To Monica, Anna and Matteo
# Contents

**Preface**

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Preface</td>
<td>vii</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Chapter</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Introduction</td>
<td>1</td>
</tr>
<tr>
<td>1.1</td>
<td>1</td>
</tr>
<tr>
<td>Graphical Models</td>
<td>1</td>
</tr>
<tr>
<td>1.2</td>
<td>2</td>
</tr>
<tr>
<td>Outline of the Book</td>
<td>2</td>
</tr>
<tr>
<td>1.2.1</td>
<td>2</td>
</tr>
<tr>
<td>Discrete Graphical Models and Their</td>
<td>2</td>
</tr>
<tr>
<td>Parameterization</td>
<td></td>
</tr>
<tr>
<td>1.2.2</td>
<td>3</td>
</tr>
<tr>
<td>Binary vs Non-binary Variables</td>
<td>3</td>
</tr>
<tr>
<td>2</td>
<td>4</td>
</tr>
<tr>
<td>Conditional Independence and Cross-product Ratios</td>
<td>4</td>
</tr>
<tr>
<td>2.1</td>
<td>4</td>
</tr>
<tr>
<td>Notation and Terminology</td>
<td>4</td>
</tr>
<tr>
<td>2.1.1</td>
<td>5</td>
</tr>
<tr>
<td>Cross-classified Tables</td>
<td>5</td>
</tr>
<tr>
<td>2.2</td>
<td>8</td>
</tr>
<tr>
<td>Conditional Independence</td>
<td>8</td>
</tr>
<tr>
<td>2.3</td>
<td>9</td>
</tr>
<tr>
<td>Establishing Independence Relationships</td>
<td>9</td>
</tr>
<tr>
<td>3</td>
<td>21</td>
</tr>
<tr>
<td>Möbius Inversion</td>
<td>21</td>
</tr>
<tr>
<td>3.1</td>
<td>21</td>
</tr>
<tr>
<td>Preliminaries</td>
<td>21</td>
</tr>
<tr>
<td>3.1.1</td>
<td>21</td>
</tr>
<tr>
<td>Notation and Terminology</td>
<td>21</td>
</tr>
<tr>
<td>3.1.2</td>
<td>22</td>
</tr>
<tr>
<td>The Zeta and the Möbius Matrices</td>
<td>22</td>
</tr>
<tr>
<td>3.2</td>
<td>24</td>
</tr>
<tr>
<td>The Möbius Inversion Formula</td>
<td>24</td>
</tr>
<tr>
<td>3.2.1</td>
<td>25</td>
</tr>
<tr>
<td>Two Basic Lemmas</td>
<td>25</td>
</tr>
<tr>
<td>3.3</td>
<td>27</td>
</tr>
<tr>
<td>Möbius Inversion and Partially Ordered Sets</td>
<td>27</td>
</tr>
<tr>
<td>4</td>
<td>34</td>
</tr>
<tr>
<td>Undirected Graph Models</td>
<td>34</td>
</tr>
<tr>
<td>4.1</td>
<td>34</td>
</tr>
<tr>
<td>Graphs</td>
<td>34</td>
</tr>
<tr>
<td>4.2</td>
<td>36</td>
</tr>
<tr>
<td>Markov Properties for Undirected Graphs</td>
<td>36</td>
</tr>
<tr>
<td>4.3</td>
<td>39</td>
</tr>
<tr>
<td>The Log-linear Parameterization</td>
<td>39</td>
</tr>
<tr>
<td>4.4</td>
<td>44</td>
</tr>
<tr>
<td>Hierarchical Log-linear Models</td>
<td>44</td>
</tr>
<tr>
<td>4.5</td>
<td>48</td>
</tr>
<tr>
<td>Log-linear Graphical Models</td>
<td>48</td>
</tr>
<tr>
<td>4.6</td>
<td>49</td>
</tr>
<tr>
<td>Data, Estimation and Testing</td>
<td>49</td>
</tr>
<tr>
<td>4.7</td>
<td>56</td>
</tr>
<tr>
<td>Graph Decomposition and Decomposable Graphs</td>
<td>56</td>
</tr>
<tr>
<td>Section</td>
<td>Title</td>
</tr>
<tr>
<td>---------</td>
<td>----------------------------------------------------------------------</td>
</tr>
<tr>
<td>4.8</td>
<td>Local Computation Properties</td>
</tr>
<tr>
<td>4.9</td>
<td>Models for Decomposable Graphs</td>
</tr>
<tr>
<td>4.10</td>
<td>Log-linear Models and the Exponential Family</td>
</tr>
<tr>
<td>4.10.1</td>
<td>Basic Facts on the Theory of the Exponential Family</td>
</tr>
<tr>
<td>4.10.2</td>
<td>The Cross-classified Bernoulli Distribution</td>
</tr>
<tr>
<td>4.10.3</td>
<td>Exponential Family Representations of the Saturated Model</td>
</tr>
<tr>
<td>4.10.4</td>
<td>Exponential Family Representation of Hierarchical Log-linear Models</td>
</tr>
<tr>
<td>4.11</td>
<td>Modular Structure of the Asymptotic Variance of ML Estimates</td>
</tr>
<tr>
<td>4.11.1</td>
<td>The Variance Function and the Asymptotic Variance of ML Estimates</td>
</tr>
<tr>
<td>4.11.2</td>
<td>Variances in the Saturated Model</td>
</tr>
<tr>
<td>4.11.3</td>
<td>Variances in Hierarchical Log-linear Models</td>
</tr>
<tr>
<td>4.11.4</td>
<td>Decompositions and Decomposable Models</td>
</tr>
<tr>
<td>5</td>
<td>Bidirected Graph Models</td>
</tr>
<tr>
<td>5.1</td>
<td>Bidirected Graphs</td>
</tr>
<tr>
<td>5.2</td>
<td>Markov Properties for Bidirected Graphs</td>
</tr>
<tr>
<td>5.3</td>
<td>The Log-mean Linear Parameterization</td>
</tr>
<tr>
<td>5.4</td>
<td>Log-mean Linear Graphical Models</td>
</tr>
<tr>
<td>5.5</td>
<td>Example: Symptoms in Psychiatric Patients</td>
</tr>
<tr>
<td>5.6</td>
<td>Parsimonious Graphical Modeling</td>
</tr>
<tr>
<td>6</td>
<td>Directed Acyclic and Regression Graph Models</td>
</tr>
<tr>
<td>6.1</td>
<td>Directed Acyclic Graphs</td>
</tr>
<tr>
<td>6.2</td>
<td>Markov Properties for Directed Acyclic Graphs</td>
</tr>
<tr>
<td>6.3</td>
<td>Regression Graphs</td>
</tr>
<tr>
<td>6.4</td>
<td>Markov Properties for Regression Graphs</td>
</tr>
<tr>
<td>6.5</td>
<td>On the Interpretation of Models defined by Regression Graphs</td>
</tr>
<tr>
<td>6.6</td>
<td>The Log-hybrid Linear Parameterization</td>
</tr>
<tr>
<td>6.7</td>
<td>Log-hybrid Linear Graphical Models</td>
</tr>
<tr>
<td>6.8</td>
<td>Inference in Regression Graph Models</td>
</tr>
<tr>
<td></td>
<td><strong>Bibliography</strong></td>
</tr>
</tbody>
</table>
Preface

This book arises out of a short course given in a Séminaires Européens de Statistiques (SemStat) meeting at the European Institute for Statistics, Probability, Stochastic Operations Research and their Applications (EURANDOM) in Eindhoven, The Netherlands, over March 7–10, 2017. This SemStat meeting was organized as a part of the COST Action “European Cooperation for Statistics of Network Data Science” (COSTNET, CA15109) with the aim of introducing early career researchers to the field of statistical network science. In this perspective, the material presented here concerns the theory of graphical models and includes well-established methodology from the early developments in this field, but also the theory of models introduced more recently in the graphical model literature. The focus is on the discrete case where all the variables involved in the analysis are categorical and, in this context, classical and more recent results are presented in a unified way.

The approach followed in this text regarding the parameterization of the models is the result of a long-standing collaboration with Luca La Rocca and Monia Lupparelli on graphical models for discrete data. My views on this research area have also benefited from interactions with people of the “graphical model community,” who shared with me their ideas and enthusiasm during these years. I would not be the author of this book were it not for the encouragement of Ernst Wit who also made valuable comments on the manuscript. Finally, I would like to thank my family, Monica, Anna and Matteo, for their support.

The material in this book is based upon work supported by the Air Force Office of Scientific Research under award number FA9550-17-1-0039.

Alberto Roverato
March 2017
1 Introduction

1.1 Graphical Models

Graphical models are an elegant framework that combines uncertainty and graph theory to represent complex phenomena. A graph is a structure consisting of a set of objects, called vertices, and a set of connections between pairs of vertices, called edges. The vertices of the graph associated with a graphical model are the variables of the model and the edges describe how the variables interact with each other. A further fundamental component of graphical models is the notion of independence or, more generally, of conditional independence. The edges missing from the graph can be interpreted as absence of interaction, in the sense that variables are conditionally independent. An appealing characteristic of graphs is that they can be represented graphically and many of their features and properties can be understood from the visual inspection of their graphical representation. The graph greatly simplifies the interpretation of the model, making its independence structure more immediate and intuitive. This also facilitates the communication of the scientific contents of the model to researchers who are not familiar with the statistical formalism. In large models the information provided by the visual inspection of the graph may be less clear but, nevertheless, the graph is useful in many ways. To mention a few, the graph structure may imply an intrinsic modularity that allows one to split the model into submodels of smaller dimensions, so that analyses of interest and statistical inference procedures can be carried out locally on marginal distributions. Furthermore, the graph is a natural object to be dealt with in the implementation of algorithms and computational procedures. All in all, graphical models constitute a very versatile methodology that has proved useful in a wide range of domains and applications. An historical overview of graphical models, with a comprehensive list of early references, can be found in Cox and
Wermuth (1996, section 2.13) and Lauritzen (1996, chapter 1), whereas we refer to Drton et al. (2017) for a recent collection of reviews.

1.2 Outline of the Book

Graphical modeling has been a very active area of research in the last years, and nowadays a wide range of families of graphical models are available; see Sadeghi and Lauritzen (2014). This book focuses on categorical data and describes the theory of some of the most relevant classes of graphical models. In this context, the book aims at providing a unified view of the theory discussed in a largely scattered literature.

The different families of models can be distinguished for the kind of graph associated with the probability distribution. The models considered in Chapters 4 and 5 deal with symmetric relationships between variables represented by undirected and bidirected graphs, respectively. Undirected graph models, also known as Markov random fields, are characterized by collections of conditional independence relationships, whereas bidirected graph models are characterized by collections of marginal independencies. Asymmetric relationships between variables are introduced in Chapter 6. Firstly, we consider the family of models associated with directed acyclic graphs (DAGs), also known as Bayesian networks, and then models associated with regression graphs. The latter family includes each of the previous families of models as special case and, therefore, regression graph models constitute a general framework that unifies and extends the theory of undirected, bidirected and directed acyclic graph models.

1.2.1 Discrete Graphical Models and Their Parameterization

The graph associated with the model allows one to abstract the conditional independence relationships between variables from the details of their parametric forms. Accordingly, the interpretation of a graphical model and the inferential questions of interest usually do not depend on the kinds of variables involved in the analysis. On the other hand, the effective specification of the statistical model and the implementation of statistical techniques require the definition of a suitable parameterization that cannot prescind from the variable types. From this perspective, the attention is restricted to categorical variables, that are variables
taking one of a finite number of possible values. Our approach to the specification of parameterizations aims at building a common framework encompassing all the statistical models considered. Chapter 2 deals with the probability distribution of a random vector in the form of a probability table, and provides a set of rules for establishing conditional independence, based on cross-product ratios. Chapter 3 gives the theory of Möbius inversion that is then used to transform probabilities into more convenient log-linear parameterizations. Jointly, Chapters 2 and 3 provide a set of tools that can be applied, almost identically, to all the families of models we consider to obtain the corresponding parameterizations and derive their properties.

1.2.2 Binary vs Non-binary Variables

In its simplest form, a categorical variable takes one of two possible values and is thus called a binary variable. Dealing with binary variables is especially convenient because, in this setting, the notation is less involved and the presentation of the material is more immediate. Furthermore, it is often the case that the properties of the binary case can be used as building blocks for the derivation of the corresponding properties for the general case. Throughout this text, we keep a clear distinction between the general case of arbitrary categorical variables and the special case of binary variables. We deem that this approach may facilitate the comprehension of the material because the treatment of the general case can often be regarded as a mere technical generalization of the binary case. For this reason, when suitable, we will first discuss the binary case in detail, and then we will consider the general case in a separate subsection, which the reader may choose to skip on first reading.
Conditional Independence and Cross-product Ratios

In this chapter we introduce the notation and the terminology used to deal with categorical variables, and give the notions of independence and conditional independence, which play a central role in the theory of graphical models.

One important issue concerning the characterization of families of graphical models is the specification of suitable parameterizations that may allow one to efficiently deal with the collection of independence relationships defining the model. Although different families of graphical models require different parameterizations, for the models we consider it is possible to construct a unified approach based on the connection between independence relationships and cross-product ratios, which can be implemented by means of a tool called Möbius inversion. One advantage of this framework is that it leads to a simplification of the theory, in the sense that several fundamental results follow immediately from the application of a few basic lemmas connecting cross-product ratios to Möbius inversion. On the other hand, this approach sheds light on some common features of different families of graphical models, which are not usually presented jointly in the literature.

The material concerning cross-product ratios is given here, whereas the theory of Möbius inversion and the derivation of its relevant properties are deferred to the next chapter.

2.1 Notation and Terminology

This section is devoted to illustrating the notation we use to deal with vectors and cross-classified tables. It is important that the readers feel comfortable with these conventions. In our notation, vectors and matrices are indexed by the elements of a finite set. For instance, for a finite set \( V \) with cardinality \( p = |V| \) we write \( Y_V = (Y_v)_{v \in V} \) to denote a vector of \( p \)
random variables indexed by the elements of $V$, and we will simply say that $Y_V$ is indexed by $V$. Furthermore, we denote by $0_V$ the vector of length $p$ such that all entries are equal to zero and indexed by an element of $V$ and, similarly, $1_V$ is a vector with all entries equal to one. Any subset $A \subseteq V$ with $A \neq \emptyset$ identifies the subvector $Y_A = (Y_v)_{v \in A}$ of $Y_V$; we regard vectors as column vectors and denote by $Y_A^T$ the transpose of $Y_A$.

2.1.1 Cross-classified Tables

Let $Y_V = (Y_v)_{v \in V}$ be a vector of categorical random variables indexed by the finite set $V$ with $p = |V|$. Categorical variables can take a finite number of values that are usually called levels. For notational convenience, we label the levels of the variables $Y_v$, for $v \in V$, as $0, 1, \ldots, d_v$ so that $\mathcal{I}_v = \{0, 1, \ldots, d_v\}$ is the state space of $Y_v$ and the Cartesian product $\mathcal{I}_V = \times_{v \in V} \mathcal{I}_v$ is the state space of $Y_V$. We will refer to $\mathcal{I}_V$ as a $p$-dimensional cross-classified table, or simply table for short. Accordingly, the elements of $\mathcal{I}_V$, denoted by $i_V = (i_v)_{v \in V}$, are called the cells of the table. The state space of the subvector $Y_A$, with $A \subseteq V$ and $A \neq \emptyset$, is the marginal table $\mathcal{I}_A = \times_{v \in A} \mathcal{I}_v$ with cells $i_A \in \mathcal{I}_A$.

A random variable is called binary or dichotomous if it takes only two possible values and we say that a variable is non-binary or polytomous when it can take an arbitrary, finite, number of values. We will start by considering vectors of binary variables, and use the subsets of $V$ to index the cells of the associated tables. Then, we will consider the general case of non-binary vectors and introduce a notation that will allow us to represent the associated table as a collection of binary subtables. This approach facilitates both the presentation of the material and the proof of the results, typically given for the special case of binary variables, and then naturally extended to the general polytomous case by repeated application of the result for the binary case to all the binary subtables.

The Binary Case

Every binary variable takes values in the set $\{0, 1\}$ and the state space of $Y_V$ is the Cartesian product $\{0, 1\}^p$, which we call a $p$-dimensional binary table. A fundamental feature of the notation we use is that the entries of $Y_V$ are indexed by the elements of $V$ so that the cells of the binary table $\mathcal{I}_V$ associated with $Y_V$ can be indexed by subsets of $V$; i.e.,
by the power set of $V$. Concretely, we use the convention that the cell associated with the subset $D \subseteq V$ is the vector indexed by $V$ with all the entries indexed by $D$ equal to one, and all the entries indexed by $V \setminus D$ equal to zero. In this way, every subset $D \subseteq V$ uniquely identifies a cell $(1_D, 0_{V\setminus D})$, and we can write the binary table associated with $Y_V$ as

$$I_V = \{(1_D, 0_{V\setminus D})\}_{D \subseteq V}. \quad (2.1)$$

Notice that vectors indexed by the empty set are void and, therefore, when $D = \emptyset$ it holds that $(1_D, 0_{V\setminus D}) = 0_V$, whereas if $D = V$ then $(1_D, 0_{V\setminus D}) = 1_V$. An instance of three-way binary table can be found in Table 2.1. In this example, every cell of the table contains the corresponding indexing set, and one should notice that the indexing sets identify the variables whose levels are equal to one, whereas all the variables not included in the indexing set take value zero.

In the remainder of this text we will use some simplifications in order to lighten the notation. First, unless we deem it useful to highlight it, we will omit the subscript when it is equal to $V$ so that, for instance, $Y = Y_V$, $I = I_V$ and $i = i_V$. Second, for a subset $D \subseteq V$ we will write $Y_D = 1$ to denote $Y_D = 1_D$, that is, that all the entries of $Y_D$ take value one. Similarly, we will write $Y_D = 0$ in place of $Y_D = 0_D$. Finally, we will omit the brackets from singleton sets, as already done in Table 2.1.

### The Non-binary Case

The case of categorical variables with arbitrary number of levels can be conveniently dealt with by introducing a version of the sample space deprived of one level for every variable. Formally, the *restricted state space* of $Y_v$ is defined as $J_v = I_v \setminus \{0\} = \{1, \ldots, d_v\}$ so that the restricted state space of $Y$ is given by $J = J_V = \times_{v \in V} J_v$. Hereafter,

### Table 2.1 Indexing sets of the cells of the cross-classified table associated with the binary vector $Y_{\{a,b,c\}}$.

<table>
<thead>
<tr>
<th>$Y_a$</th>
<th>$Y_b$</th>
<th>$Y_c = 0$</th>
<th>$Y_b$</th>
<th>$Y_c = 1$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>$\emptyset$</td>
<td>$b$</td>
<td>0</td>
<td>$c$</td>
</tr>
<tr>
<td>1</td>
<td>$a$</td>
<td>${a, b}$</td>
<td>1</td>
<td>${a, c}$</td>
</tr>
</tbody>
</table>
we will refer to “0” as the baseline level of \( Y_v \) and we remark that the choice of the level to be set as baseline is arbitrary. The elements of \( \mathcal{J} \) are denoted by \( j = j_V \) and for every \( A \subseteq V \), with \( A \neq \emptyset \), we denote by \( \mathcal{J}_A \) the restricted state space of \( Y_A \). Every pair \( j \in \mathcal{J} \) and \( D \subseteq V \) uniquely identifies the cell \((j_D, 0_{V \setminus D}) \) in \( \mathcal{I} \), with the usual convention that for \( D = \emptyset \) the vector \( j_D \) is void, and therefore \((j_{\emptyset}, 0_v) = 0_v \). Hence, every \( j \in \mathcal{J} \) uniquely identifies the subtable
\[
\{(j_D, 0_{V \setminus D})\}_{D \subseteq V},
\] (2.2)
and with a slight abuse of terminology we will refer to (2.2) as a binary subtable. Table 2.2 gives an instance of the binary subtable identified by \( j_V \in \mathcal{J}_V \) with \( V = \{a, b, c\} \). It follows that the introduction of the restricted state space allows us to write the state space of \( Y_V \) as a collection of binary subtables, one for every \( j \in \mathcal{J} \); formally,
\[
\mathcal{I} = \{(j_D, 0_{V \setminus D})\}_{j \in \mathcal{J}, D \subseteq V}.
\] (2.3)
Clearly, in the case of binary variables, (2.3) coincides with (2.1) because the restricted state space \( \mathcal{J} \) has only one element, that is the vector \( 1_V \).

The representation of \( \mathcal{I} \) as the collection of the binary subtables \( \{(j_D, 0_{V \setminus D})\}_{D \subseteq V}, \) for \( j \in \mathcal{J}, \) is useful in practice because we will see that several relevant properties that hold in the binary case can be immediately generalized to the non-binary case by their iterative application for \( j \in \mathcal{J} \). On the other hand, we remark that this representation of \( \mathcal{I} \) is not efficient because some cells are simultaneously present in more than one binary subtable. The most relevant example of this situation is given by

<table>
<thead>
<tr>
<th>( Y_c = 0 )</th>
<th>( Y_c = j_c )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( Y_a )</td>
<td>( Y_b )</td>
</tr>
<tr>
<td>( 0 )</td>
<td>( 0_{{a,b,c}} )</td>
</tr>
<tr>
<td>( j_a )</td>
<td>( (j_{{a,b}}, 0_c) )</td>
</tr>
</tbody>
</table>
the cell $0_V$ that is contained in every subtable $\{(j_D, 0_{V \setminus D})\}_{D \subseteq V}$ for $j \in J$.

### 2.2 Conditional Independence

We introduce now the important notions of independence and conditional independence of random variables. Then, we characterize a family of cross-product ratios and show how these quantities can be used to assess independence relationships.

The probability mass function of a categorical random vector $Y_V$ can be given in the form of a probability table

$$\{p(Y = i)\}_{i \in I}$$

and, throughout this text, we restrict our attention to random vectors with positive probability mass function, that is

$$p(Y = i) > 0 \quad \text{for all } i \in I.$$

The marginal probability table of the subvector $Y_A$ is $\{p(Y = i_A)\}_{i_A \in I_A}$, and if $A$ and $B$ are two disjoint subsets of $V$, that is $A, B \subseteq V$ with $A \cap B = \emptyset$, then the subvector $Y_{A \cup B}$ can also be written as $(Y_A, Y_B)$ and, therefore, the forms $p(Y_{A \cup B} = i_{A \cup B})$ and $p(Y_A = i_A, Y_B = i_B)$ are equivalent. Accordingly, if $B = \emptyset$ then $p(Y_A = i_A, Y_B = i_B) = p(Y_A = i_A)$. More generally, for every $B \subseteq V \setminus A$,

$$p(Y_A = i_A) = \sum_{i_B \in I_B} p(Y_A = i_A, Y_B = i_B),$$

and thus it makes sense to use the convention that $p(Y_A = i_A) = 1$ when $A = \emptyset$.

The notions of independence and conditional independence of random variables are central to the interpretation of graphical models.

**Definition 2.2.1** Let $Y_V$ be a vector of categorical random variables and let $A$ and $B$ be two nonempty, disjoint, subsets of $V$. We say that $Y_A$ is independent of $Y_B$ if and only if

$$p(Y_A = i_A, Y_B = i_B) = p(Y_A = i_A)p(Y_B = i_B),$$

for every $i_A \in I_A$ and $i_B \in I_B$. 


It is straightforward to see that equation (2.4) is equivalent to either of the equations
\[ p(Y_A = i_A | Y_B = i_B) = p(Y_A = i_A) \]  
(2.5)
and
\[ p(Y_B = i_B | Y_A = i_A) = p(Y_B = i_B), \]
each of which can thus replace (2.4) to obtain two alternative, commonly used, ways to define the independence of random vectors. Hereafter, we shall use the notation introduced by Dawid (1979), and write \( Y_A \perp Y_B \) to denote the independence of \( Y_A \) and \( Y_B \).

The notion of independence can be generalized to that of conditional independence, that is the independence of two random vectors \( Y_A \) and \( Y_B \) for each fixed value of a third vector \( Y_C \).

**Definition 2.2.2** Let \( Y_V \) be a vector or categorical random variables, and let \( A, B \) and \( C \) be three nonempty, pairwise disjoint, subsets of \( V \). We say that \( Y_A \) is *conditionally independent of* \( Y_B \) *given* \( Y_C \), and write \( Y_A \perp Y_B | Y_C \), if and only if
\[ p(Y_A = i_A, Y_B = i_B | Y_C = i_C) = p(Y_A = i_A | Y_C = i_C)p(Y_B = i_B | Y_C = i_C), \]  
(2.6)
for every \( i_A \in \mathcal{I}_A, i_B \in \mathcal{I}_B \) and \( i_C \in \mathcal{I}_C \).

As well as for the definition of independence, alternative ways to define conditional independence are obtained by replacing equation (2.6) by either
\[ p(Y_A = i_A | Y_B = i_B, Y_C = i_C) = p(Y_A = i_A | Y_C = i_C), \]
or
\[ p(Y_B = i_B | Y_A = i_A, Y_C = i_C) = p(Y_B = i_B | Y_C = i_C). \]

### 2.3 Establishing Independence Relationships

We now give some alternative ways to establish independence relationships that will lead to the introduction of a relevant class of cross-product ratios. We start from the binary case and then extend the results to the general case of polytomous variables.
The Binary Case

The cells of the cross-classified table associated with a binary vector $Y_V$ are indexed by the subsets of $V$ and, accordingly, the probability table can be written as

$$
\{p(Y_D = 1, Y_{V\setminus D} = 0)\}_{D \subseteq V}.
$$

There exist several alternative ways to establish the independence of two binary random vectors. For example, consider the probabilities for the binary variables $Y_a$ and $Y_b$ given in Table 2.3. It is easy to check that a necessary and sufficient condition for the independence of $Y_a$ and $Y_b$ is that the following identity is satisfied

$$
p(Y_a = 1|Y_b = 1) = p(Y_a = 1|Y_b = 0). \tag{2.7}
$$

Another equality that constitutes a necessary and sufficient condition for the independence of these two variables is

$$
p(Y_a = 1|Y_b = 1) = p(Y_a = 1). \tag{2.8}
$$

From our viewpoint, the relevant difference between equations (2.7) and (2.8) is that the former only involves the conditional distributions of $Y_a|Y_b$, whereas the latter also involves a marginal distribution of $Y_a$. We can, rather informally, say that equation (2.7) originates from a “conditional approach,” whereas (2.8) comes from a “marginal approach,” and this distinction becomes especially relevant in multivariate settings with more than two variables. The probability distribution of a categorical random vector can be given in the form of a collection of marginal and conditional probabilities. This representation is not unique and the collection of probabilities characterizing the distribution may be suitably chosen so as to facilitate dealing with the type of independence.
relationships of interest, which may be marginal independencies between subsets of variables, conditional independencies involving all the variables or a combination of these two situations. This issue can be approached with the proper level of generality by introducing the two functions $c(\cdot)$ and $m(\cdot)$ defined on the power set of $V$. More precisely, for every $D \subseteq V$, the function $c(\cdot)$ is such that $c(D) = D$, whereas $m(\cdot)$ is such that $m(D) = \emptyset$. Then, if $A$ is a subset of $V$ and $A' \subseteq A$ we define

$$p_f(A', A) = p(Y_{A'} = 1, Y_{f(A', A')} = 0),$$

where $f(\cdot)$ can be either $c(\cdot)$ or $m(\cdot)$, which we compactly write as $f \in \{c, m\}$. Notice that

$$p_m(A', A) = p(Y_{A'} = 1)$$

is a probability from the marginal distribution of the subvector $Y_{A'}$ of $Y_A$, whereas

$$p_c(A', A) = p(Y_{A'} = 1, Y_{A \setminus A'} = 0)$$

is a probability from the joint distribution of $Y_A$. Furthermore, if $B$ is a subset of $V$ such that $A \cap B = \emptyset$, and $B' \subseteq B$ then we set

$$p_{fg}(A', A : B', B) = p(Y_{A'} = 1, Y_{f(A', A')} = 0, Y_{B'} = 1, Y_{g(B \setminus B')} = 0),$$

(2.9)

so that the pair $f, g \in \{c, m\}$ can be suitably chosen to specify four different probabilities. Specifically, $p_{cc}(A', A : B', B)$ is a probability from the joint distribution of both $Y_A$ and $Y_B$, $p_{cm}(A', A : B', B)$ is a probability from the joint distribution of $Y_A$ and the marginal distribution of $Y_{B'}$, $p_{mc}(A', A : B', B)$ is a probability from the marginal distribution of $Y_{A'}$ and the joint distribution of $Y_B$ and, finally, $p_{mm}(A', A : B', B)$ is a probability from the marginal distributions of $Y_{A'}$ and $Y_{B'}$. Note also that for $B' = \emptyset$ and $g = m$ equation (2.9) simplifies to

$$p_{fm}(A', A : \emptyset, B) = p_f(A', A) \text{ for every } f \in \{c, m\},$$

whereas $p_{mg}(\emptyset, A : B', B) = p_g(B', B)$, with the usual convention that $p_{mm}(\emptyset, A : \emptyset, B) = 1$. From (2.9) we can define the conditional probability
\[ p_{fg}(A', A|B', B) = p(Y_{A'} = 1, Y_{f(A')} = 0|Y_{B'} = 1, Y_{g(B')}) = 0). \]

(2.10)

More generally, for \( f, g \in \{c, m\} \), equation (2.10) takes four distinct forms, specifically,
\[
\begin{align*}
& p_{cc}(A', A|B', B) = p(Y_{A'} = 1, Y_{A'\setminus A} = 0|Y_{B'} = 1, Y_{B'\setminus B} = 0) \\
& p_{mc}(A', A|B', B) = p(Y_{A'} = 1|Y_{B'} = 1, Y_{B'\setminus B} = 0) \\
& p_{cm}(A', A|B', B) = p(Y_{A'} = 1, Y_{A'\setminus A} = 0|Y_{B'} = 1) \\
& p_{mm}(A', A|B', B) = p(Y_{A'} = 1|Y_{B'} = 1).
\end{align*}
\]

Hence, the equalities (2.7) and (2.8) can be expressed in terms of (2.10) because (2.7) coincides with both
\[ p_{cc}(a, a|b, b) = p_{cc}(a, a|\emptyset, b) \]
and \( p_{mc}(a, a|b, b) = p_{mc}(a, a|\emptyset, b) \),
whereas (2.8) coincides with both
\[ p_{cm}(a, a|b, b) = p_{cm}(a, a|\emptyset, b) \]
and
\[ p_{mm}(a, a|b, b) = p_{mm}(a, a|\emptyset, b). \]

The following lemma provides a general rule for establishing the independence of two random vectors through (2.10). This includes both (2.7) and (2.8) as special cases.

