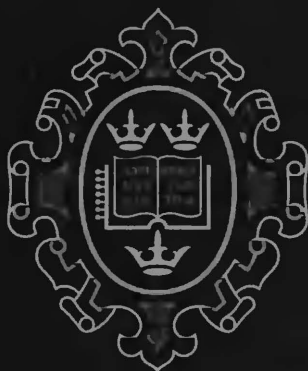


OXFORD STATISTICAL SCIENCE SERIES • 17

Graphical Models

STEFFEN L. LAURITZEN



OXFORD SCIENCE PUBLICATIONS

Meta 1023
p. 1

Graphical Models

STEFFEN L. LAURITZEN

*Department of Mathematics and Computer Science
Aalborg University*

CLARENDON PRESS • OXFORD

1996

Meta 1023
p. 2

Preface

In 1976 Terry Speed invited me to Perth, Australia where he conducted a research seminar exploring relations between statistics and statistical physics. Among other things we studied the relation between the notion of interaction as used in contingency table analysis and in thermodynamics. To our delight they were formally the same and one of the most inspiring periods in my life as a researcher was initiated. In the next couple of months we worked day and night and laid essentially the foundations for the papers Darroch *et al.* (1980) and Lauritzen *et al.* (1984).

In 1979 I was invited to lecture at the Swedish summer school in statistics. Here I met Nanny Wermuth and with her as the main source of inspiration we set out to investigate possibilities for making graphical models that simultaneously dealt with discrete and continuous variables.

In 1985 David Spiegelhalter contacted me with the purpose of discussing possibilities for using graphical models in artificial intelligence.

These three meetings have had a profound influence on my life as researcher in general and in particular on the development of the material described in the present book. I am deeply indebted to the inspiration and ideas that Terry, Nanny, David, and others have provided.

Over the years I have enjoyed being part of many research groups. Here I will particularly mention three. The Danes that have been enthusiastically interested in graphical models from their early beginning: Jens Henrik Badsberg, David Edwards, Morten Frydenberg, Svend Kreiner, and a bit later also Poul Svante Eriksen. The BAIES group working on probabilistic expert systems, consisting of Robert Cowell, Phil Dawid, and David Spiegelhalter. The ODIN group in Aalborg involving far too many to mention, but Stig Andersen, Finn Jensen, Frank Jensen, Uffe Kjærulff, and Kristian Olesen have been there all the time. The enthusiasm from members of these groups has kept up my spirits.

Countless colleagues in Denmark and throughout the world have encouraged me every time I was losing hope and energy. It is plainly impossible to mention them all, but it is clear that without these the book would never have existed. It is a privilege to belong to a scientific community with so many fine people.

Heidi Andersen, David Edwards, Jinglong Wang and Dorte Sørensen have read parts of the book in some detail and I am extremely grateful for their comments and criticism. This also applies to the students at Aalborg

University who were exposed to the somewhat compact material in the spring of 1993 and survived.

Various institutions deserve thanks for hosting me while I was writing parts of the book. This includes the Statistical Laboratory in Cambridge, where I began the writing during my sabbatical leave in 1987. Chapter 6 was largely written during a wonderful stay at the institution of San Cataldo, a former nunnery beautifully situated in Southern Italy and run as a study home for Danish researchers and artists.

The Danish Research Councils and the Carlsberg Foundation have in various ways contributed financially to the research.

Oxford University Press and my colleague Frank Jensen have been of invaluable assistance with typographical matters.

Aalborg
December 1995

S.L.L.

Contents

1	Introduction	1
1.1	Graphical models	1
1.2	Outline of book	2
2	Graphs and hypergraphs	4
2.1	Graphs	4
2.1.1	Notation and terminology	4
2.1.2	Decompositions of marked graphs	7
2.1.3	Simplicial subsets and perfect sequences	13
2.1.4	Subgraphs of decomposable graphs	19
2.2	Hypergraphs	21
2.2.1	Basic concepts	21
2.2.2	Graphs and hypergraphs	22
2.2.3	Junction trees and forests	24
2.3	Notes	26
3	Conditional independence and Markov properties	28
3.1	Conditional independence	28
3.2	Markov properties	32
3.2.1	Markov properties on undirected graphs	32
3.2.2	Markov properties on directed acyclic graphs	46
3.2.3	Markov properties on chain graphs	53
3.3	Notes	60
4	Contingency tables	62
4.1	Examples	62
4.2	Basic facts and concepts	67
4.2.1	Notation and terminology	67
4.2.2	Saturated models	70
4.2.3	Log-affine and log-linear models	71
4.3	Hierarchical models	81
4.3.1	Estimation in hierarchical log-affine models	82
4.3.2	Test in hierarchical models	85
4.3.3	Interaction graphs and graphical models	88
4.4	Decomposable models	90

4.4.1	Basic factorizations	90
4.4.2	Maximum likelihood estimation	91
4.4.3	Exact tests in decomposable models	98
4.4.4	Asymptotic tests in decomposable models	105
4.5	Recursive models	106
4.5.1	Recursive graphical models	107
4.5.2	Recursive hierarchical models	112
4.6	Block-recursive models	113
4.6.1	Chain graph models	114
4.6.2	Block-recursive hierarchical models	118
4.6.3	Decomposable block-recursive models	119
4.7	Notes	121
4.7.1	Collapsibility	121
4.7.2	Bibliographical notes	121
5	Multivariate normal models	123
5.1	Basic facts and concepts	123
5.1.1	Notation	123
5.1.2	The saturated model	124
5.1.3	Conditional independence	129
5.1.4	Interaction	131
5.2	Covariance selection models	131
5.2.1	Maximum likelihood estimation	132
5.2.2	Deviance tests	142
5.3	Decomposable models	144
5.3.1	Basic factorizations	144
5.3.2	Maximum likelihood estimation	145
5.3.3	Exact tests in decomposable models	149
5.4	Notes	153
5.4.1	Chain graph models	153
5.4.2	Lattice models	156
5.4.3	Collapsibility	156
5.4.4	Bibliographical notes	156
6	Models for mixed data	158
6.1	Basic facts and concepts	158
6.1.1	CG distributions	158
6.1.2	The saturated models	168
6.2	Graphical interaction models	173
6.2.1	CG interactions	173
6.2.2	Maximum likelihood estimation	175
6.3	Decomposable models	187
6.3.1	Basic factorizations	187
6.3.2	Maximum likelihood estimation	188

6.3.4	Exact tests in decomposable models	191
6.4	Hierarchical interaction models	199
6.4.1	General properties	199
6.4.2	Generators and canonical statistics	201
6.4.3	Maximum likelihood estimation	205
6.4.4	Mixed hierarchical model subspaces	213
6.5	Chain graph models	216
6.5.1	CG regressions	217
6.5.2	Estimation in chain graph models	218
6.6	Notes	219
6.6.1	Collapsibility	219
6.6.2	Bibliographical notes	220
7	Further topics	221
7.1	Probabilistic expert systems	221
7.1.1	Specification of the joint distribution	223
7.1.2	Local computation algorithm	226
7.1.3	Extensions	228
7.2	Model selection	229
7.3	Modelling complexity	230
7.3.1	Markov chain Monte Carlo methods	231
7.3.2	Applications	232
7.4	Missing-data problems	233
7.4.1	The EM algorithm	233
7.4.2	Hierarchical log-linear models	234
7.4.3	Recursive models	235
 Appendices		
A	Various prerequisites	237
A.1	Inequalities	237
A.2	Kullback–Leibler divergence	238
A.3	Möbius inversion	239
A.4	Iterative partial maximization	239
A.5	Sufficiency	241
B	Linear algebra and random vectors	243
B.1	Matrix results	243
B.2	Factor subspaces and interactions	246
B.3	Random vectors	250
		Meta 1023
		p. 7

C	The multivariate normal distribution	254
C.1	Basic properties	254
C.2	The Wishart distribution	258
C.3	Other derived distributions	262
C.3.1	Box-type distributions	262
C.3.2	Wilks's distribution	263
C.3.3	Test for identical covariances	264
D	Exponential models	266
D.1	Regular exponential models	266
D.1.1	Basic terminology	266
D.1.2	Analytic properties	267
D.1.3	Maximum likelihood estimation	268
D.1.4	Affine hypotheses	268
D.1.5	Iterative computational methods	269
D.2	Curved exponential models	272
D.2.1	The non-singular case	272
D.2.2	The singular case	276
	Bibliography	278
	Index	295

1

Introduction

1.1 Graphical models

Graphical models have their origin in several scientific areas. One of these is statistical physics. Here the ideas can be traced back to Gibbs (1902). The object of interest is a large system of particles, possibly the atoms of a gas or solid. Each particle is occupying a site where it can be in different states. The total energy of the system is composed by an external potential plus a potential due to *interaction* of groups of particles. It is usually assumed that only particles at sites close to each other interact. Sites that are close to each other are termed *neighbours* and an undirected graph is determined by the neighbour relationship. The total energy of the system is determining the behaviour of the system through the so-called Gibbsian distribution

$$p(x) = Z^{-1} \exp\{-E(x)\},$$

where $E(x)$ is the total energy of the system when this is in configuration x , and Z is a normalizing constant.

Another origin is in the subject of genetics, where the graphical models go back to Wright (1921, 1923, 1934), who founded the so-called path analysis. Here one is studying heritable properties of natural species and the graph relations are directed, with arrows moving from parent to child. Ideas of path analysis were later taken up in economics and social sciences (Wold 1954; Blalock 1971).

The third source has less obvious relation to graphs. Bartlett (1935) studied the notion of interaction in a three-way contingency table. It turns out that, apart from differences due to notational conventions, this notion of interaction is formally identical to the notion used in statistical physics (Darroch *et al.* 1980).

With this understanding, the graphical models are ready at hand for use in statistics. Their fundamental and universal applicability is due to a number of factors. Firstly, the graphs can visually represent the scientific content of a given model and facilitate communication between researcher and statistician. Secondly, the models are naturally modular so that complex problems can be described and handled by careful combination of

simple elements. Thirdly, the graphs are natural data structures for modern digital computers. Thus models can be efficiently communicated to these and the road is paved for exploiting their computational power.

Over the years the theory and methodology have developed and been extended in a multitude of directions. This is true to such an extent that it has not been possible to cover the area properly in the present book. Luckily a number of other books have been written that take care of some of the omissions in the present. Whittaker (1990b) focuses on the ideas behind graphical modelling and has many examples. The same holds for the book by Edwards (1995), just that here the emphasis is on models for mixed data and special attention is paid to problems that are suitable for analysis using the program MIM. Cox and Wermuth (1996) is concentrating attention on the interpretation and application of multivariate systems, in particular graphical models. Neapolitan (1990) deals with probabilistic expert systems based on graphical models, and the collection Oliver and Smith (1990) is primarily concerned with graphical models as influence diagrams, i.e. used as tools in decision analysis.

The present book is primarily concerned with the fundamental mathematical and statistical theory of graphical models. The book is mostly based on a traditional statistical approach, discussing aspects of maximum likelihood methods and significance testing in the different variety of models. It is believed that these results are basic and therefore hopefully will have interest for anyone who wants to enter the world of graphical models, independently of statistical paradigms.

The book is a research monograph and it has been written primarily for researchers and graduate students in mathematical statistics. However, I have tried to keep the mathematics at a reasonable level of abstraction and exactness. No doubt, certain sections are more difficult to read than others, but it is my hope that a relatively wide audience will be able to take advantage of the central parts of the book.

1.2 Outline of book

There are five main chapters. In Chapter 2 the basic graph theory is developed. It may not be advisable to read this chapter in detail before continuing to chapters with a more statistical content, but some familiarity with the notions is necessary to understand even the basic ideas in the subsequent developments. Some of the heavier parts can be omitted at first reading and then returned to when necessary.

Chapter 3 is concerned with conditional independence and Markov properties and these are also essential for later chapters. If desired, the reader can skip the description of directed and chain graph Markov properties at first reading, and then return to these later.

The next three chapters of the book form the core of the statistical theory. Chapter 4 is concerned with discrete variables, Chapter 5 with variables that are normally distributed, and Chapter 6 describes models for systems that contain both discrete and continuous variables. These three chapters are largely written so that they can be read independently of each other. However, as the results of Chapter 6 unify and generalize the results in Chapters 4 and 5, some readers may want to read them in the given order. On the other hand, Chapter 6 is the chapter which contains most original material that has not appeared elsewhere in the literature and others may want to attack this chapter directly.

Chapter 7 is largely a guide to the literature on topics that have not been treated in the book but constitute important and natural extensions.

Finally the book contains four appendices with some material that is used at various places and may not be sufficiently familiar to the reader. Appendix A is a collection of various little topics. Appendix B has some results from linear algebra. Appendix C treats the multivariate normal distribution in some detail, and Appendix D describes basic elements of the theory of exponential families. By having these ready at hand I hope to help readers who do not have all of this present in their minds. However, it is not advisable to learn the material in all four appendices from the given expositions alone.

3

Conditional independence and Markov properties

3.1 Conditional independence

Throughout this text a central notion is that of conditional independence of random variables, the graphs keeping track of the conditional independence relations.

Formally, if X, Y, Z are random variables with a joint distribution P , we say that X is *conditionally independent of Y given Z under P* , and write $X \perp\!\!\!\perp Y | Z [P]$, if, for any measurable set A in the sample space of X , there exists a version of the conditional probability $P(A | Y, Z)$ which is a function of Z alone. Usually P will be fixed and omitted from the notation. If Z is trivial we say that X is *independent of Y* , and write $X \perp\!\!\!\perp Y$. The notation is due to Dawid (1979) who studied the notion of conditional independence in a systematic fashion. Dawid (1980) gives a formal treatment.

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

$$X \perp\!\!\!\perp Y | Z \iff f_{XY|Z}(x, y | z) = f_{X|Z}(x | z)f_{Y|Z}(y | z), \quad (3.1)$$

where this equation is to hold almost surely with respect to P . If all densities are continuous, the equality in (3.1) must hold for all z with $f_Z(z) > 0$. Here it is understood that all functions on a discrete space are considered continuous functions. The condition (3.1) can be rewritten as

$$X \perp\!\!\!\perp Y | Z \iff f_{XYZ}(x, y, z)f_Z(z) = f_{XZ}(x, z)f_{YZ}(y, z) \quad (3.2)$$

and this equality must hold for all values of z when the densities are continuous.

The ternary relation $X \perp\!\!\!\perp Y | Z$ has the following properties, where h denotes an arbitrary measurable function on the sample space of X :

- (C1) if $X \perp\!\!\!\perp Y | Z$ then $Y \perp\!\!\!\perp X | Z$;
- (C2) if $X \perp\!\!\!\perp Y | Z$ and $U = h(X)$, then $U \perp\!\!\!\perp Y | Z$;
- (C3) if $X \perp\!\!\!\perp Y | Z$ and $U = h(X)$, then $X \perp\!\!\!\perp Y | (Z, U)$;
- (C4) if $X \perp\!\!\!\perp Y | Z$ and $X \perp\!\!\!\perp W | (Y, Z)$, then $X \perp\!\!\!\perp (W, Y) | Z$.

We leave the proof of these facts to the reader. Note that the converse to (C4) follows from (C2) and (C3).

If we use f as generic symbol for the probability density of the random variables corresponding to its arguments, the following statements are true:

$$X \perp\!\!\!\perp Y | Z \iff f(x, y, z) = f(x, z)f(y, z)/f(z) \quad (3.3)$$

$$X \perp\!\!\!\perp Y | Z \iff f(x | y, z) = f(x | z) \quad (3.4)$$

$$X \perp\!\!\!\perp Y | Z \iff f(x, z | y) = f(x | z)f(z | y) \quad (3.5)$$

$$X \perp\!\!\!\perp Y | Z \iff f(x, y, z) = h(x, z)k(y, z) \text{ for some } h, k \quad (3.6)$$

$$X \perp\!\!\!\perp Y | Z \iff f(x, y, z) = f(x | z)f(y, z). \quad (3.7)$$

The equalities above hold apart from a set of triples (x, y, z) with probability zero. If the densities are continuous functions (in particular if the state spaces are discrete), the equations hold whenever the quantities involved are well defined, i.e. when the densities of all conditioning variables are positive. We also leave the proof of these equivalences to the reader.

Another property of the conditional independence relation is often used:

- (C5) if $X \perp\!\!\!\perp Y | Z$ and $X \perp\!\!\!\perp Z | Y$ 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 . A trivial counterexample appears when $X = Y = Z$ with $P\{X = 1\} = P\{X = 0\} = 1/2$. We have however

Proposition 3.1 *If the joint density of all variables with respect to a product measure is positive and continuous, then the statement (C5) will hold true.*

Proof: Assume that the variables have a continuous density $f(x, y, z) > 0$ and that $X \perp\!\!\!\perp Y | Z$ as well as $X \perp\!\!\!\perp Z | Y$. Then (3.6) gives for all values of (x, y, z) that

$$f(x, y, z) = k(x, z)l(y, z) = g(x, y)h(y, z)$$

for suitable strictly positive functions g, h, k, l . Thus, as the density is assumed continuous, we have that for all z ,

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

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. $\square$

The proposition can be weakened to more general functions than continuous functions, but we abstain from pursuing this here.

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 Y is irrelevant for reading any chapter U of X ;
- (I3) if, knowing Z , reading Y is irrelevant for reading the book X , it remains irrelevant after having read any chapter U of X ;
- (I4) if, knowing Z , reading the book Y is irrelevant for reading X and even after having also read Y , reading W is irrelevant for reading X , then reading of both Y and W is irrelevant for reading X .

Thus one can view the relations (C1)–(C4) as pure formal properties of the notion of irrelevance. The property (C5) is slightly more subtle. In a certain sense, also the symmetry (C1) is a somewhat special property of probabilistic conditional independence, rather than general irrelevance.

It is thus tempting to use the relations (C1)–(C4) as formal axioms for conditional independence or irrelevance. A *semi-graphoid* is an algebraic structure which satisfies (C1)–(C4) where X, Y, Z are disjoint subsets of a finite set and $U = h(X)$ is replaced by $U \subseteq X$ (Pearl 1988). If also (C5) holds for disjoint subsets, it is called a *graphoid*. Below we give further examples of such structures.

Example 3.2 A very important example of a model for the irrelevance axioms above is that of *graph 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

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

Then it is not difficult to see that graph separation has the following properties:

- (S1) if $A \stackrel{\mathcal{G}}{\perp} B | C$ then $B \stackrel{\mathcal{G}}{\perp} A | C$;
- (S2) if $A \stackrel{\mathcal{G}}{\perp} B | C$ and U is a subset of A , then $U \stackrel{\mathcal{G}}{\perp} B | C$;
- (S3) if $A \stackrel{\mathcal{G}}{\perp} B | C$ and U is a subset of B , then $A \stackrel{\mathcal{G}}{\perp} B | (C \cup U)$;
- (S4) if $A \stackrel{\mathcal{G}}{\perp} B | C$ and $A \stackrel{\mathcal{G}}{\perp} D | (B \cup C)$, then $A \stackrel{\mathcal{G}}{\perp} (B \cup D) | C$.

Even the analogue of (C5) holds when all involved subsets are disjoint. Hence graph separation satisfies the graphoid axioms. $\square$

Example 3.3 As another fundamental example, consider *geometric orthogonality* in Euclidean vector spaces. Let L , M , and N be linear subspaces of a Euclidean space V and define

$$L \perp M | N \iff (L \ominus N) \perp (M \ominus N), \quad (3.8)$$

where $L \ominus N = L \cap N^\perp$. If (3.8) 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 $L \perp M | N$ and U is a linear subspace of M , then $L \perp M | (N + U)$;
- (O4) if $L \perp M | N$ and $L \perp R | (M + N)$, then $L \perp (M + R) | N$.

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

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

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

An abstract theory for conditional independence based on graphoids and semi-graphoids has recently developed (Studený 1989, 1993; Matúš 1992a). 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). See also Geiger and Pearl (1993) for a systematic study of the logical implications of conditional independence.

3.2 Markov properties

In this section we consider conditional independence in the special situation where we have a collection of random variables $(X_\alpha)_{\alpha \in V}$ taking values in probability 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. The set V is assumed to be the vertex set of a graph $\mathcal{G} = (V, E)$.

3.2.1 Markov properties on undirected graphs

Associated with an undirected graph $\mathcal{G} = (V, E)$ and a collection of random variables $(X_\alpha)_{\alpha \in V}$ as above there is 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 non-adjacent vertices

$$\alpha \perp\!\!\!\perp \beta \mid V \setminus \{\alpha, \beta\};$$

(L) *the local Markov property*, relative to $\mathcal{G}$, if for any vertex $\alpha \in V$

$$\alpha \perp\!\!\!\perp V \setminus \text{cl}(\alpha) \mid \text{bd}(\alpha);$$

(G) *the global Markov property*, relative to $\mathcal{G}$, if for any triple (A, B, S) of disjoint subsets of V such that S separates A from B in $\mathcal{G}$

$$A \perp\!\!\!\perp B \mid S.$$

The Markov properties are related as described in the proposition below.

Proposition 3.4 *For any undirected graph $\mathcal{G}$ and any probability distribution on $\mathcal{X}$ it holds that*

$$(G) \implies (L) \implies (P). \quad (3.9)$$

Proof: Firstly, (G) implies (L) because $\text{bd}(\alpha)$ separates α from $V \setminus \text{cl}(\alpha)$. Assume next that (L) holds. We have $\beta \in V \setminus \text{cl}(\alpha)$ because α and β are non-adjacent. Hence

$$\text{bd}(\alpha) \cup ((V \setminus \text{cl}(\alpha)) \setminus \{\beta\}) = V \setminus \{\alpha, \beta\},$$

and it follows from (L) and (C3) that

$$\alpha \perp\!\!\!\perp V \setminus \text{cl}(\alpha) \mid V \setminus \{\alpha, \beta\}.$$

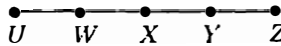
Application of (C2) then gives $\alpha \perp\!\!\!\perp \beta \mid V \setminus \{\alpha, \beta\}$ which is (P). $\square$

It is worth noting that (3.9) only depends on the properties (C1)–(C4) of conditional independence and hence holds for any semi-graphoid. The various Markov properties are different in general, as the following examples show.

Example 3.5 Define the joint distribution of five binary random variables U, W, X, Y, Z as follows: U and Z are independent with

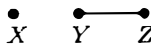
$$P(U = 1) = P(Z = 1) = P(U = 0) = P(Z = 0) = 1/2,$$

$W = U$, $Y = Z$, and $X = WY$. The joint distribution so defined is easily seen to satisfy (L) but not (G) for the graph below.



In fact, Matúš (1992b) shows that the global and local Markov properties coincide if and only if the dual graph $\tilde{\mathcal{G}}$ (defined as $\alpha \sim \beta$ if and only if $\alpha \not\sim \beta$) does not have the 4-cycle as an induced subgraph. $\square$

Example 3.6 A simple example of a probability distribution of (X, Y, Z) that satisfies the pairwise Markov property (P) with respect to the graph



but does not satisfy the local Markov property (L) can be constructed by letting $X = Y = Z$ with $P\{X = 1\} = P\{X = 0\} = 1/2$. It can be shown (Matúš 1992b) that the global and pairwise Markov properties coincide if and only if the dual graph $\tilde{\mathcal{G}}$ does not have a subset of three vertices with two or three edges in its induced subgraph. $\square$

If it holds for all disjoint subsets A, B, C , and D that

$$\text{if } A \perp\!\!\!\perp B \mid (C \cup D) \text{ and } A \perp\!\!\!\perp C \mid (B \cup D) \text{ then } A \perp\!\!\!\perp (B \cup C) \mid D, \quad (3.10)$$

then the Markov properties are all equivalent. This condition is analogous to (C5) and holds, for example, if P has a positive and continuous density with respect to a product measure μ . This is seen as in Proposition 3.1. The result is stated in the theorem below, due to Pearl and Paz (1987): see also Pearl (1988).

Theorem 3.7 (Pearl and Paz) *If a probability distribution on $\mathcal{X}$ is such that (3.10) holds for disjoint subsets A, B, C, D then*

$$(G) \iff (L) \iff (P).$$

Proof: We need to show that (P) implies (G), so assume that S separates A from B in $\mathcal{G}$ and that (P) as well as (3.10) hold. Without loss of generality we can also assume that both A and B are non-empty. The proof is then backward 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 $S \cup \{\alpha\}$ separates $A \setminus \{\alpha\}$ from B and also $S \cup A \setminus \{\alpha\}$ separates α from B . Thus by the induction hypothesis

$$A \setminus \{\alpha\} \perp\!\!\!\perp B \mid S \cup \{\alpha\} \text{ and } \alpha \perp\!\!\!\perp B \mid S \cup A \setminus \{\alpha\}.$$

Now (3.10) gives $A \perp\!\!\!\perp B \mid S$.

If $A \cup B \cup S \subset V$ we choose $\alpha \in V \setminus (A \cup B \cup S)$. Then $S \cup \{\alpha\}$ separates A and B , implying $A \perp\!\!\!\perp B \mid S \cup \{\alpha\}$. Further, either $A \cup S$ separates B from $\{\alpha\}$ or $B \cup S$ separates A from $\{\alpha\}$. Assuming the former gives $\alpha \perp\!\!\!\perp B \mid A \cup S$. Using (3.10) and (C2) we derive that $A \perp\!\!\!\perp B \mid S$. The latter case is similar. $\square$

Note that the proof only exploits (C1)–(C4) and (3.10) and therefore applies to any graphoid.

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. 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, and there exists a product

measure $\mu = \otimes_{\alpha \in V} \mu_\alpha$ on $\mathcal{X}$, such that P has density f with respect to μ where f has the form

$$f(x) = \prod_{a \text{ complete}} \psi_a(x). \quad (3.11)$$

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

$$f(x) = \prod_{c \in \mathcal{C}} \psi_c(x), \quad (3.12)$$

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 3.8 *For any undirected graph $\mathcal{G}$ and any probability distribution on $\mathcal{X}$ it holds that*

$$(F) \implies (G) \implies (L) \implies (P).$$

Proof: We only have to show that (F) implies (G) as the remaining implications are given in (3.9). 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 (3.12) 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 (3.6) gives that $\tilde{A} \perp\!\!\!\perp \tilde{B} | S$. Applying (C2) twice gives the desired independence. $\square$

In the case where P has a positive and continuous 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 in various forms by a number of authors (Speed 1979) but is usually attributed to Hammersley and Clifford (1971) who proved the result in the discrete case. The condition that the density be continuous can probably be considerably relaxed (Koster 1994), whereas the positivity is essential. More precisely, we have

Theorem 3.9 (Hammersley and Clifford) *A probability distribution P with positive and continuous 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 3.8, so we just have to show that (P) implies (F).

