langle u, v\rangle = \langle Ku, \Sigma v\rangle. \tag{C.4}$$

If Σ is not regular, any K which satisfies $\Sigma = \Sigma K \Sigma$, i.e. K is a generalized inverse to Σ, can be used as the concentration operator, and the relation (C.4) then holds for all u, v in the range of Σ, since

$$\langle K\Sigma x, \Sigma\Sigma y\rangle = \langle \Sigma K \Sigma x, \Sigma y\rangle = \langle \Sigma x, \Sigma y\rangle.$$

Note that the concentration and the covariance operator depend on the given inner product on V, and the covariance and concentration matrices further depend on a chosen orthonormal basis. A fully invariant approach to random vectors and the normal distribution on vector spaces avoids introducing the first inner product, but we have chosen not to proceed to this level of abstraction.

In most cases the space V will be $\mathbb{R}^n$ with the usual inner product and standard orthonormal basis, but we also frequently deal with the space $\mathbb{R}^{n\times p}$ of $n \times p$-matrices with inner product

$$\langle A, B\rangle = \operatorname{tr}(A^\top B) \tag{C.5}$$

and canonical basis formed by the matrices E_{ij} with ij-th entry equal to one and the remaining entries equal to zero. In the case where $n = p$, an interesting subspace is formed by the set $\mathcal{S}_p$ of symmetric $p \times p$ matrices where the transpose in (C.5) becomes unnecessary. An orthonormal basis for this space consists of the symmetric matrices

$$\tilde{E}_{ii} = E_{ii}, \quad \tilde{E}_{ij} = (E_{ij} + E_{ji})/\sqrt{2} \quad \text{for } i \neq j. \tag{C.6}$$

If $V = \mathbb{R}^n$, the mean vector is of the form $\xi = (\xi_1, \ldots, \xi_n)^\top$ and we have

$$\langle e_i, \xi\rangle = \mathbf{E}\langle e_i, X\rangle = \xi_i = \mathbf{E}X_i,$$

where $X = (X_1, \ldots, X_n)^\top$. Similarly for the covariance we get

$$\sigma_{ij} = \Sigma(e_i, e_j) = \operatorname{Cov}(\langle e_i, X\rangle, \langle e_j, X\rangle) = \operatorname{Cov}(X_i, X_j).$$

These formulae indicate how the notation conforms with that used in most statistical literature.

Suppose we have two Euclidean spaces V and W where to avoid confusion we denote their inner products by $\langle\cdot,\cdot\rangle_V$ and $\langle\cdot,\cdot\rangle_W$ respectively. Let A be a linear map from V to W and b an element of W. Then $Y = AX + b$ is a random vector in W. Its mean and covariance are given below.

Proposition C.4 *If the random vector X has mean ξ and covariance Σ, then the mean and covariance operator of Y are*

$$\mathbf{E}Y = A\xi + b, \quad \mathbf{V}Y = A\Sigma A^\top.$$

In the special case of $V = \mathbb{R}^n$ and $W = \mathbb{R}^m$, the same expressions hold for matrices.

Proof Direct calculation gives

$$\begin{aligned}\mathbf{E}\langle w, Y\rangle_W &= \mathbf{E}\langle w, AX + b\rangle_W = \mathbf{E}\langle A^\top w, X\rangle_V + \langle w, b\rangle_W \\ &= \langle A^\top w, \xi\rangle_V + \langle w, b\rangle_W = \langle w, A\xi + b\rangle_W,\end{aligned}$$

which gives the result for the mean, and similarly,

$$\begin{aligned}\operatorname{Cov}(\langle w, Y\rangle_W, \langle y, Y\rangle_W) &= \operatorname{Cov}(\langle w, AX\rangle_W, \langle y, AX\rangle_W) \\ &= \operatorname{Cov}\left(\langle A^\top w, X\rangle_V, \langle A^\top y, X\rangle_V\right) \\ &= \langle A^\top w, \Sigma A^\top y\rangle_V = \langle w, A\Sigma A^\top y\rangle_W,\end{aligned}$$

which gives the covariance operator. We abstain from repeating the calculations in the matrix case. □

Finally we mention that the distribution of a random vector is uniquely determined by the distribution of all linear functions of the vector. More precisely, the following holds.

Proposition C.5 *If X and Y are two random vectors in V and*

$$\langle v, X\rangle \stackrel{\mathcal{D}}{=} \langle v, Y\rangle \quad \textit{for all } v \textit{ in } V,$$

then $X \stackrel{\mathcal{D}}{=} Y$.

This is essentially equivalent to the fact that the characteristic function of X,

$$\psi(v) = \mathbf{E}e^{i\langle v, X\rangle},$$

determines the distribution. Here i is the complex unit, i.e. $i^2 = -1$. This result can be found in Cramér (1946). That this is equivalent to the statement in Proposition C.5 follows from the uniqueness of the Fourier transform in the case where $V = \mathbb{R}$.