**Lemma 2.1** Let \( Y_V \) be a random vector of binary variables and let \( A \) and \( B \) be two nonempty, disjoint, subsets of \( V \). Then, for any pair \( f, g \in \{c, m\} \) it holds that \( Y_A \indep Y_B \) if and only if
\[
p_{fg}(A', A|B', B) = p_{fg}(A', A|\emptyset, B),
\]
for every \( A' \subseteq A \) and \( B' \subseteq B \) such that \( A', B' \neq \emptyset \).

**Proof.** We first show that \( Y_A \indep Y_B \) implies the case \((f, g) = (m, m)\). More concretely, we have to show that \( Y_A \indep Y_B \) implies \( p(Y_{A'} = 1|Y_{B'} = 1) = p(Y_{A'} = 1) \) for every \( \emptyset \neq A' \subseteq A \) and \( \emptyset \neq B' \subseteq B \) and this follows from (2.5) because \( Y_A \indep Y_B \) implies \( Y_{A'} \indep Y_{B'} \).

Second, we show that the case \((f, g) = (m, m)\) is equivalent to the case \((f, g) = (c, m)\). The proof that the case \((f, g) = (c, m)\) implies the
case \((f, g) = (m, m)\) is straightforward because it follows by taking the sum over the subsets \(A'' \subseteq A \setminus A'\) of both sides of the equality
\[
p(Y_{A'} = 1, Y_{A''} = 1, Y_{A \setminus (A' \cup A'')} = 0 | Y_{B'} = 1) = p(Y_{A'} = 1, Y_{A''} = 1, Y_{A \setminus (A' \cup A'')} = 0)
\]
that, in the case \((f, g) = (c, m)\), holds for every \(\emptyset \neq A' \subseteq A\) and \(\emptyset \neq B' \subseteq B\). We now show the reverse implication and, more concretely, that
\[
p(Y_{A'} = 1 | Y_{B'} = 1) = p(Y_{A'} = 1) \tag{2.11}
\]
for every \(\emptyset \neq A' \subseteq A\) and \(\emptyset \neq B' \subseteq B\) implies \(p(Y_{A'} = 1, Y_{A \setminus A'} = 0 | Y_{B'} = 1) = p(Y_{A'} = 1, Y_{A \setminus A'} = 0)\) for every \(\emptyset \neq A' \subseteq A\) and \(\emptyset \neq B' \subseteq B\). The proof of this point is by induction on the cardinality of \(A \setminus A'\) which, in the rest of this proof, we will denote by \(\overline{A} = A \setminus A'\).

The result is trivially true for \(A' = A\), that is for \(|\overline{A}| = 0\), and we show that if the result is true for \(|\overline{A}| = k\) with \(k \geq 0\) then it is true for \(|\overline{A}| = k + 1\). If \(|\overline{A}| = k + 1\) then \(\overline{A} \neq \emptyset\) and we can choose an arbitrary element \(a \in \overline{A}\) to write
\[
p(Y_{A'} = 1, Y_{\overline{A}} = 0 | Y_{B'} = 1)
= p(Y_{A'} = 1, Y_{\overline{A} \setminus \{a\}} = 0, Y_a = 0 | Y_{B'} = 1)
= p(Y_{A'} = 1, Y_{\overline{A} \setminus \{a\}} = 0 | Y_{B'} = 1) - p(Y_{A'} = 1, Y_{\overline{A} \setminus \{a\}} = 0, Y_a = 1)
= p(Y_{A'} = 1, Y_{\overline{A} \setminus \{a\}} = 0) - p(Y_{A'} = 1, Y_{\overline{A} \setminus \{a\}} = 0, Y_a = 1) \tag{2.12}
= p(Y_{A'} = 1, Y_{\overline{A}} = 0)
\]
where (2.12) follows from the induction assumption because \(|\overline{A} \setminus \{a\}| = k\).

Third, we show that the case \((f, g) = (c, m)\) implies the case \((f, g) = (m, c)\). It is easy to see that if the equality (2.11) holds true, then the same equality also holds if the subsets \(A\) and \(B\) are swapped. Because the case \((f, g) = (m, m)\) is equivalent to the case \((f, g) = (c, m)\), it follows that we can also swap the subsets \(A\) and \(B\) in the latter case. Hence, we have to show that
\[
p(Y_{B'} = 1, Y_{B \setminus B'} = 0 | Y_{A'} = 1) = p(Y_{B'} = 1, Y_{B \setminus B'} = 0) \tag{2.13}
\]
for every $\emptyset \neq A' \subseteq A$ and $\emptyset \neq B' \subseteq B$ implies $p(Y_{A'} = 1|Y_{B'} = 1, Y_{B\setminus B'} = 0) = p(Y_{A'} = 1|Y_{B} = 0)$ for every $\emptyset \neq A' \subseteq A$ and $\emptyset \neq B' \subseteq B$. We first notice that the identity (2.13) also holds for $B' = \emptyset$. This can be seen by summing both sides of (2.13) over $B' \subseteq B$, with $B' \neq \emptyset$, to obtain $1 - p(Y_{B} = 0|Y_{A'} = 1) = 1 - p(Y_{B} = 0)$ and therefore $p(Y_{B} = 0|Y_{A'} = 1) = p(Y_{B} = 0)$. Hence, we can multiply both sides of (2.13) by $p(Y_{A'} = 1)$ and divide by $p(Y_{B'} = 1, Y_{B\setminus B'} = 0)$ to obtain

$$p(Y_{A'} = 1|Y_{B'} = 1, Y_{B\setminus B'} = 0) = p(Y_{A'} = 1)$$

(2.14)

for every $B' \subseteq B$ and $\emptyset \neq A' \subseteq A$. For $B' = \emptyset$ the equality (2.14) becomes $p(Y_{A'} = 1|Y_{B} = 0) = p(Y_{A'} = 1)$ and therefore we can replace $p(Y_{A'} = 1)$ by $p(Y_{A'} = 1|Y_{B} = 0)$ in (2.14) to obtain that

$$p(Y_{A'} = 1|Y_{B'} = 1, Y_{B\setminus B'} = 0) = p(Y_{A'} = 1|Y_{B} = 0)$$

(2.15)

for every $A' \subseteq A$ and $\emptyset \neq B' \subseteq B$, as required.

In order to show that the case $(f, g) = (m, c)$ implies the case $(f, g) = (c, c)$, it is sufficient to use the same induction argument used to show that the case $(f, g) = (m, m)$ implies the case $(f, g) = (c, m)$.

Finally, we show that the case $(f, g) = (c, c)$ implies $Y_{A} \perp Y_{B}$. We first notice that the equality

$$p(Y_{A'} = 1, Y_{A\setminus A'} = 0|Y_{B'} = 1, Y_{B\setminus B'} = 0)$$

$$= p(Y_{A'} = 1, Y_{A\setminus A'} = 0|Y_{B} = 0)$$

(2.16)

holds true for every $A' \subseteq A$ and $B' \subseteq B$. Indeed, under the case $(f, g) = (c, c)$ the identity (2.16) is true for every $\emptyset \neq A' \subseteq A$ and $\emptyset \neq B' \subseteq B$. Furthermore, (2.16) is trivially true in the case $B' = \emptyset$ and, finally, it can be shown that it holds true also when $A' = \emptyset$ by using the same argument used to show that the identity (2.13) holds true for $A' = \emptyset$.

Then, one can compute $p(Y_{A'} = 1, Y_{A\setminus A'} = 0)$ by applying the law of total probability to show that this probability is equal to $p(Y_{A'} = 1, Y_{A\setminus A'} = 0|Y_{B} = 0)$, thereby obtaining

$$p(Y_{A'} = 1, Y_{A\setminus A'} = 0|Y_{B'} = 1, Y_{B\setminus B'} = 0) = p(Y_{A'} = 1, Y_{A\setminus A'} = 0),$$

for every $A' \subseteq A$ and $B' \subseteq B$, and this completes the proof. $\square$

A fundamental measure of the association of two binary variables $Y_{a}$ and $Y_{b}$ is the odds ratio that is the ratio of the products
\[ p(Y_a = 0, Y_b = 0)p(Y_a = 1, Y_b = 1) \text{ and } p(Y_a = 0, Y_b = 1)p(Y_a = 1, Y_b = 0). \]

It is well-known that \( Y_a \perp Y_b \) if and only if the odds ratio is equal to one, that is
\[
\frac{p(Y_a = 0, Y_b = 0)p(Y_a = 1, Y_b = 1)}{p(Y_a = 0, Y_b = 1)p(Y_a = 1, Y_b = 0)} = 1; \tag{2.17}
\]

see Agresti (2013) for a comprehensive treatment of this issue. The odds ratio is also called the cross-product ratio because it is the ratio of the products of probabilities from diagonally opposite cells, which are marked by an “*” in Table 2.3. The odds ratio is the ratio of the cross-products of the joint probabilities in Table 2.3. We introduce three alternative cross-product ratios involving also the marginal probabilities of the two variables. Each of these cross-product ratios leads to an equality that is equivalent to the independence of the two variables.

We first consider the cross-product ratio of the probabilityes marked by a “‡” in Table 2.3 which involves the marginal distribution of \( Y_a \), but not that of \( Y_b \), and it holds that \( Y_a \perp Y_b \) if and only if
\[
\frac{p(Y_a = 0)p(Y_a = 1, Y_b = 1)}{p(Y_a = 0, Y_b = 1)p(Y_a = 1)} = 1. \tag{2.18}
\]

Second, we consider the cross-product ratio of the probabilities marked by a “§”. This involves the marginal distribution of \( Y_b \), but not that of \( Y_a \), and it holds that \( Y_a \perp Y_b \) if and only if
\[
\frac{p(Y_b = 0)p(Y_a = 1, Y_b = 1)}{p(Y_b = 1)p(Y_a = 1, Y_b = 0)} = 1. \tag{2.19}
\]

Third, we consider the cross-product ratio of the probabilities marked by a “†”. This involves the marginal distributions of both \( Y_a \) and \( Y_b \) and, also in this case, it holds that \( Y_a \perp Y_b \) if and only if
\[
\frac{1 \times p(Y_a = 1, Y_b = 1)}{p(Y_b = 1)p(Y_a = 1)} = 1. \tag{2.20}
\]

It is possible to give a unified, and more general, version of the cross-product ratios above as follows
\[
\text{cpr}_{fg}(A', A : B', B) = \frac{p_{fg}(\emptyset, A : \emptyset, B) p_{fg}(A', A : B', B)}{p_{fg}(\emptyset, A : B', B) p_{fg}(A', A : \emptyset, B)} \tag{2.21}
\]
where \( f, g \in \{ c, m \} \). Hence, the equality (2.17) can be written as \( \text{cpr}_{ce}(a, a : b, b) = 1 \) and, furthermore, the equalities (2.18), (2.19) and (2.20) can be written as \( \text{cpr}_{cm}(a, a : b, b) = 1 \), \( \text{cpr}_{mc}(a, a : b, b) = 1 \) and \( \text{cpr}_{mm}(a, a : b, b) = 1 \), respectively.

**Lemma 2.2** Let \( Y_V \) be a random vector of binary variables and let \( A \) and \( B \) be two nonempty, disjoint, subsets of \( Y \). Then, for any pair \( f, g \in \{ c, m \} \) it holds that \( Y_A \perp Y_B \) if and only if

\[
\text{cpr}_{fg}(A', A : B', B) = 1
\]

for every \( A \subseteq A \) and \( B \subseteq B \) such that \( A', B' \neq \emptyset \).

**Proof.** \( Y_A \perp Y_B \) implies \( p_{fg}(A', A : B', B) = p_f(A', A)p_g(B', B) \) for every \( A' \subseteq A \) and \( B' \subseteq B \) and therefore, in this case, \( \text{cpr}_{fg}(A', A : B', B) = 1 \) for every \( A' \subseteq A \) and \( B' \subseteq B \), as required.

We now show the reverse implication. In the case \( f = c \) then \( \text{cpr}_{cg}(A', A : B', B) = 1 \) implies that for every \( \emptyset \neq A' \subseteq A \) and \( \emptyset \neq B' \subseteq B \)

\[
\frac{p_{cg}(A', A : B', B)}{p_{cg}(\emptyset, A : B', B)} = \frac{p_{cg}(A', A : \emptyset, B)}{p_{cg}(\emptyset, A : \emptyset, B)},
\]

and therefore

\[
\frac{p_{cg}(A', A|B', B)}{p_{cg}(\emptyset, A|B', B)} = \frac{p_{cg}(A', A|\emptyset, B)}{p_{cg}(\emptyset, A|\emptyset, B)}.
\]  

(2.22)

The equality (2.22) trivially holds true for \( A' = \emptyset \) and therefore we can sum both sides over \( A' \subseteq A \) to obtain \( p_{cg}(\emptyset, A|B', B)^{-1} = p_{cg}(\emptyset, A|\emptyset, B)^{-1} \), and therefore that \( p_{cg}(\emptyset, A|B', B) = p_{cg}(\emptyset, A|\emptyset, B) \). Hence, the two denominators of (2.22) cancel out thereby giving \( p_{cg}(A', A|B', B) = p_{cg}(A', A|\emptyset, B) \) for every \( \emptyset \neq A' \subseteq A \) and \( \emptyset \neq B' \subseteq B \) that implies \( Y_A \perp Y_B \) by Lemma 2.1.

Consider now the case \( f = m \). The equality \( \text{cpr}_{mg}(A', A : B', B) = 1 \) implies

\[
1 = \frac{p_g(\emptyset, B)p_{mg}(A', A : B', B)}{p_g(B', B)p_{mg}(A', A : \emptyset, B)} = \frac{p_{mg}(A', A|B', B)}{p_{mg}(A', A|\emptyset, B)},
\]

that, in turn, implies \( p_{mg}(A', A|B', B) = p_{mg}(A', A|\emptyset, B) \). The latter, when satisfied for every \( \emptyset \neq A' \subseteq A \) and \( \emptyset \neq B' \subseteq B \), implies \( Y_A \perp Y_B \) by Lemma 2.1. This states the result. \( \Box \)
The results of Lemmas 2.1 and 2.2 are feasible of a more general formulation involving conditional independence. If \( C \subseteq V \) is such that \( C \cap A = C \cap B = \emptyset \), then, for every \( C' \subseteq C \), we can consider the conditional distribution of \( Y_{A \cup B|\{Y_{C'} = 1, Y_{C \setminus C'} = 0\}} \) and, accordingly, define the quantities \( p_{fg}(A', A : B', B|C', C) \) and \( cpr_{fg}(A', A : B', B|C', C) \). The formulation of Lemma 2.1 in terms of conditional independence is as follows.

**Corollary 2.3** Let \( Y_V \) be a random vector of binary variables and let \( A, B \) and \( C \) be three pairwise disjoint subsets of \( V \) such that \( A, B \neq \emptyset \). Then, for any pair \( f, g \in \{c, m\} \) it holds that \( Y_A \perp \!\!\!\perp Y_B|Y_C \) if and only if

\[
p_{fg}(A', A|B', B : C', C) = p_{fg}(A', A|\emptyset, B : C', C)
\]

for every \( A' \subseteq A \) and \( B' \subseteq B \) such that \( A', B' \neq \emptyset \) and for every \( C' \subseteq C \).

**Proof.** This follows from the repeated application of Lemma 2.1 for every \( C_0' \subseteq C \).

We now give the formulation of Lemma 2.2 in terms of conditional independence.

**Corollary 2.4** Let \( Y_V \) be a random vector of binary variables and let \( A, B \) and \( C \) be three pairwise disjoint subsets of \( V \) such that \( A, B \neq \emptyset \). Then, for any pair \( f, g \in \{c, m\} \) it holds that \( Y_A \perp \!\!\!\perp Y_B|Y_C \) if and only if

\[
cpr_{fg}(A', A : B', B|C', C) = 1
\]

for every \( A' \subseteq A \) and \( B' \subseteq B \) such that \( A', B' \neq \emptyset \) and for every \( C' \subseteq C \).

**Proof.** This follows from the repeated application of Lemma 2.2 for every \( C' \subseteq C \).

**The Non-binary Case**

In this section the results of the binary case are extended to the general setting of non-binary variables. We show how the quantities introduced in the binary case can be defined with respect to the binary subtables identified by the elements \( j \in J \) and, furthermore, that the results of the previous section hold in the general polytomous case when they hold for every such binary subtable.
For every \( j \in J \) we define the non-binary versions of (2.9) and (2.10) as
\[
p_{fg}(j_A', A : j_{B'}', B) = p(Y_A' = j_A', Y_{f(A \setminus A')} = 0, Y_{B'}' = j_{B'}', Y_{g(B \setminus B')} = 0)
\]
and
\[
p_{fg}(j_A', A | j_{B'}', B) = p(Y_A' = j_A', Y_{f(A \setminus A')} = 0 | Y_{B'}' = j_{B'}', Y_{g(B \setminus B')} = 0),
\]
respectively, so that the version of Lemma 2.1 for non-binary variables is as follows.

**Lemma 2.5** Let \( Y_V \) be a random vector of categorical variables and let \( A \) and \( B \) be two nonempty, disjoint, subsets of \( V \). Then, for any pair \( f, g \in \{c, m\} \) it holds that \( Y_A \perp Y_B \) if and only if
\[
p_{fg}(j_A', A | j_{B'}', B) = p_{fg}(j_A', A | \emptyset, B)
\]
for every \( A' \subseteq A \) and \( B' \subseteq B \) such that \( A', B' \neq \emptyset \) and for every \( j \in J \).

**Proof.** This proof can be obtained from the proof of Lemma 2.1 by operating the following changes. (i) Replace, everywhere, \( Y_{A'} = 1 \) and \( Y_{A'} = j_{A'} \) and \( Y_{A'} = j_{A'} \), respectively, and, similarly, \( Y_{B'} = 1 \) by \( Y_{B'} = j_{B'} \). (ii) Replace, everywhere, the condition for every \( \emptyset \neq A' \subseteq A \) and \( \emptyset \neq B' \subseteq B \) by for every \( \emptyset \neq A' \subseteq A \), \( \emptyset \neq B' \subseteq B \) and \( j \in J \). (iii) Replace, twice, the sum over \( A' \subseteq A \) by the sum over \( j_{A'} \) for \( A' \subseteq A \) and \( j_{A'} \in J_{A'} \), and once the sum over \( A'' \subseteq A \setminus A' \) by the sum over \( j_{A''} \) for \( A'' \subseteq A \) \( j_{A''} \in J_{A''} \) and \( j_{A'} \in J_{A'} \). (iv) In the part of the proof where it is shown that the case \( (f, g) = (m, m) \) implies the case \( (f, g) = (c, m) \), replace \( p(Y_{A'} = 1, Y_{A \setminus \{a\}} = 0, Y_a = 1 | Y_{B'} = 1) \) by \( \sum_{j_{A'}=1}^{d_{A'}} p(Y_{A'} = j_{A'}, Y_{A \setminus \{a\}} = 0, Y_a = j_a | Y_{B'} = j_{B'}) \) and, similarly, \( p(Y_{A'} = 1, Y_{A \setminus \{a\}} = 0, Y_a = 1) \) by \( \sum_{j_{A'}=1}^{d_{A'}} p(Y_{A'} = j_{A'}, Y_{A \setminus \{a\}} = 0, Y_a = j_a) \). \( \square \)

For every \( j \in J \) the non-binary versions of the cross-product ratio in (2.21) is
\[
cpr_{fg}(j_A', A : j_{B'}', B) = \frac{p_{fg}(\emptyset : \emptyset, B) \ p_{fg}(j_A', A : j_{B'}', B)}{p_{fg}(\emptyset : j_{B'}', B) \ p_{fg}(j_A', A : \emptyset, B)} \tag{2.23}
\]
so that we can give the version of Lemma 2.2 for non-binary variables.
Lemma 2.6  Let $Y_V$ be a random vector of categorical variables and let $A$ and $B$ be two nonempty, disjoint, subsets of $V$. Then, for any pair $f, g \in \{c, m\}$ it holds that $Y_A \perp Y_B$ if and only if

$$\text{cpr}_{fg}(j_A', A : j_B', B) = 1$$

for every $A' \subseteq A$ and $B' \subseteq B$ such that $A', B' \neq \emptyset$ and for every $j \in J$.

Proof. This proof can be obtained from the proof of Lemma 2.2 by operating the following changes. (i) Replace, everywhere, $A_0'$ by $j_A'$ and, similarly, $B_0'$ by $j_B'$. (ii) Replace, the condition for every $A' \subseteq A$ and $B' \subseteq B$ by for every $A' \subseteq A, B' \subseteq B$ and $j \in J$. (iii) Replace, twice, the condition for every $\emptyset \neq A_0' \subseteq A$ and $\emptyset \neq B_0' \subseteq B$ by for every $\emptyset \neq A' \subseteq A, \emptyset \neq B' \subseteq B$ and $j \in J$. (iv) Replace the sum over $A_0' \subseteq A$ by the sum over $j_A'$ for $A' \subseteq A$ and $j \in J$. \hfill $\square$

Next, we can consider the conditional distribution of $Y_{A \cup B} | \{Y_C = j_C', Y_{C'} = 0\}$ and, accordingly define the quantities

$$p_{fg}(j_A', A : j_B', B | j_C', C)$$
$$\text{cpr}_{fg}(j_A', A : j_B', B | j_C', C)$$

to obtain a non-binary extension of Corollary 2.3.

Corollary 2.7  Let $Y_V$ be a random vector of categorical variables and let $A$, $B$ and $C$ be three pairwise disjoint subsets of $V$ such that $A, B \neq \emptyset$. Then, for any pair $f, g \in \{c, m\}$ it holds that $Y_A \perp Y_B | Y_C$ if and only if

$$p_{fg}(j_A', A | j_B', B : j_C', C) = p_{fg}(j_A', A | \emptyset, B : j_C', C)$$

for every $A' \subseteq A, B' \subseteq B$ and $C' \subseteq C$, such that $A', B' \neq \emptyset$, and for every $j \in J$.

Proof. This follows from the repeated application of Lemma 2.5 for every $C' \subseteq C$ and $j \in J$. \hfill $\square$

Finally, we give the non-binary version of Corollary 2.4.

Corollary 2.8  Let $Y_V$ be a random vector of categorical variables and let $A$, $B$ and $C$ be three pairwise disjoint subsets of $V$ such that $A, B \neq \emptyset$. Then, for any pair $f, g \in \{c, m\}$ it holds that $Y_A \perp Y_B | Y_C$ if and only if
\[ \text{cpr}_{fg}(j_{A'}, A : j_{B'}, B | j_{C'}, C) = 1 \]

for every \( A' \subseteq A, B' \subseteq B \) and \( C' \subseteq C \), such that \( A', B' \neq \emptyset \), and for every \( j \in J \).

**Proof.** This follows from the repeated application of Lemma 2.6 for every \( C' \subseteq C \) and \( j \in J \). \( \square \)
Möbius Inversion

The general statement of Möbius inversion formula was first given by Weisner (1935) and Hall (1934) in the area of group theory. Subsequently, in a fundamental paper on Möbius functions, Rota (1964) showed the importance of this theory in combinatorial mathematics. Since then, the theory of Möbius inversion and related topics has become an active area of combinatorics. We devote one full chapter to this topic because the material given here makes it possible to present in a common framework some fundamental features of all the families of graphical models considered in the forthcoming chapters. Furthermore, Möbius inversion turns out to be central also in the exponential family representation of undirected graph models as well as for the derivation of the asymptotic theory for these models.

3.1 Preliminaries

3.1.1 Notation and Terminology

Here we extend the notation used so far to include vectors and matrices with entries indexed by the subsets of \( V \). More specifically, we write \( \mathcal{P}(V) \) to denote the power set of \( V \), \( \mathcal{P}(V) = \{ D : D \subseteq V \} \), and denote by \( \theta_{\mathcal{P}(V)} = (\theta_D)_{D \subseteq V} \) a real vector indexed by the subsets of \( V \). Hence, any subset \( A \subseteq \mathcal{P}(V) \) identifies a subvector \( \theta_A = (\theta_D)_{D \in A} \) of \( \theta_{\mathcal{P}(V)} \). We omit the subscript for vectors indexed by the whole power set of \( V \), thereby setting \( \theta = \theta_{\mathcal{P}(V)} \). Likewise, the product set \( \mathcal{P}(V) \times \mathcal{P}(V) \) indexes the entries of the matrix \( G_{\mathcal{P}(V)\mathcal{P}(V)} = (G_{D,E})_{D,E \subseteq V} \) where the first argument indexes the rows and the second the columns. Any subset \( A \subseteq V \) identifies a submatrix \( G_{\mathcal{P}(A)\mathcal{P}(A)} \) of \( G_{\mathcal{P}(V)\mathcal{P}(V)} \) and, more generally, we denote by \( G_{AB} \) the submatrix with rows and columns indexed by the elements of the subsets \( A, B \subseteq \mathcal{P}(V) \), respectively. In order to lighten notation we set \( G_A = G_{AA}, G_A = G_{\mathcal{P}(A)} \) and \( G = G_V \).

We will mostly deal with subsets of \( \mathcal{P}(V) \) that are simplicial complexes. Formally, a subset \( C \subseteq \mathcal{P}(V) \) is called an (abstract) simplicial complex if
$E \in C$ and $D \subseteq E$ implies $D \in C$. We remark that $\emptyset \in C$ and write $C_\emptyset$ to denote that $C$ has been deprived of the empty set; formally $C_\emptyset = C \setminus \emptyset$.

The elements of the power set of $V$ can be partially ordered by assuming that, for $A, B \subseteq V$, the set $A$ precedes the set $B$ whenever $A \subseteq B$. Hereafter, when we say that a matrix is either diagonal or triangular, we mean that this property is fulfilled under any total ordering of its rows and columns compatible with the above partial ordering.

### 3.1.2 The Zeta and the Möbius Matrices

The zeta matrix, denoted by $\mathbb{Z}$, and its inverse, called the Möbius matrix, denoted by $\mathbb{M}$, are almost ubiquitous in this text. They are repeatedly used in the presentation of several topics as well as in many of the proofs. It is thus worthwhile to present them in a detailed way. The relevance of these two matrices rests in the fact that they result from the implementation of the Möbius inversion formula (see e.g., Speed, 1983; Lovász, 1993; Jokinen, 2006).

We start from the $\mathbb{Z}$ and $\mathbb{M}$ matrices associated with singleton sets. Hence, for every entry $v \in V$ the $2 \times 2$ matrices $\mathbb{Z}_v$ and $\mathbb{M}_v$, indexed by $\mathcal{P}(\{v\})$ are defined as

\[
\begin{align*}
\emptyset & \{v\} \\
\{v\} & = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix} = \mathbb{Z}_v \quad \text{and} \quad \emptyset & \{v\} \\
\{v\} & \begin{pmatrix} 1 & -1 \\ 0 & 1 \end{pmatrix} = \mathbb{M}_v
\end{align*}
\]

where we have also displayed the indexes associated with every row and column. We now consider an arbitrary, nonempty, subset $A \subseteq V$. The zeta and the Möbius matrices associated with $A$ are indexed by the elements of $\mathcal{P}(A)$ and given by

\[
\mathbb{Z}_A = \bigotimes_{v \in A} \mathbb{Z}_v \quad \text{and} \quad \mathbb{M}_A = \bigotimes_{v \in A} \mathbb{M}_v,
\]

respectively, where $\bigotimes$ denotes the Kronecker product of matrices. It is easy to see that $\mathbb{Z}_v^{-1} = \mathbb{M}_v$ and, consequently, an immediate consequence of the properties of the Kronecker product is that

\[
\mathbb{Z}_A^{-1} = \bigotimes_{v \in A} \mathbb{Z}_v^{-1} = \bigotimes_{v \in A} \mathbb{M}_v = \mathbb{M}_A. \tag{3.1}
\]

Every pair of subsets $D, E \subseteq A$ uniquely identifies the entry

\[
(Z_A)_{D,E} = \prod_{v \in A} (Z_v)_{D \cap \{v\}, E \cap \{v\}} \tag{3.2}
\]
of \( \mathbb{Z}_A \), and similarly for \( \mathbb{M}_A \). On the other hand, every entry (3.2) can be obtained as a direct function of the indexing subsets.

**Proposition 3.1** Let \( \mathbb{Z}_A \) and \( \mathbb{M}_A \) be the zeta and Möbius matrices, respectively, associated with the nonempty subset \( A \subseteq V \). Then, for every \( D, E \subseteq A \)

\[
(\mathbb{Z}_A)_{D,E} = 1(D \subseteq E) \quad \text{and} \quad (\mathbb{M}_A)_{D,E} = (-1)^{|E \setminus D|} \ 1(D \subseteq E),
\]

where \( 1(\cdot) \) denotes the indicator function.