Since the density is positive, we may take logarithms on both sides of (3.11). Hence this equation can be rewritten as

$$\log f(x) = \sum_{a:a \subseteq V} \phi_a(x), \quad (3.13)$$

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.2 (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

$$\phi_a(x) = \sum_{b:b \subseteq c} (-1)^{|c \setminus b|} \{H_b - H_{b \cup \{\alpha\}} - H_{b \cup \{\beta\}} + H_{b \cup \{\alpha, \beta\}}\}. \quad (3.14)$$

Let $d = V \setminus \{\alpha, \beta\}$. Then, by the pairwise Markov property and (3.7), we have

$$\begin{aligned} H_{b \cup \{\alpha, \beta\}}(x) - H_{b \cup \{\alpha\}}(x) &= \log \frac{f(x_b, x_\alpha, x_\beta, x_{d \setminus b}^*)}{f(x_b, x_\alpha, x_\beta^*, x_{d \setminus b}^*)} \\ &= \log \frac{f(x_\alpha | x_b, x_{d \setminus b}^*) f(x_\beta, x_b, x_{d \setminus b}^*)}{f(x_\alpha | x_b, x_{d \setminus b}^*) f(x_\beta^*, x_b, x_{d \setminus b}^*)} \end{aligned}$$

$$\begin{aligned}
&= \log \frac{f(x_\alpha^* | x_b, x_{d \setminus b}^*) f(x_\beta, x_b, x_{d \setminus b}^*)}{f(x_\alpha^* | x_b, x_{d \setminus b}^*) f(x_\beta^*, x_b, x_{d \setminus b}^*)} \\
&= \log \frac{f(x_b, x_\alpha^*, x_\beta, x_{d \setminus b}^*)}{f(x_b, x_\alpha^*, x_\beta^*, x_{d \setminus b}^*)} \\
&= H_{b \cup \{\beta\}}(x) - H_b(x).
\end{aligned}$$

Thus all terms in the curly brackets in (3.14) add to zero and henceforth the entire sum is zero. This completes the proof. $\square$

The expression inside curly brackets in (3.14) is the logarithm of what is known as the *partial cross-product ratio* so that we can alternatively write

$$\phi_a(x) = \sum_{b: b \subseteq c} (-1)^{|c \setminus b|} \log \text{cpr}(x_\alpha, x_\beta; x_\alpha^*, x_\beta^* | x_b, x_{d \setminus b}^*).$$

The pairwise Markov property ensures that all these partial cross-product ratios are equal to 1.

Example 3.10 The following example is due to Moussouris (1974) and shows that the global Markov property (G) may not imply the factorization property (F) without positivity assumptions on the density.

The example is concerned with the distribution of four binary random variables, denoted by (X_1, X_2, X_3, X_4) . The following combinations are assumed to have positive probabilities, in fact each of them given a probability equal to $1/8$:

$$\begin{array}{cccc}
(0, 0, 0, 0) & (1, 0, 0, 0) & (1, 1, 0, 0) & (1, 1, 1, 0) \\
(0, 0, 0, 1) & (0, 0, 1, 1) & (0, 1, 1, 1) & (1, 1, 1, 1).
\end{array}$$

The distribution so specified satisfies the global Markov property (G) with respect to the chordless four-cycle



with the vertices identified cyclically with the random variables. This is seen as follows.

For example, if we consider the conditional distribution of (X_1, X_3) , given that $(X_2, X_4) = (0, 1)$, we find

$$P\{X_1 = 0 \mid (X_2, X_4) = (0, 1)\} = 1.$$

Since the conditional distribution of X_1 is degenerate, it is trivially independent of X_3 . All other combinations of conditions on (X_2, X_4) give in a

similar way degenerate distributions for one of the remaining variables and this picture is repeated when conditioning on (X_1, X_3) . Hence, we have

$$X_1 \perp\!\!\!\perp X_3 \mid (X_2, X_4) \text{ and } X_2 \perp\!\!\!\perp X_4 \mid (X_1, X_3),$$

which shows that the distribution is globally Markov with respect to the graph displayed.

But the density does not factorize. This is seen by an indirect argument. Assume the density factorizes. Then

$$0 \neq 1/8 = f(0, 0, 0, 0) = \phi_{\{1,2\}}(0, 0)\phi_{\{2,3\}}(0, 0)\phi_{\{3,4\}}(0, 0)\phi_{\{4,1\}}(0, 0).$$

But also

$$0 = f(0, 0, 1, 0) = \phi_{\{1,2\}}(0, 0)\phi_{\{2,3\}}(0, 1)\phi_{\{3,4\}}(1, 0)\phi_{\{4,1\}}(0, 0),$$

whereby

$$\phi_{\{2,3\}}(0, 1)\phi_{\{3,4\}}(1, 0) = 0.$$

Since now

$$0 \neq 1/8 = f(0, 0, 1, 1) = \phi_{\{1,2\}}(0, 0)\phi_{\{2,3\}}(0, 1)\phi_{\{3,4\}}(1, 1)\phi_{\{4,1\}}(1, 0),$$

this implies $\phi_{\{2,3\}}(0, 1) \neq 0$ and hence $\phi_{\{3,4\}}(1, 0) = 0$, which contradicts that

$$0 \neq 1/8 = f(1, 1, 1, 0) = \phi_{\{1,2\}}(1, 1)\phi_{\{2,3\}}(1, 1)\phi_{\{3,4\}}(1, 0)\phi_{\{4,1\}}(0, 1).$$

Hence the density cannot factorize. $\square$

In general none of the Markov properties are preserved under weak limits. This is because weak convergence of joint distributions does not imply convergence of conditional distributions. This fact is illustrated in the following example.

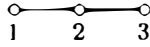
Example 3.11 Let $Y = (Y_1, Y_2, Y_3)^\top$ be a trivariate normal random variable with mean zero and covariance matrix

$$\Sigma = \begin{pmatrix} 1 & \frac{1}{\sqrt{n}} & \frac{1}{2} \\ \frac{1}{\sqrt{n}} & \frac{2}{n} & \frac{1}{\sqrt{n}} \\ \frac{1}{2} & \frac{1}{\sqrt{n}} & 1 \end{pmatrix}.$$

Using Proposition C.5 the conditional distribution of $(Y_1, Y_3)^\top$ given Y_2 is bivariate normal with covariance matrix

$$\Sigma_{13|2} = \begin{pmatrix} 1 & \frac{1}{2} \\ \frac{1}{2} & 1 \end{pmatrix} - \begin{pmatrix} \frac{1}{\sqrt{n}} \\ \frac{1}{\sqrt{n}} \end{pmatrix} \left(\frac{n}{2} \right) \begin{pmatrix} \frac{1}{\sqrt{n}} & \frac{1}{\sqrt{n}} \end{pmatrix} = \begin{pmatrix} \frac{1}{2} & 0 \\ 0 & \frac{1}{2} \end{pmatrix}$$

and hence $Y_1 \perp\!\!\!\perp Y_3 \mid Y_2$, which means that Y satisfies the global Markov property on the graph



As we shall see later in Proposition 5.2, this can alternatively be seen from inspection of the inverse covariance matrix

$$K = \Sigma^{-1} = \begin{pmatrix} 2 & -\sqrt{n} & 0 \\ -\sqrt{n} & \frac{3n}{2} & -\sqrt{n} \\ 0 & -\sqrt{n} & 2 \end{pmatrix}.$$

The distribution of Y converges weakly to the normal distribution with covariance matrix

$$\bar{\Sigma} = \begin{pmatrix} 1 & 0 & \frac{1}{2} \\ 0 & 0 & 0 \\ \frac{1}{2} & 0 & 1 \end{pmatrix}$$

which is not Markov on the graph considered, since Y_2 is degenerate and

$$\bar{\Sigma}_{13|2} = \bar{\Sigma}_{13} = \begin{pmatrix} 1 & \frac{1}{2} \\ \frac{1}{2} & 1 \end{pmatrix}.$$

Hence the Markov property is not in general stable under weak limits. $\square$

There are, however, important exceptions where the Markov property is preserved under limits and the case of discrete sample space is one.

Proposition 3.12 *Assume that a sequence of probability distributions P_n on a discrete sample space converges to P . If P_n all satisfy any of the Markov properties (P), (L), or (G), then P satisfies the same Markov property.*

Proof: The density p of the limiting distribution is given as $p = \lim p_n$, where p_n is the density of P_n . Let A, B, C be disjoint and assume that $A \perp\!\!\!\perp B \mid C$ for all P_n . If $p(i_C) > 0$ we get

$$\begin{aligned} p(i_A, i_B, i_C) &= \lim_{n \rightarrow \infty} p_n(i_A, i_B, i_C) = \lim_{n \rightarrow \infty} \frac{p_n(i_A, i_C)p_n(i_B, i_C)}{p_n(i_C)} \\ &= \frac{p(i_A, i_C)p(i_B, i_C)}{p(i_C)}, \end{aligned}$$

whereby $A \perp\!\!\!\perp B \mid C$ also in the limit. Thus we have shown that all conditional independences are preserved, as desired. $\square$

Observe that the property (F) is not preserved under weak limits, even in the discrete case, as the following example shows.

Example 3.13 Consider four binary random variables (X_1, X_2, X_3, X_4) and let

$$f_n(x_1, x_2, x_3, x_4) = \frac{n^{x_1 x_2 + x_2 x_3 + x_3 x_4 - x_1 x_4 - x_2 - x_3 + 1}}{8 + 8n}.$$

Direct calculations show that, as $n \rightarrow \infty$, f_n converges to the distribution in Example 3.10. Hence we deduce that (F) is not stable under limits. $\square$

It holds however, in the discrete case, that any probability distribution which satisfies (F) can be obtained as a limit of positive probabilities that factorize. This is true because if p factorizes we get

$$p(i) = \prod_a \psi_a(i) = \lim_{\epsilon \rightarrow 0} \frac{\prod_a \{\psi_a(i) + \epsilon\}}{\sum_j \prod_a \{\psi_a(j) + \epsilon\}} = \lim_{\epsilon \rightarrow 0} p_\epsilon(i),$$

where p_ϵ clearly is positive and factorizes.

In the discrete case we refer to limits of positive Markov probabilities as *extended Markov* probabilities and denote these by $M_E(\mathcal{G})$. The argument above shows that $M_F(\mathcal{G}) \subseteq M_E(\mathcal{G})$ but the inclusion is strict for a general graph $\mathcal{G}$.

Distributions in $M_E(\mathcal{G})$ are identified through their clique marginals when their state space is finite. This fact is a special case of the following lemma (when $\mathcal{A}$ is taken to be the set of cliques of the graph $\mathcal{G}$).

Lemma 3.14 *Let $\mathcal{A}$ be a set of subsets of a finite set Δ and P and Q be probability measures on the product space $\mathcal{I} = \times_{\delta \in \Delta} \mathcal{I}_\delta$ where, for each δ , $\mathcal{I}_\delta$ is a finite set. If P and Q have identical marginals to the sets $a \in \mathcal{A}$ and both are limits of measures whose densities with respect to a product measure μ factorize over $\mathcal{A}$, then $P = Q$.*

Proof: It is no restriction to assume that μ is the counting measure on $\mathcal{I}$ such that we have

$$p(i) = \lim_{n \rightarrow \infty} \prod_{a \in \mathcal{A}} \phi_a^n(i_a), \quad q(i) = \lim_{n \rightarrow \infty} \prod_{a \in \mathcal{A}} \psi_a^n(i_a).$$

Then we get

$$\begin{aligned} \sum_i p(i) \log q(i) &= \sum_i p(i) \log \left\{ \lim_{n \rightarrow \infty} \prod_{a \in \mathcal{A}} \psi_a^n(i_a) \right\} \\ &= \lim_{n \rightarrow \infty} \sum_i p(i) \sum_{a \in \mathcal{A}} \log \psi_a^n(i_a) \\ &= \lim_{n \rightarrow \infty} \sum_{a \in \mathcal{A}} \sum_i p(i) \log \psi_a^n(i_a) \\ &= \lim_{n \rightarrow \infty} \sum_{a \in \mathcal{A}} \sum_i q(i) \log \psi_a^n(i_a) \end{aligned}$$

$$\begin{aligned}
&= \lim_{n \rightarrow \infty} \sum_i q(i) \sum_{a \in \mathcal{A}} \log \psi_a^n(i_a) \\
&= \sum_i q(i) \log q(i),
\end{aligned}$$

where we have exploited that $\sum_i p(i) \log \psi_a^n(i_a)$ depends on P through its a -marginal only and therefore is equal to $\sum_i q(i) \log \psi_a^n(i_a)$.

Interchanging the role of p and q in the above calculations yields

$$\sum_i q(i) \log p(i) = \sum_i p(i) \log p(i).$$

Combining these identities with the information inequality (A.4) gives

$$\begin{aligned}
\sum_i p(i) \log p(i) &= \sum_i q(i) \log p(i) \leq \sum_i q(i) \log q(i) \\
&= \sum_i p(i) \log q(i) \leq \sum_i p(i) \log p(i).
\end{aligned}$$

Since (A.4) is strict unless $p(i) \equiv q(i)$, we must have $P = Q$ and the lemma is proved. $\square$

Lemma 3.14 still holds in the finite case for a general μ and a general system of subsets $\mathcal{A}$. For countable state spaces and general μ it does not hold, i.e. we may have two different probability distributions P and Q that both have factorizing densities (H. G. Kellerer, personal communication). This fact seems to contradict a remark in Csiszár (1975).

If $\mathcal{A}$ is the set of cliques of a decomposable graph, the lemma holds under quite general circumstances (Kellerer 1964a, 1964b). It is not known whether the conclusion of the lemma holds for factorizing probability distributions in the case of a product measure μ and countable (or more general) state spaces. The critical point in the proof is the interchange of the two sums.

The following example is due to Matúš and Studený (1995) and shows that not all global Markov probabilities can be obtained as limits of factorizing distributions.

Example 3.15 Let $\mathcal{G}$ be the four-cycle as in Example 3.10, but let all four variables have three possible values a , b , and c . Let P be the distribution with each of the following nine states having probability equal to $1/9$:

$$\begin{array}{lll}
(a, a, a, a) & (b, a, b, c) & (c, a, c, b) \\
(a, b, b, b) & (b, b, c, a) & (c, b, a, c) \\
(a, c, c, c) & (b, c, a, b) & (c, c, b, a)
\end{array}$$

This distribution is globally Markov with respect to $\mathcal{G}$ because the conditional distribution of (X_1, X_3) given (X_2, X_4) is always degenerate and this is also true for the conditional distribution of (X_2, X_4) given (X_1, X_3) . It cannot be obtained as a limit of factorizing distributions, as the argument below shows.

All pairs (X_i, X_{i+1}) are marginally independent in P (where we have let $X_5 = X_1$). Hence P has the same clique marginals as the distribution Q defined by $X_1, \dots, X_4$ being mutually independent and uniformly distributed on the space $\{a, b, c\}$. Clearly $Q \in M_E(\mathcal{G})$. By Lemma 3.14 there is only one element in $M_E(\mathcal{G})$ with these marginals so we must have $P \notin M_E(\mathcal{G})$. $\square$

We introduce the notation $M^+(\mathcal{G})$ for the positive Markov probabilities, $M_P(\mathcal{G})$ for the pairwise, $M_L(\mathcal{G})$ for the local, $M_G(\mathcal{G})$ for the global, and recall that we use $M_F(\mathcal{G})$ for those that factorize. We can then summarize the previous considerations as follows. In the general case we have the chain of inclusions

$$M^+(\mathcal{G}) \subset M_F(\mathcal{G}) \subseteq M_G(\mathcal{G}) \subseteq M_L(\mathcal{G}) \subseteq M_P(\mathcal{G}). \quad (3.15)$$

In the discrete case we have further that $M_G(\mathcal{G})$, $M_L(\mathcal{G})$, and $M_P(\mathcal{G})$ are all closed under limits and

$$M_F(\mathcal{G}) \subseteq M_E(\mathcal{G}) \subseteq M_G(\mathcal{G}). \quad (3.16)$$

All inclusions in (3.15) and (3.16) are strict for a general graph $\mathcal{G}$ as shown through various examples.

When (A, B, S) form a weak decomposition of $\mathcal{G}$ the Markov properties decompose accordingly. This is expressed formally in three propositions below.

Proposition 3.16 *Assume that (A, B, S) decompose $\mathcal{G} = (V, E)$ weakly. 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*

$$f(x)f_S(x_S) = f_{A \cup S}(x_{A \cup S})f_{B \cup S}(x_{B \cup S}). \quad (3.17)$$

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

$$f_{AUS}(x_{AUS}) = h(x_{AUS})\bar{k}(x_S)$$

where

$$\bar{k}(x_S) = \int k(x_{BUS})\mu_B(dx_B),$$

and similarly with the other marginals. This gives (3.17) as well as the factorizations of both marginal densities.

Conversely, assume that (3.17) holds and that f_{AUS} and f_{BUS} 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

$$\tilde{f}(x) = f_{AUS}(x_{AUS})f_{BUS}(x_{BUS})\psi_S(x_S)$$

is a density for P and P factorizes. □

The analogous result is also true for the global Markov property (G), as we shall now show.

Proposition 3.17 *Assume that (A, B, S) decompose $\mathcal{G} = (V, E)$ weakly. Then a probability distribution P is globally Markov with respect to $\mathcal{G}$ if and only if both marginal distributions P_{AUS} and P_{BUS} are globally Markov with respect to $\mathcal{G}_{AUS}$ and $\mathcal{G}_{BUS}$ respectively and the densities satisfy (3.17).*

Proof: Suppose that a probability distribution P satisfies the global Markov property with respect to $\mathcal{G}$. Then (3.17) follows because S separates A from B . To show that P_{AUS} is globally Markov with respect to $\mathcal{G}_{AUS}$ we show that for any triple $(\tilde{A}, \tilde{B}, \tilde{S})$ of disjoint subsets of $A \cup S$ with $\tilde{S}$ separating $\tilde{A}$ from $\tilde{B}$ in $\mathcal{G}_{AUS}$, $\tilde{S}$ separates $\tilde{A}$ from $\tilde{B}$ in $\mathcal{G}$. So let $(\tilde{A}, \tilde{B}, \tilde{S})$ be such a triple and let $\alpha \in \tilde{A}$ and $\beta \in \tilde{B}$. Because S is complete, at least one of them, say α , is in A . If there were a path from α to β avoiding $(A \cup S) \setminus \{\alpha, \beta\}$ it would go via B , contradicting that A is separated from B by S .

Next, assume that P_{AUS} and P_{BUS} both satisfy the global Markov property and the joint density f factorizes as in (3.17). Let U , W and C be disjoint subsets of V such that C separates U from W and assume first that $V = U \cup W \cup C$.

Since S is complete and C separates U from W , either $U \cap S$ or $W \cap S$ must be empty; we assume the former. The entire set of variables is then partitioned into eight disjoint subsets

$$U \cap A, C \cap A, W \cap A, C \cap S, W \cap S, U \cap B, C \cap B, W \cap B$$

and we denote the corresponding subsets of random variables by $X_1, \dots, X_8$; see the diagram below.

	U	C	W
A	X_1	X_2	X_3
S		X_4	X_5
B	X_6	X_7	X_8

It must hold that $C \cap (A \cup S)$ separates $U \cap A$ from $W \cap (A \cup S)$ in $\mathcal{G}_{A \cup S}$, for otherwise C would not separate U from W as assumed. Hence, from the global Markov property of $P_{A \cup S}$ and (3.6) we must have the factorization

$$f_{A \cup S}(x_{A \cup S}) = h_1(x_1, x_2, x_4)k_1(x_2, x_3, x_4, x_5).$$

By symmetry we further find

$$f_{B \cup S}(x_{B \cup S}) = h_2(x_4, x_6, x_7)k_2(x_4, x_5, x_7, x_8).$$

Combined with (3.17) this gives

$$f(x)f_S(x_S) = h_1(x_1, x_2, x_4)k_1(x_2, x_3, x_4, x_5)h_2(x_4, x_6, x_7)k_2(x_4, x_5, x_7, x_8).$$

Using that $x_S = (x_4, x_5)$ and letting $h = h_1 h_2$ and $k = k_1 k_2 / f_S$ yields

$$f(x) = h(x_1, x_2, x_4, x_6, x_7)k(x_2, x_3, x_4, x_5, x_7, x_8) = h(x_{U \cup C})k(x_{W \cup C}),$$

which shows that $U \perp\!\!\!\perp W \mid C$.

If $U \cup W \cup C$ is not the entire set of variables V we proceed as in the proof of Proposition 3.8. Details are omitted. $\square$

In the case of discrete sample spaces we have the analogous result for extended Markov probabilities.

Proposition 3.18 *Assume that (A, B, S) decompose $\mathcal{G} = (V, E)$ weakly and the sample space is discrete. Then a probability distribution P is extended Markov with respect to $\mathcal{G}$ if and only if both marginal distributions $P_{A \cup S}$ and $P_{B \cup S}$ are extended Markov with respect to $\mathcal{G}_{A \cup S}$ and $\mathcal{G}_{B \cup S}$ respectively and the densities satisfy*

$$p(i)p(i_S) = p(i_{A \cup S})p(i_{B \cup S}). \quad (3.18)$$

Proof: Assume that P is extended Markov. Then

$$p(i) = \lim_{n \rightarrow \infty} p_n(i),$$

where p_n are positive and factorize. By Proposition 3.16 we therefore have

$$p_n(i)p_n(i_S) = p_n(i_{A \cup S})p_n(i_{B \cup S}),$$

where the factors on the right-hand side factorize on $\mathcal{G}_{AUS}$ and $\mathcal{G}_{BUS}$ respectively. Hence their limits are extended Markov. Letting n tend to infinity in the equation yields thus the desired factorization.

Suppose conversely that (3.18) holds with the factors being extended Markov. Then

$$p(i_{AUS}) = \lim_{n \rightarrow \infty} q_n(i_{AUS}), \quad p(i_{BUS}) = \lim_{n \rightarrow \infty} r_n(i_{BUS}), \quad (3.19)$$

where q_n and r_n factorize appropriately. Let now

$$\pi_n(i) = \frac{q_n(i_{AUS})}{q_n(i_S)} r_n(i_{BUS}). \quad (3.20)$$

Then π_n defines a probability distribution that factorizes. We need to show that

$$p(i) = \lim_{n \rightarrow \infty} \pi_n(i).$$

This is trivial if $p(i_S) \neq 0$. If $p(i_S) = 0$ then both $p(i) = 0$ and $p(i_{BUS}) = 0$. Hence, by (3.19), the second factor in (3.20) converges to zero. The first factor in (3.20) is bounded above by one and therefore the product converges to zero as desired. $\square$

As we have seen, the various Markov properties are different for general, undirected graphs, unless the densities are all positive. However, in the special case of decomposable graphs some Markov properties coincide.

Proposition 3.19 *Let $\mathcal{G}$ be weakly decomposable. Then*

$$M_F(\mathcal{G}) = M_G(\mathcal{G}).$$

Proof: The proof is by induction on $|\mathcal{C}|$, the number of cliques in the graph $\mathcal{G}$.

The result is trivial for $|\mathcal{C}| = 1$, so let us assume the equation to be true for all graphs with at most n cliques and let $\mathcal{G}$ have $n + 1$ cliques.

We only have to show that $M_G(\mathcal{G}) \subseteq M_F(\mathcal{G})$. Assume that P is globally Markov with respect to $\mathcal{G}$ and let (A, B, S) be a proper weak decomposition of $\mathcal{G}$. By Proposition 3.17, both marginals P_{AUS} and P_{BUS} are globally Markov and (3.17) holds. By the induction assumption the marginal distributions both factorize. The full factorization of P is now obtained from (3.17) by dividing with f_S . $\square$

And a direct consequence of this and (3.16) is the corollary below.

Corollary 3.20 *Assume that $\mathcal{G}$ is weakly decomposable and the state space is discrete. Then*

$$M_F(\mathcal{G}) = M_E(\mathcal{G}) = M_G(\mathcal{G}).$$

It is an easy consequence of Example 3.10 that the converse to Proposition 3.19 holds in the sense that if the graph is not decomposable, it has a chordless cycle and one can by analogy construct a distribution which does not factorize but is globally Markov.

The global Markov property is the strongest of the Markov properties in the sense that the associated list of conditional independence statements strictly contains the statements associated with the other properties. Moreover, it cannot be further strengthened. For example it holds (Frydenberg 1990b) that if all state spaces are binary, i.e. $\mathcal{X}_\alpha = \{1, -1\}$, then

$$A \perp\!\!\!\perp B \mid S \text{ for all } P \in M_F(G) \iff S \text{ separates } A \text{ from } B. \quad (3.21)$$

In other words, if A and B are not separated by S then there is a factorizing distribution that makes them conditionally dependent. Geiger and Pearl (1993) conjecture that to any undirected graph $\mathcal{G}$ and fixed state space $\mathcal{X}$ one can find a single $P \in M_F(G)$ such that for this P it holds that

$$A \perp\!\!\!\perp B \mid S \iff S \text{ separates } A \text{ from } B.$$

This is clearly a stronger statement but no proof is known.

3.2.2 Markov properties on directed acyclic graphs

Before we proceed to the case of a general chain graph we consider the same setup as in the previous subsection, only now the graph $\mathcal{G}$ is assumed to be directed and acyclic. The Markov property on a directed acyclic graph was first studied systematically in Kiiveri *et al.* (1984) but see also for example Pearl and Verma (1987), Verma and Pearl (1990a), J.Q. Smith (1989), Geiger and Pearl (1990), Lauritzen *et al.* (1990) and other references given below.

We say that a probability distribution P admits a *recursive factorization* according to $\mathcal{G}$, if there exist non-negative functions, henceforth referred to as *kernels*, $k^\alpha(\cdot, \cdot), \alpha \in V$ defined on $\mathcal{X}_\alpha \times \mathcal{X}_{\text{pa}(\alpha)}$, such that

$$\int k^\alpha(y_\alpha, x_{\text{pa}(\alpha)}) \mu_\alpha(dy_\alpha) = 1$$

and P has density f with respect to μ , where

$$f(x) = \prod_{\alpha \in V} k^\alpha(x_\alpha, x_{\text{pa}(\alpha)}).$$

We then also say that P has property (DF). It is an easy induction argument to show that if P admits a recursive factorization as above, then the kernels $k^\alpha(\cdot, x_{\text{pa}(\alpha)})$ are densities for the conditional distribution of X_α , given $X_{\text{pa}(\alpha)} = x_{\text{pa}(\alpha)}$. Also it is immediate that if we form the undirected moral graph $\mathcal{G}^m$ (marrying parents and deleting directions) such as described towards the end of Section 2.1.1, we have

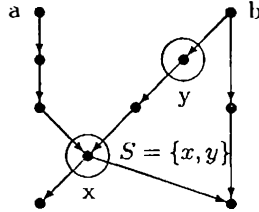


Fig. 3.1. The directed global Markov property. Is $a \perp\!\!\!\perp b \mid S$?

Lemma 3.21 *If P admits a recursive factorization according to the directed, acyclic graph $\mathcal{G}$, it factorizes according to the moral graph $\mathcal{G}^m$ and obeys therefore the global Markov property relative to $\mathcal{G}^m$.*

Proof: The factorization follows from the fact that, by construction, the sets $\{\alpha\} \cup \text{pa}(\alpha)$ are complete in $\mathcal{G}^m$ and we can therefore let $\psi_{\{\alpha\} \cup \text{pa}(\alpha)} = k^\alpha$. The remaining part of the statement follows from the fact that (F) implies (G) in the undirected case; see Proposition 3.8. $\square$

It clearly also holds that

Proposition 3.22 *If P admits a recursive factorization according to the directed, acyclic graph $\mathcal{G}$ and A is an ancestral set, then the marginal distribution P_A admits a recursive factorization according to $\mathcal{G}_A$.*

From this it directly follows that

Corollary 3.23 *Let P factorize recursively according to $\mathcal{G}$. Then*

$$A \perp\!\!\!\perp B \mid S$$

whenever A and B are separated by S in $(\mathcal{G}_{\text{An}(A \cup B \cup S)})^m$, the moral graph of the smallest ancestral set containing $A \cup B \cup S$.