APPENDIX D

THE MULTIVARIATE NORMAL DISTRIBUTION

The exposition of the multivariate normal distribution and derived distributions is close to that given in Eaton (1983). Proofs not given here can be found there or in Anderson (1984).

D.1 Basic properties

We first formally define what it means for a random vector in V to be normally distributed:

Definition D.1 A random vector X on a Euclidean space V is said to have a *normal distribution on* V if there exists an element $\xi \in V$ and a bilinear form Σ on V such that

$$\langle v, X\rangle \sim \mathcal{N}\{\langle v, \xi\rangle, \Sigma(v, v)\} \quad \text{for all } v \text{ in } V,$$

where $\mathcal{N}(\mu, \sigma^2)$ denotes the univariate normal distribution with mean μ and variance σ^2.

From Proposition C.5 it follows that the definition is unambiguous and the preceding pages show that then ξ is the mean and Σ the covariance of the random vector X. If X is normally distributed on V we write

$$X \sim \mathcal{N}_V(\xi, \Sigma).$$

In the special cases $V = \mathbb{R}^p$ and $V = \mathbb{R}^{n\times p}$ we write

$$X \sim \mathcal{N}_p(\xi, \Sigma) \quad \text{and} \quad X \sim \mathcal{N}_{n\times p}(\xi, \Sigma).$$

The mean ξ and covariance Σ determine a unique normal distribution on V.
Conversely, to any pair (ξ, Σ), where ξ is a vector in V and Σ is a bilinear form on V which is non-negative, i.e. $\Sigma(v, v) \geq 0$ for all $v \in V$, there is a normal distribution with these as mean and covariance.
If Σ is regular, the normal distribution has density with respect to the Lebesgue measure on V that gives mass 1 to a unit cube. This density is equal to

$$f_{\xi,\Sigma}(x) = (2\pi)^{-p/2}(\det \Sigma)^{-1/2} e^{-\langle x-\xi, K(x-\xi)\rangle/2}, \tag{D.1}$$

where $K = \Sigma^{-1}$ is the concentration operator of the normal distribution. If $\langle \cdot, \cdot \rangle$ is standard inner product on $\mathbb{R}^n$, we have in matrix notation that

$$\langle x - \xi, K(x - \xi)\rangle = (x - \xi)^\top K(x - \xi),$$

where K is the inverse of the covariance matrix. Note that we have then assumed an orthonormal basis in V to be chosen and K depends on this choice.

Adding two independent normal random vectors gives a normal random vector. More accurately:

Proposition D.2 *If $X_1 \sim \mathcal{N}_V(\xi_1, \Sigma_1)$ and $X_2 \sim \mathcal{N}_V(\xi_2, \Sigma_2)$ are independent, then*

$$X_1 + X_2 \sim \mathcal{N}_V(\xi_1 + \xi_2, \Sigma_1 + \Sigma_2).$$

Proof For $v \in V$ it holds that

$$\langle v, X_1 + X_2 \rangle = \langle v, X_1 \rangle + \langle v, X_2 \rangle.$$

The terms on the right-hand side are independent and univariate normal. Hence the sum is univariate normal. Definition D.1 implies that $X_1 + X_2$ is a normal random vector and the expressions for mean and covariance follow by direct calculation. □

Another important fact about the normal distribution is that an affine transformation of a normal random vector is itself a normal random vector. We consider a situation analogous to that in Proposition C.4.

Proposition D.3 *If A is a linear map from V to W, b an element of W, and $X \sim \mathcal{N}_V(\xi, \Sigma)$, then*

$$Y = AX + b \sim \mathcal{N}_W(A\xi + b, A\Sigma A^\top).$$

Proof The mean and covariance of Y have been given in Proposition C.4. What remains to be established is that Y follows a normal distribution. But for all $w \in W$ we have

$$\langle w, Y \rangle_W = \langle w, AX + b \rangle_W = \langle A^\top w, X \rangle_V + \langle w, b \rangle_W.$$

Since X has a normal distribution, $\langle A^\top w, X \rangle_V$ is univariate normally distributed. This is not changed by adding the constant $\langle w, b \rangle_W$. Definition D.1 then establishes the result. □

A special case of this result is of interest. Suppose $V = \mathbb{R}^n$ and assume the random vector X partitioned into components X_1 and X_2, where $X_1 \in \mathbb{R}^p$ and $X_2 \in \mathbb{R}^q$ with $p + q = n$. The mean vector and covariance matrix can then be partitioned accordingly into blocks as

$$\xi = \begin{pmatrix} \xi_1 \\ \xi_2 \end{pmatrix} \quad \text{and} \quad \Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}$$

such that Σ_{11} has dimensions $p \times p$ and so on. If A in Proposition D.3 is the linear map that sends X into X_2, we obtain:

Proposition D.4 *Let X be distributed as $\mathcal{N}_n(\xi, \Sigma)$, where X, ξ and Σ are partitioned as above. Then the marginal distribution of X_2 is $\mathcal{N}_q(\xi_2, \Sigma_{22})$.*

Proposition D.5 *Let X be distributed as $\mathcal{N}_n(\xi, \Sigma)$, where X, ξ and Σ are partitioned as above. Then X_1 and X_2 are independent if and only if $\Sigma_{12} = 0$. If Σ is regular, this holds if and only if $K_{12} = 0$.*

Proof The first statement follows directly from the expression for the conditional mean $\xi_{1|2}$ in Example 1.20. The second statement then follows from (1.6). □

REFERENCES

Aigner, M. (1979). *Combinatorial Theory*. Springer-Verlag, New York.

Anderson, T. W. (1984). *An Introduction to Multivariate Statistical Analysis* (2nd edn). John Wiley and Sons, New York.

Averintsev, M. B. (1970). Об одном способе описание случайных полей с дискретным аргументом (On one method of describing random fields with a discrete argument). Проблемы Передачи Информации *(Problems of Information Transmission)*, **6**, 100–108. In Russian.

Bahl, L., Cocke, J., Jelinek, F., and Raviv, J. (1974). Optimal decoding of linear codes for minimizing symbol error rate. *IEEE Transactions on Information Theory*, **20**, 284–287.

Baum, L. E. (1972). An equality and associated maximization technique in statistical estimation for probabilistic functions of Markov processes. *Inequalities*, **3**, 1–8.

Berge, C. (1973). *Graphs and Hypergraphs*. North-Holland, Amsterdam. Translated from French by E. Minieka.

Besag, J. (1972). Nearest-neighbour systems and the auto-logistic model for binary data. *Journal of the Royal Statistical Society, Series B*, **34**, 75–83.

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

Boyd, S. and Vandenberghe, L. (2004). *Convex Optimization*. Cambridge University Press, Cambridge, UK.

Buhl, S. L. (1993). On the existence of maximum likelihood estimators for graphical Gaussian models. *Scandinavian Journal of Statistics*, **20**, 263–270.

Cannings, C., Thompson, E. A., and Skolnick, M. H. (1976). Recursive derivation of likelihoods on pedigrees of arbitrary complexity. *Advances in Applied Probability*, **8**, 622–625.

Clifford, P. (1990). Markov random fields in statistics. In *Disorder in Physical Systems.* (ed. G. R. Grimmett and D. J. A. Welsh), pp. 19–32. Clarendon Press, Oxford. A volume in honour of John M. Hammersley on the occasion of his 70th birthday.

Cowell, R. G., Dawid, A. P., Lauritzen, S. L., and Spiegelhalter, D. J. (1999). *Probabilistic Networks and Expert Systems*. Springer-Verlag, New York.

Cramér, H. (1946). *Mathematical Methods of Statistics*. Princeton University Press, Princeton, N.J.

Diestel, R. (1987). Simplicial decompositions of graphs – some uniqueness results. *Journal of Combinatorial Theory, Series B*, **42**, 133–145.

Diestel, R. (1990). *Graph Decompositions*. Clarendon Press, Oxford.

Dirac, G. A. (1961). On rigid circuit graphs. *Abhandlungen Mathematisches Seminar Hamburg*, **25**, 71–76.

Eaton, M. L. (1983). *Multivariate Statistics. A Vector Space Approach*. John Wiley