**Proof.** It follows from (3.2) that the entry \( (\mathbb{Z}_A)_{D,E} \) of \( \mathbb{Z}_A \) is equal either to zero or to one and, more specifically, it is equal to one if and only if

\[
(\mathbb{Z}_v)_{D \cap \{v\}, E \cap \{v\}} = 1 \text{ for every } v \in A
\]

that is, if and only if there exists no \( v \in A \) such that both \( v \in D \) and \( v \notin E \). Clearly, this happens if and only if \( D \subseteq E \).

We now turn to \( \mathbb{M}_A \). It follows from (3.2) that (i) every entry \( (\mathbb{M}_A)_{D,E} \) is equal either to zero or to plus/minus one and, (ii) because \( \mathbb{Z}_v \) and \( \mathbb{M}_v \) have the same entries equal to zero, then also \( (\mathbb{M}_A)_{D,E} = 0 \) if and only if \( (\mathbb{Z}_A)_{D,E} = 0 \), so that we can write \( (\mathbb{M}_A)_{D,E} = \pm 1(D \subseteq E) \). Finally, for every \( v \in A \), it holds that \( (\mathbb{M}_v)_{D \cap \{v\}, E \cap \{v\}} = -1 \) if and only if \( D \cap \{v\} = \emptyset \) and \( E \cap \{v\} = \{v\} \); i.e., if and only if \( v \in E \setminus D \). It follows that if \( (\mathbb{M}_v)_{D,E} = \pm 1 \) then it is equal to \(-1\) if and only if \( |E \setminus D| \) is odd so that we can write \( (\mathbb{M}_A)_{D,E} = (-1)^{|E \setminus D|} \ 1(D \subseteq E) \) as required. \( \Box \)

For example, the matrix \( \mathbb{Z}_{\{a,b,c\}} \), displayed with its row and column indexes, can be written as

\[
\begin{pmatrix}
\emptyset & \{a\} & \{b\} & \{c\} & \{a,b\} & \{a,c\} & \{b,c\} & \{a,b,c\} \\
\emptyset & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\
\{a\} & 0 & 1 & 0 & 0 & 1 & 1 & 0 & 1 \\
\{b\} & 0 & 0 & 1 & 0 & 1 & 0 & 1 & 1 \\
\{c\} & 0 & 0 & 0 & 1 & 0 & 1 & 1 & 1 \\
\{a,b\} & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 1 \\
\{a,c\} & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 1 \\
\{b,c\} & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 \\
\{a,b,c\} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1
\end{pmatrix}.
\]
whereas $M_{\{a,b,c\}}$ is

$$
\begin{array}{cccccccc}
\emptyset & \{a\} & \{b\} & \{c\} & \{a,b\} & \{a,c\} & \{b,c\} & \{a,b,c\} \\
\emptyset & 1 & -1 & -1 & 1 & 1 & 1 & -1 \\
\{a\} & 0 & 1 & 0 & 0 & -1 & -1 & 0 & 1 \\
\{b\} & 0 & 0 & 1 & 0 & -1 & 0 & -1 & 1 \\
\{c\} & 0 & 0 & 0 & 1 & 0 & -1 & -1 & 1 \\
\{a,b\} & 0 & 0 & 0 & 0 & 1 & 0 & 0 & -1 \\
\{a,c\} & 0 & 0 & 0 & 0 & 0 & 1 & 0 & -1 \\
\{b,c\} & 0 & 0 & 0 & 0 & 0 & 0 & 1 & -1 \\
\{a,b,c\} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\
\end{array}
$$

Notice that if one respects the order relation "⊆" in the rows and columns of the matrices, then both $Z_A$ and $M_A$ are upper triangular with diagonal entries equal to one. Furthermore, it follows from Proposition 3.1 that it makes sense to set $Z_{\emptyset} = M_{\emptyset} = 1$, and that if $B \subseteq A$ then the zeta and Möbius matrices associated with $B$, i.e., $Z_B$ and $M_B$, are equal to the relevant submatrices of $Z_A$ and $M_A$, respectively. Finally, we recall that we may omit the subscript when it is equal to $V$, so that $Z = Z_V$ and $M = M_V$.

### 3.2 The Möbius Inversion Formula

Here we provide the Möbius inversion formula in a very simplified setting, whereas Section 3.3 deals with a slightly more general version of this result. Applications of Möbius inversion to the analysis of discrete graphical models can be found, among others, in Lauritzen (1996); Drton and Richardson (2008a); Drton (2009); Massam et al. (2009); La Rocca and Roverato (2017); Lupparelli and Roverato (2017).

**Theorem 3.2** Let $\theta = (\theta_D)_{D \subseteq V}$ and $\omega = (\omega_D)_{D \subseteq V}$ be two real vectors indexed by the subsets of a finite set $V$, then it holds that

$$
\omega_D = \sum_{D' \subseteq D} \theta_{D'} \quad \text{if and only if} \quad \theta_D = \sum_{D' \subseteq D} (-1)^{|D \setminus D'|} \omega_{D'},
$$

(3.3)
where identities are intended for all $D \subseteq V$. This statement can be written in matrix notation as

$$\theta = \mathbf{M}^T \omega \text{ if and only if } \omega = \mathbf{Z}^T \theta.$$ 

The proof of this theorem is straightforward if one approaches it from the matrix formulation because $\mathbf{Z} = \mathbf{M}^{-1}$, as shown in (3.1). The two equivalent formulations of the Möbius inversion formula given in Theorem 3.2 are not redundant. Indeed, they provide a powerful tool that will be used throughout this text to derive a compact matrix formulation of the quantities of interest as well as an explicit description of the individual entries which constitute such quantities.

### 3.2.1 Two Basic Lemmas

The following lemmas provide two properties of Möbius inversion, which will be useful in connection with the cross-product ratios of Corollary 2.4 (see also La Rocca and Roverato, 2017).

**Lemma 3.3** Let $\theta = (\theta_D)_{D \subseteq V}$ and $\omega = (\omega_D)_{D \subseteq V}$ be two real vectors indexed by the subsets of a finite set $V$. For a term $A, B, C$ of mutually disjoint subsets of $V$ the following statements are equivalent

(i) for every $D \subseteq A \cup B \cup C$ such that both $D \cap A \neq \emptyset$ and $D \cap B \neq \emptyset$ it holds that

$$\theta_D = 0,$$

(ii) for every $A' \subseteq A$ and $B' \subseteq B$ such that $A', B' \neq \emptyset$, and for every $C' \subseteq C$ it holds that

$$\omega_{A' \cup B' \cup C'} - \omega_{A' \cup C'} - \omega_{B' \cup C'} + \omega_{C'} = 0.$$

**Proof.** (i) $\Rightarrow$ (ii). Because for every $D \subseteq V$ it holds that $\omega_D = \sum_{D' \subset D} \theta_{D'}$, then for every $\emptyset \neq A' \subseteq A$, $\emptyset \neq B' \subseteq B$ and $C' \subseteq C$ we can write

$$\omega_{A' \cup B' \cup C'} = \sum_{D' \subset A' \cup B' \cup C'} \theta_{D'},$$

and it follows from (i) that $\theta_{D'} = 0$ for every $D'$ such that both $D' \subset A' \cup C'$ and $D' \subset B' \cup C'$. Hence, we can write (3.4) as
which implies (ii). Note that the term \(-\sum_{D' \subseteq C} \theta_{D'}\) in (3.5) is necessary because, otherwise, every term \(\theta_{D'}\) with \(D' \subseteq C\) would be considered twice, once for each of the other sums in (3.5).

(ii) \(\Rightarrow\) (i). We prove this result by induction on the cardinality of the set \(A \cup B \cup C\).

We first show that the result is true for \(|A \cup B \cup C| = 2\). In this case, \(A = \{a\}, B = \{b\}\) and \(C = \emptyset\) because \(A\) and \(B\) cannot be empty. Hence, we have to show that \(\omega_{\{a,b\}} - \omega_a - \omega_b + \omega_{\emptyset} = 0\) implies \(\theta_D = \theta_{\{a,b\}} = 0\) and the result follows immediately from the fact that

\[
\theta_{\{a,b\}} = \sum_{D' \subseteq \{a,b\}} (-1)^{|\{a,b\}| - |D'|} \omega_{D'}
= \omega_{\{a,b\}} - \omega_a - \omega_b + \omega_{\emptyset}.
\]

We now set \(k > 2\) and show that if the result holds for \(|A \cup B \cup C| < k\) then it also holds true for \(|A \cup B \cup C| = k\). More specifically, the induction assumption implies that \(\theta_{D'} = 0\) for every \(D' \subseteq A \cup B \cup C\) such that \(D' \cap A \neq \emptyset\) and \(D' \cap B \neq \emptyset\). This implies both that it is sufficient to show that (ii) implies \(\theta_D = \theta_{A \cup B \cup C} = 0\) and that

\[
\sum_{D' \subseteq A \cup B \cup C} \theta_{D'} = \theta_{A \cup B \cup C} + \sum_{D' \subseteq A \cup C} \theta_{D'} + \sum_{D' \subseteq B \cup C} \theta_{D'} - \sum_{D' \subseteq A \cup C} \theta_{D'}.
\] (3.6)

Every sum in (3.6) corresponds to an entry of \(\omega\) as follows

\[
\omega_{A \cup B \cup C} = \omega_{A \cup B \cup C} + \omega_{A \cup C} + \omega_{B \cup C} - \omega_C,
\]
so that \(\theta_{A \cup B \cup C} = \omega_{A \cup B \cup C} - \omega_{A \cup C} - \omega_{B \cup C} + \omega_C\) and because, by (ii), it holds that \(\omega_{A \cup B \cup C} - \omega_{A \cup C} - \omega_{B \cup C} + \omega_C = 0\), it follows that \(\theta_{A \cup B \cup C} = 0\) as required.

For a subset \(B \subseteq V\) we set \(A = V \setminus A\) and denote by \(s_B(\omega) = (\omega_{B \cup A'})_{A' \subseteq A}\) the subvector of \(\omega\) whose entries are indexed by the supersets of \(B\). In this context, an object of interest is the vector \(\theta_B = \mathbb{M}_A^T s_B(\omega)\), with entries...
\[ \theta_{[B] A'} = \sum_{A' \subseteq A} (-1)^{|A' \setminus A|} \omega_{B \cup A'} \text{ for every } A' \subseteq A. \] (3.7)

If \( \theta = \mathbb{M}^T \omega \) then it follows immediately from the definition of \( \theta_{[B]} \) that \( \theta_{[\emptyset]} = \theta \) but, in general, \( \theta_{[B]} \) is not a subvector of \( \theta \), and the following lemma states the connection existing between \( \theta_{[B]} \) and \( \theta \) (see also La Rocca and Roverato, 2017).

**Lemma 3.4** For a real vector \( \omega \) indexed by the power set of \( V \) let \( \theta = \mathbb{M}^T \omega \). Furthermore, for a subset \( B \subseteq V \) let \( \theta_{[B]} \) be the vector defined by (3.7), where \( A = V \setminus B \). Then it holds that

\[ \theta_{[B] A'} = \sum_{B' \subseteq B} \theta_{A' \cup B'} \text{ for every } A' \subseteq A. \] (3.8)

**Proof.** If we let \( \theta^* = (\theta^*_A)_{A \subseteq A} \) be the vector with entries \( \theta^*_A = \sum_{B' \subseteq B} \theta_A' \cup B' \) then, in order to establish (3.8), we have to show that \( \theta_{[B]} = \theta^* \). Because \( \omega = \mathbb{Z}^T \theta \), then for every \( A' \subseteq A \)

\[ \omega_{B \cup A'} = \sum_{D \subseteq B \cup A'} \theta_D = \sum_{A' \subseteq A} \sum_{B' \subseteq B} \theta_{A' \cup B'} = \sum_{A' \subseteq A} \theta^*_A, \] (3.9)

and (3.9) can be written in vector form as \( s_B(\omega) = \mathbb{Z}^T \theta^* \). On the other hand, by definition \( \theta_{[B]} = \mathbb{M}^T s_B(\omega) \), so it also holds that \( s_B(\omega) = \mathbb{Z}^T \theta_{[B]} \). Hence, it holds that \( \mathbb{Z}^T \theta_{[B]} = \mathbb{Z}^T \theta^* \) and, as \( \mathbb{Z} \) has full rank, we have \( \theta_{[B]} = \theta^* \), as required. \( \square \)

### 3.3 Möbius Inversion and Partially Ordered Sets

This section deals with more advanced topics on Möbius inversion, which are mostly used in the asymptotic theory presented in Section 4.11. Readers may prefer to skip this section on first reading and refer back to it when the relevant results are required.

Let \( (\mathcal{P}, \leq) \) be a finite partially ordered set, that is a finite set \( \mathcal{P} \) endowed with the partial ordering “\( \leq \).” Furthermore, let \( f(\cdot) \) be a real function on \( \mathcal{P} \). We consider an arbitrary total ordering of the elements of \( \mathcal{P} \), compatible with the partial ordering “\( \leq \),” so that \( f(\cdot) \) is univocally associated with the vector, indexed by \( \mathcal{P} \), \( f_\mathcal{P} = \{ f(D) \}_{D \in \mathcal{P}} \). Similarly, any real function \( g(\cdot, \cdot) \) on \( \mathcal{P} \times \mathcal{P} \) is univocally associated with a matrix \( G_\mathcal{P} = \{ g(D, H) \}_{D, H \in \mathcal{P}} \).
For any finite partially ordered set \((P, \leq)\) there are two important numerical functions defined on \(P \times P\): the zeta function \(z_P(\cdot, \cdot)\) given by \(z_P(D, H) = 1\) if \(D \leq H\), and 0 otherwise, and the Möbius function \(m_P(\cdot, \cdot)\) defined as

\[
m_P(D, H) = \begin{cases} 
1 & \text{if } D = H; \\
- \sum_{D \leq F < H} m_P(D, F) & \text{if } D < H; \\
0 & \text{otherwise.}
\end{cases}
\]

We refer to Rota (1964) for a more detailed account on the theory of Möbius functions. The matrix associated with the zeta function is called the zeta matrix and denoted by \(Z_P\). It is straightforward to see that \(Z_P\) is upper triangular with all diagonal entries equal to one, so that \(Z_P\) is invertible with determinant equal to one; \(|Z_P| = 1\). The matrix associated with the Möbius function, denoted by \(M_P\), is called the Möbius matrix. It should be clear that the zeta and Möbius matrices defined here are generalizations of the matrices of Section 3.1.2. It can be checked that, also in this more general setting, the product \(M_P \times Z_P\) is equal to the identity matrix, which implies that \(M_P = Z_P^{-1}\) and, consequently, that \(M_P\) is upper triangular with diagonal entries equal to one; see Speed (1983), Lovász (1993, p. 216) and Lütkepol (1996, p. 164). We can now give a more general version of the Möbius inversion formula that deals with arbitrary partially ordered sets.

**Theorem 3.5** Consider a finite partially ordered set \(P\) and let \(f(\cdot)\) and \(h(\cdot)\) be two real functions on \(P\). Then

(a) \(h(D) = \sum_{E \supseteq D} f(E)\)

if and only if

\(f(D) = \sum_{E \in P} m_P(D, E) h(E),\)

(b) \(h(D) = \sum_{E \subseteq D} f(E)\)

if and only if

\(f(D) = \sum_{E \in P} m_P(E, D) h(E),\)

where identities are intended for all \(D \in P\). These two statements can be rewritten in matrix notation as follows,
\( h_P = Z_P f_P \) if and only if \( f_P = M_P h_P \),

\( h_P = Z_P^T f_P \) if and only if \( f_P = M_P^T h_P \).

As well as for Theorem 3.2, also the proof of Theorem 3.5 is straightforward if one approaches it from the matrix formulation in (A) and (B) because \( M_P = Z_P^{-1} \); see also Lovász (1993, p. 218) and Rota (1964, proposition 2 and corollary 1).

We now consider a finite set \( V \) and turn to the partially ordered set \((C, \subseteq)\) where \( C \subseteq \mathcal{P}(V) \) is a simplicial complex. In this case, the zeta and Möbius functions can be written as

\[
z_C(D, H) = 1(D \subseteq H) \quad \text{and} \quad m_C(D, H) = (-1)^{|H \setminus D|} \times 1(D \subseteq H)
\]

(3.10) respectively (see Lovász, 1993, p. 216). Here, it deserves notice that if \( Z \) is the zeta matrix associated with \( \mathcal{P}(V) \), then the zeta matrix \( Z_C \) associated with a simplicial complex \( C \subseteq \mathcal{P}(V) \) is the submatrix of \( Z \) indexed by \( C \times C \). Furthermore, the Möbius matrix associated with \( C \) can be obtained in two different but equivalent ways, namely as the inverse of the zeta matrix, \( M_C = Z_C^{-1} \), and as the submatrix of \( M = Z^{-1} \) indexed by \( C \times C \). When \( \mathcal{P} \) is a simplicial complex, Theorem 3.5 can be restated as follows.

**Corollary 3.6** For a finite set \( V \) consider a simplicial complex \( C \subseteq \mathcal{P} \) and let \( f(\cdot) \) and \( h(\cdot) \) be two real functions on \( \mathcal{P} \). Then

\[
(A') \quad h(D) = \sum_{E \subseteq V \setminus D} f(F) \ 1(F \in C)
\]

if and only if

\[
f(D) = \sum_{E \subseteq V \setminus D} (-1)^{|E|} \ h(F) \ 1(F \in C),
\]

\[
(b') \quad h(D) = \sum_{E \subseteq D} f(E)
\]

if and only if

\[
f(D) = \sum_{E \subseteq D} (-1)^{|D \setminus E|} h(E);
\]

where \( F = D \cup E \) and identities are intended for all \( D \in C \). The matrix version of \((a')\) can be obtained from (A) of Theorem 3.6 by replacing \( \mathcal{P} \) with \( C \) everywhere, and the matrix version of \((b')\) can be obtained in the same way from (B).
Proof. We have to show that for the simplicial complex $C$ the four identities in (a) and (b) of Theorem 3.5 are equivalent to the corresponding identities in $(a')$ and $(b')$. Firstly, we notice that for any function $g(\cdot)$ on $C$ it holds that

$$\sum_{E \subseteq V \setminus D} g(E) = \sum_{D \subseteq E} g(D \cup E) \mathbb{1}(D \cup E \in C) = \sum_{E \subseteq V \setminus D} g(F) \mathbb{1}(F \in C)$$

(3.11)

where $F = D \cup E$. The equivalence of the identities on the left-hand side of (a) and $(a')$ is an immediate consequence of (3.11):

$$h(D) = \sum_{E \subseteq D} f(E) = \sum_{E \subseteq V \setminus D} f(F) \mathbb{1}(F \in C).$$

Consider now the identity on the right-hand side of (a). By (3.10)

$$f(D) = \sum_{E \subseteq C} m_C(D, E) h(E) = \sum_{E \subseteq C} (-1)^{|E \setminus D|} \mathbb{1}(D \subseteq E) h(E)$$

and as $D \subseteq E$ if and only if $E \supseteq D$, then it holds that

$$f(D) = \sum_{E \supseteq D} (-1)^{|E \setminus D|} h(E)$$

so that we can apply (3.11) to obtain the right-hand side of $(a')$

$$f(D) = \sum_{E \subseteq V \setminus D} (-1)^{|(D \cup E) \setminus D|} h(D \cup E) \mathbb{1}(D \cup E \in C)
= \sum_{E \subseteq V \setminus D} (-1)^{|E|} h(F) \mathbb{1}(F \in C).$$

The equivalence of the identities on the left-hand side of (b) and $(b')$ is straightforward, whereas to show the equivalence of the identities on the right-hand side (b) and $(b')$ it is sufficient to apply (3.10)

$$f(D) = \sum_{E \in \mathcal{P}} m_P(E, D) h(E) = \sum_{E \subseteq C} (-1)^{|D \setminus E|} \mathbb{1}(E \subseteq D) h(E)$$

$$= \sum_{E \subseteq D} (-1)^{|D \setminus E|} h(E);$$

and the proof is complete. □
We now turn to the bivariate case. For a simplicial complex $C$, the product space $C \times C$ is a partially ordered set naturally endowed with the ordering
\[(D, D') \leq (H, H') \text{ if and only if } D \subseteq H \text{ and } D' \subseteq H'. \quad (3.12)\]

It follows from (3.12) and (3.10) that
\[
z_{C \times C}\{(D, D'), (H, H')\} = z_{C}(D, H) \times z_{C}(D', H')
= 1(D \subseteq H) \times 1(D' \subseteq H') \quad (3.13)
\]
and therefore that the corresponding zeta matrix is $Z_{C \times C} = Z_{C} \otimes Z_{C}$. The Möbius matrix is $M_{C \times C} = Z_{C \times C}^{-1} = M_{C} \otimes M_{C}$, so that the Möbius function of $C \times C$ becomes
\[
m_{C \times C}\{(D, D'), (H, H')\} = m_{C}(D, H) \times m_{C}(D', H')
= (-1)^{|H \setminus D| + |H' \setminus D'|} \times 1(D \subseteq H)
\times 1(D' \subseteq H'), \quad (3.14)
\]
see Rota (1964, proposition 5). In this case, Theorem 3.5 can be applied with respect to two functions on $C \times C$ by setting $P = C \times C$ and of special interest is the matrix formulation of the Möbius inversion formula.

**Corollary 3.7** For a finite set $V$ consider a simplicial complex $C \subseteq P$ and let $g(\cdot, \cdot)$ and $k(\cdot, \cdot)$ be functions on $C \times C$ with associated matrices $G_{C}$ and $K_{C}$, respectively.

Then
\[
(a''') \quad k(D, D') = \sum_{(E, E') \geq (D, D')} g(E, E')
\text{ if and only if } g(D, D') = \sum_{(E, E') \in C \times C} m_{C \times C}\{(D, D'), (E, E')\} k(E, E')
\]
\[
(b'') \quad k(D, D') = \sum_{(E, E') \leq (D, D')} g(E, E')
\text{ if and only if } g(D, D') = \sum_{(E, E') \in C \times C} m_{C \times C}\{(E, E'), (D, D')\} k(E, E')
\]
where identities are intended for all \((D, D') \in C \times C\) and the zeta and Möbius functions are given in (3.13) and (3.14), respectively. These two statements can be rewritten in a matrix notation as follows,

\((A^\prime)\quad K_C = Z_C G_C Z_C^T \quad \text{if and only if} \quad G_C = M_C K_C M_C^T;\)

\((B^\prime)\quad K_C = Z_C^T G_C Z_C \quad \text{if and only if} \quad G_C = M_C^T K_C M_C.\)

**Proof.** Points \((a^\prime)\) and \((b^\prime)\) are a straightforward application of Theorem 3.5 obtained by replacing \(P\) with \(C/C2C\), \(h(\cdot,C1)\) with \(k(\cdot,C1); D\) with \((D, D')\) and, finally, \(E\) with \((E, E')\). We have now to show that \((A^\prime)\) and \((B^\prime)\) are equivalent to \((a^\prime)\) and \((b^\prime)\), respectively. To improve readability we split the proof into smaller parts.

1. Let \(\text{vec}(G_C^T)\) be the \(|C|^2 \times 1\) vector obtained by stacking the (transposed) rows of \(G_C\) into a single column, and similarly for \(\text{vec}(K_C^T)\). The entries of these two vectors follow a common ordering and here we show that such ordering is compatible with the partial ordering of \(C \times C\). Consider, without loss of generality, \(\text{vec}(G_C^T)\). Two pairs \((D, D'), (H, H') \in C \times C\) index the entries \(G_{D,D'}\) and \(G_{H,H'}\) of \(G\), respectively, and we have to show that if \((D, D') \leq (H, H')\) then \(G_{D,D'}\) either is equal or precedes \(G_{H,H'}\) in \(\text{vec}(G_C^T)\). This is an immediate consequence of the fact that \(\text{vec}(G_C^T)\) is the vector made up of taking the entries of \(G\) by row, and that \((D, D') \leq (H, H')\) implies both \(D \leq H\) and \(D' \leq H'\). Because the rows and columns of \(G_C\) follow a common ordering compatible with the partial ordering of \(C \times C\), then the row indexed by \(D\) either is equal or precede the row indexed by \(H\) and the column indexed by \(D'\) either is equal or precede the column indexed by \(H'\). Thus the ordering of the entries of \(\text{vec}(G_C^T)\) is compatible with the partial ordering of \(C \times C\).

2. Consider the matrices \(Z_C \otimes Z_C\) and \(M_C \otimes M_C\). Because the rows and columns of \(Z_C\) and \(M_C\) follow the same ordering as the rows and columns of \(G_C\) and \(K_C\), it is an immediate consequence of the definition of the Kronecker product that the rows and columns of \(Z_C \otimes Z_C\) and \(M_C \otimes M_C\) follow the same ordering as the entries of \(\text{vec}(G_C^T)\) and \(\text{vec}(K_C^T)\).

3. Points \((a^\prime)\) and \((b^\prime)\) are equivalent to

\[(i)\quad \text{vec}(K_C^T) = (Z_C \otimes Z_C) \text{ vec}(G_C^T)\]
\[\text{if and only if}\]
\[\text{vec}(G_C^T) = (M_C \otimes M_C) \text{ vec}(K_C^T).\]
(ii) \( \text{vec}(K_C^T) = (Z_C^T \otimes Z_C^T) \text{vec}(G_C^T) \)
if and only if
\( \text{vec}(G_C^T) = (M_C^T \otimes M_C^T) \text{vec}(K_C^T) \).

respectively. This can be obtained by direct application of Theorem 3.5 by setting \( P = C \times C, \ Z_P = Z_C \otimes Z_C, \ M_P = M_C \otimes M_C, \ h_P = \text{vec}(K_C^T) \) and \( f_P = \text{vec}(G_C^T) \). Points 1 and 2 above guarantee that the entries of all the matrices and vectors involved in (i) and (ii) follow a common ordering compatible with the partial ordering of \( C \times C \).

4. Finally, we show that \((A^*)\) and \((B^*)\) are equivalent to (i) and (ii) of the previous point, respectively. Consider the first identity in (i). By applying the rule for the Kronecker product described, for instance, in Lütkepol (1996, p. 17) we obtain
\[ \text{vec}(K_C^T) = (Z_C \otimes Z_C) \text{vec}(G_C^T) = \text{vec}(Z_C G_C^T Z_C^T) \]
so that \( K_C^T = Z_C G_C^T Z_C^T \) and, therefore, \( K_C = Z_C G_C Z_C^T \). The equivalence of the remaining three identities can be shown in the same way. \( \square \)
4

Undirected Graph Models

In a graphical model the interest is for the conditional independence structure of the variables, and the distinguishing feature of these models is that the collection of independence relationships characterizing the model is encoded by a graph. The first class of graphical models we consider is that of models associated with undirected graphs, which are usually referred to as undirected graphical models or Markov random fields. In the forthcoming chapters we will present the theory of models for bidirected graphs, known as bidirected graph models, and of models for directed acyclic and regression graphs, called directed acyclic graph and regression graph models, respectively. Accordingly, in order to keep the terminology consistent through this text we will slightly modify the traditional terminology and refer to models for undirected graphs as undirected graph models.

Discrete models for undirected graphs are among the very first classes of graphical models introduced in the literature. They have close connections with log-linear models and with systems investigated in other scientific areas such as statistical physics. A description of the origins of these models can be found in Cox and Wermuth (1996, section 2.13) and Lauritzen (1996, chapter 1). Here, we mention that the theory of undirected graph models presented in this chapter emerged from the seminal paper of Darroch et al. (1980), where the connection existing between log-linear models, conditional independence and undirected graphs was firstly established. Furthermore, we refer the reader to Edwards (2000, chapter 2) for a list of relevant classical monographs on this area and to Højsgaard et al. (2012) for an introduction to the available R packages (R Core Team, 2016) to deal with graphical models.

4.1 Graphs

A graph \( G = (V, E) \) is a mathematical object that consists of a finite set of vertices \( V \) and a set of edges \( E \), which are pairs of vertices. We write
a ≡ b to denote that the vertices \( a, b \in V \) form an edge in \( G \), and we say that \( a \) and \( b \) are adjacent or, alternatively, that they are joined by an edge. Conversely, we write \( a \not\equiv b \) for non-adjacent vertices. The graphs we consider are simple, i.e., they contain no multiple edges joining the same two vertices and no degenerate edge joining a vertex to itself. Furthermore, we only need three different types of edges that are typically labelled as undirected, directed and bidirected. A basic feature of graphs is that they are visual objects, and the usual way to picture a graph is drawing a dot for each vertex and joining two dots by a line \((-\)) for undirected edges, an arrow \((\rightarrow)\) for directed edges and an arrow with two heads \((\leftrightarrow)\) for bidirected edges.

Undirected Graphs

This chapter is devoted to graphical models for undirected graphs; i.e., graphs with only undirected edges, and when we want to highlight that a graph is undirected we write its edge set as \( E^{-} \). For example, Figure 4.1 gives a graphical representation of the undirected graph \( G = (V, E^{-}) \) with \( V = \{a, b, c, d, e, f\} \) and

\[
E^{-} = \{\{a, b\}, \{a, c\}, \{b, d\}, \{c, d\}, \{c, e\}, \{c, f\}, \{d, f\}, \{e, f\}\}.
\]

We write \( a \sim b \) to denote that \( a \) and \( b \) are joined by an undirected edge so that in this graph we have, for instance, \( c \sim f \) because \( \{c, f\} \in E^{-} \), but \( a \not\sim f \) because \( \{a, f\} \notin E^{-} \).