The property in Corollary 3.23 will be referred to as the *directed global Markov property* (DG). The directed global Markov property has the same role as the global Markov property has in the case of an undirected graph, in the sense that it gives the sharpest possible rule for reading conditional independence relations off the directed graph. The procedure is illustrated in the following example.

Example 3.24 Consider a directed Markov field on the graph in Fig. 3.1 and the problem of deciding whether $a \perp\!\!\!\perp b \mid S$. The moral graph of the smallest ancestral set containing all the variables involved is shown in Fig. 3.2. It is immediate that S separates a from b in this graph, implying $a \perp\!\!\!\perp b \mid S$. $\square$

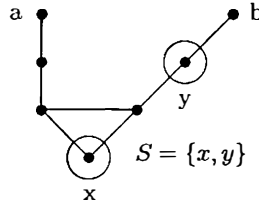


Fig. 3.2. The moral graph of the smallest ancestral set in the graph of Fig. 3.1 containing $\{a\} \cup \{b\} \cup S$. Clearly S separates a from b in this graph, implying $a \perp\!\!\!\perp b \mid S$.

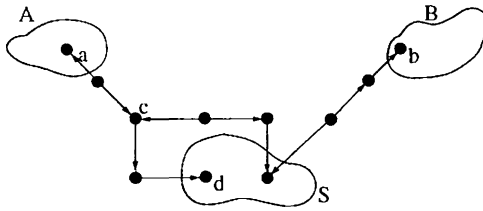


Fig. 3.3. Example of an active chain from A to B . The path from c to d is not part of the chain, but indicates that c must have descendants in S .

An alternative formulation of the directed global Markov property was given by Pearl (1986a, 1986b) with a full formal treatment in Verma and Pearl (1990a, 1990b). A chain π from a to b in a directed, acyclic graph $\mathcal{G}$ is said to be *blocked* by S , if it contains a vertex $\gamma \in \pi$ such that either

- $\gamma \in S$ and arrows of π do not meet head-to-head at γ , or
- $\gamma \notin S$ nor has γ any descendants in S , and arrows of π do meet head-to-head at γ .

A chain that is not blocked by S is said to be *active*. Two subsets A and B are now said to be *d-separated* by S if all chains from A to B are blocked by S . We then have

Proposition 3.25 *Let A , B and S be disjoint subsets of a directed, acyclic graph $\mathcal{G}$. Then S d-separates A from B if and only if S separates A from B in $(\mathcal{G}_{\text{An}(A \cup B \cup S)})^m$.*

Proof: Suppose S does not d-separate A from B . Then there is an active chain from A to B such as, for example, indicated in Fig. 3.3. All vertices in this chain must lie within $\text{An}(A \cup B \cup S)$. This follows because if the arrows

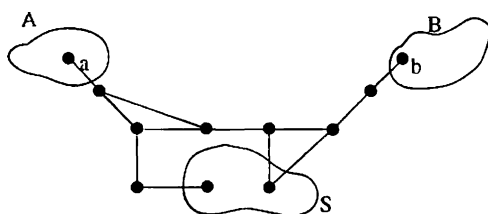


Fig. 3.4. The moral graph corresponding to the active chain in $\mathcal{G}$.

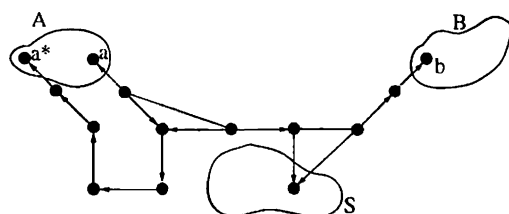


Fig. 3.5. The chain in the graph $(\mathcal{G}_{An(A \cup B \cup S)})^m$ makes it possible to construct an active chain in $\mathcal{G}$ from A to B .

meet head-to-head at some vertex γ , either $\gamma \in S$ or γ has descendants in S . And if not, either of the subpaths away from γ either meets another arrow, in which case γ has descendants in S , or leads all the way to A or B . Each of these head-to-head meetings will give rise to a marriage in the moral graph such as illustrated in Fig. 3.4, thereby creating a chain from A to B in $(\mathcal{G}_{An(A \cup B \cup S)})^m$, circumventing S .

Suppose conversely that A is not separated from B in $(\mathcal{G}_{An(A \cup B \cup S)})^m$. Then there is a chain in this graph that circumvents S . The chain has pieces that correspond to edges in the original graph and pieces that correspond to marriages. Each marriage is a consequence of a meeting of arrows head-to-head at some vertex γ . If γ is in S or it has descendants in S , the meeting does not block the chain. If not, γ must have descendants in A or B , since the ancestral set was smallest. In the latter case, a new chain can be created with one head-to-head meeting fewer, using the line of descent, such as illustrated in Fig. 3.5. Continuing this substitution process eventually leads to an active chain from A to B and the proof is complete. $\square$

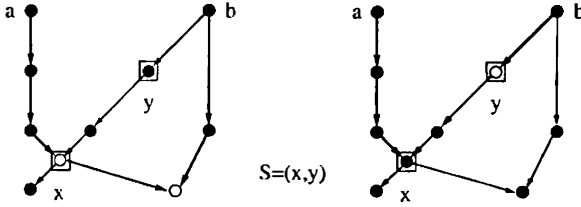


Fig. 3.6. Illustration of Pearl's separation criterion. There are two chains from a to b , drawn with bold lines. Both are blocked, but different vertices γ , indicated with open circles, play the role of blocking vertices.

We illustrate the concept of d-separation by applying it to the query of Example 3.24. As Fig. 3.6 indicates, all chains between a and b are blocked by S , whereby the global Markov property gives that $a \perp\!\!\!\perp b \mid S$.

Geiger and Pearl (1990) show in their Theorem 5 that the criterion of d-separation cannot be improved, in the sense that for any given directed acyclic graph $\mathcal{G}$, one can find state spaces $\mathcal{X}_\alpha, \alpha \in V$ and a probability P such that

$$A \perp\!\!\!\perp B \mid S \iff S \text{ d-separates } A \text{ from } B. \quad (3.22)$$

An argument analogous to that given in Frydenberg (1990b) shows that the analogue of (3.21) holds for d-separation in directed, acyclic graphs. Geiger and Pearl (1990) show that in the case of real sample spaces, a Gaussian distribution satisfying (3.22) exists and conjecture the similar result to be true also for the case where the state spaces all have two points, i.e. $\mathcal{X}_\alpha = \{1, -1\}$.

Also the directed case has analogues of the pairwise and local Markov properties. We say that P obeys the *directed pairwise Markov property* (DP) if for any pair (α, β) of non-adjacent vertices with $\beta \in \text{nd}(\alpha)$,

$$\alpha \perp\!\!\!\perp \beta \mid \text{nd}(\alpha) \setminus \{\beta\}.$$

Similarly P obeys the *directed local Markov property* (DL) if any variable is conditionally independent of its non-descendants, given its parents:

$$\alpha \perp\!\!\!\perp \text{nd}(\alpha) \mid \text{pa}(\alpha).$$

As $\text{pa}(\alpha) \subseteq \text{nd}(\alpha)$, it follows from properties (C2) and (C3) of conditional independence that the directed local Markov property (DL) implies the pairwise (DP). The converse is not true in general, as the following example shows.

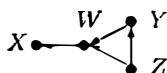


Fig. 3.7. The local and pairwise directed Markov properties are not equivalent in general. This is illustrate by the directed acyclic graph above.

Example 3.26 Let $X = Y = Z$ and W be independent of X with

$$P\{X = 1\} = P\{X = 0\} = P\{W = 1\} = P\{W = 0\} = 1/2.$$

This distribution is pairwise Markov with respect to the graph in Fig. 3.7. However, it is not locally Markov, as this would imply $X \perp\!\!\!\perp (Y, Z) | W$, which clearly is not the case. A consequence of Theorem 3.27 is that the density in this example does not admit a recursive factorization. $\square$

In contrast to the undirected case we have that the remaining three properties (DF), (DL) and (DG) are equivalent, just assuming the existence of the density f .

Theorem 3.27 *Let $\mathcal{G}$ 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:*

- (DF) P admits a recursive factorization according to $\mathcal{G}$;
- (DG) P obeys the directed global Markov property, relative to $\mathcal{G}$;
- (DL) P obeys the directed local Markov property, relative to $\mathcal{G}$.

Proof: That (DF) implies (DG) is Corollary 3.23. That (DG) implies (DL) follows by observing that $\{\alpha\} \cup \text{nd}(\alpha)$ is an ancestral set and that $\text{pa}(\alpha)$ obviously separates $\{\alpha\}$ from $\text{nd}(\alpha) \setminus \text{pa}(\alpha)$ in $(\mathcal{G}_{\{\alpha\} \cup \text{nd}(\alpha)})^m$. The final implication is shown by induction on the number of vertices $|V|$ of $\mathcal{G}$. Let α_0 be a terminal vertex of $\mathcal{G}$. Then we can let k^{α_0} be the conditional density of X_{α_0} , given $X_{V \setminus \{\alpha_0\}}$, which by (DL) can be chosen to depend on $x_{\text{pa}(\alpha_0)}$ only. The marginal distribution of $X_{V \setminus \{\alpha_0\}}$ trivially obeys the directed local Markov property and admits a factorization by the inductive assumption. Combining this factorization with k^{α_0} yields the factorization for P . This completes the proof. $\square$

In fact (DL) and (DG) are equivalent, even without assuming the existence of a density with respect to a product measure (Lauritzen *et al.* 1990).

If the probability distribution P satisfies (3.10), for example if the density is positive, then it is not difficult to see that (DP) implies (DL) and all directed Markov properties are equivalent.

Since the three conditions in Theorem 3.27 are all equivalent, we choose to speak of a *directed Markov distribution* as one where any of the conditions (DF), (DL) or (DG) is satisfied. The set of such distributions is denoted by $M(\mathcal{G})$ where $\mathcal{G}$ is a directed and acyclic graph.

It follows as in Proposition 3.12 that in the case of a discrete sample space, the directed Markov properties are preserved under weak limits. So there is no need for a special term or symbol for extended Markov probabilities on a directed acyclic graph.

In the particular case when the directed acyclic graph $\mathcal{G}$ is perfect (see Section 2.1.3) the directed Markov property on $\mathcal{G}$ and the factorization Markov property on its undirected version $\mathcal{G}^\sim$ coincide. This is contained in the following

Proposition 3.28 *Let $\mathcal{G}$ be a perfect directed acyclic graph and $\mathcal{G}^\sim$ its undirected version. Then P admits a recursive factorization with respect to $\mathcal{G}$ if and only if it factorizes according to $\mathcal{G}^\sim$.*

Proof: That the graph is perfect means that $\text{pa}(\alpha)$ is complete for all $\alpha \in V$. Hence $\mathcal{G}^m = \mathcal{G}^\sim$. From Lemma 3.21 it then follows that any $P \in M(\mathcal{G})$ also factorizes with respect to $\mathcal{G}^\sim$.

The reverse inclusion is established by induction on the number of vertices $|V|$ of $\mathcal{G}$. For $|V| = 1$ there is nothing to show. For $|V| = n + 1$ let $P \in M_F(\mathcal{G}^\sim)$ and find a terminal vertex $\alpha \in V$. This vertex has $\text{pa}_{\mathcal{G}}(\alpha) = \text{bd}_{\mathcal{G}^\sim}(\alpha)$ and, since $\mathcal{G}$ is perfect, this set is complete in both graphs as well. Hence $(V \setminus \{\alpha\}, \{\alpha\}, \text{bd}(\alpha))$ is a weak decomposition of $\mathcal{G}^\sim$ and Proposition 3.16 gives the factorization

$$f(x) = f(x_{V \setminus \{\alpha\}})f(x_{\text{cl}(\alpha)})/f(x_{\text{bd}(\alpha)}),$$

where the first factor factorizes according to $\mathcal{G}_{V \setminus \{\alpha\}}^\sim$. Using the induction assumption on this factor gives the full recursive factorization of P . $\square$

In the discrete case Proposition 3.18 implies that the above result can be strengthened to comprise the extended Markov property.

Proposition 3.29 *Let $\mathcal{G}$ be a perfect directed acyclic graph and $\mathcal{G}^\sim$ its undirected version. If all vertices are discrete and hence the total sample space, then*

$$M(\mathcal{G}) = M_F(\mathcal{G}^\sim) = M_E(\mathcal{G}^\sim).$$

In particular, any extended Markov probability factorizes.

Proof: The proof that $M(\mathcal{G}) = M_E(\mathcal{G}^\sim)$ is completely analogous to the proof of the preceding proposition. Since $M(\mathcal{G}) \subseteq M_F(\mathcal{G}^\sim) \subseteq M_E(\mathcal{G}^\sim)$ the result follows. $\square$

3.2.3 Markov properties on chain graphs

In the present section we deal with the Markov properties for a general chain graph $\mathcal{G} = (V, E)$ thereby unifying the directed and undirected cases. A detailed study of the Markov property on chain graphs can be found in Frydenberg (1990a). Here we give the most basic results and definitions.

The factorization in the case of a chain graph is more complex and involves two parts. Denote the set of chain components of the graph by $\mathcal{T}$.

Then we first assume a factorization of the density as in the directed acyclic case:

$$f(x) = \prod_{\tau \in \mathcal{T}} f(x_\tau | x_{\text{pa}(\tau)}). \quad (3.23)$$

But there is one more assumption. To explain this assumption, let τ_* denote the undirected graph that has $\tau \cup \text{pa}(\tau)$ as nodes and undirected edges between a pair (α, β) if either both of these are in $\text{pa}(\tau)$ or there is an edge, directed or undirected, between them in the chain graph $\mathcal{G}$. Thus for chain components that are singletons, τ_* is complete. This is the case for all chain components if $\mathcal{G}$ is a directed acyclic graph. In an undirected graph, τ_* is just the subgraph induced by the connected component τ .

The second requirement in the chain graph factorization is then that the factors in (3.23) can be further factorized as

$$f(x_\tau | x_{\text{pa}(\tau)}) = \prod_a \phi_a(x), \quad (3.24)$$

where a varies over all subsets of $\tau \cup \text{pa}(\tau)$ that are complete in τ_* , and $\phi_a(x)$ as usual represent functions that depend on x through x_a only.

This factorization clearly unifies the directed and undirected cases. A probability distribution P on $\mathcal{X}$ that has a density f with respect to a product measure which satisfies (3.23) and (3.24) is said to *factorize* according to $\mathcal{G}$, and we also say that P has property (FC).

The chain graph factorization has several equivalent formulations. To describe these, let the vertex set be partitioned in a dependence chain as $V = V(1) \cup \dots \cup V(T)$ such that each of the sets $V(t)$ only has lines between vertices, and arrows point from vertices in sets with lower number to those with higher number. The set of *concurrent* variables relative to this partitioning is next defined to be the set $C(t) = V(1) \cup \dots \cup V(t)$. Let $\mathcal{G}^*(t)$ be the undirected graph with vertex set $C(t)$ and α adjacent to β in $\mathcal{G}^*(t)$ if either $(\alpha, \beta) \in E$ or $(\beta, \alpha) \in E$ or if $\{\alpha, \beta\} \subseteq C(t-1)$, i.e. $C(t-1)$

is made complete in $\mathcal{G}^*(t)$ by adding all missing edges between these and directions on existing edges are ignored. Let further $B(t) = \text{pa}\{V(t)\}$ and $\mathcal{G}_*(t) = \mathcal{G}^*(t)_{V(t) \cup B(t)}$. Then we have the following equivalent formulations of the chain graph factorization. There are obviously others, for example described through conditioning. We abstain from listing them all.

Proposition 3.30 *For a probability distribution P on $\mathcal{X}$ that admits a density with respect to a product measure, the following conditions are equivalent:*

- (i) P factorizes according to $\mathcal{G}$;
- (ii) P admits a density that factorizes as

$$f(x) = \prod_{\tau \in \mathcal{T}} \frac{f(x_{\tau \cup \text{pa}(\tau)})}{f(x_{\text{pa}(\tau)})} \quad (3.25)$$

and each of the denominators factorizes on the graph τ_* ;

- (iii) for any dependence chain, P admits a density that factorizes as

$$f(x) = \prod_{t=1}^T \frac{f(x_{C(t)})}{f(x_{C(t-1)})} \quad (3.26)$$

and each of the denominators factorizes on the graph $\mathcal{G}^*(t)$;

- (iv) for any dependence chain, P admits a density that factorizes as

$$f(x) = \prod_{t=1}^T \frac{f(x_{V(t) \cup B(t)})}{f(x_{B(t)})} \quad (3.27)$$

and each of the denominators factorizes on the graph $\mathcal{G}_*(t)$;

- (v) for some dependence chain, P admits a density that factorizes as in either of (iii) or (iv).

Proof: We leave the details of this to the reader. □

Also in the case of chain graphs, there are pairwise, local and global Markov properties, in an even greater variety. We say that a probability P satisfies

- (PB) the *pairwise block-recursive Markov property* relative to the dependence chain $V(1), \dots, V(T)$, if for any pair (α, β) of non-adjacent vertices we have

$$\alpha \perp\!\!\!\perp \beta \mid C(t^*) \setminus \{\alpha, \beta\},$$

where t^* is the smallest t that has $\alpha, \beta \in C(t)$;

(PC) *the pairwise chain Markov property*, relative to $\mathcal{G}$, if for any pair (α, β) of non-adjacent vertices with $\beta \in \text{nd}(\alpha)$

$$\alpha \perp\!\!\!\perp \beta \mid \text{nd}(\alpha) \setminus \{\beta\};$$

(LC) *the local chain Markov property*, relative to $\mathcal{G}$, if for any vertex $\alpha \in V$

$$\alpha \perp\!\!\!\perp \text{nd}(\alpha) \mid \text{bd}(\alpha);$$

(GC) *the global chain Markov property*, relative to $\mathcal{G}$, if for any triple (A, B, S) of disjoint subsets of V such that S separates A from B in $(\mathcal{G}_{\text{An}(A \cup B \cup S)})^m$, the moral graph of the smallest ancestral set containing $A \cup B \cup S$, we have

$$A \perp\!\!\!\perp B \mid S.$$

In general, the property (PB) depends on the particular dependence chain.

Recently Bouckaert and Studený (1995) have given a separation criterion which is analogous to d-separation and equivalent to the global chain Markov property (GC). They also show that any conditional independence statement that is derivable from the local chain Markov property and the properties (C1)–(C5) is represented in the global chain Markov property (GC). Further, they show that probability distributions exist that satisfy the local chain Markov property and violate all conditional independence statements that are not of the form given by (GC). In this sense, the global chain Markov property is therefore the strongest possible.

Note that the Markov properties unify the corresponding properties for the directed and undirected cases. For in the undirected case we have that $\text{nd}(\alpha) = V \setminus \{\alpha\}$ and $\mathcal{G} = (\mathcal{G}_{\text{An}(A \cup B \cup S)})^m$. And in the directed case $\text{bd}(\alpha) = \text{pa}(\alpha)$.

Generally all these Markov properties are therefore different without additional assumptions, just as in the undirected case. In the remaining part of the section we assume that all probability measures have positive densities, implying that all five of the basic properties of conditional independence (C1)–(C5) hold. As we shall see, this implies that all Markov properties are equivalent. We first need a few lemmas.

Lemma 3.31 *If $V(1), \dots, V(T)$ is a dependence chain for $\mathcal{G}$ and P satisfies the pairwise block-recursive Markov property relative to $\mathcal{G}$, then $P_{C(T-1)}$ satisfies the pairwise block-recursive Markov property relative to $\mathcal{G}^*(T-1)$.*

Proof: This is trivial, as the pairwise block-recursive Markov property is defined to be identical to the combination of the pairwise undirected

Markov properties for the marginal distributions $P_{C(t)}$ of the concurrent variables $C(t)$ on the graphs $\mathcal{G}^*(t)$, for $t = 1, \dots, T$. $\square$

In the case of the pairwise chain graph Markov property, the corresponding result needs an extra condition.

Lemma 3.32 *If C is a terminal chain component in $\mathcal{G}$ and P satisfies (3.10) and the pairwise chain Markov property relative to $\mathcal{G}$, then $P_{V \setminus C}$ satisfies the pairwise chain Markov property relative to $\mathcal{G}_{V \setminus C}$.*

Proof: Let $\alpha, \beta \in V \setminus C$ be non-adjacent in $\mathcal{G}$ and assume $\beta \in \text{nd}(\alpha)$. We have to show that $\alpha \perp\!\!\!\perp \beta \mid (\text{nd}(\alpha) \setminus (C \cup \{\beta\}))$, since $\text{nd}(\alpha) \setminus C$ is the set of non-descendants of α in $\mathcal{G}_{V \setminus C}$. If $C \cap \text{nd}(\alpha) = \emptyset$ this is a direct consequence of the fact that (PC) holds for $\mathcal{G}$. Else we must have $C \subseteq \text{nd}(\alpha)$. Since no pair δ, γ with $\delta \in \text{dc}(\alpha) \cup \{\alpha\}$ and $\gamma \in C$ can be adjacent and $\text{nd}(\gamma) = V \setminus \{\gamma\}$ for all such γ , the pairwise chain Markov property implies that $\delta \perp\!\!\!\perp \gamma \mid V \setminus \{\delta, \gamma\}$ for any such pair. Repeated use of (3.10) yields

$$\text{dc}(\alpha) \cup \{\alpha\} \perp\!\!\!\perp C \mid \text{nd}(\alpha) \setminus C$$

and thus by (C2),

$$\alpha \perp\!\!\!\perp C \mid \text{nd}(\alpha) \setminus C. \quad (3.28)$$

The pairwise property gives directly that

$$\alpha \perp\!\!\!\perp \beta \mid \text{nd}(\alpha) \setminus \{\beta\}. \quad (3.29)$$

Thus (3.10) used on (3.28) and (3.29) yields

$$\alpha \perp\!\!\!\perp (C \cup \{\beta\}) \mid \text{nd}(\alpha) \setminus (C \cup \{\beta\})$$

and the result follows from (C2). $\square$

Lemma 3.33 *If P satisfies the pairwise chain Markov property (PC) with respect to $\mathcal{G}$ as well as (3.10), it satisfies the pairwise undirected Markov property (P) with respect to $\mathcal{G}^m$.*

Proof: We use induction on the number of chain components. If there is only one chain component in $\mathcal{G}$, it is undirected and connected, $\mathcal{G}^m = \mathcal{G}$ and there is nothing to show.

Assume then the lemma to hold for all graphs with n or fewer chain components and let $\mathcal{G}$ have $n + 1$ components. Let α, β be non-adjacent in $\mathcal{G}^m$. Without loss of generality we assume that $\beta \in \text{nd}(\alpha)$. If $V \setminus \{\alpha\} = \text{nd}(\alpha)$, the conditional independence $\alpha \perp\!\!\!\perp \beta \mid V \setminus \{\alpha, \beta\}$ follows directly. Else there must be a terminal chain component $C \subseteq (V \setminus (\{\alpha\} \cup \text{nd}(\alpha)))$.

Because α and β are non-adjacent in $\mathcal{G}^m$ they will also be in $(\mathcal{G}_{V \setminus C})^m$. The induction assumption together with Lemma 3.32 gives that

$$\alpha \perp\!\!\!\perp \beta \mid V \setminus (C \cup \{\alpha, \beta\}). \quad (3.30)$$

Also at least one of α and β , say the former, cannot be in $\text{bd}(C)$, since otherwise they would be adjacent in the moral graph. Thus

$$\alpha \perp\!\!\!\perp \gamma \mid V \setminus \{\alpha, \gamma\} \text{ for all } \gamma \in C. \quad (3.31)$$

Repeated use of property (3.10) of conditional independence on (3.31) yields

$$\alpha \perp\!\!\!\perp C \mid V \setminus (C \cup \{\alpha\}). \quad (3.32)$$

We can now use (C4) on (3.30) and (3.32) to obtain that

$$\alpha \perp\!\!\!\perp \{\beta\} \cup C \mid V \setminus (C \cup \{\alpha, \beta\}).$$

From (C3) and (C2) it follows that $\alpha \perp\!\!\!\perp \beta \mid V \setminus \{\alpha, \beta\}$, which was to be shown. $\square$

Clearly, from arguments analogous to the directed and undirected cases, we have in general that

$$(\text{FC}) \implies (\text{GC}) \implies (\text{PC}) \implies (\text{PB}) \text{ for any dependence chain,}$$

but if we assume (3.10), all Markov properties are equivalent.

Theorem 3.34 *Assume that P is such that (3.10) holds for disjoint subsets of V ; then*

$$(\text{GC}) \iff (\text{LC}) \iff (\text{PC}) \iff (\text{PB}) \text{ for some dependence chain.}$$

Proof: We must show that the pairwise block-recursive property for any dependence chain implies the global chain Markov property (GC). We argue first that (PB) implies (PC) and then that (PC) implies (GC).

The first part is shown by using induction on the number of chain components of $\mathcal{G}$. If there is only one chain component the statement is trivially true. Assume that (PB) implies (PC) for all chain graphs with at most n chain components and let $\mathcal{G}$ have $n+1$ chain components. Assume then that P satisfies (PB) with respect to a given dependence chain and let α and β be non-adjacent vertices.

If $t^* < T$, marginalization to $C(t^*)$ and the inductive assumption will give that

$$\alpha \perp\!\!\!\perp \beta \mid \text{nd}(\alpha) \setminus \{\beta\}.$$

If $t^* = T$ and $V(T)$ has only one chain component, the same conclusion can be drawn because then $\text{nd}(\alpha) = C(T) \setminus \{\alpha\}$.

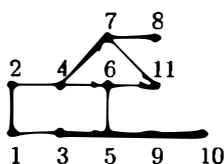


Fig. 3.8. A chain graph. The chain components are $\{1,2,3,4\}$, $\{5,6\}$, $\{7,8\}$, $\{9,10\}$, $\{11\}$. Is $3 \perp\!\!\!\perp 8 \mid \{2,5\}$? Is $3 \perp\!\!\!\perp 8 \mid 2$?

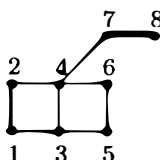


Fig. 3.9. The moral graph of the smallest ancestral set in the graph of Fig. 3.8 containing $\{2,3,5,8\}$. A connection between 3 and 4 has been introduced since these both have children in the same chain component $\{5,6\}$. We cannot conclude $3 \perp\!\!\!\perp 8 \mid \{2,5\}$.

If there is more than one chain component in $V(T)$, Theorem 3.7 yields that P is globally Markov with respect to $\mathcal{G}^*(T)$ and hence the same conclusion holds. Thus P satisfies (PC).