and Sons, New York.
Elston, R C and Stewart, J (1971). A general model for the genetic analysis of pedigree data. *Human Heredity*, **21**, 523–542.
Grimmett, G. R. (1973). A theorem about random fields. *Bulletin of the London Mathematical Society*, **5**, 81–84.
Gross, E. and Sullivant, S. (2018). The maximum likelihood threshold of a graph. *Bernoulli*, **24**, 386–407.
Hammersley, J. M. and Clifford, P. E. (1971). Markov fields on finite graphs and lattices. Unpublished manuscript.
Hansen, E. (2009). *Measure Theory* (4th edn). Department of Mathematical Sciences, University of Copenhagen.
Jensen, F. V. and Jensen, F. (1994). Optimal junction trees. In *Proceedings of the 10th Conference on Uncertainty in Artificial Intelligence* (ed. R. L. de Mantaras and D. Poole), San Francisco, CA, pp. 360–366. Morgan Kaufmann Publishers.
Jensen, F. V., Lauritzen, S. L., and Olesen, K. G. (1990). Bayesian updating in causal probabilistic networks by local computation. *Computational Statistics Quarterly*, **4**, 269–282.
Kalman, R. E. and Bucy, R. (1961). New results in linear filtering and prediction. *Journal of Basic Engineering*, **83 D**, 95–108.
Kong, A. (1986). *Multivariate Belief Functions and Graphical Models*. Ph.D. Thesis, Department of Statistics, Harvard University, Massachusetts.
Koster, J. T. A. (1994). Gibbs–factorization and the Markov property. Unpublished manuscript.
Koster, J. T. A. (2002). Marginalizing and conditioning in graphical models. *Bernoulli*, **8**(6), 817–840.
Lauritzen, S. L. (1996). *Graphical Models*. Clarendon Press, Oxford, United Kingdom.
Lauritzen, S. L. and Jensen, F. V. (1997). Local computation with valuations from a commutative semigroup. *Annals of Mathematics and Artificial Intelligence*, **21**, 51–69.
Lauritzen, S. L. and Spiegelhalter, D. J. (1988). Local computations with probabilities on graphical structures and their application to expert systems (with discussion). *Journal of the Royal Statistical Society, Series B*, **50**, 157–224.
Leimer, H.-G. (1993). Optimal decomposition by clique separators. *Discrete Mathematics*, **113**, 99–123.
Parter, S. (1961). The use of linear graphs in Gauss elimination. *SIAM Review*, **3**, 119–130.
Pearl, J. (1986). Fusion, propagation and structuring in belief networks. *Artificial Intelligence*, **29**, 241–288.
Pearl, J. (1988). *Probabilistic Reasoning in Intelligent Systems*. Morgan Kaufmann Publishers, San Mateo.
Pearl, J. and Paz, A. (1987). Graphoids: A graph based logic for reasoning about relevancy relations. In *Advances in Artificial Intelligence—II* (ed. B. D. Boulay, D. Hogg, and L. Steel), Amsterdam, pp. 357–363. North-Holland.

Rose, D. J. (1970). Triangulated graphs and the elimination process. *Journal of Mathematical Analysis and its Applications*, **32**, 597–609.

Rose, D. J., Tarjan, R. E., and Lueker, G. S. (1976). Algorithmic aspects of vertex elimination on graphs. *SIAM Journal on Computing*, **5**, 266–283.

Schilling, R. (2005). *Measures, Integrals and Martingales*. Cambridge University Press, Cambridge, UK.

Shenoy, P. P. and Shafer, G. (1986). Propagating belief functions using local propagation. *IEEE Expert*, **1**, 43–52.

Shenoy, P. P. and Shafer, G. R. (1990). Axioms for probability and belief-function propagation. In *Uncertainty in Artificial Intelligence 4* (ed. R. D. Shachter, T. S. Levitt, L. N. Kanal, and J. F. Lemmer), Amsterdam, The Netherlands, pp. 169–198. North-Holland.

Slater, M. (1950). Lagrange multipliers revisited. Cowles Commission Discussion Paper: Mathematics: 403.

Speed, T. P. and Kiiveri, H. (1986). Gaussian Markov distributions over finite graphs. *The Annals of Statistics*, **14**, 138–150.

Spitzer, F. (1971). Markov random fields and Gibbs ensembles. *American Mathematical Monthly*, **78**, 142–154.

Studený, M. (1992). Conditional independence relations have no finite complete characterization. In *Transactions of the 11th Prague Conference on Information Theory, Statistical Decision Functions and Random Processes*, Prague, pp. 377–396. Academia.

Tarjan, R. E. (1985). Decomposition by clique separators. *Discrete Mathematics*, **55**, 221–232.

Tarjan, R. E. and Yannakakis, M. (1984). Simple linear-time algorithms to test chordality of graphs, test acyclicity of hypergraphs, and selectively reduce acyclic hypergraphs. *SIAM Journal on Computing*, **13**, 566–579.

Thiele, T. N. (1880). Om Anvendelse af mindste Kvadraters Methode i nogle Tilfælde, hvor en Komplikation af visse Slags uensartede tilfældige Fejlkilder giver Fejlene en 'systematisk' Karakter. *Vidensk. Selsk. Skr. 5. Rk., naturvid. og mat. Afd.*, **12**, 381–408. French version: *Sur la Compensation de quelques Erreurs quasi-systématiques par la Méthode des moindres Carrés.* Reitzel, København, 1880.

Uhler, C. (2012). Geometry of maximum likelihood estimation in Gaussian graphical models. *Annals of Statistics*, **40**, 238–261.

Viterbi, A. J. (1967). Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. *IEEE Transactions on Information Theory*, **13**, 260–269.

Wagner, K. (1937). Über eine Eigenschaft der ebenen Komplexe. *Mathematische Annalen*, **114**, 570–590.

INDEX