A path of length \( k \) from \( a \) to \( b \) is a sequence \( a = v_0, \ldots, v_k = b \) of distinct vertices such that the vertices \( v_{r-1} \) and \( v_r \) are joined by an edge for all \( r = 1, \ldots, k \). A cycle is a path with the modification that it begins and ends in the same vertex, i.e., \( a = b \). For example, in the graph of Figure 4.1 the sequence \( (a, b, d, c, f) \) is a path from \( a \) to \( f \), whereas \( (a, b, d, f, c, a) \) is a cycle with five vertices. A subset \( C \subseteq V \) is said to separate \( a \) from \( b \) in \( G \) if all paths from \( a \) to \( b \) intersect \( C \). The subset \( C \) is said to separate the subsets \( A, B \subseteq V \) if it separates every \( a \in A \) from every \( b \in B \). Hence, both the sets \( C_1 = \{c, d\} \) and \( C_2 = \{b, c\} \) separate \( a \) from \( f \) in the graph of Figure 4.1 and, moreover, the same sets also separate \( A = \{a\} \) from \( B = \{e, f\} \).

If \( a \) and \( b \) are joined by an undirected edge we say that they are neighbors. The set of neighbors of a vertex \( a \) in \( G \) is denoted as \( \text{nb}_G(a) \), and we will omit the subscript when clear from the context which graph we are referring to. The set of neighbors of a subset \( A \subseteq V \), \( \text{nb}(A) \), is the
collection of the neighbors of the vertices in $A$ that are not themselves elements of $A$; formally, $\text{nb}(A) = (\cup_{a \in A} \text{nb}(a)) \setminus A$. Hence, in the graph of Figure 4.1 the neighbors of $a$ are $\text{nb}(a) = \{b, c\}$, whereas $\text{nb}(b) = \{a, d\}$ so that if $A = \{a, b\}$ then $\text{nb}(A) = \{c, d\}$.

4.2 Markov Properties for Undirected Graphs

A graphical model uses a graph to represent conditional independence relations holding among a set of variables. The vertices of the graph are associated with the variables, and the rules that translate properties of the graph into conditional independence statements are called Markov properties. In undirected graph models the variables of a random vector $Y_V$ are indexed by the vertices of an undirected graph $G = (V, E^\sim)$, and in this context there are three Markov properties named pairwise, local and global.

The Pairwise Markov Property

The probability distribution of a random vector $Y_V$ is said to obey the pairwise Markov property with respect to $G = (V, E^\sim)$ if for every $a, b \in V$ such that $\{a, b\} \notin E^\sim$ it holds that

$$Y_a \independent Y_b | Y_{V \setminus \{a, b\}}.$$

For example, two conditional independence relations implied by the pairwise Markov property for the graph of Figure 4.1 are $Y_a \independent Y_f | Y_{\{b, c, d, e\}}$ and $Y_b \independent Y_c | Y_{\{a, d, e, f\}}$.

The Local Markov Property

The probability distribution of a random vector $Y_V$ is said to obey the local Markov property with respect to $G = (V, E^\sim)$ if for every $a \in V$ it holds that
For example, if $Y_V$ is locally Markov with respect to the graph of Figure 4.1 then it holds that $Y_a \perp Y_{V \setminus (\text{nb}(a) \cup \{a\})} \mid Y_{\text{nb}(a)}$. The Global Markov Property

The probability distribution of a random vector $Y_V$ is said to obey the global Markov property with respect to $G = (V, E^\sim)$ if for any triple of pairwise disjoint subsets $A, B, S \subseteq V$ such that $S$ separates $A$ from $B$ in $G$ it holds that

$$ Y_A \perp Y_B \mid Y_S. $$

Examples of conditional independencies implied by the global Markov property for the graph of Figure 4.1 are $Y_a \perp Y_{\{d,e,f\}} \mid Y_{\{b,c\}}$, $Y_d \perp Y_{\{a,e\}} \mid Y_{\{b,c,f\}}$ and $Y_c \perp Y_b \mid Y_{\{a,d,e,f\}}$.

Clearly, the global Markov property implies the local Markov property because nb$(a)$ separates $a$ from $V \setminus (\text{nb}(a) \cup \{a\})$ for every $a \in V$ and, furthermore, it also implies the pairwise Markov property because $V \setminus \{a, b\}$ separates $a$ from $b$ whenever the two vertices are not joined by an edge. In addition, it can be shown that the local Markov property implies the pairwise Markov property (see Lauritzen, 1996, proposition 3.4) and therefore we can write

$$ (\text{global}) \Rightarrow (\text{local}) \Rightarrow (\text{pairwise}). $$

The reverse implications are not true in general; however, a sufficient condition for the reverse implications to hold is that the joint density of $Y_V$ with respect to a product measure is positive and continuous (Pearl and Paz, 1987); see also Lauritzen (1996, proposition 3.1 and theorem 3.7). It follows that in the case considered in this text, i.e., categorical variables with positive probability, the three Markov properties are equivalent,

$$ (\text{global}) \Leftrightarrow (\text{local}) \Leftrightarrow (\text{pairwise}), $$

and in the following we will simply say that a distribution is Markov with respect to an undirected graph to mean that it satisfies all the three (undirected) Markov properties. It is also worth mentioning that, when $Y$
admits a density with respect to a product measure, an important role is played by a relevant factorization of the density with respect to the graph (see Lauritzen, 1996, section 3.2). Such factorization implies each of the three Markov properties for undirected graphs, and in the case where the density is positive and continuous also the reverse implication holds true. This fundamental result is known as the Hammersley–Clifford Theorem (Hammersley and Clifford, 1971); see also Lauritzen (1996, theorem 3.9).

We have thus seen that, among the three Markov properties, the global Markov property provides the most detailed information on the independence structure of the random vector because it gives a general criterion for assessing if two subvectors $Y_A$ and $Y_B$ are conditionally independent given a third subvector $Y_S$. More importantly, it can be shown that the global Markov property is complete; that is, the conditional independencies encoded by any undirected graph $G$ under the global Markov property are the only conditional independencies that are simultaneously satisfied by all the distributions which are globally Markov with respect to $G$; see Geiger and Pearl (1993) and Studeny (2005, p. 45) for a more detailed account on this topic. However, to practically deal with probability distributions satisfying this property, it is also useful to notice that the global Markov property is characterized by a smaller set of independence relationships, namely those with separating sets of the form $S = V \setminus (A \cup B)$. This is formally proved below.

**Proposition 4.1** The probability distribution of $Y_V$ obeys the global Markov property with respect to $G = (V, E)$ if and only if for all pairs of nonempty disjoint subsets $A, B \subseteq V$ it holds that $Y_A \independent Y_B | Y_{V \setminus (A \cup B)}$ whenever $V \setminus (A \cup B)$ separates $A$ from $B$ in $G$.

**Proof.** If the distribution of $Y_V$ is globally Markov with respect to $G$ then every separation implies a conditional independence relationship and therefore it also holds that $Y_A \independent Y_B | Y_{V \setminus (A \cup B)}$ whenever $V \setminus (A \cup B)$ separates $A$ from $B$ in $G$. Hence, it is sufficient to show the reverse implication; i.e., that the collection of independence relationships of the form $Y_A \independent Y_B | Y_{V \setminus (A \cup B)}$ imply the global Markov property. In this case, if $A, B$ and $S$ is a triple of pairwise disjoint subsets of $V$ such that $S$ separates $A \neq \emptyset$ from $B \neq \emptyset$, then one can always find two disjoint sets $D_1$ and $D_2$ such that $D_1 \cup D_2 = V \setminus (A \cup B \cup S)$ and $S$ separates $A \cup D_1$ from $B \cup D_2$. Hence, it holds that $Y_{A \cup D_1} \independent Y_{B \cup D_2} | Y_S$ because
\( S = V \backslash (A \cup D_1 \cup B \cup D_2) \) and it follows from the properties of conditional independence (see Lauritzen, 1996, section 3.1) that \( Y_A \perp Y_B | Y_S \).

### 4.3 The Log-linear Parameterization

If \( Y \) is a random vector with state space \( \mathcal{I} \), a statistical model for \( Y \) is a family of probability distributions on \( \mathcal{I} \). Our interest is in models characterized by collections of independence relationships, and a basic prerequisite to implement, in an efficient way, the restrictions defining the model is the specification of a suitable parameterization. A straightforward way to parameterize the probability distribution of a set of categorical variables is by means of their joint probability mass function, and we say that a model is a saturated model when the cell probabilities have no restrictions, apart from the sum-to-one constraint. Probabilities are easy to interpret but have the drawback that submodels of interest typically involve nonlinear constraints on these parameters. For instance, pairwise conditional independence relationships can be specified by requiring certain factorizations of the cell probabilities. For this reason, it is useful to develop alternative parameterizations such that submodels of interest correspond to linear subspaces of the parameter space of the saturated model.

An important advance in the analysis of categorical variables was to write the probability mass function as a log-linear expansion of the cell probabilities; see Agresti (2013, chapter 9) and Whittaker (1990, chapter 7). The rest of this section is devoted to the log-linear parameterization and shows how it can be used to implement conditional independence constraints.

#### The Binary Case

Our approach to the development of the log-linear parameterization heavily relies upon the connection existing between cross-product ratios and conditional independence relationships, described in Section 2.3, and on the theory of Möbius inversion given in Chapter 3. In order to readily exploit the results on Möbius inversion, it is useful to represent the probability mass function \( \pi^V \) of a binary vector \( Y_V \) as a vector indexed by the subsets of \( V \). More concretely, we set \( \pi^V = (\pi^V_D)_{D \subseteq V} \) where
\[ \pi_D^V = p(Y_D = 1, Y_{V \setminus D} = 0) \quad \text{for every } D \subseteq V. \]

Thus we shall in the following switch between the case where \( \pi^V \) has the form of a probability table and the case where \( \pi^V \) is a vector indexed by the subsets of \( V \). The former notation is perhaps easier to read, whereas the latter allows a compact representation of the results and a straightforward generalization to categorical variables with arbitrary number of levels.

We will refer to \( \pi^V \) as the probability parameter of \( Y_V \), and the log-linear parameter \( \lambda^V \) of \( Y_V \) is defined as
\[
\lambda^V = \mathbb{M}^T \log \pi^V, \tag{4.1}
\]
where the logarithm is taken entrywise. Recall that we may omit the superscript when it is equal to \( V \) so that \( \pi = \pi^V \) and \( \lambda = \lambda^V \).

The quantity \( \mathbb{M} \) in (4.1) is the Möbius matrix of Section 3.1.2, and it follows from Theorem 3.2 that \( \lambda \) constitutes a valid parameterization of the distribution because there exists a one-to-one relationship (a smooth bijection) between \( \lambda \) and \( \pi \) such that \( \pi = \exp(\mathbb{Z}^T \lambda) \), where the exponential is also taken entrywise. Notice that the definition of the vector \( \pi \), and therefore of the log-linear parameters in (4.1), requires the choice of a reference level for every variable, that in our case is the level “0”. This is equivalent to specifying a cell to be set as baseline, and when the reference level of each variables is the value 0 the baseline cell is that indexed by the empty set. As a consequence, the value taken by \( \lambda \) is not invariant under relabeling of the two levels that are taken by the variables. However, such relabeling does not change the log-linear models we consider here. The choice of a reference cell in the specification of log-linear parameters is known as the baseline or corner constraint approach, whereas a common alternative approach is based on symmetrical constraints (see Agresti, 2013, chapter 9).

It follows from (3.3) that the entrywise version of the equality \( \log \pi = \mathbb{Z}^T \lambda \) is given by
\[
\log \pi_D = \sum_{D' \subseteq D} \lambda_{D'} \quad \text{for every } D \subseteq V,
\]
where the individual log-linear parameters can be computed as
\[
\lambda_D = \sum_{D' \subseteq D} (-1)^{|D \setminus D'|} \log \pi_{D'} \quad \text{for every } D \subseteq V. \tag{4.2}
\]
If $D$ is a singleton set, then $\lambda_D$ is called a main effect, whereas for $|D| > 1$ the parameter $\lambda_D$ is called a $|D|$-way interaction or, equivalently, an interaction of order $(|D| - 1)$.

The interpretation of the value taken by a log-linear term is not straightforward, especially for high-order interactions. To this aim, it is useful to describe the entries of $\lambda$ when $V$ is a set of low cardinality such as $V = \{a, b, c\}$. For $D = \emptyset$ the term $\lambda_D$ is the log-probability of the baseline cell of the table,

$$\lambda_\emptyset = \log \pi_\emptyset = \log p(Y_a = 0, Y_b = 0, Y_c = 0).$$

The main effect for $D = \{a\}$ is given by

$$\lambda_a = \log \frac{\pi_a}{\pi_\emptyset} = \log \frac{p(Y_a = 1, Y_b = 0, Y_c = 0)}{p(Y_a = 0, Y_b = 0, Y_c = 0)}.$$

The two-way interaction relative to $D = \{a, b\}$ is the logarithm of the odds ratio of the conditional distribution $Y_{\{a,b\}|\{Y_c = 0\}}$, introduced in Section 2.3,

$$\lambda_{\{a,b\}} = \log \frac{\pi_{\{a,b\}}}{\pi_a \pi_b} = \log \frac{p(Y_a = 1, Y_b = 1, Y_c = 0) p(Y_a = 0, Y_b = 0, Y_c = 0)}{p(Y_a = 1, Y_b = 0, Y_c = 0) p(Y_a = 0, Y_b = 1, Y_c = 0)} = \log \text{cpr}_{cc}(a, a : b, b|\emptyset, c),$$

and, therefore, $\lambda_{\{a,b\}} = 0$ is equivalent to the conditional independence relationship $Y_a \perp Y_b|\{Y_c = 0\}$. Finally, the three-way interaction for $D = \{a, b, c\}$ involves the odds ratios of $Y_{\{a,b\}|\{Y_c = 0\}}$ and of $Y_{\{a,b\}|\{Y_c = 1\}}$ as follows,

$$\lambda_{\{a,b,c\}} = \log \frac{\pi_{\{a,b,c\}}}{\pi_{\{a,b\}} \pi_{\{a,c\}} \pi_{\{b,c\}} \pi_\emptyset} = \log \frac{p(Y_a = 1, Y_b = 1|Y_c = 1) p(Y_a = 0, Y_b = 0|Y_c = 1)}{p(Y_a = 1, Y_b = 0|Y_c = 1) p(Y_a = 0, Y_b = 1|Y_c = 1)} - \log \frac{p(Y_a = 1, Y_b = 0|Y_c = 0) p(Y_a = 0, Y_b = 1|Y_c = 0)}{p(Y_a = 1, Y_b = 0|Y_c = 0) p(Y_a = 0, Y_b = 1|Y_c = 0)} = \log \text{cpr}_{cc}(a, a : b, b|\emptyset, c) - \log \text{cpr}_{cc}(a, a : b, b|\emptyset, c).$$
Hence, $\lambda_{\{a,b,c\}} = \lambda_{\{a,b\}} = 0$ is equivalent $\text{cpr}_{cc}(a, a : b, b|c, c) = \text{cpr}_{ce}(a, a : b, b|\emptyset, c) = 1$ and, therefore, to $Y_a \perp Y_b | Y_c$ by Corollary 2.4.

We now formally establish the connection between conditional independence relationships and the vanishing of log-linear parameters. In our approach, this is obtained as an immediate consequence of the joint application of Corollary 2.4 and Lemma 3.3. More concretely, Corollary 2.4 states the equivalence of certain conditional independence relationships and the fact that certain cross-product ratios are equal to one. On the other hand, Lemma 3.3 establishes the equivalence of certain vanishing entries of $\lambda$ and the fact that certain contrasts of log-probabilities are equal to zero. The connection between the two results follows because the latter contrasts of log-probabilities and (the logarithm of) the relevant cross-product ratios are exactly the same quantities (see also La Rocca and Roverato, 2017).

**Theorem 4.2**  
For a vector $Y_V$ of binary variables with probability parameter $\pi > 0$ let $\lambda = M^T \log \pi$. Then, for a pair of disjoint nonempty subsets $A$ and $B$ of $V$ the following conditions are equivalent:

(i) $Y_A \perp Y_B | Y_{V \setminus (A \cup B)}$;

(ii) for every $D \subseteq V$ such that both $D \cap A \neq \emptyset$ and $D \cap B \neq \emptyset$ it holds that $\lambda_D = 0$;

(iii) for every $A' \subseteq A$, $B' \subseteq B$, such that $A', B' \neq \emptyset$, and $C' \subseteq V \setminus (A \cup B)$ it holds that

$$\log \pi_{A' \cup B' \cup C'} - \log \pi_{A' \cup C'} - \log \pi_{B' \cup C'} + \log \pi_{C'} = 0. \quad (4.3)$$

**Proof.** The equivalence of (i) and (iii) follows from Corollary 2.4 because the identity $(4.3)$ is equal to $\log \text{cpr}_{cc}(A', A : B', B|C', C) = 0$ with $C = V \setminus (A \cup B)$. The equivalence of (ii) and (iii) is that stated in Lemma 3.3 when $\theta = \lambda$ and $\omega = \log \pi$. \qed

**The Non-binary Case**

In this section we extend the results given in the binary case to the general case of categorical variables with an arbitrary number of levels.

Let $Y$ be a vector of categorical variables with state space $\mathcal{I}$ and restricted state space $\mathcal{J}$. We now define the probability parameter and
the log-linear parameter of $Y$ as a collection of vectors for $j \in J$. More concretely, we represent $I$ by means of a collection of binary tables $\{(j_D, 0_{V \setminus D})\}_{D \subseteq V}$ for all $j \in J$ as in (2.3) and, accordingly, the probability mass function $\pi$ of $Y$ can be written as the collection of vectors

$$
\pi_j = (\pi_{j_D})_{D \subseteq V} \text{ for all } j \in J,
$$

(4.4)

where

$$
\pi_{j_D} = p(Y_D = j_D, Y_{V \setminus D} = 0).
$$

Consequently, we can define the log-linear parameter $\lambda$ of $Y$ as the collection of vectors

$$
\lambda_j = (\lambda_{j_D})_{D \subseteq V} \text{ for all } j \in J,
$$

(4.5)

where

$$
\lambda_j = M^T \log \pi_j \text{ and } \lambda_{j_D} = (\lambda_j)_D,
$$

with the convention that $j_{\emptyset} = J_{\emptyset} = \emptyset$. The representation (4.4) is useful in practice because most properties that hold in the binary case can be readily generalized to the non-binary case by their iterative application to the vectors $\pi_j$ for $j \in J$. On the other hand, it is not an efficient representation of the probability mass function of $Y$ because every pair of vectors $\pi_j$ and $\pi_{j'}$ has nonempty intersection, in the sense that they have common entries. Indeed, for every $j, j' \in J$ and $D \subseteq V$ such that $j_D = j'_D$ it holds that $\pi_{j_D} = \pi_{j'_D}$. Hence, the probability parameter of $Y$ can be more efficiently written as $\pi = (\pi_{j_D})_{D \subseteq V, j_D \in J, D}$. Similarly, we can write the log-linear parameter of $Y$ as $(\lambda_{j_D})_{D \subseteq V, j_D \in J, D}$.

We can now give the extension of Theorem 4.2 to the general case of categorical variables with arbitrary number of levels.

**Corollary 4.3** Let $Y_V$ be a vector of categorical variables with probability parameter $\pi = (\pi_{j_D})_{D \subseteq V, j_D \in J, D} > 0$, and log-linear parameter $\lambda = (\lambda_{j_D})_{D \subseteq V, j_D \in J, D}$. Then, for a pair of disjoint nonempty subsets $A$ and $B$ of $V$ the following conditions are equivalent:

(i) $Y_A \independent Y_B | Y_{V \setminus (A \cup B)}$;

(ii) for every $D \subseteq V$ such that both $D \cap A \neq \emptyset$ and $D \cap B \neq \emptyset$ it holds that
\[ \lambda_{jD} = 0, \]
for every \( j \in \mathcal{J}; \)

(iii) for every \( D \subseteq V/ (A \cup B) \) it holds that

\[
\log \pi_{j_{A'} \cup B' \cup C'} + \log \pi_{j_{A'} \cup C'} + \log \pi_{j_{B'} \cup C'} - \log \pi_{j_c} = 0, \tag{4.6}
\]

for every \( j \in \mathcal{J}. \)

Proof. The equivalence of (i) and (iii) follows from Corollary 2.8 because the identity (4.6) is equal to \( \log \text{cpr}_{cc}(j_{A'} : A, j_{B'} : B | j_{C'}, C) = 0 \) with \( C = V/ (A \cup B). \) The equivalence of (ii) and (iii) holds for every \( j \in \mathcal{J} \) and is that stated in Lemma 3.3 when \( \theta = \lambda_j \) and \( \omega = \log \pi_j. \]

4.4 Hierarchical Log-linear Models

The importance of the log-linear expansion of the cell probabilities rests on the fact that many interesting hypotheses can be specified as linear constraints on the log-linear parameters. Our interest is for models defined by the collection of conditional independence relationships encoded by an undirected graph. It follows directly from the definition of the pairwise and the local Markov properties, and from Proposition 4.1 for the global Markov property, that the conditional independencies that characterize each of the three Markov properties for undirected graphs are all of the form \( Y_A \perp Y_B | Y_{V\setminus(A \cup B)}. \) This observation, together with the results stated in Theorem 4.2 and Corollary 4.3, shows that we can focus on log-linear models defined by vanishing collections of log-linear terms.

The Binary Case

The log-linear expansion of cell probabilities for the a binary random vector \( Y \) is given by

\[ \log \pi_D = \sum_{D' \subseteq D} \lambda_{D'} \text{ for every } D \subseteq V, \]

and in the saturated model no restriction is posed on the log-linear parameters where the only fixed term is \( \lambda_{\emptyset}, \) resulting from the sum-to-one constraint of probabilities. It also deserves noticing that, apart
from \( \lambda_\emptyset \), the entries of \( \lambda \) are variation independent, in the sense that setting some parameters to particular values does not restrict the valid range of other parameters.

When the interest is for log-linear models defined by zero constraints, a relevant family of models is that of hierarchical log-linear models. A log-linear model, defined by zero restrictions on the log-linear terms, is called hierarchical if whenever an interaction term is fixed to zero then all higher-order interaction terms involving the same variables are also fixed to zero. More formally, this means that in a hierarchical log-linear model, for every \( D \subseteq V \) with \( D \neq \emptyset \)

the restriction \( \lambda_D = 0 \) implies the restriction \( \lambda_E = 0 \) for all \( E \supseteq D \)
or, equivalently,
\[
\lambda_D \text{ unrestricted implies } \lambda_{D'} \text{ unrestricted for all } D' \subseteq D \text{ with } D' \neq \emptyset.
\]

Nonhierarchical models are rarely used in practice because, in most applications, it is not meaningful to include higher-order interaction terms without incorporating the lower-order interactions composed from the variables. Furthermore, the family of hierarchical log-linear models is sufficiently general to encompass all the models defined by the conditional independencies encoded by an undirected graph, under any of the three Markov properties. This can be seen by noticing that the constraints on the log-linear terms given in point (ii) of Theorem 4.2 are of the hierarchical type. As an example, we consider the case \( V = \{a, b, c\} \) so that \( \lambda \) is made up of as few as eight terms that are all present in the linear expansion of \( \log \pi_V \),

\[
\log \pi_{\{a,b,c\}} = \lambda_\emptyset + \lambda_a + \lambda_b + \lambda_c + \lambda_{\{a,b\}} + \lambda_{\{a,c\}} + \lambda_{\{b,c\}} + \lambda_{\{a,b,c\}}.
\]

Hence, the hierarchical model with no second-order interaction, i.e., \( \lambda_{\{a,b,c\}} = 0 \), is such that,

\[
\log \pi_{\{a,b,c\}} = \lambda_\emptyset + \lambda_a + \lambda_b + \lambda_c + \lambda_{\{a,b\}} + \lambda_{\{a,c\}} + \lambda_{\{b,c\}},
\]

whereas in the hierarchical model with one first-order interaction set to zero, say \( \lambda_{\{a,b\}} = 0 \), also \( \lambda_{\{a,b,c\}} \) must be fixed to zero,

\[
\log \pi_{\{a,b,c\}} = \lambda_\emptyset + \lambda_a + \lambda_b + \lambda_c + \lambda_{\{a,c\}} + \lambda_{\{b,c\}}.
\]
There exists a close connection between simplicial complexes of subsets of $V$ and hierarchical log-linear model. For any hierarchical model the collection of sets indexing its non-vanishing terms forms a simplicial complex $C$ and, conversely, any simplicial complex $C$ of subsets of $V$ can be used to define a hierarchical model. However, the most parsimonious way to specify a hierarchical model is by means of the collection of maximal sets of $C$, with respect to inclusion, that is called the generating class of the model, and denoted by $C^\uparrow$. Accordingly, we call $C$ the expanded generating class and denote the relative model by either $M(C)$ or $M(C^\uparrow)$. Hence, for $V = \{a, b, c\}$ the generating class of the saturated model is $C^\uparrow = \{\{a, b, c\}\}$ whereas the expanded generating class for the model in (4.8) is

$$C = \{\emptyset, \{a\}, \{b\}, \{c\}, \{a, b\}, \{a, c\}, \{b, c\}\},$$

and therefore $C^\uparrow = \{\{a, b\}, \{a, c\}, \{b, c\}\}$. For the model in (4.9) it holds that $C^\uparrow = \{\{a, c\}, \{b, c\}\}$. Furthermore, for $C^\uparrow = \{\{a\}, \{b, c\}\}$ we have

$$\log \pi_{\{a, b, c\}} = \lambda_{\emptyset} + \lambda_a + \lambda_b + \lambda_c + \lambda_{\{a, c\}}, \quad (4.10)$$

whereas the generating class $\{\{a\}, \{b\}, \{c\}\}$ gives the main effects model with

$$\log \pi_{\{a, b, c\}} = \lambda_{\emptyset} + \lambda_a + \lambda_b + \lambda_c. \quad (4.11)$$

We now turn to the use of undirected graphs for describing the independence structure of $Y$. By Theorem 4.2, a necessary and sufficient condition for the conditional independence of two variables given the remaining variables, say $Y_a \perp \! \perp Y_b | Y_{V \setminus \{a, b\}}$, is that $\lambda_D = 0$ whenever $D$ is a superset of $\{a, b\}$. Hence, if $G = (V, E)$ is an undirected graph with vertex set $V$, and the log-linear parameter $\lambda$ is such that for every pair $a, b \in V$ such that $a \neq b$, it holds that $\lambda_{\{a, b\} \cup D'} = 0$ for every $D' \subseteq V \setminus \{a, b\}$, then the distribution of $Y$ is pairwise Markov with respect to $G$. Furthermore, when the distribution is positive, it also obeys both the local and the global Markov properties with respect to $G$.

We can therefore associate any hierarchical log-linear model with an undirected graph, which we call the independence graph of $M(C^\uparrow)$. This is the graph $G$, with vertex set $V$, constructed by joining two vertices $a, b \in V$ by an edge if and only if $\lambda_{\{a, b\}}$ is unconstrained in $M(C^\uparrow)$ or, equivalently, if and only if there exists a $C \in C^\uparrow$ such that $\{a, b\} \subseteq C$. By
construction, all the distributions in \( M(C) \) obey the pairwise Markov property with respect to \( G \), and also the local and global Markov properties in the case, considered here, of positive distributions. For instance, when \( V = \{ a, b, c \} \), the saturated model is associated with the graph with no missing edge of Figure 4.2A. The same independence graph is associated with the no-second-order interaction model in (4.8) because, although \( \lambda = 0 \), all first-order interactions are unconstrained. The model with \( \lambda_{ab} = \lambda_{ac} = \lambda_{bc} = 0 \) in (4.9) leads to the graph of Figure 4.2B where \( a \Perp b \) implies \( Y_a \Perp Y_b | Y_c \). The model with \( \lambda_{ab} = \lambda_{ac} = \lambda_{bc} = 0 \) in (4.10) leads to the graph of Figure 4.2C where \( a \Perp b \) implies \( Y_a \Perp Y_b | Y_c \) and \( a \Perp c \) implies \( Y_a \Perp Y_c | Y_b \). Furthermore, the global Markov property also implies the independencies \( Y_a \Perp Y_{\{b,c\}} \), \( Y_a \Perp Y_b \) and \( Y_a \Perp Y_c \). Finally, the independence graph corresponding to the model in (4.11) is that given in Figure 4.2D, which encodes all the pairwise conditional independence relationships and, under the global Markov property, the mutual independence of the three variables, \( Y_a \Perp Y_b \Perp Y_c \).

We close this section by noticing that we can use the expanded generating class \( C \) of a hierarchical model to extract the subvector \( \lambda_C \).
of $\lambda$. Furthermore, if we denote by $\mathcal{P}(V)$ the expanded generating class of the saturated model and set $\overline{C} = \mathcal{P}(V) \setminus \mathcal{C}$ then we can write $\lambda = (\lambda_C, \lambda_{\overline{C}})$ where, under $M(C)$, $\lambda_{\overline{C}} = 0$. Correspondingly, we can write $\pi = (\pi_C, \pi_{\overline{C}})$ and, if $Z_C$ is the submatrix of $Z$ whose entries are indexed by $C \times C$ then it holds that $\log \pi_C = Z_C^T \lambda_C$ and thus, by Corollary 3.6, that $\lambda_C = M_C^T \log \pi_C$. This makes it clear that for a log-linear model with expanded generating class $C$, the probability parameter $\pi$ can be partitioned into $\pi = (\pi_C, \pi_{\overline{C}})$, where $\pi_{\overline{C}}$ can be obtained as a function of $\lambda_C$ and, therefore, of $\pi_C$. Consequently, the cell probabilities identified by $C$, i.e., $\pi_C$, are sufficient to characterize the probability distribution of $Y$.

**The Non-binary Case**

The extension of the theory presented in the above subsection for the binary case to categorical variables with arbitrary number of levels is straightforward, and it is mainly a matter of notation.