Assume next that P satisfies (PC). If R is any ancestral set in $\mathcal{G}$, it can be obtained from V by stepwise removal of terminal chain components. By Lemma 3.32 each removal preserves the pairwise property. Combining with Lemma 3.33 we find that P_R is pairwise Markov with respect to $(\mathcal{G}_R)^m$. Theorem 3.7 yields that P_R is globally Markov relative to the same graph. Thus, the result follows by letting $R = \text{An}(A \cup B \cup S)$. $\square$

Example 3.35 As an illustration of the global chain Markov property, consider the graph in Fig. 3.8 and the question of deciding whether it holds that $3 \perp\!\!\!\perp 8 \mid \{2,5\}$. The smallest ancestral set containing these variables is the set $\{1,2,3,4,5,6,7,8\}$. The moral graph of this adds an edge between 3 and 4, because these both have children in the chain component $\{5,6\}$. Thus the graph in Fig. 3.9 appears. Since there is a path between 3 and 8 circumventing 2 and 5 in this graph, we cannot conclude that $3 \perp\!\!\!\perp 8 \mid \{2,5\}$.

If we instead consider the question whether $3 \perp\!\!\!\perp 8 \mid 2$, the smallest ancestral set becomes $\{1,2,3,4,7,8\}$, no edge has to be added between 3 and 4 and Fig. 3.10 reveals that $3 \perp\!\!\!\perp 8 \mid 2$. $\square$

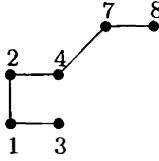


Fig. 3.10. The moral graph of the smallest ancestral set in the graph of Fig. 3.8 containing $\{2,3,8\}$. We conclude that $3 \perp\!\!\!\perp 8 \mid 2$.

We cannot expect factorization results to be more general for chain graphs than for undirected graphs, since the chain graphs contain these as special cases. But there is a result analogous to Theorem 3.9.

Theorem 3.36 *A probability distribution with strictly positive and continuous density f satisfies the pairwise block-recursive Markov property with respect to $\mathcal{G}$ if and only if it factorizes according to $\mathcal{G}$.*

Proof: That any chain graph Markov density factorizes as in (3.26) is immediate. That the denominators also factorize appropriately is seen as follows. Since the density is assumed positive and continuous, (3.10) holds and the different Markov properties are equivalent. From the pairwise block-recursive Markov property it follows that any two variables α and β that are not adjacent in $\mathcal{G}^*(t)$ are conditionally independent given the remaining concurrent variables $C(t)$. But since $V(t) \perp\!\!\!\perp C(t-1) \mid B(t)$ we have that

$$\mathcal{L}(X_\alpha, X_\beta \mid X_{C(t) \setminus \{\alpha, \beta\}}) = \mathcal{L}(X_\alpha, X_\beta \mid X_{(V(t) \cup B(t)) \setminus \{\alpha, \beta\}})$$

and hence the marginal distribution of the variables in $V(t) \cup B(t)$ is pairwise Markov with respect to the undirected graph $\mathcal{G}^*(t)$. Hence the densities factorize by Theorem 3.9.

Conversely, assume a factorization given with the stated properties and let α, β be non-adjacent in $\mathcal{G}$. Let t^* be the smallest t with $\alpha, \beta \in C(t)$. Clearly, if the density factorizes, so does the marginal density of $C(t)$ for all t and hence we can assume that $t^* = T$. If α and β are not both in $V(T) \cup B(T)$, the expression (3.27) directly gives that α and β are independent given the remaining variables in $C(t^*) = C(T) = V$. Else if $\alpha, \beta \in V(T) \cup B(T)$ they are not both in $B(T)$ since $t^* = T$. Therefore they are also non-adjacent in $\mathcal{G}^*(T)$. Hence $f(x_{V(T) \cup B(T)})$ is a product of functions, one of which does not depend on β and one of which does not depend on α . Since this is the only factor in (3.27) containing both variables, the same is true for the full joint density. Hence α and β are conditionally independent given the remaining variables, and the pairwise block-recursive Markov property has been established. $\square$

If P is not strictly positive but extended Markov, i.e. P is the limit of positive chain graph Markov distributions, there is a modified version of Theorem 3.36:

Corollary 3.37 *A probability distribution on a discrete sample space satisfies the extended Markov property with respect to a chain graph $\mathcal{G}$ if and only if it factorizes as*

$$p(i) = \prod_{t=1}^T \frac{p(i_{V(t) \cup B(t)})}{p(i_{B(t)})}, \quad (3.33)$$

where $0/0 = 0$ and each of the terms in the denominator are extended Markov with respect to the graph $\mathcal{G}^*(t)$.

Proof: This follows as in the undirected case from Theorem 3.36 by taking limits. $\square$

Also this factorization can be written in a number of equivalent ways.

3.3 Notes

There are several aspects of graphs and conditional independence that have not been covered here. Cox and Wermuth (1993) discuss alternative ways of encoding conditional independence assumptions through graphs. In Andersson and Perlman (1993), conditional independence models for the multivariate normal distribution are determined through certain distributive lattices of subspaces. The precise relation between such conditional independence restrictions and those given by graphs is discussed in Andersson *et al.* (1995b). Basically, conditional independence restrictions given by lattices correspond to those given by directed acyclic graphs that are transitive, i.e. they satisfy that $\alpha \mapsto \beta$ implies $\alpha \rightarrow \beta$.

An important question is related to the notion of *Markov equivalence*. Two graphs are called Markov equivalent if they induce the same conditional independence restrictions. Conditions for two chain graphs to be Markov equivalent were given by Frydenberg (1990a) and further results related to this notion are given in Andersson *et al.* (1995a, 1996).

Recently, Koster (1996) has studied Markov properties of *reciprocal graphs*. Those generalize chain graphs but permit directed cycles and appear therefore to be natural objects for studying systems with feedback, such as encountered in connection with structural equation models (Goldberger and Duncan 1973; Jöreskog 1977). Paz and Geva (1996) study *annotated graphs* that are even more general.

Another recent development is the systematic use of directed acyclic graphs for the interpretation, conjecture and discovery of causal relations

(Spirtes *et al.* 1993; Pearl 1995; Shafer 1996). This development is a natural extension of earlier work by Wright (1921), Wold (1954, 1960), Blalock (1971), and others.

Appendix A

Various prerequisites

A.1 Inequalities

Some elementary inequalities will be of frequent use. First recall that for $x > 0$

$$\log x \leq x - 1 \quad (\text{A.1})$$

and (A.1) is strict unless $x = 1$. Here, and throughout the text, $\log$ denotes the natural logarithm. This follows from Taylor's formula, expanding around $x = 1$:

$$\log x = x - 1 - \frac{(x - 1)^2}{2\theta^2}$$

for some θ between 1 and x . It follows that for $x, y \geq 0$ we have

$$x^y e^{-x} \leq y^y e^{-y} \quad (\text{A.2})$$

and (A.2) is strict unless $x = y$. This is seen by first considering the cases where at least one of x and y is zero, and using (A.1) to see that if x and y are both positive, we have

$$y \log \left(\frac{x}{y} \right) \leq y \left(\frac{x}{y} - 1 \right) = x - y.$$

Taking exponentials on both sides yields (A.2). Also we will make much use of the following inequality.

Lemma A.1 *Let $(a(i))_{i \in \mathcal{I}}$ and $(b(i))_{i \in \mathcal{I}}$ be vectors of non-negative numbers with $\sum_i a(i) = \sum_i b(i) < \infty$. Then*

$$\prod_{i \in \mathcal{I}} a(i)^{b(i)} \leq \prod_{i \in \mathcal{I}} b(i)^{b(i)}. \quad (\text{A.3})$$

The inequality is strict unless $a(i) = b(i)$ for all $i \in \mathcal{I}$.

Proof: From (A.2) we get

$$\prod_{i \in \mathcal{I}} a(i)^{b(i)} e^{-a(i)} \leq \prod_{i \in \mathcal{I}} b(i)^{b(i)} e^{-b(i)}.$$

The inequality (A.3) now follows upon dividing both sides with the common positive constant $\exp\{-\sum_i a(i)\}$. $\square$

The inequality (A.3) is a version of the so-called *information inequality*. Taking logarithms on both sides of (A.3) yields the version

$$\sum_{i \in \mathcal{I}} b(i) \log a(i) \leq \sum_{i \in \mathcal{I}} b(i) \log b(i), \quad (\text{A.4})$$

holding for non-negative a and b with $\sum_i a(i) = \sum_i b(i)$. It should be noticed that $0 \log 0 = 0$ and the inequality is strict unless $a(i) \equiv b(i)$.

A.2 Kullback–Leibler divergence

The Kullback–Leibler *divergence* $D(P, Q)$ of the probability measure P from Q is defined as

$$D(P, Q) = \int p \log \frac{p}{q} d\mu,$$

where P and Q have densities p and q with respect to the σ -finite measure μ . If P has density with respect to Q this may also be written as

$$D(P, Q) = \int \frac{dP}{dQ} \log \frac{dP}{dQ} dQ.$$

If not, $D(P, Q) = \infty$. If also Q has density with respect to P , the divergence can alternatively be expressed as

$$D(P, Q) = \int \log \frac{dP}{dQ} dP.$$

The notion was introduced by Kullback and Leibler (1951) although they use the term divergence for the symmetrized version $D(P, Q) + D(Q, P)$.

The divergence is always non-negative and $D(P, Q) = 0$ if and only if $P = Q$. In the discrete case this is a direct consequence of the inequality (A.4). Even though neither D nor its symmetrized version is a true metric as they do not satisfy the triangle inequality, the divergence D has a clear geometric interpretation as a non-symmetric distance measure, and a version of Pythagoras's theorem applies to the divergence (Chentsov 1972; Csiszár 1975).

When in the discrete case P is taken to be the empirical distribution and Q is varying in a family specified by a statistical model, then

$$D(P, Q) = -\frac{1}{n} \log L(Q),$$

where L denotes the likelihood function. Hence minimizing $D(P, Q)$ over Q corresponds to maximum likelihood estimation. Reversing the role of

P and Q leads to minimum discriminant information estimation (Kullback 1959; Ireland *et al.* 1969). See also Brown (1986) and E.S. Christensen (1989).

A.3 Möbius inversion

An important combinatorial trick is contained in the following

Lemma A.2 (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. $\square$

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 V , the vector space of linear maps on V or the vector space of symmetric $k \times k$ 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.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 $\Theta_i(\theta^*)$, $i = 1, \dots, k$ 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 $i = 1, \dots, k$,

$$L\{T_i(\theta^*)\} > L(\theta), \quad \text{if } \theta \neq T_i(\theta^*).$$

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.3 *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

$$L\{S(\theta^*)\} = \lim_{k \rightarrow \infty} L\{S(\theta_{n_k})\} \leq \lim_{k \rightarrow \infty} L(\theta_{n_{k+1}}) = L(\theta^*).$$

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. $\square$

The above result is the basis of a class of algorithms used in this book 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.

A.5 Sufficiency

If the random variable X has distribution P which is unknown apart from belonging to a class of distributions $\mathcal{P}$, a statistic $T = t(X)$ is said to be *sufficient* for $\mathcal{P}$ if the conditional distributions

$$P^{t_0} = \mathcal{L}\{X | t(X) = t_0, P\}$$

can all be chosen so that they do not depend on $P \in \mathcal{P}$.

If the measures $P \in \mathcal{P}$ have densities $f(x|P)$ with respect to a σ -finite measure μ which factorize as

$$f(x|P) = g(x)h\{t(x), P\}, \quad (\text{A.5})$$

i.e. into a term that does not depend on P and a term that depends on x through $t(x)$ only, then T is sufficient. A suitable converse is also true. The criterion (A.5) is known as *Neyman's factorization criterion*.

We say that a statistic T as above is (boundedly) *complete* with respect to $\mathcal{P}$ if it holds for any bounded and measurable function f that

$$[\mathbf{E}\{f(T)|P\} = 0 \text{ for all } P \in \mathcal{P}] \implies [f(t) = 0 \text{ (} Pt^{-1}\text{-a.s. for all } P \in \mathcal{P})].$$

The following important result is Theorem 2 of Basu (1955) and will be used extensively.

Theorem A.4 (Basu) *If T is sufficient and complete with respect to $\mathcal{P}$ and the distribution of $Y = s(X)$ does not depend on $P \in \mathcal{P}$, then T and Y are independent.*

Proof: We have for all $P \in \mathcal{P}$,

$$g(A) = P(Y \in A) = \mathbf{E}\{\chi_A(Y)|P\} = \mathbf{E}[\mathbf{E}\{\chi_A(Y)|T, P\} | P].$$

Since T is sufficient, the inner conditional expectation does not depend on P and we can define a function f as

$$f(t) = \mathbf{E}\{\chi_A(Y)|T = t, P\} - g(A) = P(Y \in A | T = t) - P(Y \in A).$$

The first equation then reads

$$\int_{\mathcal{X}} f\{t(x)\}P(dx) = 0.$$

Since this holds for all $P \in \mathcal{P}$, we conclude that

$$P(Y \in A | T = t) = P(Y \in A)$$

and hence Y and T are independent. □

Complete sufficient statistics exist typically in exponential models; see Appendix D.

An important notion related to sufficiency is that of Y being a *cut* in (X, Y) (Barndorff-Nielsen 1978), which means that the family $\mathcal{P}$ can be parametrized by a pair (κ, λ) where κ and λ are variation-independent and the joint density of (X, Y) factorizes as

$$f(x, y | P) = f(x | y, \kappa) f(y | \lambda).$$

That Y is a cut implies that for any fixed value of κ , $t(x, y) = y$ is a sufficient statistic for λ , but many other interesting properties follow. In particular the likelihood function factorizes, so that the joint likelihood can be maximized by maximizing each factor separately.

Similarly we say that Z is a *split* in (X, Y, Z) if $\mathcal{P}$ can be parametrized with the triple (κ, λ, μ) with all variation-independent and

$$f(x, y, z | P) = f(x | z, \kappa) f(y | z, \lambda) f(z | \mu).$$

Hence in particular X and Y are conditionally independent given Z , written as $X \perp\!\!\!\perp Y | Z$; see Chapter 3.

Appendix B

Linear algebra and random vectors

B.1 Matrix results

Lemma B.1 *A symmetric block diagonal matrix Σ with*

$$\Sigma = \begin{pmatrix} A & B \\ B^\top & D \end{pmatrix},$$

where D is assumed to be regular, is positive (non-negative) definite if and only if $E = A - BD^{-1}B^\top$ and D are both positive (non-negative) definite.

Proof: If we let

$$u = \begin{pmatrix} x \\ y - D^{-1}B^\top x \end{pmatrix},$$

we find

$$\begin{aligned} u^\top \Sigma u &= x^\top A x + 2x^\top B(y - D^{-1}B^\top x) \\ &\quad + (y - D^{-1}B^\top x)^\top D(y - D^{-1}B^\top x) \\ &= x^\top (A - BD^{-1}B^\top)x + y^\top D y. \end{aligned}$$

Hence the result follows. $\square$

The determinant of a partitioned $n \times n$ matrix can be factorized as

$$\det \begin{pmatrix} A & B \\ C & D \end{pmatrix} = (\det E)(\det D), \quad (\text{B.1})$$

where A and D are supposed to be $p \times p$ and $q \times q$ matrices respectively, and $E = A - BD^{-1}C$ if D is non-singular. The correctness of (B.1) follows from the calculation

$$\det \begin{pmatrix} A & B \\ C & D \end{pmatrix} = \det \begin{pmatrix} A & B \\ C & D \end{pmatrix} \det \begin{pmatrix} I_p & 0 \\ -D^{-1}C & I_q \end{pmatrix} =$$

$$\begin{aligned}
&= \det \left\{ \begin{pmatrix} A & B \\ C & D \end{pmatrix} \begin{pmatrix} I_p & 0 \\ -D^{-1}C & I_q \end{pmatrix} \right\} \\
&= \det \begin{pmatrix} E & B \\ 0 & D \end{pmatrix} = (\det E)(\det D).
\end{aligned}$$

Finally, the inverse of a partitioned matrix is given as

$$\begin{pmatrix} A & B \\ C & D \end{pmatrix}^{-1} = \begin{pmatrix} E^{-1} & -E^{-1}G \\ -FE^{-1} & D^{-1} + FE^{-1}G \end{pmatrix}. \quad (\text{B.2})$$

The inverse on the left-hand side exists if and only if the inverses on the right-hand side exist. Here $E = A - BD^{-1}C$, $F = D^{-1}C$, and $G = BD^{-1}$. Performing the multiplication shows the correctness.

The Kronecker product of matrices

Let A be an $m \times n$ and B a $p \times q$ matrix. We define the Kronecker product $A \otimes B$ as the $mp \times nq$ matrix with entries

$$(A \otimes B)_{ir,js} = a_{ij}b_{rs},$$

where $A = \{a_{ij}\}$ and $B = \{b_{rs}\}$. When displaying $A \otimes B$ it looks slightly different, depending on how we choose to arrange the indices ij and rs . For example, we can arrange it as

$$\begin{pmatrix} a_{11}B & \cdots & a_{1\nu}B & \cdots & a_{1n}B \\ \vdots & & \vdots & & \vdots \\ a_{\mu 1}B & \cdots & a_{\mu\nu}B & \cdots & a_{\mu n}B \\ \vdots & & \vdots & & \vdots \\ a_{m1}B & \cdots & a_{m\nu}B & \cdots & a_{mn}B \end{pmatrix}, \quad (\text{B.3})$$

i.e. as a block matrix with $m \times n$ blocks of size $p \times q$, or as $p \times q$ blocks of size $m \times n$, as

$$\begin{pmatrix} Ab_{11} & \cdots & Ab_{1\kappa} & \cdots & Ab_{1q} \\ \vdots & & \vdots & & \vdots \\ Ab_{\pi 1} & \cdots & Ab_{\pi\kappa} & \cdots & Ab_{\pi q} \\ \vdots & & \vdots & & \vdots \\ Ab_{p1} & \cdots & Ab_{p\kappa} & \cdots & Ab_{pq} \end{pmatrix}. \quad (\text{B.4})$$

The two ways of writing are equivalent and the important point is how to use the Kronecker product in the calculations.

As A and B can be interpreted as matrices of linear maps from $\mathcal{R}^n$ to $\mathcal{R}^m$ and $\mathcal{R}^q$ to $\mathcal{R}^p$, the product $A \otimes B$ can be interpreted as a linear map from $R^{n \times q}$ to $R^{m \times p}$ as follows. An element x in $R^{n \times q}$ is an $n \times q$ matrix and we simply let

$$(A \otimes B)(x) = Ax B^\top, \quad (\text{B.5})$$

where the expression on the right-hand side involves usual matrix products. The correctness of this interpretation follows from the calculation

$$\begin{aligned} (Ax B^\top)_{ir} &= \sum_{\nu} \sum_{\kappa} a_{i\nu} x_{\nu\kappa} b_{\kappa r}^\top \\ &= \sum_{\nu} \sum_{\kappa} a_{i\nu} b_{r\kappa} x_{\nu\kappa} = \sum_{\nu\kappa} (A \otimes B)_{ir, \nu\kappa} x_{\nu\kappa}. \end{aligned}$$

Expressions such as $(A \otimes B)(U)$ will sometimes refer to the usual matrix product and sometimes to the map $(A \otimes B)$ evaluated at U . The dimensions of the elements will always identify which.

Proposition B.2 *The Kronecker product satisfies:*

T1: $0 \otimes A = A \otimes 0 = 0$

T2: $A \otimes (B_1 + B_2) = A \otimes B_1 + A \otimes B_2$

T3: $(A_1 + A_2) \otimes B = A_1 \otimes B + A_2 \otimes B$

T4: $(\lambda A) \otimes (\mu B) = (\lambda\mu)(A \otimes B)$ if $\lambda, \mu \in \mathcal{R}$

T5: $(A \otimes B)^\top = A^\top \otimes B^\top$

T6: $(A_1 \otimes B_1)(A_2 \otimes B_2) = (A_1 A_2) \otimes (B_1 B_2)$

T7: $(A \otimes B)^- = A^- \otimes B^-$

T8: $\text{tr}(A \otimes B) = (\text{tr } A)(\text{tr } B)$ if A and B are symmetric matrices

T9: $\det(A \otimes B) = (\det A)^p (\det B)^n$ if A is $n \times n$ and B is $p \times p$

T10: If Φ and Σ are both positive (semi)definite matrices, then so is $\Phi \otimes \Sigma$.

Proof: This is straightforward algebraic manipulation. $\square$

Here A^- denotes any generalized inverse of A , i.e. a matrix satisfying $AA^-A = A$, and **T7** should be read: if A^- and B^- are g-inverses of A and B , then $A^- \otimes B^-$ is a g-inverse of $A \otimes B$. If Φ and Σ are symmetric matrices of dimensions $n \times n$ and $p \times p$ respectively, then $\Phi \otimes \Sigma$ is a symmetric matrix. The corresponding bilinear form is given as

$$\langle x, y \rangle_{\Phi \otimes \Sigma} = \langle x, (\Phi \otimes \Sigma)(y) \rangle = \text{tr}(x^\top \Phi y \Sigma).$$

The outer product

Consider two Euclidean spaces V and W with inner products $\langle \cdot, \cdot \rangle_V$ and $\langle \cdot, \cdot \rangle_W$. For vectors $v \in V$ and $w \in W$ we define the *outer product* $w \odot v$ as the linear map from V to W that has

$$(w \odot v)(u) = \langle u, v \rangle_V w. \quad (\text{B.6})$$

Thus the range of the map $w \mapsto w \odot v$ is the subspace $\text{span}\{w\}$.

B.2 Factor subspaces and interactions

We spend some time here considering interaction spaces and generating classes from an algebraic point of view, inspired by Darroch and Speed (1983).

Let Δ be a finite set and assume that for each $\delta \in \Delta$ a finite set $\mathcal{I}_\delta$ of possible levels is given. We often refer to the elements of Δ as *variables*. The elements $i = (i_\delta)_{\delta \in \Delta}$ of the product $\mathcal{I}$ of the level sets

$$i \in \mathcal{I} = \times_{\delta \in \Delta} \mathcal{I}_\delta$$

are denoted *cells* of the *table* $\mathcal{I}$.

The vector space $L = \mathcal{R}^{\mathcal{I}}$ of real-valued functions defined on $\mathcal{I}$. L is a Euclidean space with inner product

$$\langle x, y \rangle = \sum_{i \in \mathcal{I}} x(i)y(i).$$

An *a*-marginal table is obtained for $a \subset \Delta$ by only considering the variables in a . It has *marginal cells*

$$i_a \in \mathcal{I}_a = \times_{\delta \in a} \mathcal{I}_\delta.$$

For an arbitrary $x \in \mathcal{R}^{\mathcal{I}}$ and an arbitrary subset a of Δ , the *a*-marginal of x is defined as

$$x(i_a) = \sum_{j: j_a = i_a} x(j) = \sum_{j_{\Delta \setminus a} \in \mathcal{I}_{\Delta \setminus a}} x(i_a \cdot j_{\Delta \setminus a})$$

and $|x| = \sum_{i \in \mathcal{I}} |x(i)|$. For $a = \emptyset$ we write $x(i_\emptyset) = \sum_{i \in \mathcal{I}} x(i)$, so that if $x(i) \geq 0$ for all i then we also have that $|x| = x(i_\emptyset)$.

If a is a subset of Δ , we consider the *factor subspace* F_a of L of functions that only depend on i through i_a , i.e.

$$x \in F_a \iff x(i) = x(j) \quad \text{for all } i, j \text{ with } i_a = j_a.$$

The space F_Δ is L itself and $F_\emptyset$ is the space of constants. The orthogonal projections onto F_a are given as

$$\Pi_a x(i) = x(i_a) / |\mathcal{I}_{\Delta \setminus a}|. \quad (\text{B.7})$$

From this it directly follows that

$$\Pi_a \Pi_b = \Pi_b \Pi_a = \Pi_{a \cap b}. \quad (\text{B.8})$$

Subspaces of the form

$$H_{\mathcal{A}} = \sum_{a \in \mathcal{A}} F_a,$$

where $\mathcal{A}$ is a class of subsets of Δ , are called *hierarchical model subspaces*. Denote the orthogonal projection onto $H_{\mathcal{A}}$ by $\Pi_{\mathcal{A}}$. We then have the following useful lemma.

Lemma B.3 *For $x, y \in L$ the equation*

$$\Pi_{\mathcal{A}} x = \Pi_{\mathcal{A}} y \quad (\text{B.9})$$

is equivalent to the system of equations

$$\Pi_a x = \Pi_a y \quad \text{for all } a \in \mathcal{A}. \quad (\text{B.10})$$

Proof: It is obvious that the equation (B.9) implies the system of equations (B.10), since $\Pi_a \Pi_{\mathcal{A}} = \Pi_a$. Conversely, assume that the pair (x, y) satisfies (B.10). To establish (B.9) we just have to show that

$$\langle x - \Pi_{\mathcal{A}} y, u \rangle = 0$$

or, equivalently, that

$$\langle y, u \rangle = \langle x, u \rangle$$

for an arbitrary $u \in H_{\mathcal{A}}$. If $u = \sum_a u_a$ with $u_a \in F_a$, we have

$$\begin{aligned} \langle y, u \rangle &= \sum_{a \in \mathcal{A}} \langle y, u_a \rangle = \sum_{a \in \mathcal{A}} \langle \Pi_a y, u_a \rangle \\ &= \sum_{a \in \mathcal{A}} \langle \Pi_a x, u_a \rangle = \sum_{a \in \mathcal{A}} \langle x, u_a \rangle = \langle x, u \rangle, \end{aligned}$$

and the result is established. $\square$

Introduce the subspaces for $a \subseteq \Delta$:

$$E_a = F_a \cap \left(\bigcap_{b: b \subset a} F_b^\perp \right).$$

The space E_a is called the *space of interactions* among criteria in a and its elements are called $|a|$ -*factor interactions* or *interactions of order* $|a| - 1$. For $|a| = 1$ we also speak of *main effects*. Note that (B.7) implies that

$$x \in E_a \iff x \in F_a \text{ and } x(i_b) = 0 \text{ for all } b \subset a, \quad (\text{B.11})$$

which is the familiar relation that 'summing an interaction term over any index gives 0'.

We assume in the following that $F_\delta \subseteq H_{\mathcal{A}}$ for all $\delta \in \Delta$, i.e. all main effects are represented in $\mathcal{A}$.