In the general case where $Y$ is a vector of polytomous variables the log-linear expansion of cell probabilities is

$$\log \pi_{jd} = \sum_{D' \subseteq D} \lambda_{jD'} \quad \text{for every } D \subseteq V \text{ and } j \in J,$$

and the constraint $\lambda_D = 0$ in the binary case needs to be reformulated as $\lambda_{jd} = 0$ for every $j \in J$. Then, the definition of hierarchical log-linear model is generalized accordingly, so that also in the non-binary case every hierarchical model is uniquely associated with an (extended) generating class. The connection between vanishing log-linear parameters and conditional independence follows from Corollary 4.3, which generalizes Theorem 4.2. This implies that, in the non-binary case, the independence graph associated with a hierarchical log-linear model can be constructed from the generating class of the model in the same way as in the binary case.

**4.5 Log-linear Graphical Models**

Every hierarchical log-linear model has an independence graph associated with it. However, different hierarchical models may have the same independence graph. For example, both the saturated model $\{\{a, b, c\}\}$
and the no-second-order interaction model with generating class $$\{\{a,b\}, \{a,c\}, \{b,c\}\}$$ are associated with the complete graph of Figure 4.2A. This is due to the fact that the two models differ for the constraint $$\lambda_{\{a,b,c\}} = 0$$ which does not imply any conditional independence. Therefore, an independence graph does not in general provide a complete description of the structure of a hierarchical model. The family of graphical log-linear models is the subclass of hierarchical models characterized by their independence graph (Darroch et al., 1980; Edwards and Kreiner, 1983).

**Definition 4.5.1** An undirected graph model with graph $$G$$ is the hierarchical log-linear model whose generating class is the set of the cliques of $$G$$. Thus, the graph $$G$$ defines the model, and we write $$M(G)$$ to denote the graphical model with independence graph $$G$$.

Undirected graph models are therefore models that can be interpreted solely in terms of conditional independence relationships, and the latter can be completely derived from the independence graph through graph separation. For example, the graphical model defined by the graph of Figure 4.1 is the hierarchical log-linear model with generating class $$\{\{a,b\}, \{a,c\}, \{b,d\}, \{c,d,f\}, \{c,e,f\}\}$$. In the binary case this model has 17 non-vanishing log-linear terms, including $$\lambda_{\varnothing}$$. The smallest hierarchical log-linear model with the same independence graph has generating class $$\{\{a,b\}, \{a,c\}, \{b,d\}, \{c,d\}, \{c,e\}, \{c,f\}, \{d,f\}, \{e,f\}\}$$. This model is not graphical, and it can be obtained from the relevant graphical model by imposing the additional constraints $$\lambda_{\{c,d,f\}} = \lambda_{\{c,e,f\}} = 0$$. It has the advantage that it is more parsimonious than the graphical model with the same independence graph but, on the other hand, the additional constraints have no conditional independence interpretation.

### 4.6 Data, Estimation and Testing

Let $$Y^{(1)} = y^{(1)}, \ldots, Y^{(n)} = y^{(n)}$$ be a random sample of $$n$$ independent and identically distributed observations from a distribution with probability mass function $$\pi$$. Samples from a vector of categorical variables can be conveniently presented as a contingency table $$\{n(Y = i)\}_{i \in \mathcal{I}}$$. 
where the sampling units are classified according to the levels of the variables, i.e., the cells of the table. For instance, Table 4.1 gives the contingency table of the infant survival data set provided by Bishop et al. (1975) (see also Bishop et al., 2007; Whittaker, 1990, section 1.2) that gives information on the survival of 715 infants attending two clinics and the amount of care received by the mother.

The entry $n(Y = i)$ is called the count of the cell $i$, and it is the number of observations falling in the cell $i$, formally computed as $n(Y = i) = \sum_{r=1}^{n} 1(Y^{(r)} = i)$; recall that $1(\cdot)$ is the indicator function. It is also useful to consider tables of marginal counts that are tables produced by cross-classifying the observations on the basis of a subset of the variables. The table of marginal counts relative to a nonempty subset $A \subseteq V$ is denoted by $\{n(Y_A = i_A)\}_{i_A \in \mathcal{I}_A}$ where $n(Y_A = i_A) = \sum_{r=1}^{n} 1(Y^{(r)}_A = i_A)$. For instance, Table 4.2 gives the marginal contingency tables of the pairs of variables \{Clinic, Survival\} and \{Care, Clinic\} for the infant survival data.

We remark that the marginal counts can also be obtained from a larger contingency table by marginalizing over the variables not in $A$ so that $n(Y_A = i_A) = \sum_{Y_{V \setminus A} = i_{V \setminus A}} n(Y_A = i_A, Y_{V \setminus A} = i_{V \setminus A})$; likewise, $n = \sum_{i \in \mathcal{I}} n(Y = i)$.

We now consider the problem of computing the maximum likelihood (ML) estimate of the parameters of a hierarchical log-linear model. To keep the presentation at a simpler level, we assume that the cell counts are all strictly positive and refer to Lauritzen (1996, chapter 4) and Agresti (2013, section 10.6) for an extension of the results of this section to the case of empty cells and sparse contingency tables.
Under the multinomial sampling scheme considered here, the sample size $n$ is fixed and the observations are independent and identically distributed. Thus, the sampling distribution of counts follows a multinomial distribution with probability mass function

$$p\{n(Y = i)\} = \frac{n!}{\prod_{i \in I} n(Y = i)!} \prod_{i \in I} p(Y = i)^{n(Y = i)} \text{ for all } i \in I.$$  

(4.12)

The expected count of the cell $i$, $E\{n(Y = i)\}$, is denoted by $m(Y = i)$ and it can be computed by multiplying by $n$ the relevant cell probability,

$$m(Y = i) = np(Y = i) \text{ for all } i \in I.$$  

It follows that the ML estimate can be equivalently computed with respect to expected counts because

$$\hat{p}(Y = i) = \frac{\hat{m}(Y = i)}{n} \text{ for every } i \in I.$$  

Equation (4.12) leads to the log-likelihood function

$$l(\pi) = \sum_{i \in I} n(Y = i) \log p(Y = i),$$  

(4.13)

and it follows immediately from (4.13) that for the saturated model, where $C^\uparrow = \{V\}$, the sufficient statistic is the table of observed counts $\{n(Y = i)\}_{i \in I}$ and, furthermore, it can be shown that the observed counts are the ML estimates of the expected counts,

$$\{\hat{m}(Y = i)\}_{i \in I} = \{n(Y = i)\}_{i \in I}.$$  

<table>
<thead>
<tr>
<th>Clinic</th>
<th>Survival</th>
<th>Care</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>Clinic A</td>
<td>7</td>
<td>469</td>
</tr>
<tr>
<td>Clinic B</td>
<td>19</td>
<td>220</td>
</tr>
</tbody>
</table>

Table 4.2 Infant survival data: marginal tables of counts for the pairs of variables \{Clinic, Survival\} and \{Care, Clinic\}.

Graphical Models for Categorical Data

Downloaded from https://www.cambridge.org/core, IP address: 54.191.40.80, on 30 Jul 2017 at 12:12:58, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms.
https://www.cambridge.org/core/product/F9D4523D3559BC2BA9251CEF3077F20D
The rules to compute the ML estimates for hierarchical log-linear models where first derived by Birch (1963) and are usually referred to as the Birch’s results. They state that for an arbitrary hierarchical log-linear model with generating class $C^\uparrow = \{C_1, \ldots, C_k\}$ the minimal sufficient statistic is the set of the marginal tables of counts corresponding to the maximal interactions of the model, i.e.,

$$\{n(Y_C = i_C)\}_{i_C \in I_C} \text{ for every } C \in C^\uparrow,$$

and the ML estimates of the marginal distributions relative to the maximal interactions of the models are directly obtained from the sufficient statistics as follows

$$\{\hat{m}(Y_C = i_C)\}_{i_C \in I_C} = \{n(Y_C = i_C)\}_{i_C \in I_C} \text{ for every } C \in C^\uparrow. \quad (4.14)$$

Hence, the ML estimate of $\{m(Y = i)\}_{i \in I}$ must satisfy both equations (4.14) and the restrictions defining the model. This is formally stated below.

**Proposition 4.4** The maximum likelihood estimate of $\{m(Y = i)\}_{i \in I}$, under the hierarchical log-linear model with generation class $C^\uparrow$, based on a random sample from the multinomial distribution, exists and it is the unique $\{\hat{m}(Y = i)\}_{i \in I}$ that satisfies both the system of equation (4.14) and the restrictions defining $M(C^\uparrow)$.

A proof of this result can be found, for instance, in Lauritzen (1996, theorem 4.8). Here, it deserves noticing that log-linear models are regular exponential families and the result of Proposition 4.4 follows from the general properties of this class of models, where the ML estimates can be found by equating the sufficient statistics to their expectation. Although Proposition 4.4 is stated assuming a multinomial sampling scheme, it can be shown that it is also valid under the Poisson sampling scheme. Moreover, in the case of independent multinomial sampling, where certain marginal distributions are fixed by the sampling design, Proposition 4.4 holds in the case where the fixed marginals correspond to interactions that are present in the model; see Agresti (2013) for further details.

Many log-linear models have no closed-form ML estimate, and the solutions to the likelihood equations are computed by means of iterative procedures. A simple algorithm commonly used in this context is called iterative proportional fitting (IPF), and it works by successively
adjusting the ML estimates of the marginal tables corresponding to the sufficient statistics (Deming and Stephan, 1940; Darroch and Ratcliff, 1972). The IPF algorithm for computing the ML estimate of a hierarchical model with generating class $C^\uparrow = \{C_1, \ldots, C_k\}$ has the following steps:

1. Start with an initial estimate $\{\tilde{m}_0(Y = i)\}_{i \in I}$ having association structure not more complex than $M(C^\uparrow)$. A common choice is $\tilde{m}_0(Y = i) = 1$ for all $i \in I$.

2. For every $s = 1, \ldots, k$ set $C = C_s$ and apply the updating step

$$\tilde{m}_s(Y = i) = \frac{n(Y_C = i_C)}{\tilde{m}_{s-1}(Y_C = i_C)} \times \tilde{m}_{s-1}(Y = i)$$

for every $i \in I$.

3. If the maximal difference between the sufficient statistics $\{n(Y_C = i_C)\}_{i_C \in I_C}$ and their fitted values $\{\tilde{m}_k(Y_C = i_C)\}_{i_C \in I_C}$ is sufficiently close to zero for all $C \in C^\uparrow$ then set $\{\tilde{m}(Y = i)\}_{i \in I} = \{\tilde{m}_k(Y = i)\}_{i \in I}$ and stop. Otherwise, set $\{\tilde{m}_0(Y = i)\}_{i \in I} = \{\tilde{m}_k(Y = i)\}_{i \in I}$ and go to the step 2 above.

For details on the convergence of the IPF algorithm to the ML estimate we refer to Whittaker (1990, section 4.7) and references therein. Here we remark that the IPF works even in the case when there exists an explicit expression for calculating the ML estimate, and in this case the algorithm has the nice property that it converges to the ML estimate after a finite number of cycles.

As an example, we show the application of the IPF algorithm to the infant survival data. We consider the model with $C^\uparrow = \{\{\text{Clinic, Survival}\}, \{\text{Care, Clinic}\}\}$ corresponding to $Y_{\text{Care}} \perp Y_{\text{Survival}} | Y_{\text{Clinic}}$ and therefore to the independence graph of Figure 4.2B where Clinic is $c$ and Care and Survival are $a$ and $b$, respectively. The tables below give the estimates of the expected counts and the implied marginal tables for the pairs $\{\text{Clinic, Survival}\}$ and $\{\text{Care, Clinic}\}$, which should be compared with the marginal counts in Table 4.2 to monitor the convergence of the procedure. We set the initial values constant and equal to 1, thereby specifying an initial model with no interaction terms, and therefore with simpler structure than the model being fitted.
Next, we apply step 2 of the algorithm to adjust the margin $C = \{\text{Clinic, Survival}\}$, thereby obtaining the updated estimate,
Subsequently, we apply step 2 of the algorithm to adjust the margin \( C = \{\text{Care, Clinic}\} \) and this operation gives the following estimate:

<table>
<thead>
<tr>
<th>Clinic</th>
<th>Less</th>
<th>More</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>238.0</td>
<td>238.0</td>
</tr>
<tr>
<td>B</td>
<td>119.5</td>
<td>119.5</td>
</tr>
</tbody>
</table>

<p>| Survival |
|---|---|</p>
<table>
<thead>
<tr>
<th>Clinic</th>
<th>Care</th>
<th>No</th>
<th>Yes</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>Less</td>
<td>2.63</td>
<td>176.37</td>
</tr>
<tr>
<td></td>
<td>More</td>
<td>4.37</td>
<td>292.63</td>
</tr>
<tr>
<td>B</td>
<td>Less</td>
<td>17.01</td>
<td>196.99</td>
</tr>
<tr>
<td></td>
<td>More</td>
<td>1.99</td>
<td>23.01</td>
</tr>
</tbody>
</table>

<p>| Survival |
|---|---|</p>
<table>
<thead>
<tr>
<th>Clinic</th>
<th>No</th>
<th>Yes</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>7</td>
<td>469</td>
</tr>
<tr>
<td>B</td>
<td>19</td>
<td>220</td>
</tr>
</tbody>
</table>

<p>| Care |
|---|---|---|</p>
<table>
<thead>
<tr>
<th>Clinic</th>
<th>Less</th>
<th>More</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>179</td>
<td>297</td>
</tr>
<tr>
<td>B</td>
<td>214</td>
<td>25</td>
</tr>
</tbody>
</table>

As both the \( \{\text{Clinic, Survival}\} \)-marginal and the \( \{\text{Care, Clinic}\} \)-marginal satisfy the likelihood equations the algorithm has converged, and there is no need for further iterations. The fitted expected counts can then be divided by \( n = 715 \) to obtain the ML estimates of the cell probabilities. Notice that convergence has required a single cycle because, as we will see later, this model admits an explicit expression for the ML estimate.
We now turn to model fitting. The basic statistic used in this context is the deviance which, for a model \( M(C) \), is computed as twice the difference between the log-likelihood of the saturated model and the log-likelihood taken over \( M(C) \),

$$\text{dev}(C) = 2 \sum_{i \in I} n(Y = i) \log \frac{n(Y = i)}{\hat{m}(Y = i)}.$$  

For hierarchical log-linear models the deviance has the same form as the likelihood-ratio statistic for testing a log-linear model against the corresponding saturated model. When \( M(C) \) holds, the deviance has asymptotic chi-squared distribution with degrees of freedom given by the number of constraints of \( M(C) \), i.e., \( 2^{|V|} - |C| \). For example, if for the infant survival data we consider the model with generation class \( C^\uparrow = \{\text{Clinic, Survival}, \{\text{Care, Clinic}\} \) we can use the ML estimate computed above to see that \( \text{dev}(C) = 0.08 \). Because the number of constraints of this model is equal to two, then it makes sense to compare the value of the deviance with the chi-squared quantile \( \chi^2_{2,0.95} = 5.99 \), thereby showing that this model provides a very good fit of the data (\( p \)-value = 0.97).

For two models \( M_0 \) and \( M_1 \) we write \( M_0 \subseteq M_1 \) if \( M_0 \) can be obtained from \( M_1 \) by imposing additional restrictions. It is easy to see that if \( M_0 \) and \( M_1 \) are hierarchical log-linear models with expanded generating classes \( C_0 \) and \( C_1 \), respectively, then \( M_0 \subseteq M_1 \) if and only if \( C_0 \subseteq C_1 \). The deviance statistic to compare the two models is the difference

$$\text{dev}(C_0|C_1) = \text{dev}(C_0) - \text{dev}(C_1) = 2 \sum_{i \in I} n(Y = i) \log \frac{\hat{m}_1(Y = i)}{\hat{m}_0(Y = i)}.$$  

(4.15)

When \( M(C_0) \) holds the deviance difference has asymptotic chi-squared distribution with degrees of freedom given by the number of additional constrains of \( M(C_0) \) compared to \( M(C_1) \), i.e., \( |C_1| - |C_0| \).

### 4.7 Graph Decomposition and Decomposable Graphs

This section introduces some definitions and results from the theory of undirected graphs that can be used to identify relevant features of graphical models.
If $G = (V, E)$ is an undirected graph, the subgraph of $G$ induced by the subset $A \subseteq V$, denoted by $G_A = (A, E_A)$, is the graph with vertex set $A$ and all those edges which join two vertices that are both in $A$. An undirected graph $G = (V, E)$ is complete if all its vertices are mutually adjacent and we say that a subset $A \subseteq V$ is complete when $G_A$ is complete. We denote by $\mathcal{V} = \mathcal{V}(G)$ the collection of all the complete subsets of $V$, and we remark that $\mathcal{V}$ is a simplicial complex. This can be seen by noticing that both all singleton subsets of $V$ and the empty set are complete because they induce subgraphs which have exactly one vertex and no vertices, respectively. Accordingly, in the following we will refer to $\mathcal{V}$ as the simplicial complex relative to $G$. A subset $C \subseteq V$ is called a clique if it is maximally complete, i.e., $C$ is complete, and if $C \subset D$, then $D$ is not complete. For example, in the graph of Figure 4.1 the set $\{c, d\}$ is complete but it is not maximally complete because it is a subset of the complete set $\{c, d, f\}$. In parallel with the generating class of a hierarchical log-linear model that is formed by the maximal sets of a simplicial complex, also the cliques of $G$ are the maximal sets of $\mathcal{V}$, and thus we denote the set of the cliques of $G$ by $\mathcal{V}^\uparrow = \{C_1, \ldots, C_k\}$. An undirected graph can be identified by the set of its cliques, so that, for instance, the graph in Figure 4.1 is identified by $\mathcal{V}^\uparrow = \{\{a, b\}, \{a, c\}, \{b, d\}, \{c, d, f\}, \{c, e, f\}\}$.

**Definition 4.7.1** The pair $(A, B)$ of subsets of $V$ such that $A \cup B = V$ forms a decomposition of $G = (V, E)$ if it holds that

1. $A \cap B$ is a complete subset of $V$;
2. $A \cap B$ separates $A \setminus B$ from $B \setminus A$ in $G$.

The subgraphs $G_A$ and $G_B$ are two components of $G$. A decomposition is proper if both $A \setminus B \neq \emptyset$ and $B \setminus A \neq \emptyset$, and a graph is said to be prime if it does not admit any decomposition. For example, the pair $(\{a, b, c, d\}, \{c, d, e, f\})$ forms a decomposition of the graph in Figure 4.1. The components associated with this decomposition, given in Figure 4.3, are $G_{\{a, b, c, d\}}$, that is prime, and $G_{\{c, d, e, f\}}$, that is not prime because it admits the decomposition $(\{c, d, f\}, \{c, e, f\})$. Any undirected graph can be recursively decomposed into its prime components (Diestel, 1990). Figure 4.4 displays the prime components of the graph in Figure 4.1. In this example, the prime components $G_{\{c, d, f\}}$ and $G_{\{c, e, f\}}$ are complete subgraphs, and their vertex sets are cliques of $G$. 

---

**Graphical Models for Categorical Data**

---

Downloaded from https://www.cambridge.org/core, IP address: 54.191.40.80, on 30 Jul 2017 at 12:12:58, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://www.cambridge.org/core/product/F9D4523D3559BC2BA9251CEF3077F20D
Conversely, the prime component $G_{\{a,b,c,d\}}$ is not complete, but it cannot be further decomposed.

The recursive decomposition of a graph into its prime components can be used to define the distinguished subclass of decomposable graphs.

**Definition 4.7.2** An undirected graph $G$ is said to be decomposable if either (i) it is complete, or (ii) there exists a proper decomposition $(A, B)$ of $G$ such that both $G_A$ and $G_B$ are decomposable.

This definition is recursive, but it is well posed because the decompositions are required to be proper. To check whether a graph is decomposable it is sufficient to recursively decompose it into its prime components, and the graph is decomposable if the prime components are all complete, otherwise it is nondecomposable. It follows that a graph is decomposable if and only if the vertex sets of its prime components are the cliques of the graph. Thus, the graph in Figure 4.1 is nondecomposable, whereas it is not difficult to check that the graph in Figure 4.5 is decomposable.

There exist a number of alternative ways to characterize the class of decomposable graphs (see Cowell *et al.*, 1999, chapter 4). Every alternative characterization identifies a useful property of decomposable graphs.
graphs. For example, the fact that the class of decomposable graphs coincides with the class of chordal graphs provides a useful way to visually check decomposability in small graphs. A chord is a pair of non-consecutive vertices in a cycle that are joined by an edge. A cycle is chordless if it has no chords, and an undirected graph is said to be chordal or triangulated if it has no chordless cycle with four or more vertices. For example, in the graph of Figure 4.1 the cycle $(c, d, f, e, c)$ possesses the chord $c \sim f$, whereas $(a, b, d, c, a)$ is a chordless cycle with four vertices. On the other hand, the graph in Figure 4.5 has no chordless cycles. It can be shown that an undirected graph is chordal if and only if it is decomposable (see Cowell et al., 1999, theorem 4.4). A visual inspection of the graphs in Figures 4.5 and 4.1 allows one immediately to assess that the former is triangulated, and therefore decomposable, whereas the latter is not.

Let $(C_1, \ldots, C_k)$ be a sequence of complete subsets of the vertex set $V$. For all $r = 2, \ldots, k$ let

$$S_r = (C_1 \cup \ldots \cup C_{r-1}) \cap C_r.$$  

The sequence $(C_1, \ldots, C_k)$ is said to be perfect if for every $r = 2, \ldots, k$ there is an $s < r$ such that $S_r \subseteq C_s$. This condition is known as the running intersection property, and the proof of the following result can be found in Lauritzen (1996, proposition 2.17).

**Proposition 4.5** An undirected graph is decomposable if and only if its cliques can be numbered to form a perfect sequence.

For instance, the sequence $(\{a, b\}, \{b, c\}, \{c, d, e\}, \{d, e, f\}, \{d, e, g\})$ is a perfect sequence of the cliques of the graph in Figure 4.5. It is easy to see that any decomposable graph can be recursively decomposed by starting from a perfect sequence $(C_1, \ldots, C_k)$ of its cliques, and then by decomposing $G_{C_1 \cup \ldots \cup C_r}$ into the components $G_{C_1 \cup \ldots \cup C_{r-1}}$ and $G_{C_r}$ for every $r = k, (k-1), \ldots, 2$. Every step of this procedure is a
decomposition and has a separating set associated with it, which we denote by $S_r$. Accordingly, every perfect sequence of cliques is associated with a sequence $(S_2, \ldots, S_k)$ of separators of $G$. The sequence of separators, for the graph in Figure 4.5, associated with the perfect sequence of cliques given above is $\langle \{b\}, \{c\}, \{d, e\}, \{d, e\} \rangle$. As this example shows, separators need not be distinct. Furthermore, apart from their ordering, separators are the same for all perfect sequences of cliques. For example, another perfect sequence of the cliques of the same graph is $\langle \{e, d, f\}, \{d, e, g\}, \{c, d, e\}, \{b, c\}, \{a, b\} \rangle$ and the associated sequence of separators is $\langle \{d, e\}, \{d, e\}, \{c\}, \{b\} \rangle$.

Formal algorithms have been developed for deciding whether an undirected graph is chordal or not. A convenient algorithm was provided by Tarjan and Yannakakis (1984) and it is called maximum cardinality search; see also Cowell et al. (1999, section 4.4). If a graph is chordal this procedure provides additional useful information such as the cliques of the graphs as well as a perfect ordering of the cliques.

For two undirected graphs with common vertex set, $G = (V, E)$ and $G' = (V, E')$, we say that $G$ is larger than $G'$, and write $G' \subseteq G$, if $E' \subseteq E$; when the inclusion is strict, i.e., $E' \subset E$, we write $G' \subset G$. Frydenberg and Lauritzen (1989) showed that it is possible to explore the space of decomposable graphs by moving between neighboring decomposable models which differ exactly by one edge.

**Proposition 4.6** Let $G = (V, E)$ and $G' = (V, E')$ be two decomposable graphs such that $G' \subset G$ and $|E \setminus E'| = k$. Then there is sequence $G' = G_0 \subset \ldots \subset G_k = G$ of decomposable graphs that differ by exactly one edge.

Frydenberg and Lauritzen (1989) also gave the following result concerning neighboring decomposable graphs that differ by one edge.

**Proposition 4.7** Let $G = (V, E)$ and $G' = (V, E')$ be two decomposable graphs such that $G' \subset G$ and $G$ having exactly one more edge than $G'$, i.e., $|E \setminus E'| = 1$. Then this edge is contained in exactly one clique $C$ of $G$.

### 4.8 Local Computation Properties

One key feature of graphical models is modularity; i.e., they exploit conditional independence relations to split complex systems into
simpler modules. Consider the undirected graph model $M(G)$ for $Y$ with independence graph $G = (V, E)$. If the pair of subsets $A, B \subseteq V$ is such that $A \cup B = V$ and $S = A \cap B$ separates $A \setminus B \neq \emptyset$ from $B \setminus A \neq \emptyset$ then it holds that $Y_{A \setminus B} \perp Y_{B \setminus A} | Y_S$, and it follows from the definition of conditional independence (2.6) that the probability mass function of $Y$ factorizes as

$$p(Y = i) = \frac{p(Y_A = i_A)p(Y_B = i_B)}{p(Y_S = i_S)} \text{ for all } i \in \mathcal{I}. \quad (4.16)$$

It is straightforward to see that this factorization can be equivalently stated in terms of expected cell counts $m(Y = i) = np(Y = i)$, and thus also the estimated expected cell counts factorize accordingly,

$$\hat{m}(Y = i) = \frac{\hat{m}(Y_A = i_A)\hat{m}(Y_B = i_B)}{\hat{m}(Y_S = i_S)} \text{ for all } i \in \mathcal{I}, \quad (4.17)$$

where \( \{\hat{m}(Y_A = i_A)\}_{i_A \in \mathcal{I}_A} \) is obtained by marginalizing \( \{\hat{m}(Y = i)\}_{i \in \mathcal{I}} \) over the variables $Y_{B \setminus A}$, and similarly for \( \{\hat{m}(Y_B = i_B)\}_{i_B \in \mathcal{I}_B} \) and \( \{\hat{m}(Y_S = i_S)\}_{i_S \in \mathcal{I}_S} \). Consider now the graph $G_A$, i.e., the subgraph of $G$ induced by $A$. This graph can be used to define the undirected graph model $M(G_A)$ for $Y_A$, and we denote by \( \{\hat{m}^A(Y_A = i_A)\}_{i_A \in \mathcal{I}_A} \) the ML estimates of the expected cell counts under $M(G_A)$. In general, the marginal table of estimates \( \{\hat{m}(Y_A = i_A)\}_{i_A \in \mathcal{I}_A} \) is different from the estimate \( \{\hat{m}^A(Y_A = i_A)\}_{i_A \in \mathcal{I}_A} \) computed under the model $M(G_A)$. One relevant exception is the case where $(A, B)$ forms a decomposition of $G$ because in this case the ML estimate in $M(G)$ can be obtained from the marginal vectors $Y_A$ and $Y_B$ by plugging in (4.17) the ML estimates computed under $M(G_A)$ and $M(G_B)$.

**Theorem 4.8** Let \( \{\hat{m}(Y = i)\}_{i \in \mathcal{I}} \) be the ML estimates of the expected cell counts in the undirected graph model $M(G)$. If $(A, B)$ forms a proper decomposition of $G$ then it holds that

$$\hat{m}(Y = i) = \frac{\hat{m}(Y_A = i_A)\hat{m}(Y_B = i_B)}{\hat{m}(Y_S = i_S)} \text{ for all } i \in \mathcal{I},$$

and, furthermore, for every $i_A \in \mathcal{I}_A$ and $i_B \in \mathcal{I}_B$,

$$\hat{m}(Y_A = i_A) = \hat{m}^A(Y_A = i_A) \text{ and } \hat{m}(Y_B = i_B) = \hat{m}^B(Y_B = i_B),$$

where \( \{\hat{m}^A(Y_A = i_A)\}_{i_A \in \mathcal{I}_A} \) and \( \{\hat{m}^B(Y_B = i_B)\}_{i_B \in \mathcal{I}_B} \) are the ML estimates under $M(G_A)$ and $M(G_B)$, respectively.
This theorem is a special case of a more general result proved in Lauritzen (1996, theorem 4.4). Theorem 4.8 states one of the most relevant features of graphical models, that is the local computation of several statistical quantities marginally on the component models \( M(G_A) \) and \( M(G_B) \). Consider the case where \( G^1 \) and \( G^0 \) are two graphs on the same vertex set such that \( G^0 \subset G^1 \). To see and example of this situation, let \( G^1 \) be the graph in Figure 4.1 and \( G^0 \) the graph obtained by removing from \( G^1 \) the edge \( e \sim f \). Then the deviance difference (4.15) for comparing these two models is

\[
\text{dev}(G^0|G^1) = 2 \sum_{i \in \mathcal{I}} n(Y = i) \log \frac{\hat{m}_1(Y = i)}{\hat{m}_0(Y = i)}
\]

where \( \{\hat{m}_0(Y = i)\}_{i \in \mathcal{I}} \) and \( \{\hat{m}_1(Y = i)\}_{i \in \mathcal{I}} \) are the ML estimates under \( M(G^0) \) and \( M(G^1) \), respectively. Assume now that \((A, B)\) forms a proper decomposition of both \( G^0 \) and \( G^1 \), and that \( G_A^0 = G_A^1 \). More concretely, in the example above, the pair \((A, B)\) that forms a decomposition of both graphs is \( \{\{a, b, c, d\}, \{c, d, e, f\}\} \). In this case, \( G_{\{a,b,c,d\}}^0 = G_{\{a,b,c,d\}}^1 \) is the graph in Figure 4.3A, whereas \( G_{\{c,d,e,f\}}^1 \) is the graph in Figure 4.3B and \( G_{\{c,d,e,f\}}^0 \) can be obtained from the latter by removing the edge \( e \sim f \). It follows that we can apply Theorem 4.8 to obtain

\[
\text{dev}(G^0|G^1) = 2 \sum_{i \in \mathcal{I}} n(Y = i) \log \left( \frac{\hat{m}_1(Y_A = i_A)\hat{m}_1(Y_B = i_B)}{\hat{m}_0(Y_A = i_A)\hat{m}_0(Y_B = i_B)} \right)
\]

\[
= 2 \sum_{i \in \mathcal{I}} n(Y = i) \log \frac{\hat{m}_1(Y_A = i_A)\hat{m}_1(Y_B = i_B)}{\hat{m}_0(Y_A = i_A)\hat{m}_0(Y_B = i_B)}
\]

However, because \( G_A^0 = G_A^1 \) then also \( M(G_A^0) = M(G_A^1) \) and therefore \( \hat{m}_0(Y_A = i_A) = \hat{m}_1(Y_A = i_A) \) for every \( i_A \in \mathcal{I}_A \). Hence,

\[
\text{dev}(G^0|G^1) = 2 \sum_{i_B \in \mathcal{I}_B} n(Y_B = i_B) \log \frac{\hat{m}_1(Y_B = i_B)}{\hat{m}_0(Y_B = i_B)}
\]

and we can conclude that the deviance difference can be computed locally with respect to the marginal distribution of \( Y_B \); i.e.,

\[
\text{dev}(G^0|G^1) = \text{dev}(G_B^0|G_B^1). \quad (4.18)
\]
In fact, in the example described above a further simplification is possible because one can exploit the decomposition \(\{c, d, f\}, \{c, e, f\}\) of both \(G^0_{\{c, d, e, f\}}\) and \(G^1_{\{c, d, e, f\}}\) to localize the computation of the deviance on the submodels defined by \(G^0_{\{c, e, f\}}\) and \(G^1_{\{c, e, f\}}\).

The local computation in (4.18) has the advantage that it simplifies the computation of the deviance and, furthermore, it involves marginal tables with larger expected cell counts with respect to the full table and this has typically a positive effect on the quality of the approximation of the distribution of the deviance to its asymptotic chi-squared distribution.

Local Computation of the Log-linear Terms

We now turn to the log-linear parameters of the model and to the local computation of these quantities. The following theorem gives the decomposition of the log-linear terms implied by conditional independence.

**Theorem 4.9**  
Let \(\lambda^V\) be the log-linear parameter of the distribution of \(Y_V\) with probability parameter \(\pi > 0\). Consider a pair of subsets \(A\) and \(B\) such that \(A \cup B = V\), with both \(A \setminus B\) and \(B \setminus A\) nonempty. Furthermore, let \(\lambda^A, \lambda^B\) and \(\lambda^{A \cap B}\) be the log-linear parameters of the distributions of \(Y_A\), \(Y_B\) and \(Y_{A \cap B}\), respectively. If \(Y_{A \setminus B} \independent Y_{B \setminus A} \mid Y_{A \cap B}\), then in the binary case it holds that for all \(D \subseteq V\)

\[
\lambda_D = \lambda^A_D + \lambda^B_D - \lambda^{A \cap B}_D
\]  
where we use the convention that \(\lambda_D^A = 0\) whenever \(D \subseteq A\), and similarly for \(\lambda^B\) and \(\lambda^{A \cap B}\). In the general, non-binary, case (4.19) has the form

\[
\lambda_D = \lambda^{A \setminus D}_D + \lambda^{B \setminus D}_D - \lambda^{A \cap B \setminus D}_D\]

for all \(D \subseteq V\) and \(j \in J\).

**Proof.**  
We first consider the binary case. Notice that if \(D\) is such that both \(D \cap A \neq \emptyset\) and \(D \cap B \neq \emptyset\) then the result is trivially true because \(\lambda_D = 0\) by Theorem 4.2, whereas \(\lambda^A_D = \lambda^B_D = \lambda^{A \cap B}_D = 0\) by assumption. Hence, in the rest of this proof we can assume, without loss of generality, that \(D \subseteq A\). The conditional independence relationship \(Y_{A \setminus B} \independent Y_{B \setminus A} \mid Y_{A \cap B}\) implies that \(\pi_D = \pi^{A \setminus D}_D \pi^{B \setminus D}_D \pi^{A \cap B \setminus D}_D / \pi^S_D\), for every \(D \subseteq V\), and therefore that

\[
\lambda_D = \sum_{E \subseteq D} (-1)^{|D \setminus E|} \log \pi_E
\]

\[
= \sum_{E \subseteq D} (-1)^{|D \setminus E|} \left( \log \pi^A_E + \log \pi^B_{E \cap B} - \log \pi^S_{E \cap S} \right)
\]
where $E \cap A = E$ because $E \subseteq D \subseteq A$. For the same reason $E \cap B = E \cap S$ so we can write

$$
\lambda_D = \left\{ \sum_{E \subseteq D} (-1)^{|D \setminus E|} \log \pi_E^A \right\} + \sum_{E \subseteq D} (-1)^{|D \setminus E|} \left( \log \pi_{E \cap S}^B - \log \pi_{E \cap S}^S \right)
$$

$$
= \lambda_D^A + \sum_{E \subseteq D} (-1)^{|D \setminus E|} \left( \log \pi_{E \cap S}^B - \log \pi_{E \cap S}^S \right).
$$

Hence,

$$
\lambda_D = \lambda_D^A + \sum_{E_1 \subseteq D \setminus S} \sum_{E_2 \subseteq D \cap S} (-1)^{|D \setminus (E_1 \cup E_2)|} \left( \log \pi_{E_2}^B - \log \pi_{E_2}^S \right)
$$

where we have used the fact that $(E_1 \cup E_2) \cap S = E_2$. We can now factorize $(-1)^{|D \setminus (E_1 \cup E_2)|}$ as $(-1)^{|(D \cap S) \setminus E_1|} \times (-1)^{|(D \setminus S) \setminus E_2|}$ to obtain

$$
\lambda_D = \lambda_D^A + \left\{ \sum_{E_1 \subseteq D \setminus S} (-1)^{|(D \setminus S) \setminus E_1|} \right\}
$$

$$
\times \sum_{E_2 \subseteq D \cap S} (-1)^{|(D \cap S) \setminus E_2|} \left( \log \pi_{E_2}^B - \log \pi_{E_2}^S \right). \tag{4.20}
$$

At this point there are two possible cases, i.e., either $D \setminus S \neq \emptyset$ or $D \setminus S = \emptyset$. If $D \setminus S \neq \emptyset$ then the first quantity within brackets is equal to zero because any finite, nonempty set has the same number of subsets of even as of odd cardinality. Hence, $\lambda_D = \lambda_D^A$ and the result follows because in this case both $\lambda_D^B = \lambda_D^S = 0$ by assumption because both $D \subseteq B$ and $D \subseteq S$. On the other hand, if $D \setminus S = \emptyset$ then the first quantity within brackets is equal to one and, furthermore, in this case $D \cap S = D$ so we can write (4.20) as

$$
\lambda_D = \lambda_D^A + \sum_{E_2 \subseteq D} (-1)^{|D \setminus E_2|} \log \pi_{E_2}^B - \sum_{E_2 \subseteq D} (-1)^{|D \setminus E_2|} \log \pi_{E_2}^S
$$

$$
= \lambda_D^A + \lambda_D^B - \lambda_D^S.
$$

We have thus shown the assertion of the theorem for the binary case. The extension to the non-binary case can be obtained by repeating the proof above for $\lambda_j$ for every $j \in \mathcal{J}$. \square

Let $G$ be an undirected graph with vertex set $V$ so that $\mathcal{V}$ is the simplicial complex relative to $G$. If the subsets $A, B \subseteq V$ are such that
\( A \cup B = V \), then the triple \((A \setminus B, S, B \setminus A)\), where \( S = A \cap B \), forms a partition of \( V \). The following lemma shows that if \( S \) separates \( A \setminus B \) from \( B \setminus A \) in \( G \) then there exists a partition of \( V \) that resembles that of \( V \).

**Lemma 4.10** Consider the simplicial complex \( V \) relative to the graph \( G = (V, E) \). For the subsets \( A, B \subseteq V \) set \( S = A \cap B \) and let \( A, B \) and \( S \) be the simplicial complexes relative to \( G_A, G_B \) and \( G_S \), respectively. Then, if \( S \) separates \( A \setminus B \) from \( B \setminus A \) in \( G \) it holds that

\[
A \cup B = V \quad \text{and} \quad A \cap B = S
\]

so that the triple \((A \setminus B, S, B \setminus A)\) forms a partition of \( V \).

**Proof.** We first notice that if a pair of subsets \( Q, E \subseteq V \) is such that if \( Q \subseteq E \) then it also holds that \( Q \subseteq \mathcal{E} \), where \( Q \) and \( \mathcal{E} \) are the simplicial complexes relative \( G_Q \) and \( G_E \), respectively. By definition \( A, B \subseteq V \) so that \( A, B \subseteq V \) and therefore \( A \cup B \subseteq V \). Hence, to show the equality \( A \cup B = V \) it is sufficient to show the reverse inclusion, i.e., that if \( D \in V \) then \( D \in A \cup B \). Any set \( D \in V \) is complete in \( G \), and the fact that \( A \cap B = S \) separates \( A \setminus B \) from \( B \setminus A \) in \( G \) implies that at least one of the two inclusions \( D \subseteq A \) and \( D \subseteq B \) is satisfied. This implies, in turn, that at least one of the two inclusions \( D \in A \) and \( D \in B \) is satisfied and therefore that \( D \in A \cup B \), as required. We now show that \( A \cap B = S \).

Because \( S \subseteq A \) and \( S \subseteq B \) then \( S \subseteq A \) and \( S \subseteq B \) and therefore \( S \subseteq A \cap B \). We now show the reverse implication, that is that if \( D \in A \cap B \) then \( D \in S \). Any set \( D \in A \cap B \) is such that both \( D \in A \) and \( D \in B \) which, in turn, imply both \( D \subseteq A \) and \( D \subseteq B \). Hence, \( D \subseteq A \cap B = S \), and as \( D \) is complete in \( G \) it is also complete in \( G_S \) and therefore \( D \in S \), and this completes the proof. \( \square \)

The partition of \( V \) in Lemma 4.10 provides a way to write the result of Theorem 4.9 in vector form. If \( \lambda \) is the log-linear parameter of a random vector under \( M(G) \) then we can write \( \lambda = (\lambda_V, \lambda_\mathcal{V}) \) where \( \lambda_\mathcal{V} = 0 \). Furthermore, by Lemma 4.10, \( \lambda_V \) can be partitioned into \( \lambda_V = (\lambda_{A \setminus B}, \lambda_S, \lambda_{B \setminus A}) \) and the result of Theorem 4.9 can be written as

\[
\lambda_V = \begin{pmatrix} \lambda_{A \setminus B} \\ \lambda_S \\ \lambda_{B \setminus A} \end{pmatrix} = \begin{pmatrix} \lambda^A_{A \setminus B} \\ \lambda^S \\ \lambda^B_{B \setminus A} \end{pmatrix} = \begin{pmatrix} \lambda^A_{A \setminus B} \\ \lambda^S \\ \lambda^B_{B \setminus A} \end{pmatrix}.
\]

(4.21)
In graphical models for random variables with multivariate Gaussian distribution the role of \( \lambda \) is played by the inverse of the variance and covariance matrix, and this parameter can be locally computed in a way that resembles the local computation of \( \lambda \) in (4.21) as shown, for example, in Roverato and Whittaker (1998).

When \((A, B)\) forms a proper decomposition of \( G \) then also the ML estimate \( \hat{\lambda} \) can be computed locally.

**Corollary 4.11** Let \( \hat{\lambda} \) be the ML estimate of the log-linear parameter of the distribution of the binary vector \( Y_V \), with probability parameter \( \pi > 0 \), in the undirected graph model \( M(G) \). If \((A, B)\) forms a proper decomposition of \( G \) then it holds that

\[
\hat{\lambda}_V = \begin{pmatrix}
\hat{\lambda}_{A|B} \\
\hat{\lambda}_S \\
\hat{\lambda}_{B\setminus A}
\end{pmatrix} = \begin{pmatrix}
\hat{\lambda}_{A|B}^A \\
\hat{\lambda}_S^A + \hat{\lambda}_S^B - \hat{\lambda}_S^S \\
\hat{\lambda}_{B\setminus A}^B
\end{pmatrix},
\]

(4.22)

where \( \hat{\lambda}_A \), \( \hat{\lambda}_B \) and \( \hat{\lambda}_S \) are the ML estimates of the log-linear parameters under \( M(G_A) \), \( M(G_B) \) and \( M(G_S) \), respectively.

**Proof.** This follows immediately from Theorems 4.8 and 4.9, and from equation (4.21). \( \square \)

The results given in equations (4.21) and (4.22) can be immediately extended to the case of categorical variables with arbitrary number of levels.

**4.9 Models for Decomposable Graphs**

Undirected graph models defined by decomposable graphs are called *decomposable*. Decomposable models are important for various reasons, and one is that many theoretical and computational aspects are more tractable with decomposable models than with general models. One reason for that is that the modules of a decomposable graph are all complete and this confers additional advantages. In this section, we focus on the decomposition of the ML estimate of the model and refer to Cowell *et al.* (1999) for a more comprehensive treatment of this topic.

A fundamental property of decomposable graphs is that they can be recursively decomposed into their cliques, which are complete subsets.
This recursive decomposition into the cliques of the graph implies that decomposable graphical models can themselves be recursively decomposed into saturated models for subsets of variables. As shown in Section 4.7, if $G$ is a decomposable graph then its cliques can be numbered to form a perfect sequence $(C_1, \ldots, C_k)$. For every $r = 2, \ldots, k$ the pair of subsets $C_1 \cup \ldots \cup C_{r-1}$ and $C_r$ forms a decomposition of the subgraph $G_{C_1 \cup \ldots \cup C_r}$, giving the sequence of separators $(S_2, \ldots, S_k)$. The repeated application of the factorization (4.16) to this sequence of decompositions produces a factorization of the probability mass function of $Y$ into terms relative to the marginal distributions for cliques and separators,

$$p(Y = i) = \frac{\prod_{r=1}^k p(Y_{C_r} = i_{C_r})}{\prod_{r=2}^k p(Y_{S_r} = i_{S_r})}.$$

Accordingly, the repeated application of (4.17) gives the formula

$$\hat{m}(Y = i) = \frac{\prod_{r=1}^k \hat{m}(Y_{C_r} = i_{C_r})}{\prod_{r=2}^k \hat{m}(Y_{S_r} = i_{S_r})}.$$

Because the set of the cliques of $G$ are the generating class of the model, then it follows from (4.14) that the ML estimates of the expected cell counts can be computed from the minimal sufficient statistics as follows (see also Lauritzen, 1996, Proposition 4.18).

**Proposition 4.12** Let $\{\hat{m}(Y = i)\}_{i \in \mathcal{I}}$ be the ML estimates of the expected cell counts in the undirected graph model $M(G)$. If $G$ is decomposable then it holds that

$$\hat{m}(Y = i) = \frac{\prod_{r=1}^k n(Y_{C_r} = i_{C_r})}{\prod_{r=2}^k n(Y_{S_r} = i_{S_r})}$$

where $\{C_1, \ldots, C_k\}$ is the set of cliques of $G$ and $\{S_2, \ldots, S_k\}$ the set of separators.

Proposition 4.12 specifies the result of Theorem 4.8 to the decomposable case and shows that there exists an explicit formula for the ML estimate of a decomposable model. When applied to a decomposable model, the IPF will converge to the ML estimate after a finite number of steps, as this is the case for the application of the IPF to the infant survival data in Section 4.6.
Consider now the comparison of two nested decomposable models through the deviance statistic. As stated in Proposition 4.6, there exists a path of decomposable models from the larger to the smaller model, formed by removing one edge at the time. Hence, we can restrict the attention to the comparison of the decomposable models with graphs $G^1 = (V, E^1)$ and $G^0 = (V, E^0)$ where $G^1$ has exactly one more edge than $G^0$. To see an example of this framework, assume that $G^1$ is the graph in Figure 4.5 and that $G^0$ is obtained from the latter by removing the edge $c \sim e$. As stated in Proposition 4.7, the edge removed from $G$ belongs to a unique clique $C$ of $G$ and to no separator. As a consequence, the comparison of the two models can be localized on the marginal distribution of $Y_C$ so that

$$\text{dev}(G^0|G^1) = \text{dev}(G^0_C|G^1_C).$$ (4.23)

The local computation of the deviance statistic in (4.23) follows immediately from the factorization of the two probability mass functions with respect to the cliques and separators of $G^1$, but a formal proof can be found in Dawid and Lauritzen (1993, section 6). Consider the example with the graph $G^1$ in Figure 4.5. The removed edge $c \sim e$ belongs to the clique $\{c, d, e\}$ and, thus, the comparison of the two models can be localized on the marginal distribution of $Y_{\{c, d, e\}}$ to obtain

$$\text{dev}(G^0|G^1) = \text{dev}(G^0_{\{c, d, e\}}|G^1_{\{c, d, e\}})$$

$$= 2 \sum_{i_{\{c, d, e\}} \in \mathcal{I}_{\{c, d, e\}}} n(Y_{\{c, d, e\}} = i_{\{c, d, e\}}) \log \frac{\hat{m}_1(Y_{\{c, d, e\}} = i_{\{c, d, e\}})}{\hat{m}_0(Y_{\{c, d, e\}} = i_{\{c, d, e\}})}.$$

However, if one notices that the graph in Figure 4.5 has cliques $\{\{a, b\}, \{b, c\}, \{c, d, e\}, \{d, e, f\}, \{d, e, g\}\}$ and separators $\{\{b\}, \{c\}, \{d, e\}\}$, and that the graph $G^0$ has cliques $\{\{a, b\}, \{b, c\}, \{c, d\}, \{d, e, f\}, \{d, e, g\}\}$ and separators $\{\{b\}, \{c\}, \{d\}, \{d, e\}\}$, then, by exploiting the factorization of the ML estimate of Proposition 4.12, the above deviance can expressed in terms of marginal cell counts

$$\text{dev}(G^0|G^1) = 2 \sum_{i_{\{c, d, e\}} \in \mathcal{I}_{\{c, d, e\}}} n(Y_{\{c, d, e\}} = i_{\{c, d, e\}})$$

$$\times \log \frac{n(Y_{\{c, d\}} = i_{\{c, d\}})n(Y_{\{d\}} = i_{\{d\}})}{n(Y_{\{c, d\}} = i_{\{c, d\}})n(Y_{\{d, e\}} = i_{\{d, e\}})}.$$
More generally, if we denote by \( u \sim v \) the edge removed from \( G^1 \) to obtain \( G^0 \), and the removed edge belongs to the clique \( C \) of \( G^1 \), then the deviance statistic for comparing the corresponding decomposable models has form

\[
\text{dev}(G^0|G^1) = 2 \sum_{i_C \in IC} n(Y_C = i_C) \log \left( \frac{n(Y_C = i_C)n(Y_{C \setminus \{u,v\}} = i_C \setminus \{u,v\})}{n(Y_C \setminus \{u\})n(Y_{C \setminus \{v\}} = i_C \setminus \{v\})} \right).
\]

As well as the deviance statistic, other inferential quantities can be locally computed in a similar fashion on saturated marginal models. For instance, Dawid and Lauritzen (1993) gave the local computation of the Bayes factor for model comparison in Bayesian inference, whereas Roverato and Whittaker (1998) considered the local computation of the Laplace approximation to the Bayes factor.

We now turn to the log-linear parameterization of decomposable models. Similarly to Proposition 4.12, which reproduces the result of Theorem 4.8 in the special case of decomposable model, we give below a proposition that specifies the results of Theorem 4.9 and Corollary 4.11. For notational convenience, this is presented for the binary case; the generalization to the polytomous case is straightforward.

**Proposition 4.13**  Let \( \hat{\lambda} \) be the ML estimate, under \( M(G) \), of the log-linear parameter of the distribution of the binary vector \( Y_V \), with probability parameter \( \pi > 0 \). If \( G \) is decomposable with cliques \( \{C_1, \ldots, C_k\} \) and separators \( \{S_2, \ldots, S_k\} \) then

\[
\lambda_D = \sum_{r=1}^{k} \lambda_{CD}^r - \sum_{r=2}^{k} \lambda_{SD}^r \quad \text{and} \quad \hat{\lambda}_D = \sum_{r=1}^{k} \hat{\lambda}_{CD}^r - \sum_{r=2}^{k} \hat{\lambda}_{SD}^r
\]

where for every \( r = 1, \ldots, k \), \( \lambda_{CD}^r \) is the log-linear parameter of the distributions of \( Y_C \), and \( \hat{\lambda}_{CD}^r \) its ML estimate in the saturated model, and similarly for the separators. Here, we use the convention that \( \lambda_{CD}^r = 0 \) whenever \( D \subseteq C \), and similarly for \( \lambda_{SD}^r \).

**Proof.** This follows from the recursive application of Theorem 4.9 and equation (4.22) with respect to a perfect sequence of cliques of \( G \).
4.10 Log-linear Models and the Exponential Family

The exponential family is a convenient and widely used unified family of distributions whose members share many important properties. Hierarchical log-linear models belong to the exponential family (see Geiger and Meek, 1998; Koski and Noble, 2009; Lauritzen, 1996, p. 76) and in the next section we will exploit some general properties of this family of distributions in the asymptotic analysis of log-linear models. Much of the power of the exponential family formalism derives from the insights that are obtained from considering different parameterizations for a given family and, in this context, a central role is played by the connection between the canonical parameter and the mean parameter of the family. For notational convenience we restrict the attention to the binary case, and in this section we provide an overview of the alternative exponential family representations of the joint distribution of a vector of binary variables. Then, we derive an exponential family representation of log-linear models in which the canonical parameter is the collection of log-linear terms. In this way, we obtain a framework that naturally suits the subclass of graphical models, and thus the derivation of the results can be obtained by applying the properties of the regular exponential family together with the tools provided by the theory of Möbius inversion.

4.10.1 Basic Facts on the Theory of the Exponential Family

We first review the exponential family theory as required for this text, and refer to Barndorff-Nielsen (1978) for a full account on exponential families; see also Gutiérrez-Peña and Smith (1997) and Barndorff-Nielsen (2014).

Let $Y$ be a random vector whose probability distribution belongs to a regular exponential family, then the density function of $Y$, with respect to an appropriate $\sigma$-finite measure, can be written in the form

$$f(y ; \theta) = b(y)\exp\{\theta^T t(y) - M(\theta)\}$$

for some nonnegative measurable function $b(\cdot)$. The vectors $\theta \in \Theta$ and $T = t(Y)$ are the canonical parameter and the canonical statistic of the family, respectively, and $M(\theta)$ is known as the cumulant function. The mean parameter of the family is

$$\mu = \mu(\theta) = E_\theta(T) = \frac{\delta M(\theta)}{\delta \theta},$$
whereas the function

\[ V(\mu) = \text{var}_\theta(T) = \frac{\delta^2 M(\theta)}{\delta \theta \delta \theta^T} = \frac{\delta \mu(\theta)}{\delta \theta^T} \]

defined for \( \mu \in \mu(\Theta) \) is called the variance function of the family; see Davison (2003, pp. 169 and 170) and Gutiérrez-Peña and Smith (1997, sections 2.1 and 2.2). The importance of the variance functions stems from the fact that, together with its domain \( \mu(\Theta) \), it characterizes the family within the class of natural exponential families (Morris, 1982, 1983).

A general result for the regular exponential family (Barndorff-Nielsen, 1978, p. 150) is that the asymptotic variance of the ML estimate of the canonical parameter can be obtained as the inverse of the variance of the canonical statistic,

\[ \text{avar}(\hat{\theta}) = \text{var}_\theta(T)^{-1} = V(\mu)^{-1} = \frac{\delta \theta(\mu)}{\delta \mu^T}, \]

where \( \theta(\cdot) \) denotes the inverse of the mapping \( \mu(\cdot) \); see also Davison (2003, p. 171) and Gutiérrez-Peña and Smith (1997, section 2.2.1).

4.10.2 The Cross-classified Bernoulli Distribution

A variable that may take only the values zero and one is said to follow a Bernoulli distribution. When dealing with the binary case we have assumed that the state space of \( Y_v \) is given by \( \{0, 1\} \), and in the literature the distribution of this random vector is known as the multivariate or cross-classified Bernoulli. Here we introduce some useful quantities and recall a common way, based on the entries of \( \pi \), to write the probability function of a cross-classified Bernoulli distribution. This will be elaborated in the forthcoming subsections to obtain the required exponential family representations.

For a set \( D \subseteq V \) consider the following two Bernoulli random variables,

\[ \tilde{Y}_D = \prod_{v \in D} Y_v \quad \text{and} \quad \tilde{Y}_D^c = \prod_{v \in D} (1 - Y_v), \]

(4.24)

with the convention that \( \tilde{Y}_\emptyset = \tilde{Y}_\emptyset^c = 1 \); see also Consonni and Leucari (2006). We remark that the quantity \( \tilde{Y}_D \) depends on \( Y \) only through the subvector \( Y_D \), and similarly for \( \tilde{Y}_D^c \). Notice also that, for any pair of
subsets $D, E \subseteq V$, it holds that both $\bar{Y}_{D \cup E} = \bar{Y}_D \times \bar{Y}_E$ and $\bar{Y}_{D \cup E}^c = \bar{Y}_D^c \times \bar{Y}_E^c$. We now introduce the further Bernoulli random variable

$$U_{D}^V = \bar{Y}_D \times \bar{Y}_E^c$$

so that, in summary, we have introduced three vectors of binary variables, namely $\bar{Y} = (\bar{Y}_D)_{D \subseteq V}$, $\bar{Y}^c = (\bar{Y}_D^c)_{D \subseteq V}$ and $U = U^V = (U_{D}^V)_{D \subseteq V}$. It is clear from (4.25) that $U_{D}^V = 1$ if and only if both $\bar{Y}_D = 1$ and $\bar{Y}_E^c = 1$, and therefore if and only if the observed value of $Y$ belongs to the cell indexed by $D$. This implies that $p(U_D = 1) = \pi_D$, and therefore $E(U_D) = \pi_D$ or, more generally, that $E(U) = \pi$. Furthermore, the vector $U$ can be used to write the density mass function of $Y$ as

$$p(Y_D = 1, Y_{V \setminus D} = 0) = \prod_{E \subseteq V} \pi^u_E \text{ for every } D \subseteq V,$$

where the entry $\pi_\emptyset$ is determined by the affine constraint

$$\pi_\emptyset = 1 - \sum_{D \subseteq V : D \neq \emptyset} \pi_D.$$

### 4.10.3 Exponential Family Representations of the Saturated Model

When the parameter $\pi$ is unrestricted, apart from the sum-to-one constraint (4.27), the cross-classified Bernoulli distribution belongs to the regular exponential family, and its density mass function is typically represented in two alternative ways. These are given below, and we refer to Brown (1986, p. 4), Lauritzen (1996, p. 76), Gutiérrez-Peña and Smith (1997, sections 2.1 and 2.2) and Davison (2003, p. 169) for a more detailed account on this issue.

The formulation of the probability distribution of $Y$ in (4.26) leads to a first possible exponential family representation of the cross-classified Bernoulli,

$$p(Y_D = 1, Y_{V \setminus D} = 0) = \exp \left( \sum_{E \subseteq V} u_E \log \pi_E \right) \text{ for every } D \subseteq V,$$

where $U$ and $\log \pi$ are the canonical statistic and parameter respectively. This can be found, for instance, in Brown (1986, p. 4) and Lauritzen
However, this exponential family representation is not full because of the affine constraint (4.27); see Brown (1986, p. 4). A full regular exponential family representation is typically obtained by setting \( \phi_D = \log \pi_D - \log \pi_\emptyset \), for all \( D \subseteq V \), and writing

\[
p(Y_D = 1, Y_{V \setminus D} = 0) = \exp \left\{ \sum_{E \subseteq V} \phi_E u_E - M(\phi) \right\}
\]  

(4.29)

where

\[
M(\phi) = -\log \pi_\emptyset = \log \left\{ 1 + \sum_{D \subseteq V; D \neq \emptyset} \exp(\phi_D) \right\}
\]

(see, e.g., Brown (1986, p. 5) and Davison (2003, p. 175). Notice that in (4.29) the canonical parameter is \( \phi_\emptyset \) whereas \( U_\emptyset \) is the canonical statistic because \( \phi_\emptyset = 0 \) and, accordingly, the mean parameter is \( \pi_\emptyset = E(U_\emptyset) \). Recall that we are using the notation introduced in Section 3.1.1 so that, for instance, \( \pi_\emptyset \) is the subvector obtained from \( \pi \) by removing the entry indexed by the empty set. For the computation of the variance function \( V(\pi_\emptyset) \), and of the asymptotic variance of the ML estimate of the canonical parameter \( \text{avar}(\hat{\phi}_\emptyset) = V(\pi_\emptyset)^{-1} \), it is convenient to introduce the matrix \( \mathbb{P} \) that can be regarded as a bivariate representation of the probability parameter \( \pi \). Formally, for a nonempty subset \( A \subseteq V \) we set \( \mathbb{P}^A = \text{diag}(\pi^A) \) so that \( \mathbb{P}^A \) is the matrix with zero off-diagonal entries and diagonal entries given by the probabilities

\[
\mathbb{P}^A_{D, D} = \pi^A_D = p(Y_D = 1, Y_{A \setminus D} = 0) \quad \text{for every } D \subseteq A,
\]

and it is straightforward to see that \( |\mathbb{P}^A| = \prod_{D \subseteq A} \pi^A_D \). According with the notation introduced in Section 3.1.1, \( \mathbb{P}^A_\emptyset \) is the submatrix of \( \mathbb{P}^A \) obtained by removing the row and the column indexed by the empty set; namely, \( \mathbb{P}^A_\emptyset = \text{diag}(\pi^A_\emptyset) \). Furthermore, we adopt here the usual convention that we may omit the superscript when it is equal to \( V \), so that \( \mathbb{P} = \mathbb{P}^V \).