If we let Q_a denote the orthogonal projection onto the interaction subspace E_a , we get

$$\Pi_d Q_a = \begin{cases} Q_a & \text{if } d \supseteq a \\ 0 & \text{otherwise} \end{cases} \quad (\text{B.12})$$

since the first of these assertions is obvious and the second is a consequence of (B.11) and the calculation

$$\Pi_d Q_a = \Pi_d \Pi_a Q_a = \Pi_{d \cap a} Q_a = 0.$$

Since

$$\bigcap_{b: b \subset a} F_b^\perp = \bigcap_{\delta: \delta \in a} F_{a \setminus \{\delta\}}^\perp,$$

we also have

$$E_a = F_a \cap \left(\bigcap_{\delta: \delta \in a} F_{a \setminus \{\delta\}}^\perp \right).$$

From this we obtain

$$Q_a = \Pi_a \prod_{\delta: \delta \in a} (I - \Pi_{a \setminus \{\delta\}}), \quad (\text{B.13})$$

which in particular implies that all Q_a and Π_b commute. As a consequence we derive

$$Q_a Q_b = Q_b Q_a = 0 \quad \text{if } a \neq b, \quad (\text{B.14})$$

since $a \neq b$ implies either $b \not\subseteq a$ or the converse, whereby

$$Q_a Q_b = \Pi_a Q_a Q_b = Q_a \Pi_a Q_b = 0.$$

Thus the interaction subspaces are mutually orthogonal. By direct multiplication we obtain

$$\Pi_a \prod_{\delta: \delta \in a} (I - \Pi_{a \setminus \{\delta\}}) = \sum_{b: b \subseteq a} (-1)^{|b|} \Pi_{a \setminus b} = \sum_{b: b \subseteq a} (-1)^{|a \setminus b|} \Pi_b.$$

A consequence is the formula

$$Q_a = \sum_{b: b \subseteq a} (-1)^{|a \setminus b|} \Pi_b. \quad (\text{B.15})$$

Using Lemina A.2 (Möbius inversion) on (B.15) we obtain

Proposition B.4 *The factor subspaces F_a are direct orthogonal sums of the interaction subspaces $(E_b)_{b \subseteq a}$ and for all $a \subseteq \Delta$ we have*

$$\Pi_a = \sum_{b: b \subseteq a} Q_b.$$

In particular, any vector in $\mathcal{R}^{\mathcal{I}}$ has a unique expansion into its interactions, in symbols expressed as

$$\mathcal{R}^{\mathcal{I}} = \oplus_{a: a \subseteq \Delta} E_a. \quad (\text{B.16})$$

Letting $\tilde{\mathcal{A}}$ denote the permissible interactions

$$c \in \tilde{\mathcal{A}} \iff c \subseteq a \text{ for some } a \in \mathcal{A},$$

we further have the expressions

$$H_{\mathcal{A}} = \oplus_{c \in \tilde{\mathcal{A}}} E_c, \quad \Pi_{\mathcal{A}} = \sum_{c \in \tilde{\mathcal{A}}} Q_c. \quad (\text{B.17})$$

If $\mathcal{A}$ is a collection of subsets of Δ , we let $\text{red } \mathcal{A}$ denote the elements of $\mathcal{A}$ that are maximal with respect to inclusion, i.e. $\text{red } \mathcal{A}$ is obtained from $\mathcal{A}$ by throwing away all subsets that are contained in other subsets in $\mathcal{A}$. Since

$$H_{\mathcal{A}} = H_{\text{red } \mathcal{A}},$$

this reduced collection is the most economical way of specifying the model space. If no sets in $\mathcal{A}$ are subsets of other sets, $\mathcal{A}$ is reduced already, and the collection is called a *generating class*.

As shown in Section 2.2, the set of generating classes forms a distributive lattice with join and meet operations given as

$$\begin{aligned} \mathcal{A} \vee \mathcal{B} &= \text{red}(\mathcal{A} \cup \mathcal{B}) \\ \mathcal{A} \wedge \mathcal{B} &= \text{red}\{a \cap b \mid a \in \mathcal{A}, b \in \mathcal{B}\}. \end{aligned}$$

We then have

Proposition B.5 *Let $\mathcal{A}$ and $\mathcal{B}$ be generating classes. Then the following relations hold for their hierarchical model subspaces:*

$$H_{\mathcal{A} \vee \mathcal{B}} = H_{\mathcal{A}} + H_{\mathcal{B}}, \quad H_{\mathcal{A} \wedge \mathcal{B}} = H_{\mathcal{A}} \cap H_{\mathcal{B}}$$

and correspondingly for the orthogonal projections

$$\Pi_{\mathcal{A} \vee \mathcal{B}} = \Pi_{\mathcal{B}} \Pi_{\mathcal{A}} = \Pi_{\mathcal{A} \wedge \mathcal{B}}. \quad (\text{B.18})$$

Proof: It is trivial that $H_{\mathcal{A} \wedge \mathcal{B}} \subseteq H_{\mathcal{A}} \cap H_{\mathcal{B}}$. Let $u \in H_{\mathcal{A}} \cap H_{\mathcal{B}}$. Then, from (B.17) we get

$$u \in H_{\mathcal{A}} \implies u = \Pi_{\mathcal{A}} u = \sum_{c \in \tilde{\mathcal{A}}} Q_c u.$$

Since also $u \in H_{\mathcal{B}}$,

$$u = \Pi_{\mathcal{B}} u = \sum_{d \in \tilde{\mathcal{B}}} Q_d \sum_{c \in \tilde{\mathcal{A}}} Q_c u.$$

From (B.14) it now follows that $u \in H_{\mathcal{A} \wedge \mathcal{B}}$. That the projections commute follows directly from (B.14) and (B.17). $\square$

In the special case where the generating class $\mathcal{C}$ is the direct join of generating classes $\mathcal{A}$ and $\mathcal{B}$ with $\mathcal{A} \wedge \mathcal{B} = \{\mathcal{A} \cap \mathcal{B}\}$, we find

$$\Pi_{\mathcal{A}} \Pi_{\mathcal{B}} = \Pi_{\mathcal{B}} \Pi_{\mathcal{A}} = \Pi_{\mathcal{A} \cap \mathcal{B}}. \quad (\text{B.19})$$

B.3 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 $V = \mathcal{R}^n$ 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 B.6 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 B.7 A bilinear form Σ on V is said to be the *covariance* of X if it holds that

$$\text{Cov}(\langle u, X \rangle, \langle v, X \rangle) = \Sigma(u, v) \quad \text{for all } u, v \in V.$$

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

$$\Sigma(u, v) = \frac{1}{4} \{ \Sigma(u + v, u + v) - \Sigma(u - v, u - v) \}.$$

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 $(e_1, \dots, e_p)$ is an orthonormal basis of V , we let

$$\sigma_{ij} = \Sigma(e_i, e_j) = \langle e_i, \Sigma e_j \rangle = \text{Cov}(\langle e_i, X \rangle, \langle e_j, X \rangle).$$

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

$$\langle u, v \rangle = \langle Ku, \Sigma v \rangle. \quad (\text{B.20})$$

If Σ is not regular, any K which satisfies $\Sigma = K\Sigma K$, i.e. K is a generalized inverse to Σ , can be used as the concentration operator, and the relation (B.20) 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 $\mathcal{R}^n$ with the usual inner product and standard orthonormal basis, but we also frequently deal with the space $\mathcal{R}^{n \times p}$ of $n \times p$ -matrices with inner product

$$\langle A, B \rangle = \text{tr}(A^\top B) \quad (\text{B.21})$$

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 (B.21) becomes unnecessary. An orthonormal basis for this space consists of the symmetric matrices

$$\tilde{E}_{ii} = E_{ii}, \quad \tilde{E}_{ij} = (E_{ij} + E_{ji})/\sqrt{2} \quad \text{for } i \neq j. \quad (\text{B.22})$$

If $V = \mathcal{R}^n$, the mean vector is of the form $\xi = (\xi_1, \dots, \xi_n)^\top$ and we have

$$\langle e_i, \xi \rangle = \mathbf{E}\langle e_i, X \rangle = \xi_i = \mathbf{E}X_i,$$

where $X = (X_1, \dots, X_n)^\top$. Similarly for the covariance we get

$$\sigma_{ij} = \Sigma(e_i, e_j) = \text{Cov}(\langle e_i, X \rangle, \langle e_j, X \rangle) = \text{Cov}(X_i, X_j).$$

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 B.8 *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 $V = \mathcal{R}^n$ and $W = \mathcal{R}^m$, 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,

$$\begin{aligned} \text{Cov}(\langle w, Y \rangle_W, \langle y, Y \rangle_W) &= \text{Cov}(\langle w, AX \rangle_W, \langle y, AX \rangle_W) \\ &= \text{Cov}(\langle A^\top w, X \rangle_V, \langle A^\top y, X \rangle_V) \\ &= \langle A^\top w, \Sigma A^\top y \rangle_V = \langle w, A\Sigma A^\top y \rangle_W, \end{aligned}$$

which gives the covariance operator. We abstain from repeating the calculations in the matrix case. $\square$

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 B.9 *If X and Y are two random vectors in V and*

$$\langle v, X \rangle \stackrel{\mathcal{D}}{=} \langle v, Y \rangle \quad \text{for all } v \text{ in } V,$$

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 B.9 follows from the uniqueness of the Fourier transform in the case where $V = \mathcal{R}$.

Appendix C

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).

C.1 Basic properties

We first formally define what it means for a random vector in V to be normally distributed:

Definition C.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 B.9 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 $V = \mathcal{R}^p$ and $V = \mathcal{R}^{n \times p}$ 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

$$f_{\xi, \Sigma}(x) = (2\pi)^{-p/2} (\det \Sigma)^{-1/2} e^{-\langle x - \xi, K(x - \xi) \rangle / 2}, \quad (\text{C.1})$$

where $K = \Sigma^{-1}$ is the concentration operator of the normal distribution. If $\langle \cdot, \cdot \rangle$ is standard inner product on $\mathcal{R}^n$, we have in matrix notation that

$$\langle x - \xi, K(x - \xi) \rangle = (x - \xi)^T K(x - \xi),$$

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 C.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 C.1 implies that $X_1 + X_2$ is a normal random vector and the expressions for mean and covariance follow by direct calculation. $\square$

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 B.8.

Proposition C.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*

$$Y = AX + b \sim \mathcal{N}_W(A\xi + b, A\Sigma A^T).$$

Proof: The mean and covariance of Y have been given in Proposition B.8. What remains to be established is that Y follows a normal distribution. But for all $w \in W$ we have

$$\langle w, Y \rangle_W = \langle w, AX + b \rangle_W = \langle A^T w, X \rangle_V + \langle w, b \rangle_W.$$

Since X has a normal distribution, $\langle A^T w, X \rangle_V$ is univariate normally distributed. This is not changed by adding the constant $\langle w, b \rangle_W$. Definition C.1 then establishes the result. $\square$

A special case of this result is of interest. Suppose $V = \mathcal{R}^n$ and assume the random vector X partitioned into components X_1 and X_2 , where $X_1 \in \mathcal{R}^p$

and $X_2 \in \mathcal{R}^q$ 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 C.3 is the linear map that sends X into X_2 , we obtain:

Proposition C.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 $\mathcal{N}_q(\xi_2, \Sigma_{22})$.*

The conditional distribution of X_1 given $X_2 = x_2$ can be found as follows.

Proposition C.5 *Let X be distributed as $\mathcal{N}_n(\xi, \Sigma)$, where X , ξ and Σ are partitioned as above. Then the conditional distribution of X_1 given $X_2 = x_2$ is $\mathcal{N}_p(\xi_{1|2}, \Sigma_{1|2})$, where*

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

Here Σ_{22}^- is an arbitrary generalized inverse to Σ_{22} .

Proof: We abstain from giving the proof of this well-known result in the general case and assume Σ to be regular. Then let $K = \Sigma^{-1}$ denote the concentration matrix and assume this to be partitioned in the same fashion as Σ . 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

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

The linear term involving x_1 has coefficient equal to

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

Using (B.2) we find that

$$K_{11}^{-1} = \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21} \quad (\text{C.3})$$

and further that

$$K_{11}^{-1}K_{12} = -\Sigma_{12}\Sigma_{22}^{-1}, \quad (\text{C.4})$$

which then gives

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

and the result follows. $\square$

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 (B.1) that we have

$$\det \Sigma = \det \Sigma_{1|2} \det \Sigma_{22} = \frac{\det \Sigma_{22}}{\det K_{11}}. \quad (\text{C.5})$$

Note also the identities (C.3) and (C.4), 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. This gives a formulation of Proposition C.5 which is symmetric to that of Proposition C.4.

Corollary C.6 *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 Proposition C.5. The second statement then follows from (C.4). $\square$

Consider $X^1, \dots, X^n$ independent random vectors in $\mathcal{R}^p$ and assume X^i to be distributed as $\mathcal{N}_p(\xi^i, \Sigma)$ such that the covariances are identical. If these random vectors are arranged in an $n \times p$ array with $(X^i)^\top$ as rows, the resulting random matrix will be distributed as $\mathcal{N}_{n \times p}(\xi, I_n \otimes \Sigma)$, where ξ has $(\xi^i)^\top$ as its rows. For this reason, it is of special interest to consider normal distributions with covariance structure of the form $\Phi \otimes \Sigma$.

Proposition C.7 *Let A and B be $m \times n$ and $q \times p$ matrices and $X \sim \mathcal{N}_{n \times p}(\xi, \Phi \otimes \Sigma)$. Then*

$$(A \otimes B)(X) = AXB^\top \sim \mathcal{N}_{m \times q}\{A\xi B^\top, (A\Phi A^\top) \otimes (B\Sigma B^\top)\}.$$

Proof: The result follows directly from Proposition C.3 by noting that

$$(A \otimes B)(\xi) = A\xi B^\top$$

and that

$$(A \otimes B)(\Phi \otimes \Sigma)(A \otimes B)^\top = (A\Phi A^\top) \otimes (B\Sigma B^\top),$$

where we have used T6 and T5 of Proposition B.2. $\square$

Next we partition the matrix X into $n \times r$ and $n \times s$ matrices X_1 and X_2 with $p = r + s$ and partition the mean ξ and covariance matrix Σ accordingly. Further, let $\Sigma_{1|2}$ be given as in (C.2).

Proposition C.8 *The conditional distribution of X_1 given $X_2 = x_2$ is*

$$\mathcal{N}_{n \times r}(\xi_1 + (x_2 - \xi_2)\Sigma_{22}^{-1}\Sigma_{21}, \Phi \otimes \Sigma_{1|2}).$$

Proof: This follows directly from translation of Proposition C.5. $\square$

Finally we mention that in the case when Φ and Σ are both regular, the density becomes

$$f_{\xi, \Phi \otimes \Sigma}(x) = (2\pi)^{-np/2} (\det \Phi)^{-p/2} (\det \Sigma)^{-n/2} e^{-\text{tr}\{(x-\xi)^\top \Phi^{-1}(x-\xi)\Sigma^{-1}\}/2}. \quad (\text{C.6})$$

This follows from the properties of the direct product, the expression (C.1) and the calculation

$$\langle u, (\Phi^{-1} \otimes \Sigma^{-1})(u) \rangle = \text{tr}(u^\top \Phi^{-1} u \Sigma^{-1}).$$

C.2 The Wishart distribution

The Wishart distribution is the sampling distribution of the empirical covariance matrix when the variables themselves follow a multivariate normal distribution.

Definition C.9 A random $p \times p$ matrix S is said to follow a *p-dimensional Wishart distribution* with parameter Σ and n degrees of freedom if

$$S \stackrel{\mathcal{D}}{=} X^\top X, \quad \text{where } X \sim \mathcal{N}_{n \times p}(0, I_n \otimes \Sigma).$$

We then write $S \sim \mathcal{W}_p(n, \Sigma)$. If $X_i^\top$ denote the rows of X , then

$$X^\top X = \sum_{i=1}^n X_i X_i^\top$$

such that the empirical covariance matrix is equal to $n^{-1}S$. In the case $p = 1$ the Wishart distribution is identical to the χ^2 distribution:

$$W_1(n, \sigma^2) = \sigma^2 \chi^2(n).$$

Below we list some well-known properties of the Wishart distribution. Proofs not given here can be found in Eaton (1983).

Proposition C.10 *If $S \sim \mathcal{W}_p(n, \Sigma)$, its mean and covariance operator with respect to trace inner product are equal to*

$$\mathbf{E}(S) = n\Sigma, \quad \text{Cov}(S) = 2n\Sigma \otimes \Sigma.$$

The inverse S^{-1} has mean and covariance operator equal to

$$\begin{aligned} \mathbf{E}(S^{-1}) &= \frac{1}{n-p-1}K \quad \text{for } n > p+1 \\ \text{Cov}(S^{-1}) &= \frac{2 \left(K \otimes K + \frac{1}{n-p-1} K \odot K \right)}{(n-p)(n-p-1)(n-p-3)} \quad \text{for } n > p+3, \end{aligned}$$

where $K = \Sigma^{-1}$.

It is important that the covariance operators are interpreted with respect to the trace inner product

$$\langle A, B \rangle = \text{tr}(AB)$$

on the space $\mathcal{S}_p$ of symmetric $p \times p$ matrices. In particular, the outer product is determined as in (B.6) such that

$$(K \odot K)(A) = \text{tr}(AK)K.$$

If for example we wish to calculate the variance of S_{ij} for $i \neq j$ we must first write

$$S_{ij} = \text{tr} \left\{ (\sqrt{2}/2) \tilde{E}_{ij} S \right\} = \langle \tilde{E}_{ij}, S \rangle / \sqrt{2},$$

where $\tilde{E}_{ij}$ is defined in (B.22), and then its variance is calculated as

$$\begin{aligned} \mathbf{V}(S_{ij}) &= \langle \tilde{E}_{ij}, 2n(\Sigma \otimes \Sigma)(\tilde{E}_{ij}) \rangle / 2 = n \langle \tilde{E}_{ij}, \Sigma \tilde{E}_{ij} \Sigma \rangle \\ &= n \text{tr}(\tilde{E}_{ij} \Sigma \tilde{E}_{ij} \Sigma) = n(\sigma_{ij}^2 + \sigma_{ii}\sigma_{jj}). \end{aligned}$$

The general expression for the covariance of the entries in S becomes

$$\text{Cov}(S_{ij}, S_{uv}) = n(\sigma_{ij}\sigma_{uv} + \sigma_{iu}\sigma_{jv}). \quad (\text{C.7})$$

Similarly for the inverse we have

$$\mathbf{V} \{ \text{tr}(AS^{-1}) \} = \frac{2 \left(\text{tr}(AKAK) + \frac{\{\text{tr}(AK)\}^2}{(n-p-1)} \right)}{(n-p)(n-p-1)(n-p-3)}$$

and for the single elements of the inverse matrix we find

$$\text{Cov}(S^{ij}, S^{uv}) = \frac{k_{ij}k_{uv} + k_{iu}k_{jv} + \frac{2k_{ij}k_{uv}}{(n-p-1)}}{(n-p)(n-p-1)(n-p-3)}.$$

A direct consequence of the definition of the Wishart distribution is the convolution property.

Proposition C.11 *If S_1 and S_2 are independent and Wishart-distributed as $S_i \sim \mathcal{W}_p(n_i, \Sigma)$, then*

$$S_1 + S_2 \sim \mathcal{W}_p(n_1 + n_2, \Sigma).$$

Next, a transformation property:

Proposition C.12 *If A is a $q \times p$ matrix and $S \sim \mathcal{W}_p(n, \Sigma)$, then*

$$ASA^\top \sim \mathcal{W}_q(n, A\Sigma A^\top).$$

In particular, for $q = 1$ we get that when $S \sim \mathcal{W}_p(n, \Sigma)$, then for any $v \in R^p$,

$$v^\top S v \sim \sigma_v^2 \chi^2(n),$$

where $\sigma_v^2 = v^\top \Sigma v$. The next fact relates the Wishart distribution to a slightly different normal distribution.

Proposition C.13 *If $X \sim \mathcal{N}_{n \times p}(0, \Phi \otimes \Sigma)$ and Φ^- is a generalized inverse to Φ , then*

$$X^\top \Phi^- X \sim \mathcal{W}_p(f, \Sigma),$$

where $f = \text{rank } \Phi$.

And there is a partitioning theorem analogous to that for the χ^2 distribution:

Proposition C.14 *Assume that $X \sim \mathcal{N}_{n \times p}(\xi, I_n \otimes \Sigma)$ and that $A_1, \dots, A_k$ are $n \times n$ matrices for orthogonal projections onto subspaces $L_1, \dots, L_k$ of R^n , that is,*

$$A_i A_j = \delta_{ij} A_i \quad \text{and} \quad A_i^\top = A_i.$$

Then $A_1 X, \dots, A_k X$ are independent and

$$A_i X \sim \mathcal{N}_{n \times p}(A_i \xi, A_i \otimes \Sigma).$$

Further, if $A_i \xi = 0$, then

$$S_i = X^\top A_i X \sim \mathcal{W}_p(f_i, \Sigma),$$

where $f_i = \dim L_i = \text{rank } A_i = \text{tr } A_i$. Further, $S_1, \dots, S_k$ are independent.

A special case of interest is when A represents the projection onto the $n-1$ -dimensional subspace of R^n of vectors that are orthogonal to the constant. Then under the assumption that $X \sim \mathcal{N}_{n \times p}(\mu, I_n \otimes \Sigma)$ with $\mu_{ij} = \xi_j$ for all i, j , we have that

$$SSD = \sum_{i=1}^n (X^i - \bar{X})(X^i - \bar{X})^\top = X^\top A X \sim \mathcal{W}_p(n-1, \Sigma). \quad (\text{C.8})$$

The above properties are all natural extensions of similar results for the χ^2 distribution. The next result is genuinely for the Wishart distribution. We assume Σ to be partitioned as

$$\Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix},$$

where Σ_{11} has dimensions $r \times r$, $\Sigma_{12} = \Sigma_{21}^T$ is $r \times s$ and Σ_{22} is an $s \times s$ matrix. Further assume S to be partitioned similarly and let

$$S_{1|2} = S_{11} - S_{12}S_{22}^{-1}S_{21}, \quad S_{1|2} = S_{11} - S_{12}S_{22}^{-1}S_{21},$$

provided Σ_{22} and S_{22} are both regular.

Proposition C.15 *Let $S \sim \mathcal{W}_p(n, \Sigma)$ with Σ regular and $n > s$. Then S_{22} is regular with probability one and*

- (a) $S_{1|2}$ is independent of (S_{21}, S_{22}) ;
- (b) $S_{1|2} \sim \mathcal{W}_r(n - s, \Sigma_{1|2})$;
- (c) $S_{22} \sim \mathcal{W}_s(n, \Sigma_{22})$;
- (d) $\mathcal{L}(S_{21} | S_{22} = s_{22}) = \mathcal{N}_{s \times r}(s_{22}\Sigma_{22}^{-1}\Sigma_{21}, s_{22} \otimes \Sigma_{1|2})$.

In particular, if we consider the statement for the case where $p = s$ and thus $S = S_{22}$, we deduce

Corollary C.16 *If $S \sim \mathcal{W}_p(n, \Sigma)$, where Σ is regular, then S is regular with probability one if and only if $n \geq p$.*

If the condition of the corollary is satisfied, then the distribution has density equal to

$$w_{p,n,\Sigma}(s) = c(p, n)^{-1} (\det \Sigma)^{-n/2} (\det s)^{(n-p-1)/2} \exp \left\{ -\text{tr}(\Sigma^{-1}s)/2 \right\} \quad (\text{C.9})$$

with respect to Lebesgue measure on the set of positive definite matrices. When the Lebesgue measure giving mass one to the unit cube in the space $\mathcal{S}_p$ of symmetric $p \times p$ matrices with inner product $\langle A, B \rangle = \text{tr}(AB)$, the Wishart constant $c(p, n)$ is

$$c(p, n) = 2^{np/2} (2\pi)^{p(p-1)/4} \prod_{i=1}^p \Gamma\{(n+1-i)/2\}.$$

Finally the result in Proposition C.15 can be strengthened in a special case:

Corollary C.17 *Suppose that in Proposition C.15 we further have that $\Sigma_{12} = 0$. Then $S_{1|2}$, $S_{12}S_{22}^{-1}S_{21}$ and S_{22} are mutually independent and distributed as*

$$S_{1|2} \sim \mathcal{W}_r(n - s, \Sigma_{11}), \quad S_{12}S_{22}^{-1}S_{21} \sim \mathcal{W}_r(s, \Sigma_{11}), \quad S_{22} \sim \mathcal{W}_s(n, \Sigma_{22}).$$

C.3 Other derived distributions

C.3.1 Box-type distributions

Important distributions associated with the Wishart distribution have a common structure (Box 1949). We say that the distribution of a positive random variable X is of *Box type* if it has Laplace transform (defined for sufficiently small t) equal to

$$\begin{aligned}\phi(t) &= \mathbf{E} e^{tX} \\ &= e^{2tz} \prod_{i=1}^m \frac{\Gamma\{u_i(1-2t) + \xi_i\}}{\Gamma\{u_i + \xi_i\}} \prod_{j=1}^n \frac{\Gamma\{v_j + \eta_j\}}{\Gamma\{v_j(1-2t) + \eta_j\}}, \quad (\text{C.10})\end{aligned}$$

where $\sum_i u_i = \sum_j v_j$, $z = \sum_i u_i \log u_i - \sum_j v_j \log v_j$, and further u_i , $u_i + \xi_i$, v_j , $v_j + \eta_j$ are all positive. If $X_1, \dots, X_k$ are independent and have Box-type distributions, so does $a_1 X_1 + \dots + a_k X_k$ for any non-negative numbers $a_1, \dots, a_k$ which are not all zero. This follows from straightforward direct verification as the Laplace transforms multiply.

If B has a beta distribution with parameters (α, β) , then the distribution of $X = -\log B$ is of Box type, because we get for the Laplace transform that

$$\phi(t) = \mathbf{E} B^{-t} = \frac{\int_0^1 s^{\alpha-t-1} (1-s)^{\beta-1} ds}{\int_0^1 s^{\alpha-1} (1-s)^{\beta-1} ds} = \frac{\Gamma(\alpha-t)\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\alpha-t+\beta)}.$$

This is of Box type with $m = n = 1$, $u = v = 1/2$, $\xi = \alpha - 1/2$ and $\eta = \alpha + \beta - 1/2$. The negative logarithm of a beta distribution is probably the simplest distribution of Box type. But most distributions associated with testing problems in the multivariate normal distribution are of Box type.

An important feature of Box-type distributions is that they are a unified technique can be used to approximate them. The approximations given below in (C.13) are simple to use and more accurate than the version of the original approximations given by Box (1949) used in Anderson (1984). The new approximations are due to J.L. Jensen (1991, 1995), to whom we refer for further details.

The simplest approximation uses a gamma distribution with correct mean and variance. The mean and variance of X can be found as the first and second derivatives of the logarithm of the Laplace transform, evaluated at $s = 0$. From (C.10) we find the logarithmic derivatives as

$$\mu(t) = 2z - 2 \sum_{i=1}^m u_i \psi\{u_i(1-2t) + \xi_i\} + 2 \sum_{j=1}^n v_j \psi\{v_j(1-2t) + \eta_j\} \quad (\text{C.11})$$

and

$$\sigma^2(t) = 4 \sum_{i=1}^m u_i \psi' \{u_i(1-2t) + \xi_i\} - 4 \sum_{j=1}^n v_j \psi' \{v_j(1-2t) + \eta_j\}, \quad (\text{C.12})$$

where ψ is the digamma function and ψ' the trigamma function; see for example Abramowitz and Stegun (1965). The digamma and trigamma functions are the first and second logarithmic derivatives of the gamma function. Algorithms for their computation are given in Bernardo (1976) and Schneider (1978). Abbreviate $\mu(0) = \mu$ and $\sigma^2(0) = \sigma^2$. Then the distribution of X is approximated by a gamma distribution $\Gamma(\mu^2/\sigma^2, \mu/\sigma^2)$, where $\Gamma(\alpha, \beta)$ is the gamma distribution with density

$$f(x) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x}.$$

A very accurate approximation to $P(X \geq x)$ can be obtained for a fixed x as follows. First determine the value t^* such that $\mu(t^*) = x$ by the iteration $t_{n+1} = t_n + \{x - \mu(t_n)\}/\sigma^2(t_n)$. Next let $\lambda^* = \mu(t^*)^2/\sigma^2(t^*) = x^2/\sigma^2(t^*)$. Then it holds to a high degree of accuracy that

$$P(X \geq x) \approx \phi(t^*) \left(\frac{\lambda^*}{xt^* + \lambda^*} \right)^{\lambda^*} \gamma(xt^* + \lambda^*, \lambda^*), \quad (\text{C.13})$$

where $\gamma(s, \lambda)$ is the tail of the gamma distribution

$$\gamma(s, \lambda) = \int_s^\infty \Gamma(\lambda)^{-1} u^{\lambda-1} e^{-u} du.$$

This can be evaluated from the gamma function and the incomplete gamma function (Bhattacharjee 1970). Note that the expression (C.13) differs slightly from (2.3) in J.L. Jensen (1991), partly because the latter contains a misprint (J.L. Jensen 1995), partly because we have reduced the expression exploiting that $\mu(t^*) = x$. The approximation is particularly accurate for large values of x , as the relative error tends to zero when x tends to ∞ .

C.3.2 Wilks's distribution

In multivariate testing problems concerned with linear hypotheses for the mean and block independence for the covariance, the following distributional result is important (Wilks 1932, 1935).

Proposition C.18 *If $S_1 \sim \mathcal{W}_p(f_1, \Sigma)$ and $S_2 \sim \mathcal{W}_p(f_2, \Sigma)$, where S_1 and S_2 are independent, Σ is regular, and $f_1 \geq p$, then*

$$\Lambda = \frac{\det S_1}{\det(S_1 + S_2)} \stackrel{D}{=} \prod_{i=1}^p B_i, \quad (\text{C.14})$$

where $B_1, \dots, B_p$ are independent and $B_i \sim \mathcal{B}\{(f_1 + 1 - i)/2, f_2/2\}$.

The distribution so obtained does obviously not depend on Σ and is called *Wilks's distribution* with parameters p, f_1, f_2 , and is denoted by $\Lambda(p, f_1, f_2)$. The distribution of $-\log \Lambda$ is clearly of Box type with $u_i = v_j = 1/2$, $\xi_i = (f_1 - i)/2$, and $\eta_j = (f_1 + f_2 - j)/2$, hence the approximations described above apply. Wilks's distribution satisfies the identity

$$\Lambda(p, f_1, f_2) = \Lambda(f_2, f_1 + f_2 - p, p).$$

C.3.3 Test for identical covariances

If $S_1, \dots, S_k$ are independent with $S_i \sim \mathcal{W}_p(f_i, \Sigma_i)$, where $p \geq f_i$ and Σ_i are all regular, then the deviance for testing identical covariances is

$$d = p \sum_{i=1}^k f_i \log f_i - pf \log f + f \log \det s - \sum_{i=1}^k f_i \log \det s_i, \quad (\text{C.15})$$

where $f = \sum_i f_i$ and $s = \sum_i s_i$. If $\Sigma_1 = \dots = \Sigma_k = \Sigma$, then the distribution of D does not depend on Σ and it has moment-generating function equal to

$$\begin{aligned} \phi(t) &= \exp(2tz) \prod_{i=1}^m \prod_{j=1}^k \frac{\Gamma\{f_j(1-2t)/2 + (1-i)/2\}}{\Gamma\{(f_j+1-i)/2\}} \\ &\times \prod_{l=1}^k \frac{\Gamma\{(f+1-l)/2\}}{\Gamma\{f(1-2t)/2 + (1-l)/2\}} \end{aligned}$$

with

$$z = \frac{p}{2} \left(\sum_{j=1}^k f_j \log f_j - f \log f \right).$$

Hence the distribution is of Box type with $u_{ij} = f_j/2$, $\xi_{ij} = (1-i)/2$, $v_l = f/2$, and $\eta_l = (1-l)/2$.

If $S_1, \dots, S_k$ are sums of squares and products of residuals from independent linear models each with n_i observations and dimensions q_i of the linear expectation spaces, then $f_i = n_i - q_i$ and the deviance for testing identical covariances becomes

$$d_w = p \sum_{i=1}^k n_i \log n_i - pn \log n + n \log \det s - \sum_{i=1}^k n_i \log \det s_i, \quad (\text{C.16})$$

where $n = \sum_i n_i$. If $\Sigma_1 = \dots = \Sigma_k = \Sigma$, then the distribution of D_w does not depend on Σ and it has moment-generating function equal to

$$\phi(t) = \exp(2tz_w) \prod_{i=1}^m \prod_{j=1}^k \frac{\Gamma\{n_j(1-2t)/2 + (1-q_j-i)/2\}}{\Gamma\{(f_j+1-i)/2\}}.$$

$$\times \prod_{l=1}^k \frac{\Gamma\{(f+1-l)/2\}}{\Gamma\{n(1-2t)/2 + (1-q-l)/2\}}$$

with $q = \sum q_j$ and

$$z_w = \frac{p}{2} \left(\sum_{j=1}^k n_j \log n_j - n \log n \right).$$

So this distribution is also of Box type with $u_{ij} = n_j/2$, $\xi_{ij} = (1 - q_j - v_l = n/2$, and $\eta_l = (1 - q - l)/2$.

Appendix D

Exponential models

In this appendix we give a brief survey of some important results in the exact theory of regular exponential models and the asymptotic theory of curved exponential models.

The exact theory is described in detail in Barndorff-Nielsen (1978) and important parts of it also in Johansen (1979), Brown (1986) and S.T. Jensen (1989). The asymptotic theory was first described by A.H. Andersen (1969) and Berk (1972). The description given here of both cases is mainly based upon Johansen (1979) and S.T. Jensen (1989).

All the results will be stated without proof. The reader should consult the above references for documentation.

D.1 Regular exponential models

D.1.1 Basic terminology

Let $\mathcal{X}$ be a sample space. In the present text $\mathcal{X}$ is typically either a discrete set, $\mathcal{R}^k$, or a product of these. Also let t be a statistic defined on $\mathcal{X}$ taking values in a real Euclidean vector space V with inner product $\langle \cdot, \cdot \rangle$. Further, let μ be a fixed σ -finite measure on $\mathcal{X}$ and

$$\Theta = \left\{ \theta \in V \mid \int e^{\langle \theta, t(x) \rangle} \mu(dx) < \infty \right\}.$$

For $\theta \in \Theta$ we let

$$\psi(\theta) = \log \int e^{\langle \theta, t(x) \rangle} \mu(dx).$$

The *exponential model* generated by (t, μ) has parameter space Θ and densities for P_θ with respect to μ equal to

$$f(x, \theta) = \exp \{ \langle \theta, t(x) \rangle - \psi(\theta) \}. \quad (\text{D.1})$$

The statistic t is referred to as a *canonical statistic*, θ is a *canonical parameter*, μ is a *base measure*, and ψ is the *cumulant function* of the model.

In the following we assume that Θ is not contained in an affine proper subspace of V . If the set Θ is open, we say that the exponential model is

regular. The *convex support* of the statistic t is the convex hull of the support of the measure μt^{-1} and is denoted by C . The representation (D.1) of a regular exponential model is *minimal* if no affine proper subspace of V contains $t(X)$ with P_θ -probability one, i.e. if no affine proper subspace of V contains C . In general we will allow representations to be non-minimal and possible over-parametrization of the model. The case where the representation is not minimal is referred to as the *singular case*.

The *order* of an exponential model is the dimension of V in any minimal representation of the model.

The *relative interior* of C is the interior of C when C is considered as a subset of the smallest affine subspace containing it. Thus if the representation is minimal, the relative interior is just the interior. We denote the relative interior of C by $\text{ri } C$.

D.1.2 Analytic properties

Much of the structure in a regular exponential model is contained in the analytic properties of the function ψ . We list some of them here. First let $\tau(\theta)$ and $v(\theta)$ denote the gradient vector and Hessian map of $\psi(\theta)$. That is, for $\eta, \rho \in V$ we have

$$\left. \frac{\partial}{\partial h} \psi(\theta + h\eta) \right|_{h=0} = \langle \tau(\theta), \eta \rangle$$

and

$$\left. \frac{\partial^2}{\partial h \partial k} \psi(\theta + h\eta + k\rho) \right|_{h=k=0} = \langle \eta, v(\theta)\rho \rangle.$$

Both these are well defined in regular exponential models and they determine the mean and covariance of the statistic $t(X)$ as

$$\mathbf{E}_\theta \langle \eta, t(X) \rangle = \langle \eta, \tau(\theta) \rangle, \quad \mathbf{V}_\theta \{ \langle \eta, t(X) \rangle, \langle \rho, t(X) \rangle \} = \langle \eta, v(\theta)\rho \rangle. \quad (\text{D.2})$$

If a basis is chosen for the vector space V and everything is expressed in coordinates, these will be the usual vector of means and covariance matrix of $t(X)$. We therefore allow ourselves to write

$$\tau(\theta) = \mathbf{E}_\theta \{ t(X) \}, \quad v(\theta) = \mathbf{V}_\theta \{ t(X) \}$$

instead of (D.2). The covariance $v(\theta)$ will be positive definite if and only if the model is minimally represented. In a minimally represented exponential model the map τ is strictly increasing in the sense that

$$(\theta_1 - \theta_2)^T \{ \tau(\theta_1) - \tau(\theta_2) \} > 0. \quad (\text{D.3})$$

The mean value map τ can be used to parametrize the model, since one can show that

$$\tau(\theta) = \tau(\theta^*) \iff P_\theta = P_{\theta^*}.$$

Such a parametrization is called a *mean value parametrization* of the model.

D.1.3 Maximum likelihood estimation

Suppose an observation $X = x_0$ from a regular exponential model is given. From (A.5), $t(X)$ is sufficient, so we let $t_0 = t(x_0)$. If the model is minimally represented, $t(X)$ is complete with respect to $\mathcal{P} = \{P_\theta, \theta \in \Theta\}$. The main result about maximum likelihood estimation is the following.

Theorem D.1 *The likelihood function attains its maximum in Θ if and only if $t_0 \in \text{ri } C$. The maximum is then attained exactly at points that solve the equation*

$$\tau(\theta) = t_0. \quad (\text{D.4})$$

If $\hat{\theta}$ is any particular solution to this equation, any other solution $\hat{\theta}^*$ satisfies $\hat{\theta}^* = \hat{\theta} + u$ where $\langle u, v \rangle$ is constant for $v \in C$ and thus $P_{\hat{\theta}^*} = P_{\hat{\theta}}$.

In particular we note that in the non-singular case there is a unique solution to the equation (D.4).

A consequence of the above result is that the parameter space $\tau(\Theta)$ for the mean value parameter in a regular exponential model is equal to $\text{ri } C$. Also, in the non-singular case, we have the following useful lemma.

Lemma D.2 *Suppose $t_0 \in \text{ri } C$ in a minimally represented, regular exponential model. Then for any $\theta_0 \in \Theta$ the set*

$$\Theta^* = \{\theta \mid L(\theta) \geq L(\theta_0)\} \quad (\text{D.5})$$

is compact.

D.1.4 Affine hypotheses

Consider a regular exponential model in the situation where we want to test the hypothesis that $\theta \in \Theta_0$, where

$$\Theta_0 = \Theta \cap R_0$$

for some affine subspace $R_0 = r + V_0$ of V . Such a hypothesis is called an *affine hypothesis*. The calculation

$$\begin{aligned} \langle \theta, t(x) \rangle &= \langle \theta - r, t(x) \rangle + \langle r, t(x) \rangle \\ &= \langle \theta - r, \Pi_0 t(x) \rangle + \langle r, t(x) \rangle \\ &= \langle \theta, \Pi_0 t(x) \rangle + \langle r, t(x) - \Pi_0 t(x) \rangle, \end{aligned}$$

where Π_0 denotes the orthogonal projection onto V_0 , shows that the reduced model $\{P_\theta, \theta \in \Theta_0\}$ is again a regular exponential model with $\Pi_0 t(x)$ as its canonical statistic, $\theta \in \Theta_0$ as its canonical parameter and μ_r as its base measure, where

$$\mu_r(dx) = e^{\langle r, t(x) - \Pi_0 t(x) \rangle} \mu(dx).$$

When the original model is regular and minimally represented, so is the submodel. Further, if the maximum likelihood estimate exists in the original model, it will also exist in the smaller model. The standard test statistic to use is the *deviance* given as

$$d(x) = -2 \log \frac{L(\hat{\hat{\theta}})}{L(\hat{\theta})} = 2 \left\{ \langle \hat{\theta} - \hat{\hat{\theta}}, t(x) \rangle + \psi(\hat{\hat{\theta}}) - \psi(\hat{\theta}) \right\},$$

where $\hat{\hat{\theta}}$ denotes the maximum likelihood estimate of θ in the reduced model. Large deviances speak against the hypothesis.

Other test statistics can be of interest for a variety of reasons, in particular to obtain high power against specific alternatives.

In most cases, asymptotic theory is needed to evaluate test probabilities for the deviance and other statistics; cf. Section D.2. Occasionally, however, exact results can be obtained by considering the distribution of the test statistic conditionally on $\Pi_0 t(X) = \Pi_0 t_0$. By considering this conditional distribution, the sufficiency of $\Pi_0 t(X)$ under the hypothesis ensures that the test probabilities are independent of the unknown parameter $\theta \in \Theta_0$, thus leading to similar rejection regions. There are other reasons for choosing to evaluate the test probability in this conditional distribution, but it is outside the scope of the present text to discuss general principles of inference in any detail.

In some cases the unconditional distribution of the deviance D does not depend on θ for $\theta \in \Theta_0$. Then, since the canonical statistic $\Pi_0 t(X)$ is complete with respect to the submodel, Basu's theorem (Theorem A.4) ensures that D and $\Pi_0 t(X)$ are independent. It therefore makes no difference whether the conditional or unconditional distribution is considered.

Tests based on probabilities in the conditional distribution given $\Pi_0 t(X)$ will be referred to as *exact conditional tests* or simply *exact tests*.

D.1.5 Iterative computational methods

None of the results so far give direct guidance as to how to compute the maximum likelihood estimates in a full and regular exponential model if it exists. Here iterative methods are in general necessary. The most common is Newton's method applied to the logarithm of the likelihood function. This begins at a value θ_0 and then lets

$$\theta_{n+1} = \theta_n - v(\theta_n)^-(\tau(\theta_n) - t_0),$$

where $v(\theta)^-$ is a generalized inverse to $v(\theta)$, i.e. it satisfies $vv^-v = v$. The convergence is rapid when the starting value is sufficiently close to $\hat{\theta}$.

It is often convenient and more stable to use Newton's method on the reciprocal likelihood function, which leads to the iteration

$$\theta_{n+1} = \theta_n - \bar{v}(\theta_n)^-(\tau(\theta_n) - t_0),$$

where

$$\tilde{v}(\theta) = v(\theta) + \{\tau(\theta) - t_0\}^{\otimes 2},$$

i.e. $\tilde{v}(\theta)$ is determined as

$$\langle \eta, \tilde{v}(\theta) \rho \rangle = \langle \eta, v(\theta) \rho \rangle + \langle \eta, \tau(\theta) - t_0 \rangle \langle \rho, \tau(\theta) - t_0 \rangle.$$

In fact it has been shown by S.T. Jensen *et al.* (1991) that the latter iteration is always convergent in regular exponential models of order one.

One problem with Newton's method is that it can involve much storage space and processor time to calculate v^- or $\tilde{v}^-$ in models of high order. Another is that its behaviour may depend on where the iteration is started. If the starting value is far from the true maximum, it may not converge or approach the true maximum slowly. Often a combination of Newton's method and the method of iterative partial maximization (see Section A.4) is appropriate. When suitable functions of θ are held fixed, sections in Θ^* of (D.5) appear that satisfy the conditions necessary for Proposition A.3 to apply.

A typical variant of the algorithm of iterative partial maximization is as follows. Consider a full, regular and minimally represented exponential model of order k . Further, assume there is chosen a basis such that $V = \mathcal{R}^k$ and $\theta = (\theta^1, \dots, \theta^k)$ and choose a fixed element θ_0 in the canonical parameter set Θ . Also let $\mathcal{A}$ denote a class of subsets of $\{1, \dots, k\}$ that covers all coordinates, i.e. $\cup_{a \in \mathcal{A}} a = \{1, \dots, k\}$. We next define sections for $\theta^* \in \Theta$:

$$\Theta_a(\theta^*) = \left\{ \theta \in \Theta \mid \theta^i = \theta^{*i}, i \notin a \right\}.$$

These are affine subspaces of Θ and therefore identify full and regular submodels of the original model with canonical statistics $t^a(x) = (t^i(x), i \in a)$. If we now cyclically maximize the likelihood function over the sections $\Theta_a(\theta^*)$, we obtain a convergent algorithm. If we let $\tau^a(\theta)$ denote the expected value of $t^a(X)$ under θ , the problem is at each stage to solve the equation $\tau^a(\theta) = t^a(x)$ for $\theta \in \Theta_a(\theta^*)$.

But this computation may itself be too involved. We can then instead do the following. First suppose that θ_0 can be chosen so that we can easily maximize the likelihood function over $\Theta_a(\theta_0)$, no matter what the observed value of the statistic t^a is. Let $\tilde{\theta}_a$ denote the value that maximizes the likelihood function over $\Theta_a(\theta_0)$ had the statistic been observed to be $\tau^a(\theta^*)$, and $\check{\theta}_a$ that which maximizes over $\Theta_a(\theta_0)$ for the actual observed value $t^a(x)$, i.e.

$$\tau^a(\check{\theta}_a) = t^a(x), \quad \tau^a(\tilde{\theta}_a) = \tau^a(\theta^*). \quad (\text{D.6})$$

Further, define the sections

$$\tilde{\Theta}_a(\theta^*) = \Theta \cap \left\{ \theta^* + \lambda (\check{\theta}_a - \tilde{\theta}_a), \lambda \in \mathcal{R} \right\}.$$

If $\check{\theta}_a = \bar{\theta}_a$, these sections consist of the single point θ^* . Otherwise the sections $\bar{\Theta}_a(\theta^*)$ are affine hypotheses in Θ and form full and regular one-dimensional exponential models. These sections are called *line sections*. We can then maximize the likelihood function over $\bar{\Theta}_a(\theta^*)$, for example using Newton's method applied to the reciprocal likelihood function as described earlier. This process of maximization is referred to as *line search* and we denote the maximizing value by $T_a(\theta^*)$, i.e.

$$L\{T_a(\theta^*)\} \geq L(\theta) \text{ for all } \theta \in \bar{\Theta}_a(\theta^*).$$

By repeating these line searches cyclically over the sets $a \in \mathcal{A}$, we obtain another variant of the algorithm of iterative partial maximization. To show that the modification gives a convergent algorithm, we must ensure that if a value θ' maximizes the likelihood function $L(\theta)$ over all line sections, then θ' maximizes $L(\theta)$ globally, and the maximum over any section depends continuously on θ^* . But if we let $\theta_\lambda^* = \theta^* + \lambda(\check{\theta}_a - \bar{\theta}_a)$, we get

$$\left. \frac{\partial}{\partial \lambda} \log L(\theta_\lambda^*) \right|_{\lambda=0} = (\check{\theta}_a - \bar{\theta}_a)^\top \{t^a(x) - \tau^a(\theta^*)\}.$$

Since $\check{\theta}_a$ and $\bar{\theta}_a$ solve the equations (D.6), this expression is by (D.3) strictly positive unless $\check{\theta}_a = \bar{\theta}_a$, which again is equivalent to $t^a(x) = \tau^a(\theta^*)$. If this happens, the sections consist of a single point and consequently θ^* trivially maximizes L over the full section $\Theta_a(\theta^*)$. Otherwise $\log L(\theta_\lambda^*)$ is strictly increasing in λ at $\lambda = 0$ and its maximum is not at $\theta_0^* = \theta^*$.

Hence θ^* maximizes L over all line sections $\bar{\Theta}_a(\theta^*)$ if and only if it holds that $t^a(x) = \tau^a(\theta^*)$ for all $a \in \mathcal{A}$. This system of conditions is equivalent to having $\tau(\theta^*) = t(x)$, which implies that L is globally maximized at θ^* .

Finally, we must realize that the maximizing value $T_a(\theta^*)$ depends continuously on θ^* . We abstain from giving the details of this.

In the special case where $t^a(X)$ is a cut with $\lambda = (\theta_i, i \in a)$ (see Section A.5) we have from Corollary 10.4 of Barndorff-Nielsen (1978) that the difference $\check{\theta}_a - \bar{\theta}_a$ does not depend on the chosen value of θ_0 . Hence we could for a moment assume that $\theta_0 = \theta^*$, which would imply that $\bar{\theta}_a = \theta^*$, and thus it must hold that

$$\check{\theta}_a - \bar{\theta}_a = \check{\theta}_a - \theta^*,$$

whereby

$$\tau^a(\theta^* + \check{\theta}_a - \bar{\theta}_a) = \tau^a(\check{\theta}_a) = t^a(x).$$

Hence $\theta^* + \check{\theta}_a - \bar{\theta}_a$ maximizes L not only over the line section $\bar{\Theta}_a(\theta^*)$ but also over the full section $\Theta_a(\theta^*)$.

When t^a is a cut, the iterative step for a fixed subset a changes the joint density as

$$f\{x, T_a(\theta^*)\} = f(x, \theta^*) \frac{f(x, \check{\theta}_a)}{f(x, \bar{\theta}_a)} = f(x, \theta^*) \frac{f_a(t^a, \check{\theta}_a)}{f_a(t^a, \bar{\theta}_a)},$$

where $f_a(t^a, \theta)$ denotes the marginal density of t^a . This only depends on θ through coordinates in θ_i with $i \in a$, because t^a is a cut. The algorithm therefore properly deserves the name of *general iterative proportional scaling* if all groups $a \in \mathcal{A}$ make t^a a cut.

When t^a is not a cut, it is sometimes favourable to let θ_0 vary with θ^* or $t(x)$. This can make t^a nearly a cut, which typically speeds up computation.

In practice, it is in general difficult and unnecessary to make a full line search over $\tilde{\Theta}_a(\theta^*)$. Instead we just look for a new value of θ in the section that has higher likelihood. For example, inspired by the situation where we have a cut, one can choose

$$\lambda^* = \max \{ \lambda \in \{1, 1/2, 1/4, 1/8, \dots\} \mid \theta_\lambda^* \in \Theta \text{ and } L(\theta_\lambda^*) > L(\theta^*) \}.$$

If we have a cut as above, we have $\lambda^* = 1$ and this would lead to the full maximum over the large section. Otherwise the procedure is called *modified iterative proportional scaling* (Frydenberg and Edwards 1989). This variant of the algorithm is not theoretically known to be convergent, as the maximizing value at each stage is not a continuous function of θ^* . But, as implemented for mixed hierarchical models in the program MIM (Edwards 1995), it has so far appeared to converge in all examples.

D.2 Curved exponential models

There is no essential difference between the asymptotic theory of regular exponential models and that of curved models, so we might as well treat the curved case immediately. However, there are some technicalities in the singular case that could obscure the basic ideas. Therefore the non-singular case will be treated first.

D.2.1 The non-singular case

We consider a regular exponential base model of order k as described in the previous section and assume that the corresponding representation is minimal such that $\dim V = k$. Without loss of generality we can then assume a basis to be chosen such that $V = \mathcal{R}^k$.

Consider now a submodel $\Theta_0 \subseteq \Theta$ given as $\Theta_0 = \phi(B)$, where B is an open subset of $\mathcal{R}^m$ for some m with $1 \leq m \leq k$. Assume that the map ϕ satisfies

- (i) ϕ is a homeomorphism onto Θ_0 ,
- (ii) ϕ is twice continuously differentiable,
- (iii) the matrix $\partial\phi/\partial\beta$ has full rank m for all $\beta \in B$,

where

$$\left(\frac{\partial \phi}{\partial \beta}\right)_{rs} = \frac{\partial}{\partial \beta_s} \phi_r(\beta), \quad r = 1, \dots, k, s = 1, \dots, m$$

is the matrix of partial derivatives. The condition (i) says that ϕ is continuous, globally one-to-one and has a continuous inverse. Condition (ii) is a smoothness condition that can be replaced by weaker conditions, and the last condition ensures that the hypothesis everywhere locally resembles $\mathcal{R}^m$, i.e. the local dimension is constant.

A hypothesis of this kind is called a *curved hypothesis* and the corresponding reduced exponential model a *curved exponential model*. Its *dimension* is the local dimension m of its parameter space.

Note in particular that if $\Theta_0 = \Theta$ and ϕ is the identity, the conditions are fulfilled, such that a regular exponential model is a special curved model. Also, affine hypotheses are special cases of curved hypotheses. The dimension is then simply the order of the corresponding regular model.

When we have observed $X = x$, the *observed Fisher information* about β under Θ_0 is the matrix function j_0 with entries

$$j_0(\beta; x)_{rs} = -\frac{\partial^2 \log f(x, \phi(\beta))}{\partial \beta_r \partial \beta_s}, \quad r, s = 1, \dots, m.$$

Denoting as usual the corresponding random variable with the capital letter J_0 , the *expected Fisher information* is its expected value $i_0 = \mathbf{E}J_0$ and can be expressed as

$$i_0(\beta) = \frac{\partial \phi^\top}{\partial \beta} v(\phi(\beta)) \frac{\partial \phi}{\partial \beta}. \quad (\text{D.7})$$

The condition (iii) is equivalent to $i_0(\beta)$ being positive definite for all $\beta \in B$.

Now let $X_1, \dots, X_n$ be independent and identically distributed with distribution given by a member of a curved exponential model with true parameter $\theta_0 \in \Theta_0$. Multiplying the densities, we again obtain an exponential model with the same parameter space as before and $t(X_1) + \dots + t(X_n)$ as canonical statistic. This holds in the base model Θ as well as in the reduced model Θ_0 . Let

$$\bar{t}_n = \{t(X_1) + \dots + t(X_n)\}/n$$

and let $\hat{\beta}_n$ denote the maximum likelihood estimate of β in the reduced model Θ_0 based upon these observations.

The basic asymptotic result is the following:

Proposition D.3 *As $n \rightarrow \infty$, the probability that $\hat{\beta}_n$ exists uniquely tends to 1. The estimate is then given as the solution to the equation*

$$\tau(\phi(\beta))^\top \frac{\partial \phi}{\partial \beta}(\beta) = \bar{t}_n^\top \frac{\partial \phi}{\partial \beta}(\beta) \quad (\text{D.8})$$

and

$$\hat{\beta}_n \stackrel{a}{\sim} \mathcal{N}(\beta, i_0(\beta)^{-1}/n).$$

Here $X_n \stackrel{a}{\sim} \mathcal{N}(\xi_n, a_n \Sigma)$ means that the distribution of $a_n^{-1/2}(X_n - \xi_n)$ converges weakly to a normal distribution with mean zero and covariance matrix Σ .

The asymptotic distribution simplifies when the parameters and distribution decompose simultaneously. If Y is a *cut* in (X, Y) (see Section A.5) the joint density of (X, Y) factorizes as

$$f(x, y | \beta) = f(x | y, \kappa) f(y | \lambda).$$

In this case it follows directly from the definition of the Fisher information that this is block-diagonal such that $\hat{\kappa}$ and $\hat{\lambda}$ are asymptotically independent, since all mixed second derivatives vanish in the observed information and hence also in the expected. Similarly if Z is a *split* in (X, Y, Z) , we have

$$f(x, y, z | \beta) = f(x | z, \kappa) f(y | z, \lambda) f(z | \mu).$$

When there is a split, the estimates $\hat{\kappa}$, $\hat{\lambda}$, and $\hat{\mu}$ are also asymptotically mutually independent, which is seen as in the case of a cut.