The variance function for the exponential family representation (4.29) and its determinant are

\[
V(\pi_\emptyset) = \text{var}(U_\emptyset) = \mathbb{P}_\emptyset - \pi_\emptyset \pi^T_\emptyset \quad \text{and} \quad |V(\pi_\emptyset)| = \prod_{D \subseteq V} \pi_D, \quad (4.30)
\]
The asymptotic variance of $\hat{\phi}_e$ and its determinant are therefore

$$\text{avar}(\hat{\phi}_e) = V(\pi_e)^{-1} = \mathbb{P}_e^{-1} + \frac{1}{\pi_\emptyset} 1_e 1_e^T$$

(4.31)

and

$$|\text{avar}(\hat{\phi}_e)| = |\mathbb{P}^{-1}| = \prod_{D \subseteq \mathcal{V}} \frac{1}{\pi_D},$$

(4.31a)

where $1_e$ is the vector indexed by the nonempty subsets of $\mathcal{V}$ with all entries equal to one. The computations of the determinant in (4.30), as well as of the inverse matrix in (4.31) and of its determinant, are based on standard results of matrix algebra, which can be found in Graybill (1983, theorems 8.3.3 and 8.4.3). Notice that $|\text{avar}(\hat{\phi}_e)| = |\mathbb{P}^{-1}|$, although $\text{avar}(\hat{\phi}_e) \neq \mathbb{P}^{-1}$.

To facilitate the connection between the different parameterizations and statistics given in this section, Table 4.3 details the relevant quantities involved in equations (4.26), (4.28) and (4.29) for the case $\mathcal{V} = \{a, b, c\}$.

### 4.10.4 Exponential Family Representation of Hierarchical Log-linear Models

The parametric restrictions implied by a log-linear model do not correspond to the vanishing of some elements of either log $\pi$ or $\phi$ and, consequently, for an arbitrary hierarchical log-linear model the representations (4.28) and (4.29) are not full. It is thus of interest to give an exponential family representation of hierarchical log-linear models in which the canonical parameter is the collection of log-linear terms. This can be obtained by exploiting the fact that there exists a linear relationship between $\tilde{Y}$ and $U$, which can be written in terms of a Möbius inversion. The interested reader may refer to Gutiérrez-Peña and Smith (1997) for a review on the role of linear transformations in regular exponential families.

**Lemma 4.14** Let $\tilde{Y} = (\tilde{Y}_D)_{D \subseteq \mathcal{V}}$ and $U = (U_D)_{D \subseteq \mathcal{V}}$ the vectors defined in (4.24) and (4.25), respectively. Then for all $D \subseteq \mathcal{V}$,

$$\tilde{Y}_D = \sum_{E \subseteq \mathcal{V} \setminus D} U_{D \cup E} \quad \text{and} \quad U_D = \sum_{E \subseteq \mathcal{V} \setminus D} (-1)^{|E|} \tilde{Y}_{D \cup E}$$
Table 4.3 Quantities involved in equations (4.26), (4.28) and (4.29) for $V = \{a, b, c\}$.

<table>
<thead>
<tr>
<th>$D$</th>
<th>$\tilde{y}_D$</th>
<th>$\tilde{y}_{V\setminus D}$</th>
<th>$u_D$</th>
<th>$\pi_D$</th>
<th>$\phi_D$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\emptyset$</td>
<td>1</td>
<td>$(1 - y_a)(1 - y_b)(1 - y_c)$</td>
<td>$(1 - y_a)(1 - y_b)(1 - y_c)$</td>
<td>$p(0,0,0)$</td>
<td>0</td>
</tr>
<tr>
<td>${a}$</td>
<td>$y_a$</td>
<td>$(1 - y_b)(1 - y_c)$</td>
<td>$y_a(1 - y_b)(1 - y_c)$</td>
<td>$p(1,0,0)$</td>
<td>log $p(1,0,0)$ (p(0,0,0))</td>
</tr>
<tr>
<td>${b}$</td>
<td>$y_b$</td>
<td>$(1 - y_a)(1 - y_c)$</td>
<td>$(1 - y_a)y_b(1 - y_c)$</td>
<td>$p(0,1,0)$</td>
<td>log $p(0,1,0)$ (p(0,0,0))</td>
</tr>
<tr>
<td>${c}$</td>
<td>$y_c$</td>
<td>$(1 - y_a)(1 - y_b)$</td>
<td>$(1 - y_a)(1 - y_b)y_c$</td>
<td>$p(0,0,1)$</td>
<td>log $p(0,0,1)$ (p(0,0,0))</td>
</tr>
<tr>
<td>${a, b}$</td>
<td>$y_ay_b$</td>
<td>$(1 - y_c)$</td>
<td>$y_ay_b(1 - y_c)$</td>
<td>$p(1,1,0)$</td>
<td>log $p(1,1,0)$ (p(0,0,0))</td>
</tr>
<tr>
<td>${a, c}$</td>
<td>$y_ay_c$</td>
<td>$(1 - y_b)$</td>
<td>$y_a(1 - y_b)y_c$</td>
<td>$p(1,0,1)$</td>
<td>log $p(1,0,1)$ (p(0,0,0))</td>
</tr>
<tr>
<td>${b, c}$</td>
<td>$y_by_c$</td>
<td>$(1 - y_a)$</td>
<td>$(1 - y_a)y_by_c$</td>
<td>$p(0,1,1)$</td>
<td>log $p(0,1,1)$ (p(0,0,0))</td>
</tr>
<tr>
<td>${a, b, c}$</td>
<td>$y_ay_by_c$</td>
<td>1</td>
<td>$y_ay_by_c$</td>
<td>$p(1,1,1)$</td>
<td>log $p(1,1,1)$ (p(0,0,0))</td>
</tr>
</tbody>
</table>
that in matrix form can be written as
\[ \widetilde{Y} = ZU \text{ and } U = M\widetilde{Y}. \]

Proof. We first show that \( \widetilde{Y}_D^c = \sum_{E \subseteq D} (-1)^{|E|} \widetilde{Y}_E \) for all \( D \subseteq V \). This part of the proof is by induction on the cardinality of \( D \). For \( |D| = 0 \) the result is trivially true because \( D = \emptyset \) and \( \widetilde{Y}_\emptyset = \widetilde{Y}_\emptyset = 1 \). We now show that if the result is true for a subset \( D \subset V \) with \( 0 \leq |D| < |V| \) then it is also true for \( D \cup \{v\} \) where \( v \in V \setminus D \). It follows immediately from the definition of \( \widetilde{Y}^c \) that \( \widetilde{Y}^c_{D \cup \{v\}} = \widetilde{Y}^c_D \times (1 - Y_v) \). Hence,

\[
\begin{align*}
\widetilde{Y}^c_{D \cup \{v\}} &= \widetilde{Y}^c_D \times (1 - Y_v) \\
&= \widetilde{Y}^c_D - Y_v \times \widetilde{Y}^c_D \\
&= \sum_{E \subseteq D} (-1)^{|E|} \widetilde{Y}_E - Y_v \times \sum_{E \subseteq D} (-1)^{|E|} \widetilde{Y}_E \\
&= \sum_{E \subseteq D} (-1)^{|E|} \widetilde{Y}_E + \sum_{E \subseteq D} (-1)^{|E \cup \{v\}|} \widetilde{Y}_{E \cup \{v\}} \\
&= \sum_{E \subseteq D \cup \{v\}; v \notin E} (-1)^{|E|} \widetilde{Y}_E + \sum_{E \subseteq D \cup \{v\}; v \in E} (-1)^{|E|} \widetilde{Y}_E \\
&= \sum_{E \subseteq D \cup \{v\}} (-1)^{|E|} \widetilde{Y}_E
\end{align*}
\] (4.32)

where in (4.32) we have applied the induction assumption, whereas in (4.33) we have used the facts that \( (-1) \times (-1)^{|E|} = (-1)^{|E \cup \{v\}|} \) because \( v \notin E \) and that \( Y_v \times \widetilde{Y}_E = \widetilde{Y}_{E \cup \{v\}} \).

By definition, \( U_D = \widetilde{Y}_D \widetilde{Y}_{V \setminus D} \) and by applying (4.34) we obtain

\[
U_D = \widetilde{Y}_D \sum_{E \subseteq V \setminus D} (-1)^{|E|} \widetilde{Y}_E = \sum_{E \subseteq V \setminus D} (-1)^{|E|} \widetilde{Y}_{D \cup E}
\]

which is one of the four stated equalities. The remaining three equalities follow immediately from Corollary 3.6.

The linear transformation in Lemma 4.14 can be used in (4.28) to transform the canonical parameter log \( \pi \) into \( \lambda \). This is achieved by exploiting the fact that \( MZ \) is the identity matrix so that

\[
u^T \log \pi = u^T Z^TM^T \log \pi = (Zu)^T (M^T \log \pi) = \tilde{y}^T \lambda.\]
Proposition 4.15  If the distribution of the cross-classified Bernoulli random vector $Y_V$ belongs to the hierarchical log-linear model $M(C)$, then

$$p(Y_D = 1, Y_{V \setminus D} = 0) = \exp \left\{ \sum_{E \in C} \lambda_E \tilde{y}_E - M(\lambda) \right\}$$  \hspace{1cm} (4.35)$$

where $M(\lambda) = -\lambda_\emptyset = -\log \pi_\emptyset$.

Proof. By recalling that $Y = ZU$ by Lemma 4.14, that $\lambda = M^T \log \pi$ and that $Z^T M^T$ is the identity matrix, it follows from (4.26) that

$$\log p(Y_D = 1, Y_{V \setminus D} = 0) = \sum_{E \subseteq V} u_E \log \pi_E$$

$$= u^T \log \pi$$

$$= u^T Z^T M^T \log \pi$$

$$= \tilde{y}^T \lambda$$

$$= \sum_{E \subseteq V} \tilde{y}_E \lambda_E.$$  \hspace{1cm} (4.36)$$

As $\lambda_E$ is equal to zero whenever $E \notin C$, then (4.36) simplifies to

$$\log p(Y_D = 1, Y_{V \setminus D} = 0) = \sum_{E \in C} \tilde{y}_E \lambda_E = \sum_{E \in \mathcal{C}_e} \lambda_E \tilde{y}_E + \lambda_\emptyset$$

because $\tilde{y}_\emptyset = 1$, and this completes the proof. $\square$

The canonical statistic of the exponential family in (4.35) is $\tilde{Y}_e$ and we now turn to the mean parameter $E(\tilde{Y}_e)$. Let $\mu = (\mu_D)_{D \subseteq V}$ be the vector with entries $\mu_D = p(Y_D = 1)$ for $D \subseteq V$ with $D \neq \emptyset$ and $\mu_D = 1$ for $D = \emptyset$. Clearly, $E(\tilde{Y}_D) = \mu_D$ so that $\mu_{\mathcal{C}_e}$ is the mean parameter for the exponential family representation in (4.35). There is a close connection between $\pi$ and $\mu$. Indeed, it is easy to see that $\mu_D = \pi_D$ for any $D \neq \emptyset$, but, more interestingly, $\pi$ and $\mu$ are connected through a Möbius inversion formula.

Proposition 4.16  Let $\pi = (\pi_D)_{D \subseteq V}$ be the probability parameter of a binary random vector $Y_V$ and let $\mu = (\mu_D)_{D \subseteq V}$ be such that $\mu_D = p(Y_D = 1)$ for $D \subseteq V$ with $D \neq \emptyset$ and $\mu_D = 1$ for $D = \emptyset$. Then, for all $D \subseteq V$
\[
\mu_D = \sum_{E \subseteq V \setminus D} \pi_{D \cup E} \quad \text{and} \quad \pi_D = \sum_{E \subseteq V \setminus D} (-1)^{|E|} \mu_{D \cup E},
\]

that in matrix form can be written as

\[
\mu = Z \pi \quad \text{and} \quad \pi = M \mu.
\]

**Proof.** This proposition is an immediate consequence of Lemma 4.14. Each of the four equalities stated in the proposition can be obtained from the corresponding equalities in Lemma 4.14 by taking the expectation of both sides of the identity. \(\Box\)

### 4.11 Modular Structure of the Asymptotic Variance of ML Estimates

Initially, local computation properties and modularity of graphical models were developed for probability distributions satisfying the Markov property with respect to an undirected graph. Subsequently, for the decomposable case, Dawid and Lauritzen (1993) introduced the *hyper Markov properties* for probability distributions defined at a *hyper* level, that is over the set of probability distributions forming a graphical model. They considered, for instance, the distribution of the ML estimates for the parameters of a graphical model as well as prior and posteriors distributions.

In this section, we consider the binary case and describe the asymptotic theory of hierarchical log-linear models in a formulation that allows one to fully exploit the specific features of graphical models. For hierarchical log-linear models this approach provides an alternative way to derive certain classical results of asymptotic theory given, for example, in Christensen (1997, section 12.3) and Agresti (2013, chapter 16). For the subclass of undirected graph models this approach allows us to describe the asymptotic theory of these models by giving explicit rules for the local computation of the asymptotic variance of ML estimates of the log-linear parameters.

Although our main focus is on the asymptotic distribution of the ML estimates, and more specifically on the asymptotic variance of \(\hat{\lambda}\), it is worth recalling that the material presented here can be applied to obtain a modular representation of other relevant inferential quantities. In a Bayesian approach to inference, the Bayes factor (Jeffreys, 1961) is a
fundamental tool for comparing competing hypotheses. However, the Bayes factor can be locally computed, in a similar fashion as the deviance difference in Section 4.5, only in the special case where the prior distribution fulfills the strong hyper Markov property (Dawid and Lauritzen, 1993). A widely used asymptotic version of the Bayes factor is its Laplace approximation (Kass and Raftery, 1995), and in this case local computations are possible also under a less stringent requirement for the prior distribution, known as the weak hyper Markov property (Dawid and Lauritzen, 1993; Roverato and Whittaker, 1998). On the other hand, the local computation of the Laplace approximation to the Bayes factor is based on a modular representation of the determinant of the observed information matrix. Because log-linear models are regular exponential families, the observed information matrix corresponds to the expected information matrix evaluated at $\hat{\lambda}$. Furthermore, it can be obtained as the inverse of the asymptotic variance of the ML estimate of $\lambda$ for a single observation, and therefore the modular representation of the determinant of the information matrix can be obtained by direct application of the material described here. This modular representation of the determinant is also useful when the interest is in the noninformative Jeffreys prior (Jeffreys, 1946) because it leads to a factorization of this quantity in a way similar to that of hyper Markov laws.

In the rest of this section we assume, without loss of generality, that the sample size is equal to one, so that the above asymptotic variance coincides with the inverse of the expected Fisher information matrix for the canonical parameter.

4.11.1 The Variance Function and the Asymptotic Variance of ML Estimates

This section is devoted to the variance function $V(\mu_{C_{n}})$ and the asymptotic variance of $\hat{\lambda}_{C_{n}}$. Special attention is paid to the description of the interesting mathematical relationship existing between these two quantities.

We have considered three alternative parameterizations of the cross-classified Bernoulli distribution, namely, $\mu_{n}, \lambda_{n}$ and $\pi_{n}$. There is a one-to-one relationship between every pair of these parameters. However, in the exponential family theory a key role is played by the functional relationship between the mean parameter $\mu_{n}$ and the canonical
parameter $\lambda_e$ because this determines the structure both of the variance function $V(\mu_e)$ and of the asymptotic variance of $\hat{\lambda}_e$, which is a quantity of fundamental importance in statistical inference. The theory developed in the previous sections allows us to split the functional relationship between $\mu_e$ and $\lambda_e$ into easily manageable building blocks. This is achieved by considering the vectors $\mu$ and $\lambda$ and using equation (4.1) and Propositions 4.16 to obtain $\lambda$ from $\mu$ by means of a sequence of remarkably simple bijective functions, mainly involving Möbius inversion operations. The probability parameter $\pi$ turns out to be the intermediate step of this decomposition,

$$
\mu \xrightarrow{\text{Möb. inv.}} \pi \xrightarrow{\log} \lambda.
$$

Each of the vectors involved in (4.37) is indexed by the subsets of $V$, i.e., by the elements of $\mathcal{P}(V)$. We now turn to matrices indexed by the elements of $\mathcal{P}(V) \times \mathcal{P}(V)$. More specifically, we consider the matrix $\mathbb{P}$ introduced in Section 4.10.3 and the matrix $\mathbb{V}$ with entries

$$
\mathbb{V}_{D,D'} = \mu_{D \cup D'} \quad \text{for all } D, D' \subseteq V.
$$

Notice that the main diagonal of $\mathbb{P}$ is $\pi$ and the main diagonal of $\mathbb{V}$ is $\mu$. We show below that the inverse $\mathbb{V}^{-1}$ can be computed by means of a sequence of simple bijective functions that closely resembles (4.37), namely,

$$
\mathbb{V} \xrightarrow{\text{Möb. inv.}} \mathbb{P} \xrightarrow{\text{Möb. inv.}} \mathbb{P}^{-1} \xrightarrow{\text{Möb. inv.}} \mathbb{V}^{-1},
$$

where the Möbius inversion formula used here is that of Corollary 3.7.

There is an interesting parallelism between (4.37) and (4.38): a dimensionality reduction of $\mu$ leads to the mean parameter $\mu_e$, and Proposition 4.19 shows that a dimensionality reduction of $\mathbb{V}$ leads to the variance function $V(\mu_e)$. On the other hand, a dimensionality reduction of $\lambda$ leads to the canonical parameter $\lambda_e$, and Theorem 4.21 shows that a dimensionality reduction of $\mathbb{V}^{-1}$ leads to $\text{avar}(\hat{\lambda}_e)$, the asymptotic variance of $\hat{\lambda}_e$. The decomposition of the functional relationship between $\mathbb{V}$ and its inverse given in (4.38) is of theoretical interest as well as of practical usefulness because, by linking both $\mathbb{V}$ and $\mathbb{V}^{-1}$ with the diagonal matrix $\mathbb{P}$, it allows to derive an explicit formulations of the entries and of the determinant of both $V(\mu_e)$ and $\text{avar}(\hat{\lambda}_e)$.

As shown in Proposition 4.16, $\mu$ can be obtained from $\pi$ by applying the Möbius inversion ($a'$) of Corollary 3.6 with respect to $\mathcal{P}(V)$. As well
as $\mathbb{P}$ can be regarded as a two-dimensional representation of $\pi$, the matrix $V$ can be regarded as a two-dimensional representation of the vector $\mu$ because $V$ can be obtained from $\mathbb{P}$ by applying the same Möbius inversion operation with respect to $P(V) \times P(V)$.

Proposition 4.17  
The functional relationship between the matrices $V$ and $\mathbb{P}$ is as follows

$$V = \mathbb{Z} \mathbb{P} \mathbb{Z}^T$$  
and  
$$\mathbb{P} = MV M^T.$$  

Proof.  
By point (a″) of Corollary 3.7 it follows that the equalities $V = \mathbb{Z} \mathbb{P} \mathbb{Z}^T$ is satisfied if and only if

$$V_{D,D'} = \sum_{(E,E') \geq (D,D')} \mathbb{M}_{E,E'}$$  
for every $D, D' \subseteq V$  
(4.39)

and we now prove that (4.39) holds. By definition, $P_{E,E'} = \pi_E$ whenever $E = E'$, and it is equal to zero otherwise; hence, for every $D, D' \subseteq V$ it holds that

$$\sum_{(E,E') \geq (D,D')} P_{E,E'} = \sum_{(E,E') \geq (D,D')} \pi_E.$$  

Furthermore, by noticing that $(E,E') \geq (D,D')$ if and only if $E \geq D \cup D'$, and therefore if and only if $D \cup D' \subseteq E$, for every $D, D' \subseteq V$ we can write

$$\sum_{(E,E') \geq (D,D')} \pi_E = \sum_{F \subseteq V \setminus (D \cup D')} \pi_{D \cup D' \cup F} = \mu_{D \cup D'}$$

where the last identity follows from Proposition 4.16. This proves (4.39) because $V_{D,D'} = \mu_{D \cup D'}$ by definition.

The equality $\mathbb{P} = MV M^T$ follows from point (A″) of Corollary 3.7. 

We have thus proved the first Möbius inversion in (4.38) that, interestingly, coincides with the Cholesky decomposition of $V$. The second Möbius inversion is an immediate consequence of Proposition 4.17 and coincides with the Cholesky decomposition of $V^{-1}$.

Corollary 4.18  
The functional relationship between the matrices $V^{-1}$ and $\mathbb{P}^{-1}$ is as follows

$$V^{-1} = M^T \mathbb{P}^{-1} M$$  
and  
$$\mathbb{P}^{-1} = \mathbb{Z}^T V^{-1} \mathbb{Z}.$$
Proof. This is an immediate consequence of Proposition 4.17 because \( \mathbf{V} = \mathbf{ZPZ}^T \) implies \( \mathbf{V}^{-1} = (\mathbf{ZPZ}^T)^{-1} = \mathbf{M}^T \mathbf{P}^{-1} \mathbf{M} \), and similarly for the second identity. \( \square \)

We have thus shown the validity of (4.38) and in the following we describe the connection between \( \mathbf{V} \) and the variance function \( V(\mu_e) \) and between \( \mathbf{V}^{-1} \) and the asymptotic variance of \( \hat{\lambda}_e \).

### 4.11.2 Variances in the Saturated Model

The computation of the variance function and of the asymptotic variance of the ML estimate requires the notion of partial matrix. A matrix \( \mathbf{G} \) indexed by the subsets of \( V \) can be partitioned with respect to any nonempty subset \( \mathcal{C} \subseteq \mathcal{P}(V) \),

\[
\mathbf{G} = \begin{bmatrix}
\mathbf{G}_{\mathcal{C}\mathcal{C}} & \mathbf{G}_{\mathcal{C}\bar{\mathcal{C}}} \\
\mathbf{G}_{\bar{\mathcal{C}}\mathcal{C}} & \mathbf{G}_{\bar{\mathcal{C}}\bar{\mathcal{C}}}
\end{bmatrix}
\]

where \( \bar{\mathcal{C}} = \mathcal{P}(V) \setminus \mathcal{C} \). If \( \mathbf{G} \) has full rank then the partial matrix of \( \mathbf{G} \) with respect to \( \mathcal{C} \) is defined as

\[
\mathbf{G}_{\mathcal{C}\mathcal{C}} = \mathbf{G}_{\mathcal{C}\mathcal{C}} - \mathbf{G}_{\bar{\mathcal{C}}\mathcal{C}} \mathbf{G}_{\mathcal{C}\bar{\mathcal{C}}}^{-1} \mathbf{G}_{\bar{\mathcal{C}}\mathcal{C}}.
\]

The saturated model only involves partial matrices where \( \mathcal{C} = \mathcal{P}(V) \setminus \emptyset \) and we use the shorthand \( \mathbf{G}_{\mathcal{V}_e} \) to denote them.

**The Variance Function**

We can now give the variance function for the saturated model.

**Proposition 4.19** Let \( Y_V \) be a cross-classified Bernoulli random vector with probability parameter \( \pi > 0 \). If the probability distribution of \( Y_V \) belongs to the saturated model, then the variance function in the exponential family form (4.35) is

\[
V(\mu_e) = \mathbf{V}_{\mathcal{V}_e}
\]

and its entries are

\[
V(\mu_e)_{D,D'} = \mu_{D \cup D'} - \mu_D \mu_D'
\]

for all \( D, D' \subseteq V \) with \( D, D' \neq \emptyset \). Furthermore, the determinant of \( V(\mu_e) \) is given by
\[ |V(\mu_e)| = |\mathcal{V}| = |\mathcal{P}| = \prod_{D \subseteq V} \pi_D. \]

**Proof.** We first compute the entries of the variance function. By definition, \( V(\mu_e) = \text{var}(Y_e) \) so that the entry of \( V(\mu_e) \) indexed by \( D, D' \subseteq V \) with \( D, D' \neq \emptyset \) can be computed as follows,

\[
V(\mu_e)_{D,D'} = \text{cov}(\tilde{Y}_D, \tilde{Y}_{D'})
= E(\tilde{Y}_D \tilde{Y}_{D'}) - E(\tilde{Y}_D)E(\tilde{Y}_{D'})
= E(\tilde{Y}_{D \cup D'}) - E(\tilde{Y}_D)E(\tilde{Y}_{D'})
= \mu_{D \cup D'} - \mu_D \mu_{D'}. \tag{4.40}
\]

We now show that \( V(\mu_e) = V|_e \). The first row and column of \( V \) are indexed by the empty set and, therefore, the matrix \( V \) can be written in a partitioned form as

\[
V = \begin{bmatrix}
\mathcal{V}_{\emptyset, \emptyset} & \mu_e^T \\
\mu_e & \mathcal{V}_e
\end{bmatrix}
\]

and, as \( \mathcal{V}_{\emptyset, \emptyset} = \mu_{\emptyset} = 1 \), it holds that \( \mathcal{V}|_e = \mathcal{V}_e - \mu_e \mu_e^T \). From the latter, it is easy to check that the entry of \( \mathcal{V}|_e \) indexed by \( D, D' \subseteq V \) with \( D, D' \neq \emptyset \) is equal to (4.40) and this implies that \( V(\mu_e) = \mathcal{V}|_e \). Finally, we turn to the determinant of \( V(\mu_e) \). The determinant of \( V \) can be factorized as \( |V| = |\mathcal{V}| |\mathcal{V}_{\emptyset, \emptyset} \) (see Lütkepol, 1996, p. 147, eqn. 4) so that \( |V| = |\mathcal{V}| \) because \( \mathcal{V}_{\emptyset, \emptyset} = 1 \). Hence, by recalling that \( |Z| = 1 \), we obtain

\[
|V(\mu_e)| = |\mathcal{V}| = |\mathcal{V}| = |Z\mathcal{P}Z^T| = |\mathcal{P}| = \prod_{D \subseteq V} \pi_D. \]

□

As a consequence of Proposition 4.19 it is also possible to give an alternative, very compact, proof of the determinants in (4.30) and (4.31a).

**Corollary 4.20** *The variance function of the exponential family in (4.29) has determinant \( |V(\pi_e)| = |V(\mu_e)| = \prod_{D \subseteq V} \pi_D \).*

**Proof.** By definition \( V(\pi_e) = \text{var}(U_e) \) and because \( U_e = M_e Y_e \) then

\[
|V(\pi_e)| = \text{var}(M_e Y_e) = |M_e| \text{var}(Y_e) M_e^T = |M_e| V(\mu_e) M_e^T
= |V(\mu_e)| = \prod_{D \subseteq V} \pi_D. \]

□
The Asymptotic Variance of the ML Estimate

The asymptotic variance of \( \hat{\lambda}_e \) can be computed as the inverse of the variance function, \( \text{avar}(\hat{\lambda}_e) = (V_{|e})^{-1} \). Consequently, by using the results for the inverse of a partitioned matrix it follows immediately that \( \text{avar}(\hat{\lambda}_e) \) is a submatrix of \( V^{-1} \). We show in Theorem 4.21 that this has two important implications for the asymptotic variance. Firstly, it relates this quantity to a triangular decomposition involving the matrix \( M \), which is upper triangular with unit diagonal, and the diagonal matrix \( P^{-1} \). Secondly, it allows one to exploit the Möbius inversion in Corollary 4.18 to obtain an explicit formulation of the entries of \( \text{avar}(\hat{\lambda}_e) \).

**Theorem 4.21**  
Let \( Y \) be a cross-classified Bernoulli random vector with probability parameter \( \pi > 0 \). Then the asymptotic variance of the ML estimate of \( \lambda_e \) in the saturated model is

\[
\text{avar}(\hat{\lambda}_e) = (M^T P^{-1} M)_e,  
\]  
and has entries

\[
\text{avar}(\hat{\lambda}_e)_{D, D'} = \sum_{E \subseteq D \cap D'} (-1)^{|D \backslash E| + |D' \backslash E|} \frac{1}{\pi_E} \]  

for all \( D, D' \subseteq V \) with \( D, D' \neq \emptyset \). Furthermore, its determinant is

\[
|\text{avar}(\hat{\lambda}_e)| = |P^{-1}| = \prod_{D \subseteq V} \frac{1}{\pi_D}.  
\]  

**Proof.** Equation (4.41) can be derived by recalling that \( \text{avar}(\hat{\lambda}_e) = V(\mu_e)^{-1} \) and then applying, in turn, Proposition 4.19, the rules for the inverse of a partitioned matrix (see Lütkepol, 1996, p. 29, eqn. 1) and, finally, Corollary 4.18 as follows

\[
\text{avar}(\hat{\lambda}_e) = V(\mu_e)^{-1} = (V_{|e})^{-1} = (V^{-1})_e = (M^T P^{-1} M)_e.  
\]

In order to show (4.42) it is sufficient to compute the entries of \( V^{-1} \) because (4.41) implies that \( \text{avar}(\hat{\lambda}_e)_{D, D'} = (V^{-1})_{D, D'} \) for every \( D, D' \subseteq V \) with \( V \neq \emptyset \). As \( (V^{-1})_e = M^T P^{-1} M \) by Corollary 4.18, we can apply Corollary 3.7 to write the entry of \( V^{-1} \) indexed by \( D, D' \subseteq V \) as
\[
(V^{-1})_{D,D'} = \sum_{(E,E') \in \mathcal{P}(V) \times \mathcal{P}(V)} m_{\mathcal{P}(V) \times \mathcal{P}(V)} \{(E,E'), (D,D')\} (P^{-1})_{E,E'}
\]

\[
= \sum_{E \subseteq V} m_{\mathcal{P}(V) \times \mathcal{P}(V)} \{(E,E), (D,D')\} \frac{1}{\pi_E}
\]  
(4.44)

\[
= \sum_{E \subseteq V} (-1)^{|D \setminus E| + |D' \setminus E|} \mathbb{1}(E \subseteq D') \mathbb{1}(E \subseteq D) \frac{1}{\pi_E}
\]  
(4.45)

\[
= \sum_{E \subseteq V} (-1)^{|D \setminus E| + |D' \setminus E|} \frac{1}{\pi_E}
\]  
(4.46)

where (4.44) follows from the fact that \((P^{-1})_{E,E'} = \pi_{E}^{-1}\) whenever \(E = E'\) and \((P^{-1})_{E,E'} = 0\) otherwise. Furthermore, (4.45) follows from (3.14), whereas (4.46) follows from the fact that \(\mathbb{1}(E \subseteq D') \mathbb{1}(E \subseteq D) = 1\) if \(E \subseteq D \cap D'\) and it is equal to zero otherwise.