Consider next a further subhypothesis $\Theta_1 \subseteq \Theta_0$ with $\Theta_1 = \phi(b(A))$, where A is an open subset of $\mathcal{R}^l$ for some l with $1 \leq l \leq m$ and b satisfies similar conditions as did ϕ . Let us simultaneously consider the hypotheses

$$H : \theta \in \Theta, \quad H_0 : \theta \in \Theta_0, \quad H_1 : \theta \in \Theta_1$$

or, in an equivalent formulation,

$$H : \theta \in \Theta, \quad H_0 : \theta = \phi(\beta), \beta \in B, \quad H_1 : \theta = \phi(b(\alpha)), \alpha \in A.$$

A third equivalent formulation of H_1 is that H_0 holds and $\beta = b(\alpha)$.

We now omit the index n and let $\hat{\tau} = \bar{t}$ be the average statistic, $\hat{\theta}$ the estimate of θ under H , $\hat{\beta}$ the estimate of β under H_0 and $\hat{\alpha}$ the estimate of α under H_1 . Further, we let

$$\hat{\theta}_0 = \phi(\hat{\beta}), \hat{\tau}_0 = \tau(\hat{\theta}_0), \quad \hat{\theta}_1 = \phi(b(\hat{\alpha})), \hat{\tau}_1 = \tau(\hat{\theta}_1).$$

Finally let i_1 denote the information about α under H_1 .

Assume now that H_1 holds and fix α at the true value of the parameter. Then the matrices

$$\Pi_0 = \frac{\partial \phi}{\partial \beta} i_0^{-1} \frac{\partial \phi}{\partial \beta}^\top v, \quad \Pi_1 = \frac{\partial \phi}{\partial \beta} \frac{\partial b}{\partial \alpha} i_1^{-1} \frac{\partial b}{\partial \alpha}^\top \frac{\partial \phi}{\partial \beta}^\top v,$$

where expressions are evaluated at the value corresponding to α , are matrices for the orthogonal projections with respect to the inner product determined by v onto the tangent spaces of Θ_0 and Θ_1 at $\theta = \phi(b(\alpha))$.

If we further let $C_r = \tau(\Theta_r)$ for $r = 0, 1$, their transposes $\Pi_r^\top$ are projection matrices with respect to v^{-1} onto the tangent spaces of C_r at $\tau(b(\alpha))$. Alternatively, $\Pi_r^\top = \partial \hat{\tau}_r / \partial \hat{\tau}$.

Letting $X_n \stackrel{a}{=} Y_n$ denote that $X_n - Y_n$ converges to zero in probability, we moreover have the fundamental

Theorem D.4 *Under the assumptions above,*

$$\sqrt{n}(\hat{\theta} - \hat{\theta}_0, \hat{\theta}_0 - \hat{\theta}_1, \hat{\theta}_1 - \theta) \stackrel{a}{=} (I - \Pi_0, \Pi_0 - \Pi_1, \Pi_1)Y \quad (\text{D.9})$$

$$\sqrt{n}(\hat{\tau} - \hat{\tau}_0, \hat{\tau}_0 - \hat{\tau}_1, \hat{\tau}_1 - \tau) \stackrel{a}{=} (I - \Pi_0^\top, \Pi_0^\top - \Pi_1^\top, \Pi_1^\top)Z, \quad (\text{D.10})$$

where $Y \sim \mathcal{N}(0, v^{-1})$ and $Z \sim \mathcal{N}(0, v)$. Thus the three components on the left-hand side of the asymptotic equations are asymptotically independent and normally distributed.

It follows that if $\theta \in \Theta_0$ we have

$$\hat{\theta}_0 \stackrel{a}{\sim} \mathcal{N}(\theta, u/n), \quad \hat{\tau}_0 \stackrel{a}{\sim} \mathcal{N}(\tau, w/n), \quad (\text{D.11})$$

where u and w are found by direct calculation using the expression for Π_0 and (D.7) for the Fisher information. We obtain

$$u = \Pi_0 v^{-1} \Pi_0^\top = \frac{\partial \phi}{\partial \beta} i_0^{-1} \frac{\partial \phi}{\partial \beta}^\top$$

and

$$w = \Pi_0^\top v \Pi_0 = v \frac{\partial \phi}{\partial \beta} i_0^{-1} \frac{\partial \phi}{\partial \beta}^\top v = v u v.$$

Note that u and w are mutual generalized inverses, i.e. we have that $w u w = w$ and $u w u = u$. Another consequence of Theorem D.4 is that if we let

$$\begin{aligned} K_0 &= n(\hat{\theta} - \hat{\theta}_0)^\top v(\theta)(\hat{\theta} - \hat{\theta}_0) \\ K_{01} &= n(\hat{\theta}_0 - \hat{\theta}_1)^\top v(\theta)(\hat{\theta}_0 - \hat{\theta}_1) \\ K &= n(\hat{\theta}_1 - \theta)^\top v(\theta)(\hat{\theta}_1 - \theta), \end{aligned}$$

we have

Proposition D.5 *The approximately quadratic forms K_0 and K_{01} defined above are asymptotically independent and χ^2 -distributed with $k - m$ and $m - l$ degrees of freedom and independent of $\hat{\theta}_1$. Finally K is asymptotically distributed as χ^2 with l degrees of freedom.*

The importance of this result lies in the fact that the above quadratic forms approximate the deviance statistics. More precisely, let

$$D_0 = -2 \log \frac{L(\hat{\theta}_0)}{L(\hat{\theta})}, \quad D_{01} = -2 \log \frac{L(\hat{\theta}_1)}{L(\hat{\theta}_0)}, \quad D = -2 \log \frac{L(\theta)}{L(\hat{\theta}_1)}$$

and

$$\begin{aligned}\tilde{D}_0 &= n(\hat{\tau} - \hat{\tau}_0)^\top v(\theta)^{-1}(\hat{\tau} - \hat{\tau}_0) \\ \tilde{D}_{01} &= n(\hat{\tau}_0 - \hat{\tau}_1)^\top v(\theta)^{-1}(\hat{\tau}_0 - \hat{\tau}_1) \\ \tilde{D} &= n(\hat{\tau}_1 - \tau)^\top v(\theta)^{-1}(\hat{\tau}_1 - \tau),\end{aligned}$$

as well as

$$\tilde{K}_{01} = n(\hat{\beta} - b(\hat{\alpha}))^\top i_0(b(\hat{\alpha}))(\hat{\beta} - b(\hat{\alpha})), \quad \tilde{K} = (\hat{\alpha} - \alpha)^\top i_1(\hat{\alpha})(\hat{\alpha} - \alpha).$$

Then we have

Proposition D.6 *The following asymptotic equalities hold:*

$$D_0 \stackrel{a}{=} \tilde{D}_0 \stackrel{a}{=} K_0, \quad D_{01} \stackrel{a}{=} \tilde{D}_{01} \stackrel{a}{=} K_{01} \stackrel{a}{=} \tilde{K}_{01}, \quad D \stackrel{a}{=} \tilde{D} \stackrel{a}{=} K \stackrel{a}{=} \tilde{K}.$$

These statistics are therefore all asymptotically χ^2 -distributed with appropriate degrees of freedom.

Thus the deviance test of a curved hypotheses assuming a larger one can asymptotically be performed by using any of the quadratic approximations and an appropriate χ^2 -distribution. Note that the asymptotic independence implies that it does not matter asymptotically whether the test is conditional on the sufficient statistic under the hypothesis, should this be affine.

Note also that $v(\theta)$ in any of the quadratic expressions can be replaced with either of $v(\hat{\theta})$, $v(\hat{\theta}_0)$ or $v(\hat{\theta}_1)$ without altering the asymptotic distribution as long as the true value of θ is in Θ_1 .

D.2.2 The singular case

In the case where the model is not minimally represented, most of the asymptotic results carry over with minor modifications. Essentially one only has to modify at the following points. Inverses have to be replaced with generalized inverses where appropriate and all expressions directly involving the canonical parameter – which is not uniquely defined – should be avoided or at least interpreted with care. Finally the formulation of the smoothness condition is replaced by

- (a) $\tau \circ \phi$ is a homeomorphism onto C_0 ;
- (b) $\tau \circ \phi$ is twice continuously differentiable;
- (c) the matrix $\partial(\tau \circ \phi)/\partial\beta = v(\theta)\partial\theta/\partial\beta$ has full rank m for all $\beta \in B$.

The main reason for this is that τ is always in a unique correspondence with the distribution.

The Fisher information matrices are defined exactly as in the non-singular case and condition (c) ensures that these are regular. The expected Fisher information can be calculated by either of the expressions

$$i_0(\beta) = \frac{\partial \tau \circ \phi^\top}{\partial \beta} v(\phi(\beta))^- \frac{\partial \tau \circ \phi}{\partial \beta} = \frac{\partial \phi^\top}{\partial \beta} v(\phi(\beta)) \frac{\partial \phi}{\partial \beta},$$

where $v(\theta)^-$ is any generalized inverse to $v(\theta)$.

Then the basic asymptotic result Proposition D.3 holds verbatim. In Theorem D.4 some care has to be taken to make sense out of (D.9), whereas (D.10) carries over directly as it stands. The same is true for the independence of the components.

Of the various approximate quadratic forms, those involving θ should be treated carefully. Also v^{-1} should be replaced by an arbitrary generalized inverse v^- in the $\tilde{D}$ -versions. Otherwise both Propositions D.5 and D.6 hold. Just keep in mind that k is the order of the model rather than the dimension of V .

Bibliography

- Abramowitz, M. and Stegun, I. (1965). *Handbook of mathematical functions*. Dover, New York.
- Aigner, M. (1979). *Combinatorial theory*. Springer-Verlag, New York.
- Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, **19**, 716–23.
- Almond, R. (1995). *Graphical belief modelling*. Chapman and Hall, London.
- Andersen, A. H. (1969). Asymptotic results for exponential families. *Bulletin of the International Statistical Institute*, **43**, 241–2.
- Andersen, A. H. (1974). Multidimensional contingency tables. *Scandinavian Journal of Statistics*, **1**, 115–27.
- Andersen, E. B. (1977). Multiplicative Poisson models with unequal cell rates. *Scandinavian Journal of Statistics*, **4**, 153–8.
- Andersen, H. H., Højbjerg, M., Sørensen, D., and Eriksen, P. S. (1995). *Linear and graphical models for the multivariate complex normal distribution*, Lecture Notes in Statistics 101. Springer-Verlag, New York.
- Andersen, S. K., Olesen, K. G., Jensen, F. V., and Jensen, F. (1989). HUGIN — a shell for building Bayesian belief universes for expert systems. In *Proceedings of the 11th international joint conference on artificial intelligence*, pp. 1080–5. Morgan Kaufmann, San Mateo. Also reprinted in Shafer and Pearl (1990).
- Anderson, T. W. (1984). *An introduction to multivariate statistical analysis*, (2nd edn). John Wiley and Sons, New York.
- Andersson, S. A. and Perlman, M. D. (1993). Lattice models for conditional independence in a multivariate normal distribution. *Annals of Statistics*, **21**, 1318–58.
- Andersson, S. A., Madigan, D., and Perlman, M. D. (1995a). A characterization of Markov equivalence classes for acyclic digraphs. Technical Report 287, Department of Statistics, University of Washington, Seattle.
- Andersson, S. A., Madigan, D., Perlman, M. D., and Triggs, C. M. (1995b). On the relation between conditional independence models determined by finite distributive lattices and by directed acyclic graphs. *Journal of Statistical Planning and Inference*, **48**, 25–46.

- Andersson, S. A., Madigan, D., and Perlman, M. D. (1996). On the Markov equivalence of chain graphs, undirected graphs, and acyclic digraphs. *Scandinavian Journal of Statistics*, **26**, To appear.
- Asmussen, S. and Edwards, D. (1983). Collapsibility and response variables in contingency tables. *Biometrika*, **70**, 567–78.
- Badsberg, J. H. (1991). A guide to CoCo. Technical Report R-91-43, Department of Mathematics and Computer Science, Aalborg University.
- Badsberg, J. H. (1992). Model search in contingency tables by CoCo. In *Computational statistics*, (ed. Y. Dodge and J. Whittaker), pp. 251–6. Physica Verlag, Heidelberg.
- Badsberg, J. H. (1995). An environment for graphical models. Unpublished PhD thesis, Aalborg University.
- Barndorff-Nielsen, O. E. (1978). *Information and exponential families in statistical theory*. John Wiley and Sons, New York.
- Bartlett, M. S. (1935). Contingency table interactions. *Journal of the Royal Statistical Society, Supplement*, **2**, 248–52.
- Bartlett, M. S. (1937). Properties of sufficiency and statistical tests. *Proceedings of the Royal Society of London, Series A*, **160**, 268–82.
- Basu, D. (1955). On statistics independent of a complete sufficient statistic. *Sankhyā*, **15**, 377–80.
- Beeri, C., Fagin, R., Maier, D., Mendelzon, A., Ullman, J., and Yannakakis, M. (1981). Properties of acyclic database schemes. In *Proceedings of 13th annual ACM symposium on the theory of computing, Milwaukee*. Association of Computing Machinery, New York.
- Beeri, C., Fagin, R., Maier, D., and Yannakakis, M. (1983). On the desirability of acyclic database schemes. *Journal of the Association of Computing Machinery*, **30**, 479–513.
- Berge, C. (1973). *Graphs and hypergraphs*. North-Holland, Amsterdam. Translated from French by E. Minieka.
- Berk, R. H. (1972). Consistency and asymptotic normality of maximum likelihood estimates for exponential models. *Annals of Mathematical Statistics*, **43**, 193–204.
- Bernardo, J. M. (1976). Psi (digamma) function. Algorithm AS 103. *Applied Statistics*, **25**, 315–17.
- Besag, J. and Clifford, P. (1989). Generalized Monte Carlo significance tests. *Biometrika*, **76**, 633–42.
- Besag, J. and Green, P. (1993). Spatial statistics and Bayesian computation. *Journal of the Royal Statistical Society, Series B*, **55**, 25–37.
- Besag, J., York, J., and Mollié, A. (1991). Bayesian image restoration, with two applications in spatial statistics (with discussion). *Annals of the Institute of Statistical Mathematics*, **43**, 1–59.

- Bhattacharjee, G. P. (1970). Algorithm AS 32. The incomplete gamma integral. *Applied Statistics*, **19**, 285–7.
- Birch, M. W. (1963). Maximum likelihood in three way contingency tables. *Journal of the Royal Statistical Society, Series B*, **25**, 220–33.
- Bishop, Y. M. M. (1971). Effects of collapsing multidimensional contingency tables. *Biometrics*, **27**, 545–62.
- Bishop, Y. M. M., Fienberg, S. E., and Holland, P. W. (1975). *Discrete multivariate analysis: theory and practice*. MIT Press, Cambridge, Massachusetts.
- Blalock, H. M., Jr (ed.) (1971). *Causal models in the social sciences*. Aldine-Atherton, Chicago.
- Borosh, I. and Fraenkel, A. S. (1966). Exact solutions of linear equations with rational coefficients by congruence techniques. *Mathematics of Computation*, **20**, 107–12.
- Bouckaert, R. R. and Studený, M. (1995). Chain graphs: Semantics and expressiveness. In *Symbolic and qualitative approaches to reasoning and uncertainty*, (ed. C. Froideveaux and J. Kohlas), pp. 69–76. Lecture Notes in Artificial Intelligence 946. Springer-Verlag, Berlin.
- Box, G. E. P. (1949). A general distribution theory for a class of likelihood criteria. *Biometrika*, **36**, 317–46.
- Brown, L. D. (1986). *Fundamentals of statistical exponential families*, IMS-monographs, Vol. IX. Institute of Mathematical Statistics, Hayward.
- Buhl, S. L. (1993). On the existence of maximum likelihood estimators for graphical Gaussian models. *Scandinavian Journal of Statistics*, **20**, 263–70.
- Chen, T. and Fienberg, S. E. (1974). Two-dimensional contingency tables with both completely and partially cross-classified data. *Biometrics*, **30**, 629–42.
- Chen, T. and Fienberg, S. E. (1976). The analysis of contingency tables with incompletely classified data. *Biometrics*, **32**, 133–44.
- Chentsov, N. N. (1972). *Statistical decision rules and optimal conclusions*. Nauka, Moscow. In Russian.
- Christensen, E. S. (1989). Statistical properties of I -projections within exponential families. *Scandinavian Journal of Statistics*, **16**, 307–18.
- Christensen, R. (1990). *Log-linear models*. Springer-Verlag, New York.
- Clemmensen, J., Hansen, G., Nielsen, A., Røjel, J., Steensberg, J., Sørensen, S., and Toustrup, J. (1974). Lung cancer and air pollution in Fredericia. *Ugeskrift for Læger*, **136**, 2260–8.
- Cooper, G. and Herskovits, E. (1992). A Bayesian method for the induction of probabilistic networks from data. *Machine Learning*, **9**, 309–47.

- Cox, D. R. and Wermuth, N. (1993). Linear dependencies represented by chain graphs (with discussion). *Statistical Science*, **8**, 204–218; 247–277.
- Cox, D. R. and Wermuth, N. (1996). *Multivariate dependencies*. Chapman and Hall, London.
- Cramér, H. (1946). *Mathematical methods of statistics*. Princeton University Press, Princeton, N.J.
- Csiszár, I. (1975). I-divergence geometry of probability distributions and minimization problems. *Annals of Probability*, **3**, 146–58.
- Darroch, J. N. and Ratcliff, D. (1972). Generalized iterative scaling for log-linear models. *Annals of Mathematical Statistics*, **43**, 1470–80.
- Darroch, J. N. and Speed, T. P. (1983). Additive and multiplicative models and interactions. *Annals of Statistics*, **11**, 724–38.
- Darroch, J. N., Lauritzen, S. L., and Speed, T. P. (1980). Markov fields and log-linear models for contingency tables. *Annals of Statistics*, **8**, 522–39.
- Dawid, A. P. (1979). Conditional independence in statistical theory (with discussion). *Journal of the Royal Statistical Society, Series B*, **41**, 1–31.
- Dawid, A. P. (1980). Conditional independence for statistical operations. *Annals of Statistics*, **8**, 598–617.
- Dawid, A. P. (1984). Statistical theory. The prequential approach (with discussion). *Journal of the Royal Statistical Society, Series A*, **147**, 278–92.
- Dawid, A. P. (1990). Contribution to the discussion of David Edwards: Hierarchical interaction models. *Journal of the Royal Statistical Society, Series B*, **52**, 55.
- Dawid, A. P. (1992). Applications of a general propagation algorithm for probabilistic expert systems. *Statistics and Computing*, **2**, 25–36.
- Dawid, A. P. and Lauritzen, S. L. (1993). Hyper Markov laws in the statistical analysis of decomposable graphical models. *Annals of Statistics*, **21**, 1272–317.
- Deming, W. E. and Stephan, F. F. (1940). On a least squares adjustment of a sampled frequency table when the expected marginal totals are known. *Annals of Mathematical Statistics*, **11**, 427–44.
- Dempster, A. P. (1972). Covariance selection. *Biometrics*, **28**, 157–75.
- Dempster, A. P. (1973). Aspects of the multinomial logit model. In *Proceedings of the 3rd symposium on multivariate analysis, Dayton, Ohio*, (ed. P. R. Krishnaiah), pp. 129–42. Academic Press, New York.
- Dempster, A. P., Laird, N., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm (with discussion). *Journal of the Royal Statistical Society, Series B*, **39**, 1–38.

- Diaconis, P. and Sturmfels, B. (1994). Algebraic algorithms for generating from conditional distributions. Unpublished manuscript.
- Diestel, R. (1987). Simplicial decompositions of graphs – some uniqueness results. *Journal of Combinatorial Theory, Series B*, **42**, 133–45.
- Diestel, R. (1990). *Graph decompositions*. Clarendon Press, Oxford.
- Dirac, G. A. (1961). On rigid circuit graphs. *Abhandlungen Mathematisches Seminar Hamburg*, **25**, 71–6.
- Eaton, M. L. (1983). *Multivariate statistics. A vector space approach*. John Wiley and Sons, New York.
- Edwards, D. (1990). Hierarchical interaction models (with discussion). *Journal of the Royal Statistical Society, Series B*, **52**, 3–20 and 51–72.
- Edwards, D. (1995). *Introduction to graphical modelling*. Springer-Verlag, New York.
- Edwards, D. and Havránek, T. (1985). A fast procedure for model search in multidimensional contingency tables. *Biometrika*, **72**, 339–51.
- Edwards, D. and Havránek, T. (1987). A fast model selection procedure for large families of models. *Journal of the American Statistical Association*, **82**, 205–11.
- Edwards, D. and Kreiner, S. (1983). The analysis of contingency tables by graphical models. *Biometrika*, **70**, 553–62.
- Eriksen, P. S. (1996). Tests in covariance selection models. *Scandinavian Journal of Statistics*, **23**. To appear.
- Fisher, R. A. (1925). *Statistical methods for research workers*. Harner Press, New York.
- Frydenberg, M. (1990a). The chain graph Markov property. *Scandinavian Journal of Statistics*, **17**, 333–53.
- Frydenberg, M. (1990b). Marginalization and collapsibility in graphical interaction models. *Annals of Statistics*, **18**, 790–805.
- Frydenberg, M. and Edwards, D. (1989). A modified iterative proportional scaling algorithm for estimation in regular exponential families. *Computational Statistics and Data Analysis*, **8**, 143–53.
- Frydenberg, M. and Jensen, J. L. (1989). Is the ‘improved likelihood ratio statistic’ really improved in the discrete case. *Biometrika*, **76**, 655–61.
- Frydenberg, M. and Lauritzen, S. L. (1989). Decomposition of maximum likelihood in mixed interaction models. *Biometrika*, **76**, 539–55.
- Fuchs, C. (1982). Maximum likelihood estimation and model selection in contingency tables with missing data. *Journal of the American Statistical Association*, **77**, 270–8.
- Gavril, T. (1972). Algorithms for minimum coloring, maximum clique, minimum coloring by cliques and maximum independent set of a graph. *SIAM Journal of Computing*, **1**, 180–7.

- Geiger, D. and Pearl, J. (1990). On the logic of causal models. In *Uncertainty in artificial intelligence IV*, (ed. R. D. Schachter, T. S. Levitt, L. N. Kanal, and J. F. Lemmer), pp. 136–47. North-Holland, Amsterdam.
- Geiger, D. and Pearl, J. (1993). Logical and algorithmic properties of conditional independence and graphical models. *Annals of Statistics*, **21**, 2001–21.
- Gelfand, A. E. and Smith, A. F. M. (1990). Sampling-based approaches to calculating marginal densities. *Journal of the American Statistical Association*, **85**, 398–409.
- Geman, S. and Geman, D. (1984). Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **6**, 721–41.
- Geng, Z. (1989). Algorithm AS 244. Decomposability and collapsibility for log-linear models. *Applied Statistics*, **38**, 189–97.
- Geng, Z. (1992). Collapsibility of relative risk in contingency tables with a response variable. *Journal of the Royal Statistical Society, Series B*, **54**, 585–93.
- Geng, Z. and Asano, C. (1988). Recursive procedures for hierarchical log-linear models on high-dimensional contingency tables. *Journal of the Japanese Society for Computational Statistics*, **1**, 17–26.
- Geng, Z. and Asano, C. (1993). Strong collapsibility of association measures in linear models. *Journal of the Royal Statistical Society, Series B*, **55**, 741–7.
- Geyer, C. J. and Thompson, E. A. (1992). Constrained Monte Carlo maximum likelihood for dependent data (with discussion). *Journal of the Royal Statistical Society, Series B*, **54**, 657–99.
- Gibbs, W. (1902). *Elementary principles of statistical mechanics*. Yale University Press, NewHaven, Connecticut.
- Gilks, W. R., Clayton, D. G., Spiegelhalter, D. J., Best, N. G., McNeil, A. J., Sharples, L. D., and Kirby, A. J. (1993). Modelling complexity: application of the Gibbs sampler in medicine. *Journal of the Royal Statistical Society, Series B*, **55**, 39–52.
- Gilks, W. R., Thomas, A., and Spiegelhalter, D. J. (1994). A language and a program for complex Bayesian modelling. *The Statistician*, **43**, 169–78.
- Glonek, G. (1987). Some aspects of log-linear models. Unpublished PhD thesis, Flinders University.
- Goldberger, A. S. and Duncan, O. D. (ed.) (1973). *Structural equation models in the social sciences*. Seminar Press, New York.

- Golumbic, M. C. (1980). *Algorithmic graph theory and perfect graphs*. Academic Press, London.
- Goodman, L. A. (1970). The multivariate analysis of qualitative data: interaction among multiple classifications. *Journal of the American Statistical Association*, **65**, 226–56.
- Goodman, L. A. (1971). Partitioning of chi-square, analysis of marginal contingency tables, and estimation of expected frequencies in multi-dimensional contingency tables. *Journal of the American Statistical Association*, **66**, 339–44.
- Goodman, L. A. (1973). The analysis of multidimensional contingency tables when some variables are posterior to others. A modified path analysis approach. *Biometrika*, **60**, 179–92.
- Green, P. J. (1990). On use of the EM algorithm for penalized likelihood estimation. *Journal of the Royal Statistical Society, Series B*, **52**, 443–52.
- Haberman, S. J. (1974). *The analysis of frequency data*. University of Chicago Press, Chicago.
- Hammersley, J. M. and Clifford, P. E. (1971). Markov fields on finite graphs and lattices. Unpublished manuscript.
- Hastings, W. K. (1970). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, **57**, 97–109.
- Heckerman, D., Geiger, D., and Chickering, D. M. (1994). Learning Bayesian networks: the combination of knowledge and statistical data. In *Proceedings of the 10th conference on uncertainty in artificial intelligence*, (ed. R. L. de Mantaras and D. Poole), pp. 293–301. Morgan Kaufmann, San Mateo.
- Hocking, R. R. and Oxspring, H. H. (1974). The analysis of partially categorized contingency data. *Biometrics*, **30**, 469–83.
- Højsgaard, S. and Thiesson, B. (1995). BIFROST — Block recursive models Induced From Relevant knowledge, Observations, and Statistical Techniques. *Computational Statistics & Data Analysis*, **19**, 155–75.
- Ireland, C. T., Ku, H. H., and Kullback, S. (1969). Symmetry and marginal homogeneity of an $r \times r$ contingency table. *Journal of the American Statistical Association*, **64**, 1323–41.
- Jensen, C. S., Kong, A., and Kjærulff, U. (1995). Blocking–Gibbs sampling in very large probabilistic expert systems. *International Journal of Human–Computer Studies*, **42**, 647–66.
- Jensen, F., Jensen, F. V., and Dittmer, S. L. (1994). From influence diagrams to junction trees. In *Proceedings of the 10th conference on uncertainty in artificial intelligence*, (ed. R. L. de Mantaras and D. Poole), pp. 367–73. Morgan Kaufmann, San Mateo.

- Jensen, F. V. (1996). *An introduction to Bayesian networks*. University College London Press, London.
- 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), pp. 360–6. Morgan Kaufmann, San Mateo.
- Jensen, F. V., Lauritzen, S. L. and Olesen, K. G. (1990). Bayesian updating in causal probabilistic networks by local computation. *Computational Statistics Quarterly*, **4**, 269–82.
- Jensen, J. L. (1991). A large deviation type approximation for the Box class of likelihood ratio criteria. *Journal of the American Statistical Association*, **86**, 437–40.
- Jensen, J. L. (1995). Correction to: A large deviation type approximation for the Box class of likelihood ratio criteria. *Journal of the American Statistical Association*, **90**, 812.
- Jensen, S. T. (1989). *Exponentielle familier*, Lecture notes. Institute of Mathematical Statistics. University of Copenhagen. In Danish.
- Jensen, S. T., Johansen, S. and Lauritzen, S. L. (1991). Globally convergent algorithms for maximizing a likelihood function. *Biometrika*, **78**, 867–77.
- Johansen, S. (1979). *Introduction to the theory of regular exponential families*, Lecture notes. Institute of Mathematical Statistics, University of Copenhagen.
- Jöreskog, K. G. (1973). Analysis of covariance structures. In *Proceedings of the 3rd symposium on multivariate analysis, Dayton, Ohio*, (ed. P. R. Krishnaiah), pp. 263–85. Academic Press, New York.
- Jöreskog, K. G. (1977). Structural equation models in the social sciences: specification, estimation and testing. In *Applications of statistics*, (ed. P. R. Krishnaiah), pp. 267–87. North-Holland, Amsterdam.
- Jöreskog, K. G. (1981). Analysis of covariance structures. *Scandinavian Journal of Statistics*, **8**, 65–92.
- Kellerer, H. G. (1964a). Maßtheoretische Marginalprobleme. *Mathematische Annalen*, **153**, 168–98.
- Kellerer, H. G. (1964b). Verteilungsfunktionen mit gegebenen Marginalverteilungen. *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, **3**, 247–70.
- Kiiveri, H. and Speed, T. P. (1982). Structural analysis of multivariate data: a review. In *Sociological methodology*, (ed. S. Leinhardt). Jossey-Bass, San Francisco.
- Kiiveri, H., Speed, T. P. and Carlin, J. B. (1984). Recursive causal models. *Journal of the Australian Mathematical Society, Series A*, **36**, 30–52.

- Kjærulff, U. (1990). Graph triangulation — algorithms giving small total state space. Technical Report R 90-09, Aalborg University.
- Kjærulff, U. (1992). Optimal decomposition of probabilistic networks by simulated annealing. *Statistics and Computing*, **2**, 19–24.
- Koster, J. T. A. (1994). Gibbs-factorization and the Markov property. Unpublished manuscript.
- Koster, J. T. A. (1996). Markov properties for non-recursive causal models. *Annals of Statistics*, **24**, To appear.
- Kreiner, S. (1987). Analysis of multidimensional contingency tables by exact conditional tests: techniques and strategies. *Scandinavian Journal of Statistics*, **14**, 97–112.
- Kreiner, S. (1989). User's guide to DIGRAM - a program for discrete graphical modelling. Technical Report 89-10, Statistical Research Unit, University of Copenhagen.
- Krzanowski, W. J. (1975). Discrimination and classification using both binary and continuous variables. *Journal of the American Statistical Association*, **70**, 782–90.
- Krzanowski, W. J. (1976). Canonical representation of the location model for discrimination and classification. *Journal of the American Statistical Association*, **71**, 845–8.
- Kullback, S. (1959). *Information theory and statistics*. John Wiley and Sons, New York.
- Kullback, S. and Leibler, R. A. (1951). On information and sufficiency. *Annals of Mathematical Statistics*, **22**, 79–86.
- Lancaster, H. O. (1949). The derivation and partition of χ^2 in certain discrete distributions. *Biometrika*, **36**, 117–29.
- Lauritzen, S. L. (1985). Test of hypotheses in decomposable mixed interaction models. Technical Report R 85-11, Institute for Electronic Systems, Aalborg University.
- Lauritzen, S. L. (1989a). Lectures on contingency tables, (3rd edn). Technical Report R-89-29, Institute for Electronic Systems, Aalborg University.
- Lauritzen, S. L. (1989b). Mixed graphical association models (with discussion). *Scandinavian Journal of Statistics*, **16**, 273–306.
- Lauritzen, S. L. (1992). Propagation of probabilities, means and variances in mixed graphical association models. *Journal of the American Statistical Association*, **87**, 1098–108.
- Lauritzen, S. L. (1995). The EM algorithm for graphical association models with missing data. *Computational Statistics & Data Analysis*, **19**, 191–201.

- 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.
- Lauritzen, S. L. and Wernuth, N. (1984). Mixed interaction models. Technical Report R 84-8. Institute for Electronic Systems, Aalborg University.
- Lauritzen, S. L. and Wernuth, N. (1989). Graphical models for associations between variables, some of which are qualitative and some quantitative. *Annals of Statistics*, **17**, 31–57.
- Lauritzen, S. L., Speed, T. P., and Vijayan, K. (1984). Decomposable graphs and hypergraphs. *Journal of the Australian Mathematical Society, Series A*, **36**, 12–29.
- Lauritzen, S. L., Dawid, A. P., Larsen, B. N., and Leimer, H.-G. (1990). Independence properties of directed Markov fields. *Networks*, **20**, 491–505.
- Lauritzen, S. L., Thiesson, B., and Spiegelhalter, D. J. (1994). Diagnostic systems created by model selection methods: a case study. In *Selecting models from data: AI and statistics IV*, (ed. P. Cheeseman and R. Oldford), pp. 143–52. Lecture Notes in Statistics 89. Springer-Verlag, New York.
- Leimer, H.-G. (1985). Strongly decomposable graphs and hypergraphs, PhD thesis. Berichte zur Stochastik und verwandte Gebiete 85-1, University of Mainz.
- Leimer, H.-G. (1989). Triangulated graphs with marked vertices. In *Graph theory in memory of G. A. Dirac*, (ed. L. D. Andersen, C. Thomassen, B. Toft, and P. D. Vestergaard), pp. 311–24. Annals of Discrete Mathematics 41. Elsevier Science Publishers B.V. (North-Holland). Amsterdam.
- Leimer, H.-G. (1993). Optimal decomposition by clique separators. *Discrete Mathematics*, **113**, 99–123.
- Madigan, D. and Mosurski, K. (1990). An extension of the results of Asmussen and Edwards on collapsibility in contingency tables. *Biometrika*, **77**, 315–19.
- Madigan, D. and Raftery, A. (1994). Model selection and accounting for model uncertainty using Occam's window. *Journal of the American Statistical Association*, **89**, 1535–46.
- Madigan, D. and York, J. (1995). Bayesian graphical models for discrete data. *International Statistical Review*, **63**, 215–32.
- Martin-Löf, P. (1970). *Statistiska modeller*, Lecture notes by Rolf Sundberg. University of Stockholm. In Swedish.

- Matúš, F. (1992a). Ascending and descending conditional independence relations. In *Transactions of the 11th Prague conference on information theory, statistical decision functions and random processes*, pp. 189–200. Academia, Prague.
- Matúš, F. (1992b). On equivalence of Markov properties over undirected graphs. *Journal of Applied Probability*, **29**, 745–9.
- Matúš, F. and Studený, M. (1995). Conditional independences among four random variables I. *Combinatorics, Probability and Computing*, **4**, 269–78.
- Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., and Teller, E. (1953). Equations of state calculations by fast computing machines. *Journal of Chemical Physics*, **21**, 1087–92.
- Morgan, W. M. and Blumenstein, B. A. (1991). Exact conditional tests for hierarchical models in multidimensional contingency tables. *Applied Statistics*, **40**, 435–42.
- Moussouris, J. (1974). Gibbs and Markov random systems with constraints. *Journal of Statistical Physics*, **10**, 11–33.
- Neapolitan, E. (1990). *Probabilistic reasoning in expert systems*. John Wiley and Sons, New York.
- Oliver, R. M. and Smith, J. Q. (1990). *Influence diagrams, belief nets and decision analysis*. John Wiley and Sons, Chichester.
- Olkin, I. and Tate, R. F. (1961). Multivariate correlation models with mixed discrete and continuous variables. *Annals of Mathematical Statistics*, **32**, 448–65.
- Parter, S. (1961). The use of linear graphs in Gauss elimination. *SIAM Review*, **3**, 119–30.
- Patefield, W. M. (1981). Algorithm AS 159. An efficient method of generating random $r \times c$ tables with given row and column totals. *Applied Statistics*, **30**, 91–7.
- Paz, A. and Geva, R. (1996). Representation of irrelevance relations by annotated graphs. *Fundamenta Informaticae*. To appear.
- Pearl, J. (1986a). A constraint propagation approach to probabilistic reasoning. In *Uncertainty in artificial intelligence*, (ed. L. M. Kanal and J. Lemmer), pp. 357–70. North-Holland, Amsterdam.
- Pearl, J. (1986b). Fusion, propagation and structuring in belief networks. *Artificial Intelligence*, **29**, 241–88.
- Pearl, J. (1988). *Probabilistic reasoning in intelligent systems*. Morgan Kaufmann, San Mateo.
- Pearl, J. (1995). Causal diagrams for empirical research (with discussion). *Biometrika*, **82**, 669–710.

- 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), pp. 357–63. North-Holland, Amsterdam.
- Pearl, J. and Verma, T. (1987). The logic of representing dependencies by directed graphs. In *Proceedings of the 6th conference of American Association of Artificial Intelligence*, pp. 374–9. American Association of Artificial Intelligence.
- Porteous, B. T. (1985a). Improved likelihood ratio statistics for covariance selection models. *Biometrika*, **72**, 473–5.
- Porteous, B. T. (1985b). Properties of log-linear and covariance selection models. Unpublished PhD thesis, University of Cambridge.
- Porteous, B. T. (1989). Stochastic inequalities relating a class of log-likelihood ratio statistics to their asymptotic χ^2 distribution. *Annals of Statistics*, **17**, 1723–34.
- Range, P. R. (1979). Will he be the first? *The New York Times Magazine*. 11th March, 72–82.
- Read, T. R. C. and Cressie, N. A. C. (1988). *Goodness-of-fit statistics for discrete multivariate data*. Springer-Verlag, New York.
- Ripley, B. (1987). *Stochastic simulation*. John Wiley and Sons, New York.
- Rissanen, J. (1987). Stochastic complexity (with discussion). *Journal of the Royal Statistical Society, Series B*, **49**, 223–239 and 253–265.
- 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 of Computing*, **5**, 266–83.
- Sakamoto, Y. and Akaike, H. (1978). Analysis of cross-classified data by AIC. *Annals of the Institute of Statistical Mathematics*, **30**, 185–97.
- Schneider, B. E. (1978). The trigamma function. Algorithm AS 121. *Applied Statistics*, **27**, 97–9.
- Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics*, **6**, 461–4.
- Shafer, G. R. (1976). *A mathematical theory of evidence*. Princeton University Press, Princeton, N.J.
- Shafer, G. R. (1996). *The art of causal conjecture*. MIT Press, Cambridge, Massachusetts.
- Shafer, G. R. and Pearl, J. (ed.) (1990). *Readings in uncertain reasoning*. Morgan Kaufmann, San Mateo.
- Sheehan, N. and Thomas, A. (1993). On the irreducibility of a Markov chain defined on a space of genotype configurations by a sampling scheme. *Biometrics*, **49**, 163–75.

- Shenoy, P. P. and Shafer, G. R. (1990). Axioms for probability and belief-function propagation. In *Uncertainty in artificial intelligence IV*, (ed. R. D. Shachter, T. S. Levitt, L. N. Kanal, and J. F. Lemmer), pp. 169–98. North-Holland, Amsterdam.
- Simpson, E. H. (1951). The interpretation of interaction in contingency tables. *Journal of the Royal Statistical Society, Series B*, **13**, 238–41.
- Smith, A. F. M. and Roberts, G. O. (1993). Bayesian computation via the Gibbs sampler and related Markov chain Monte Carlo methods. *Journal of the Royal Statistical Society, Series B*, **55**, 3–23.
- Smith, J. Q. (1989). Influence diagrams for statistical modelling. *Annals of Statistics*, **17**, 654–72.
- Smith, P. W. F. (1992). Assessing the power of model selection procedures used when graphical modelling. In *Computational statistics*, (ed. Y. Dodge and J. Whittaker), pp. 275–80. Physica Verlag, Heidelberg.
- Speed, T. P. (1979). A note on nearest-neighbour Gibbs and Markov probabilities. *Sankhyā, Series A*, **41**, 184–97.
- Speed, T. P. and Kiiveri, H. (1986). Gaussian Markov distributions over finite graphs. *Annals of Statistics*, **14**, 138–50.
- Spiegelhalter, D. J. and Lauritzen, S. L. (1990). Sequential updating of conditional probabilities on directed graphical structures. *Networks*, **20**, 579–605.
- Spiegelhalter, D. J., Dawid, A. P., Lauritzen, S. L., and Cowell, R. G. (1993). Bayesian analysis in expert systems (with discussion). *Statistical Science*, **8**, 219–247 and 204–283.
- Spirtes, P., Glymour, C., and Scheines, R. (1993). *Causation, prediction and search*. Lecture Notes in Statistics 81. Springer-Verlag, New York.
- Stone, M. (1974a). Cross-validation and multinomial prediction. *Biometrika*, **61**, 509–15.
- Stone, M. (1974b). Cross-validatory choice and assessment of statistical predictions. *Journal of the Royal Statistical Society, Series B*, **36**, 111–33.
- Studený, M. (1989). Multiinformation and the problem of characterization of conditional independence relations. *Problems of Control and Information Theory*, **18**, 3–16.
- 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*, pp. 377–96. Academia, Prague.
- Studený, M. (1993). Structural semigraphoids. *International Journal of General Systems*, **22**, 207–17.

- Sundberg, R. (1975). Some results about decomposable (or Markov-type) models for multidimensional contingency tables: Distribution of marginals and partitioning of tests. *Scandinavian Journal of Statistics*, **2**, 71–9.
- Tarjan, R. E. (1985). Decomposition by clique separators. *Discrete Mathematics*, **55**, 221–32.
- 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 of Computing*, **13**, 566–79.
- Thiesson, B. (1991). (G)EM algorithms for maximum likelihood in recursive graphical association models. Unpublished Master's thesis, Aalborg University.
- Thompson, E. A. and Guo, S. W. (1991). Evaluation of likelihood ratios in complex genetic models. *IMA Journal of Mathematics Applied in Medicine and Biology*, **8**, 149–69.
- Verma, T. and Pearl, J. (1990a). Causal networks: semantics and expressiveness. In *Uncertainty in artificial intelligence IV*, (ed. R. D. Schachter, T. S. Levitt, L. N. Kanal, and J. F. Lemmer), pp. 69–76. North-Holland, Amsterdam.
- Verma, T. and Pearl, J. (1990b). Equivalence and synthesis of causal models. In *Proceedings of the 6th conference on uncertainty in artificial intelligence*, (ed. P. Bonissone, M. Henrion, L. N. Kanal, and J. F. Lemmer), pp. 255–70. North-Holland, Amsterdam.
- Vorobev, N. N. (1962). Consistent families of measures and their extensions. *Theory of Probability and its Applications*, **7**, 147–63.
- Vorobev, N. N. (1963). Markov measures and Markov extensions. *Theory of Probability and its Applications*, **8**, 420–9.
- Vorobev, N. N. (1967). Coalition games. *Theory of Probability and its Applications*, **12**, 250–66.
- Wagner, K. (1937). Über eine Eigenschaft der ebenen Komplexe. *Mathematische Annalen*, **114**, 570–90.
- Wedelin, D. (1993). Efficient algorithms for probabilistic inference, combinatorial optimization, and the discovery of causal structure from data. Unpublished PhD thesis, University of Gothenburg.
- Wermuth, N. (1976a). Analogies between multiplicative models in contingency tables and covariance selection. *Biometrics*, **32**, 95–108.
- Wermuth, N. (1976b). Model search among multiplicative models. *Biometrics*, **32**, 253–63.
- Wermuth, N. (1980). Linear recursive equations, covariance selection and path analysis. *Journal of the American Statistical Association*, **75**, 963–72.

- Wermuth, N. (1987). Parametric collapsibility and the lack of moderating effects in contingency tables. *Journal of the Royal Statistical Society, Series B*, **49**, 353–64.
- Wermuth, N. (1989a). Moderating effects in multivariate normal distributions. *Methodika*, **3**, 74–93.
- Wermuth, N. (1989b). Moderating effects of subgroups in linear models. *Biometrika*, **76**, 81–92.
- Wermuth, N. (1992). Block-recursive regression equations (with discussion). *Revista Brasileira de Probabilidade e Estatística*, **6**, 1–56.
- Wermuth, N. and Lauritzen, S. L. (1983). Graphical and recursive models for contingency tables. *Biometrika*, **70**, 537–52.
- Wermuth, N. and Lauritzen, S. L. (1990). On substantive research hypotheses, conditional independence graphs and graphical chain models (with discussion). *Journal of the Royal Statistical Society, Series B*, **52**, 21–72.
- Whittaker, J. (1984). Fitting all possible decomposable and graphical models to multiway contingency tables. In *COMPSTAT 84*, (ed. T. Havránek), pp. 98–108. Physica Verlag, Wien.
- Whittaker, J. (1990a). Contribution to the discussion of David Edwards: Hierarchical interaction models. *Journal of the Royal Statistical Society, Series B*, **52**, 54.
- Whittaker, J. (1990b). *Graphical models in applied multivariate statistics*. John Wiley and Sons. Chichester.
- Whittemore, A. (1978). Collapsibility of multidimensional contingency tables. *Journal of the Royal Statistical Society, Series B*, **40**, 328–40.
- Wilks, S. S. (1932). Certain generalizations in the analysis of variance. *Biometrika*, **24**, 471–94.
- Wilks, S. S. (1935). On the independence of k sets of normally distributed statistical variables. *Econometrica*, **3**, 309–26.
- Wold, H. D. A. (1954). Causality and econometrics. *Econometrica*, **22**, 162–77.
- Wold, H. D. A. (1960). A generalization of causal chain models. *Econometrica*, **28**, 443–63.
- Wright, S. (1921). Correlation and causation. *Journal of Agricultural Research*, **20**, 557–85.
- Wright, S. (1923). The theory of path coefficients: a reply to Niles' criticism. *Genetics*, **8**, 239–55.
- Wright, S. (1934). The method of path coefficients. *Annals of Mathematical Statistics*, **5**, 161–215.

- Yule, G. U. (1903). Notes on the theory of association of attributes in statistics. *Biometrika*, **2**, 121–34.
- Zadeh, L. A. (1965). Fuzzy sets. *Information and Control*, **8**, 338–53.
- Zadeh, L. A. (1978). Fuzzy sets as a basis for a theory of possibility. *Fuzzy Sets and Systems*, **1**, 3–28.

Index

- adjacent 5
- affine hypothesis 132, 268
- ancestor 6
- ancestral set 6

- Basu's theorem 241
- beta distribution 262
- BIFROST 230
- block-recursive
 - Markov property 54
- block-recursive model 113
 - decomposable 119
 - graphical 114
 - hierarchical 118
- boundary 5
- Box type 171, 192, 262
- BUGS 233

- canonical parameter 132, 266
- cells 67, 158, 246
 - marginal 68, 246
- CG distribution 158, 159
 - canonical characteristics 159
 - conditional 164
 - canonical characteristics 164
 - moment characteristics 165
 - homogeneous 160
 - marginal 160, 161
 - moment characteristics 161, 162
 - weak 162, 165
 - mixed characteristics 160
 - moment characteristics 159
- CG regression 165, 217
 - homogeneous 165
- chain 6
 - active 48
 - blocked 48
 - dependence 7, 113
- chain component 7
 - terminal 7
- chain graph 7
 - factorization 53
 - Markov property 54
- chain graph model 114
 - estimation 115, 154, 218
- characteristic function 253
- child 5
- chord 9
- classification criteria 67
- clique 5
- closure 6
- CoCo 84, 101, 105, 230
- collapsibility 121, 156, 219
- complete
 - graph 5
 - statistic 268, 269
 - subset 5
- component
 - chain 7
 - connectivity 6
- concentration 251
 - matrix 129, 251
 - operator 251
- conditional independence 28, 32
 - normal distribution 129
- contingency table 67
 - asymptotic distributions 76
- cells 67
- counts 68
- dimension 68
- levels 67
- log-affine model 72
 - deviance 79
 - estimation 74
 - extended 72
- log-linear model 72
- marginal 68
 - cells 68
 - counts 68
- sampling schemes 69, 70

- contingency table (*cont.*)
 - saturated model 70
 - estimation 71
 - slice 68
 - structural zeros 72
- counts 68
- covariance 250
 - matrix 251
 - operator 251
- covariance selection model 131
 - decomposable 144
 - estimation 145
 - test 149, 151, 152
 - decomposition 136
 - direct join 136
 - estimation 133, 135, 138
 - likelihood equations 133
 - likelihood function 132
 - test 142
- cross-product ratio
 - partial 37
- curved exponential model 273
- cut 204, 242, 274
 - mixed model 183
- cycle 6
 - directed 6
- d-separation 48
- decomposable
 - graph 8
 - hypergraph 23
- decomposable model 90
 - asymptotics 95
 - estimation 91
 - sampling distribution 93
 - test 98
- decomposition 7
 - proper 8
 - strong 7
 - weak 8
- dependence chain 7, 53, 113
- descendant 6
- DIGRAM 101
- direct join 22
 - covariance selection model 136
 - hierarchical model 84
- divergence 162, 238
- edge 4
 - directed 4
 - undirected 4
- EM algorithm 233
- exact test 269
- expectation 250
- expert system 221
- exponential model 266
 - base measure 266
 - canonical parameter 266
 - canonical statistic 266
 - convex support 267
 - curved 272, 273
 - order 267
 - regular 267
- factor subspace 74, 246
- factorization
 - chain graph 53
 - recursive 46
- Fisher information 273
- forest 7
 - junction 24
- gamma distribution 263
- Gaussian graphical model 131
- generalized inverse 245, 251
- generating class 82, 249
 - graphical 89
 - mixed 202
 - graphical 203
- Gibbs sampler 231
- graph 4
 - chain 7
 - complete 5
 - decomposable 8
 - directed 5
 - directed version 5
 - dual 33
 - interaction 88
 - marked 4
 - moral 7
 - perfect 7
 - simple 4
 - star 10
 - triangulated 9
 - undirected 5

- undirected version 5
- weakly decomposable 9
- graphical hypergraph 22, 89
- graphical model 64, 89
 - decomposable 90
 - recursive 64
- graphoid 30
- hierarchical model 81
 - dimension 86
 - direct join 84
 - estimation 82
 - graphical 89
 - subspace 247
 - mixed 213, 215
 - test 85, 87
- hierarchical models
 - likelihood equations 82
- history 15
- HUGIN 227, 236
- hyper Markov 230
 - directed 109
 - property 85, 94, 148
- hyperedge 21
- hypergraph 21
 - clique 21
 - decomposable 23
 - graphical 22, 89
 - reduced 22
 - simple 21
- information inequality 238
- interaction 1, 82, 131, 248
 - CG 173
 - discrete 174
 - linear 174
 - mixed 174
 - quadratic 174
- interaction graph 88, 203
- IPS-algorithm 83, 134
- iterative partial maximization 135, 239, 271
- iterative proportional scaling 82
 - covariance selection model 134
 - general 272
 - modified 206, 272
- join
 - direct 22
- junction
 - forest 24
 - property 24
 - tree 24, 227
- Kronecker product 244
- Kullback–Leibler divergence 162, 238
- Lancaster components 101
- learning 228
- levels 67
- line search 271
- log–affine model 67, 72
- log–linear model 64, 72
- logarithm 237
- Möbius inversion 36, 239
- main effect 174, 248
 - discrete 174
 - linear 174
 - quadratic 174
- marginal
 - cells 246
 - table 68, 246
- Markov chain Monte Carlo 231
- Markov equivalence 60
- Markov property
 - chain graph
 - block-recursive 54
 - global 55
 - local 55
 - pairwise 55
 - directed
 - global 47
 - local 50
 - pairwise 50
 - hyper 85, 94, 148
 - undirected 131
 - decomposition 42–44
 - extended 40
 - factorization 34, 36
 - global 32
 - local 32
 - pairwise 32, 36
 - positive 36

- MCMC 231
- mean 250
- MIM 101, 201, 230
- MIM model 201, 203
- mixed interaction model
 - decomposable 187
 - estimation 188
 - test 191
- graphical 173
 - decomposition 179
 - estimation 175
 - likelihood equations 178
- hierarchical 199
 - estimation 205
 - likelihood equations 206
- mixed model
 - cut 183, 204
 - dimension 172
 - generating class 202
 - graphical 203
 - likelihood function 172
 - sampling distributions 168
 - saturated 168
 - estimation 169
 - homogeneous 168
 - test 170
- neighbour 5
- non-adjacent 5
- non-descendant 6
- normal distribution 254
 - concentration 255
 - conditional 256, 258
 - conditional independence 129
 - covariance 254
 - density 254, 258
 - marginal 256
- outer product 246, 259
- parent 5
- partial correlation coefficient 130
- partial regression coefficient 130
- path 6
- perfect
 - directed version 18
 - graph
 - numbering 15
 - sequence 14
 - probability propagation 227, 235
- recursive graphical model
 - estimation 108
- recursive hierarchical model
 - estimation 112
- residual 15
- running intersection property 15, 25
- saturated model
 - contingency table 70
 - Gaussian 124
 - estimation 125
 - likelihood function 124
- mixed 168
 - estimation 169
 - recursive 107
- semi-graphoid 30
- separate 6
- separator 6, 15
- simplicial 13
- split 242, 274
- statistic
 - complete 241, 268, 269
 - sufficient 241
- sufficiency 241
- tree 6
 - junction 24, 227
- vertex 4
 - continuous 4
 - discrete 4
 - simplicial 13
 - terminal 7
- Wilks's distribution 263
- Wishart distribution 258
 - constant 261
 - covariance 258
 - density 261
 - mean 258
- Yule-Simpson paradox 65

The idea of modelling systems using graph theory has its origin in several scientific areas: in statistical physics (the study of large particle systems), in genetics (studying inheritable properties of natural species), and in interactions in contingency tables. The use of graphical models in statistics has increased considerably over recent years and the theory has been greatly developed and extended. This book provides the first comprehensive and authoritative account of the theory of graphical models and is written by a leading expert in the field. It contains the fundamental graph theory required and a thorough study of the Markov properties associated with various types of graphs. The statistical theory of log-linear and graphical models for contingency tables and for mixed discrete-continuous variables, is developed in detail. Special topics, such as the application of graphical models to probabilistic expert systems, are described briefly, and appendices give details of the multivariate normal distribution and of the theory of regular exponential families.

ALSO AVAILABLE IN THIS SERIES

Analysis of longitudinal data

Peter J. Diggle, Kung-Yee Liang, and Scott L. Zeger

Practical methods for reliability data analysis

J. I. Ansell and M. J. Phillips

Modelling frequency and count data

J. K. Lindsey

Saddlepoint approximations

J. L. Jensen

OXFORD UNIVERSITY PRESS

ISBN 0-19-852219-3



9 780198 522195 >

Meta 1023

p. 107