Finally, (4.43) is an immediate consequence of Proposition 4.19 because \(|\text{a} \lambda| = |V(\mu_{c})|^{-1} = |V(\mu_{e})|^{-1}\).  

\[\square\]

### 4.11.3 Variances in Hierarchical Log-linear Models

We now turn to an arbitrary log-linear model with expanded generating class \(C\).

**The Variance Function**

We first consider the variance function.

**Proposition 4.22**  
Let \(Y_{V}\) be a cross-classified Bernoulli random vector with probability parameter \(\pi > 0\). If the probability distribution of \(Y_{V}\) belongs to the log-linear model \(M(C)\), then the variance function in the exponential family form (4.35) is

\[
V(\mu_{c}) = [V_{c}]|_{e}.
\]

and, furthermore, \(|V(\mu_{c})| = |V_{c}|\).

**Proof.**  
The first row and column of \(V_{c}\) are indexed by the empty set and, therefore, the matrix \(V\) can be written in a partitioned form as

\[
V_{c} = \begin{bmatrix}
V_{\emptyset, \emptyset} & V_{\emptyset, c}^T \\
V_{c, \emptyset} & V_{c, c}
\end{bmatrix}
\]
so that, by recalling that $\mathbb{V}_{\emptyset, \emptyset} = \mu_{\emptyset} = 1$ we obtain $[\mathbb{V}_C]_e = \mathbb{V}_C - \mu_C \mu_T^C$. The result follows because $V(\mu_C) = \text{var}(Y_C) = [\text{var}(Y_e)]_C = [\mathbb{V}_e]_C = [\mathbb{V}_e - \mu_C \mu_T^C]_C = \mathbb{V}_C - \mu_C \mu_T^C$. The determinant of $\mathbb{V}_C$ can be factorized as $|\mathbb{V}_C| = |[\mathbb{V}_C]_e| |\mathbb{V}_{\emptyset, \emptyset}|$ (see Lütkepol, 1996, p. 147, eqn. 4) so that $|\mathbb{V}_C| = |[\mathbb{V}_C]_e| |\mathbb{V}_{\emptyset, \emptyset}|$ because $\mathbb{V}_{\emptyset, \emptyset} = 1$.

The Asymptotic Variance of the ML Estimate

The asymptotic variance of $\hat{\lambda}_C$ can be easily derived from Proposition 4.22 by applying the rule for the inversion of a partitioned matrix.

**Proposition 4.23** Let $Y_C$ be a cross-classified Bernoulli random vector with probability parameter $\pi > 0$. Then the asymptotic variance of the ML estimate of $\lambda_C$ under the log-linear model $M(C)$ is

$$\text{avar}(\hat{\lambda}_C) = \left[ (M^T \Pi^{-1} M)_{CC} \right]_{ee},$$

and, furthermore, $|\text{avar}(\hat{\lambda}_C)| = |(M^T \Pi^{-1} M)_{CC}|$.

**Proof.** We first note that, by applying Proposition 4.17, Corollary 4.18 and the rules for the inverse of a partitioned matrix (see Lütkepol, 1996, p. 29, eqn. 1), we can write

$$\mathbb{V}_{CC} = [Z^T \Pi Z]_{CC} = \left[ (M^T \Pi^{-1} M)_{CC} \right]^{-1} = \left[ (M^T \Pi^{-1} M)_{CC} \right]_{ee}^{-1} = \left[ (M^T \Pi^{-1} M)_{CC} \right]_{ee}^{-1}.$$

(4.47)

Hence, the stated result can be derived by recalling that $\text{avar}(\hat{\lambda}_C) = V(\mu_C)^{-1}$ and then applying, in turn, Proposition 4.22, the rules for the inverse of a partitioned matrix and, finally, (4.47)

$$\text{avar}(\hat{\lambda}_C) = V(\mu_C)^{-1} = ([\mathbb{V}_C]_e)^{-1} = \left[ (\mathbb{V}_C)^{-1} \right]_e = \left[ (M^T \Pi^{-1} M)_{CC} \right]_{ee}^{-1}.$$

Furthermore, as a consequence of Proposition 4.22 and of (4.47) we obtain $|\text{avar}(\hat{\lambda}_C)| = |V(\mu_C)^{-1}| = |(\mathbb{V}_C)^{-1}| = |(M^T \Pi^{-1} M)_{CC}|$. □

We remark that alternative formulae for $\text{avar}(\hat{\lambda}_C)$ are available in the literature. For instance, $\text{avar}(\hat{\lambda}_C)$ can be obtained from $V(\mu_C)$ as follows

$$V(\mu_C) = \text{var}(\hat{Y}_C) = \text{var}(Z_{C_x, P(V)} U) = Z_{C_x, P(V)} (P - \pi \pi^T) Z_{C_x, P(V)}^T.$$
so that
\[
\text{avar}(\hat{\lambda}_e) = V(\mu_e)^{-1} = \left\{ Z_{\mathcal{C}_e,\mathcal{P}(V)} \left( \mathbb{P} - \pi \pi^T \right) Z_{\mathcal{C}_e,\mathcal{P}(V)}^T \right\}^{-1}; \tag{4.48}
\]
see also Agresti (2013, eqn. 16.22). In Proposition 4.23 we give an alternative formulation of \( \text{avar}(\hat{\lambda}_e) \) thereby providing, in explicit form, the matrix inversion required in (4.48). Such formulation, based on a Cholesky decomposition involving a Möbius matrix and a diagonal matrix of probabilities, is mathematically appealing, and we deem that it will be useful in the development of other aspects of these models. Indeed, we have exploited such Cholesky decomposition in Theorem 4.21 to give, in the saturated case, an explicit formulation of the entries of \( \text{avar}(\hat{\lambda}_e) \) as well as of its determinant. The next section provides generalisation of Theorem 4.21 to the case in which the log-linear model considered is a decomposable graphical model.

### 4.11.4 Decompositions and Decomposable Models

We consider now the case where the distribution of \( Y_V \) belongs to the graphical model \( M(G) \). If \( (A, B) \) is a decomposition of \( G \), then by the collapsibility of graphical models (Asmussen and Edwards, 1983) it holds that the distribution of \( Y_A \) belongs to \( M(G_A) \) and, similarly, the distribution of \( Y_B \) belongs to \( M(G_B) \). The corresponding log-linear parameters are \( \lambda^A_A \) and \( \lambda^B_B \), respectively, and Proposition 4.23 allows us to write the asymptotic variances of the ML estimates under these two models as

\[
\text{avar}(\lambda^A_A) = \left[ (M_A^T \mathbb{P}^{-A} M_A)_{AA|\overline{A}} \right]_e, \tag{4.49}
\]

and

\[
\text{avar}(\lambda^B_B) = \left[ (M_B^T \mathbb{P}^{-B} M_B)_{BB|\overline{B}} \right]_e. \tag{4.49a}
\]

where, in this case, \( \overline{A} = \mathcal{P}(A) \setminus A \) and \( \overline{B} = \mathcal{P}(B) \setminus B \). The separator \( S = A \cap B \) is complete so that the marginal distribution of \( Y_S \) belongs to the saturated model. The asymptotic variance of the ML estimate of \( \lambda^S_S \) follows thus by Theorem 4.21 and has form,

\[
\text{avar}(\hat{\lambda}^S_S) = (M_S^T \mathbb{P}^{-S} M_S)_{e}. \tag{4.50}
\]
If \( \lambda \) is the log-linear parameter of the distribution of \( Y_{V} \) under \( M(G) \), then \( \lambda = (\lambda_{V}, \lambda) \) where \( \lambda_{V} = 0 \), and it is shown in equation (4.21) that 

\[
\lambda_{A|B} = \lambda_{A|B}^A \quad \text{and} \quad \lambda_{B|A} = \lambda_{B|A}^B.
\]

Furthermore, because \( (A, B) \) forms a decomposition of \( G \), it follows from Corollary 4.11 that 

\[
\hat{\lambda}_{A|B} = \hat{\lambda}_{A|B}^A \quad \text{and} \quad \hat{\lambda}_{B|A} = \hat{\lambda}_{B|A}^B.
\]

The following lemma shows that \( \hat{\lambda}_{A|B} \) and \( \hat{\lambda}_{B|A} \) are asymptotically uncorrelated, and therefore independent in their joint asymptotic normal distribution.

**Lemma 4.24** Let \( \lambda \) be the log-linear parameter of the cross-classified Bernoulli random vector \( Y_{V} \) with probability parameter \( \pi > 0 \). Furthermore, let \( G = (V, E^{\sim}) \) be an undirected graph and assume that the probability distribution of \( Y_{V} \) belongs to \( M(G) \). If \( (A, B) \) forms a decomposition of \( G \) then 

\[
\text{avar}(\hat{\lambda}_{e})_{A|B, B\setminus A} = 0.
\]

**Proof.** We have to show that 

\[
\text{avar}(\hat{\lambda}_{e})_{A|B, B\setminus A} = \frac{\partial \lambda_{A|B}}{\partial \mu_{B\setminus A}^T} = 0.
\]

where \( \lambda_{A|B} = \lambda_{A|B}^A \) by Theorem 4.9. By Lemma 4.10, \( \mu_{V_{e}}^{T} \) can be partitioned into \( (\mu_{A_{e}}^{T}, \mu_{B\setminus A}^{T}) \); recall that \( \mu(\emptyset) = 1 \). As \( (A, B) \) is a decomposition of \( G \), then the graphical model with graph \( G \) is collapsible onto \( A \) (Asmussen and Edwards, 1983) and this implies that the distribution of \( Y_{A} \) is Markov respect to \( G_{A} = (A, E_{A}) \). The corresponding mean parameter is \( \mu_{A_{e}} \), whereas the canonical parameter is \( \lambda_{A_{e}}^A \). This implies that \( \lambda_{A|B}^A \) can be written as a function of \( \mu_{A_{e}} \) and, in turn, that we can apply the chain rule to obtain 

\[
\frac{\partial \lambda_{A|B}}{\partial \mu_{B\setminus A}^T} = \frac{\partial \lambda_{A|B}^A}{\partial \mu_{A_{e}}^T} \times \frac{\partial \mu_{A_{e}}^T}{\partial \mu_{B\setminus A}^T}.
\]
The vector $\mu_{\mathcal{V}_e}$ is the mean parameter of a regular exponential family in a full representation which take values in its domain $\mu(\Theta_{\mathcal{V}_e})$ and there are no functional relationships between its elements. Hence, because the subvectors $\mu_{\mathcal{A}_e}$ and $\mu_{\mathcal{B}' \setminus \mathcal{A}}$ have no common elements, it holds that $\partial \mu_{\mathcal{A}_e} / \partial \mu_{\mathcal{B}' \setminus \mathcal{A}}^T = 0$ which implies the desired result. □

We remark that Lemma 4.24 is a consequence of the connection between decomposition of the graph and cut in $\mathcal{Y}_{\mathcal{V}_e}$; see Lauritzen (1996, p. 242) and Piccioni (2000). Furthermore, for the decomposable case this result is implicit in the theory of meta Markov models introduced by Dawid and Lauritzen (1993, section 4); see also Consonni and Leucari (2006, Proposition 4).

As a consequence of Lemma 4.24 we obtain the local computation of the asymptotic variance of $\hat{\lambda}_{\mathcal{V}_e}$.

**Theorem 4.25** In the setting of Lemma 4.24 it holds that

$$ \text{avar}(\hat{\lambda}_{\mathcal{V}_e}) = [\text{avar}(\hat{\lambda}_{\mathcal{A}_e})]^0 + [\text{avar}(\hat{\lambda}_{\mathcal{B}_e})]^0 - [\text{avar}(\hat{\lambda}_{\mathcal{S}_e})]^0 $$

and

$$ |\text{avar}(\hat{\lambda}_{\mathcal{V}_e})| = \frac{|\text{avar}(\hat{\lambda}_{\mathcal{A}_e})||\text{avar}(\hat{\lambda}_{\mathcal{B}_e})|}{|\text{avar}(\hat{\lambda}_{\mathcal{S}_e})|} $$

where $\mathcal{S} = \mathcal{A} \cap \mathcal{B}$ and for $\mathcal{Q} \subseteq \mathcal{V}$, $[\text{avar}(\hat{\lambda}_{\mathcal{Q}_e})]^0$ denotes the matrix indexed by $\mathcal{V}_e \times \mathcal{V}_e$ obtained by padding zero entries to $\text{avar}(\hat{\lambda}_{\mathcal{Q}_e})$ to obtain full dimension. Explicit forms for $\text{avar}(\hat{\lambda}_{\mathcal{A}_e})$ and $\text{avar}(\hat{\lambda}_{\mathcal{B}_e})$ are given in (4.49) and (4.49a), whereas for $\text{avar}(\hat{\lambda}_{\mathcal{S}_e})$ see (4.50).

**Proof.** By Lemma 4.10, $(\mathcal{A} \setminus \mathcal{B}, \mathcal{S}_e, \mathcal{B} \setminus \mathcal{A})$ forms a partition of $\mathcal{V}_e$ and this, in turn, induces a block partition in $\text{avar}(\hat{\lambda}_{\mathcal{V}_e})$ where the block $\text{avar}(\hat{\lambda}_{\mathcal{e}})_{\mathcal{A} \setminus \mathcal{B} \setminus \mathcal{A}}$ is equal to zero by Lemma 4.24 and we can apply Lauritzen (1996, Lemma 5.5) to write

$$ \text{avar}(\hat{\lambda}_{\mathcal{V}_e}) = V(\mu_{\mathcal{V}_e})^{-1} = \left[ \{ V(\mu_{\mathcal{V}_e})_{\mathcal{A}_e} \}^{-1} \right]^0 + \left[ \{ V(\mu_{\mathcal{V}_e})_{\mathcal{B}_e} \}^{-1} \right]^0 - \left[ \{ V(\mu_{\mathcal{V}_e})_{\mathcal{S}_e} \}^{-1} \right]^0. $$

It follows from the definition of variance function that if $\mathcal{C} \subseteq \mathcal{V}$ is an expanded generating class then $V(\mu_{\mathcal{V}_e})_{\mathcal{C}_e} = V(\mu_{\mathcal{C}_e})$ and, therefore,
avar(\hat{\lambda}_{V_e}) = [V(\mu_{A_e})^{-1}]^0 + [V(\mu_{B_e})^{-1}]^0 - [V(\mu_{S_e})^{-1}]^0

and the result follows because \(V(\mu_{A_e})^{-1} = avar(\hat{\lambda}_{A_e})\), \(V(\mu_{B_e})^{-1} = avar(\hat{\lambda}_{B_e})\) and \(V(\mu_{S_e})^{-1} = avar(\hat{\lambda}_{S_e})\). The factorization of the determinant can be shown in the same way.

Finally, we consider the decomposable case and show that the asymptotic variance can be locally computed with respect to the saturated models corresponding to cliques and separators. Furthermore, in this case an explicit form for the entries of the asymptotic variance is available.

**Theorem 4.26** In the setting of Lemma 4.24, assume that G is a decomposable graph with cliques \(C_1, \ldots, C_k\) and separators \(S_2, \ldots, S_k\). Then
\[
avar(\hat{\lambda}_{V_e}) = \sum_{i=1}^{k} [avar(\hat{\lambda}_{C_i})]^0 - \sum_{i=2}^{k} [avar(\hat{\lambda}_{S_i})]^0
\]

and the entry indexed by \(D, D' \in V_e\) has the form
\[
avar(\hat{\lambda}_{e})_{D,D'} = \sum_{E \subseteq D \cap D'} (-1)^{|D \setminus E| + |D' \setminus E|} \left\{ \sum_{i=1}^{k} \frac{1(D \cup D' \subseteq C_i)}{\pi^{C_i}(E)} - \sum_{i=2}^{k} \frac{1(D \cup D' \subseteq S_i)}{\pi^{S_i}(E)} \right\}
\]

Furthermore,
\[
|avar(\hat{\lambda}_{V_e})| = \prod_{i=1}^{k} |avar(\hat{\lambda}_{C_i})| = \prod_{i=2}^{k} |avar(\hat{\lambda}_{S_i})| = \prod_{i=1}^{k} |P^{C_i}| = \prod_{i=2}^{k} |P^{S_i}|
\]

**Proof.** The formulae for \(avar(\hat{\lambda}_{V_e})\) and \(|avar(\hat{\lambda}_{V_e})|\) follow from the recursive application of Theorem 4.25 with respect to a perfect sequence of cliques of \(G\) and then by applying (4.43).

Consider now the entries of \(avar(\hat{\lambda}_{V_e})\). If \(A\) is a complete subset of \(V\), then the entry of \([avar(\hat{\lambda}_{e})]^0\) indexed by \(D, D' \in V_e\) is given in Theorem 4.21 if \(D, D' \subseteq A\) and is equal to zero otherwise. Because \(D, D' \subseteq A\) if and only if \(1(D \cup D' \subseteq A) = 1\) we can write
\[
[avar(\hat{\lambda}_{e})]^0_{D,D'} = \sum_{E \subseteq D \cap D'} (-1)^{|D \setminus E| + |D' \setminus E|} \frac{1(D \cup D' \subseteq A)}{\pi^{A}(E)}
\]

(4.51)
and the expression for the entries of $\text{avar}(\hat{\lambda}_{V_e})$ can be obtained by applying (4.51) with respect to the cliques and separators.

We close this section by noticing that the factorization of the determinant in Theorem 4.25 implies a corresponding factorizations of the determinant of the variance function because $|V(\mu_{V_e})| = |\text{avar}(\hat{\lambda}_{V_e})|^{-1}$, $|V(\mu_{A_e})| = |\text{avar}(\hat{\lambda}_{A_e})|^{-1}$, $|V(\mu_{B_e})| = |\text{avar}(\hat{\lambda}_{B_e})|^{-1}$ and, finally, $|V(\mu_{S_e})| = |\text{avar}(\hat{\lambda}_{S_e})|^{-1}$. Clearly, a similar result also holds for the factorization of the determinant in Theorem 4.26.
Undirected graph models put all the variables of the relevant random vector $Y_V$ on an equal standing. Accordingly, the edges of the graph are of a symmetric type to encode a likewise symmetric relationship between variables. When an edge is missing the corresponding variables are conditionally independent given all the remaining variables in $Y_V$. More generally, each of the three Markov properties for undirected graphs is characterized by a collection of conditional independence relationships of the type $Y_A \indep Y_B | Y_{V \setminus (A \cup B)}$, for suitable subsets $A$ and $B$ of $V$. In other words, undirected graph models are defined through independence relationships fulfilled by conditional distributions involving all the variables in $Y_V$, and in this sense we can say that they are “conditional” independence models. This chapter is devoted to models defined by bidirected graphs, i.e., graphs whose edges are arrows with two heads ($\leftrightarrow$). Like undirected edges, bidirected edges encode symmetric relationships between variables and, in the corresponding independence models, the variables are considered to be on an equal standing. The distinguishing feature of these models is that every missing edge implies the marginal independence of the corresponding variables, rather than the conditional independence. There are three distinct Markov properties for bidirected graphs, and each is characterized by a collection of marginal independencies of the type $Y_A \indep Y_B$, where $A$ and $B$ are suitable subsets of $V$. For this reason, bidirected graph models can be regarded as “marginal” independence models. The interest for the marginal independence restrictions encoded by a bidirected graph is due to the fact that they may arise through confounding effects of unobserved variables (Cox and Wermuth, 1993, 1996; Richardson and Spirtes, 2002; Drton and Richardson, 2008a).

Graphical models of marginal independence were first introduced for Gaussian random vectors. Marginal independencies in Gaussian distributions correspond to zero covariances, and Cox and Wermuth (1993,
1996) used a *covariance graph* to represent the zero pattern of a covariance matrix. In this context a relevant role is also played by the seminal work by Anderson (1969, 1973) on models defined by linear restrictions on covariances. In a covariance graph, every missing edge implies the marginal independence of the corresponding variables. Models for categorical variables with analogous independence structure were first developed by Kauermann (1997), and a comprehensive set of references on this area can be found in Drton and Richardson (2008a) and La Rocca and Roverato (2017). Notice that dashed undirected lines are sometimes used in place of bidirected arrows in the graphical representation of these models. We adopt the latter convention because it is consistent with the notation used in the path diagrams of Wright (1921).

In the following, we will present the theory concerning the Markov properties for bidirected graphs and the relative statistical models for categorical variables. We will deal with inferential issues in the next chapter. The latter gives the theory of regression graph models which include models for bidirected graphs as a special case.

### 5.1 Bidirected Graphs

We provide here some additional notions of graph theory required to deal with models defined by bidirected graphs.

A *bidirected graph* $G = (V, E)$ is a graph with only bidirected edges, and when we want to highlight that a graph is bidirected we write its edge set as $E^\leftrightarrow$. Figure 5.1 gives an example of bidirected graph that is the bidirected version of the graph in Figure 4.1. Several notions introduced for undirected graphs can be defined in the same way for bidirected graphs, and we will apply to bidirected graphs the concepts of *induced subgraph*, *path* and *separation*. Hence, similarly to what happens in the graph of Figure 4.1, also in the graph of Figure 5.1 the

![Figure 5.1 Bidirected version of the graph in Figure 4.1.](https://www.cambridge.org/core/product/F9D4523D3559BC2BA9251CEF3077F20D)
sequence \((a, b, d, c, f)\) is a path from \(a\) to \(f\). Furthermore, the subsets \(\{c, d\}\) and \(\{b, c\}\) both separate the vertex \(a\) from the vertex \(f\), and also the subset \(A = \{a\}\) from the subset \(B = \{e, f\}\). We will also apply to bidirected graphs the notion of neighbor that, for a subset \(A \subseteq V\), is denoted by \(\text{nb}(A)\) and defined just as in the undirected case. Although we do not use it here, we remark that in the literature of bidirected graph models the notion of neighbor is often replaced by that of spouse. In our notation, the spouse of a subset \(A \subseteq V\), as defined, e.g., in Drton and Richardson (2008a), is the set of vertices \(A \cup \text{nb}(A)\).

A key role in the theory of bidirected graph models is played by the connected subsets of the vertex set \(V\). A graph \(G = (V, E)\) is connected when every pair of distinct vertices in \(V\) is joined by a path. A nonempty subset \(A \subseteq V\) is a connected set in \(G\) if the induced subgraph \(G_A\) is connected, and it is disconnected otherwise. Every nonempty subset \(A \subseteq V\) can be partitioned uniquely into maximal connected sets, \(A = C_1 \cup \ldots \cup C_r\). The sets \(C_1, \ldots, C_r\) are called the connected components of \(A\). For example, in the graph of Figure 5.1 the set \(\{a, b, c, e, f\}\) is connected, whereas the set \(\{a, b, e, f\}\) is disconnected with connected components \(C_1 = \{a, b\}\) and \(C_2 = \{e, f\}\). Also the set \(\{a, d, e\}\) is disconnected and its connected components are the singletons \(C_1 = \{a\}\), \(C_2 = \{d\}\) and \(C_3 = \{e\}\).

### 5.2 Markov Properties for Bidirected Graphs

There are three distinct Markov properties for bidirected graphs, each being the dual of one of the Markov properties for undirected graphs (Kauermann, 1996).

**The Dual Pairwise Markov Property**

The probability distribution of a random vector \(Y_V\) is said to obey the dual pairwise Markov property with respect to the bidirected graph \(G = (V, E^{\leftrightarrow})\) if for every \(a, b \in V\) such that \(\{a, b\} \notin E^{\leftrightarrow}\) it holds that

\[
Y_a \perp \perp Y_b.
\]

For example, two independence relations implied by this Markov property for the graph of Figure 5.1 are \(Y_a \perp \perp Y_f\) and \(Y_b \perp \perp Y_c\). It is useful to compare these marginal independencies with the corresponding conditional independencies implied by the pairwise Markov property for the
The Dual Local Markov Property

The probability distribution of a random vector $Y_V$ is said to obey the dual local Markov property with respect to the bidirected graph $G = (V, E^\leftrightarrow)$ if for every vertex $a \in V$ it holds that

$$Y_a \perp Y_{V\setminus (\text{nb}(a) \cup \{a\})}.$$  

For example, if the distribution of $Y_V$ is locally Markov with respect to the graph of Figure 5.1 then it holds that $Y_a \perp Y_{\{d,e,f\}}$, $Y_d \perp Y_{\{a,e\}}$ and $Y_c \perp Y_b$. Also in this case, the analogy with the local Markov property for undirected graphs is straightforward.

The Dual Global Markov Property

The probability distribution of a random vector $Y_V$ is said to obey the dual global Markov property with respect to the bidirected graph $G = (V, E^\leftrightarrow)$ if for any triple of pairwise disjoint subsets $A, B, S \subseteq V$ such that $S$ separates $A$ from $B$ in $G$ it holds that

$$Y_A \perp Y_B | Y_{V\setminus (A \cup B \cup S)}.$$  

For this Markov property, duality with the relative Markov property for undirected graphs is more subtle. Like in the undirected case, every separation implies a conditional independence relationship, but the conditioning set is the complement of the separating set $S$ with respect to $V \setminus (A \cup B)$ rather than $S$ itself. Examples of conditional independencies implied by the dual global Markov property for the graph of Figure 5.1 are $Y_a \perp Y_{\{e,f\}} | Y_b$ and $Y_{\{b,d\}} \perp Y_e | Y_a$.

The above definition of global Markov property for bidirected graph is useful because it provides a rule for reading a wide range of independence relationships from the graph, including conditional ones. More specifically, as shown in Richardson and Spirtes (2002), the dual global Markov property is complete; see also Drton and
Richardson (2008a). We recall that this means that the conditional independencies encoded by any bidirected graph $G$ under the dual global Markov property are the only conditional independencies that are simultaneously satisfied by all the distributions which are (dual) globally Markov with respect to $G$. On the other hand, this Markov property has the disadvantage of being formulated in terms of conditional independencies, that may hide its “marginal” nature. In this sense, of special interest is the case where $V \backslash (A \cup B)$ separates $A$ from $B$ in $G$ because the dual global Markov property implies the marginal independence of $Y_A$ and $Y_B$. As shown in Proposition 4.1, the global Markov property for undirected graphs is characterized by the independence relationships between subvectors $Y_A$ and $Y_B$ implied by separating sets of the type $V \backslash (A \cup B)$. This is also true for the dual global Markov property and it has led to the introduction of a further, but equivalent, Markov property based on marginal independencies.

The Connected Set Markov Property

For a bidirected graph $G = (V, E^{\leftrightarrow})$ consider two disjoint subsets of vertices $A$ and $B$. It is not difficult to see that the subset $V \backslash (A \cup B)$ separates $A$ from $B$ in $G$ if and only if there is no edge joining a vertex in $A$ with a vertex in $B$, that is, if and only if $A$ and $B$ are not connected to each other in $G_{A \cup B}$. In this case the set $A \cup B$ is disconnected in $G$ and the dual global Markov property implies $Y_A \indep Y_B$. This observation suggested the introduction of a Markov property that focuses on the connected subsets of $V$. Formally, the probability distribution of a random vector $Y_V$ is said to obey the connected set Markov property with respect to the bidirected graph $G = (V, E^{\leftrightarrow})$ if for every disconnected set $A$ of $G$ it holds that

$$Y_{C_1} \indep \ldots \indep Y_{C_r},$$

where $C_1, \ldots, C_r$ are the connected components of $A$. For example, in the graph of Figure 5.1 the subset $\{a, b, e, f\}$ is disconnected with connected components $\{a, b\}$ and $\{e, f\}$ and therefore under the connected set Markov property it holds that $Y_{\{a, b\}} \indep Y_{\{e, f\}}$. In the same graph, the subset $\{a, d, e\}$ is disconnected with connected components $\{a\}$, $\{d\}$ and $\{e\}$ and therefore the connected set Markov property implies $Y_a \indep Y_d \indep Y_e$. 
The connected set Markov property is equivalent to the dual global Markov property as formally stated by the following proposition proved in Richardson (2003); see also Drton and Richardson (2008a).

**Proposition 5.1**  The probability distribution of $Y_V$ obeys the dual global Markov property with respect to a bidirected graph $G = (V, E^\leftrightarrow)$ if and only if it obeys the connected set Markov property with respect to $G$.

**Proof.** Assume that the distribution of $Y$ is dual globally Markov with respect to $G = (V, E^\leftrightarrow)$, and let $A \subseteq V$ be an arbitrary disconnected set with connected components $C_1, \ldots, C_r$. If we let $A_2 = C_2 \cup \ldots \cup C_r$, then $V \setminus A$ separates $C_1$ from $A_2$ in $G$, and therefore by the dual global Markov property $Y_{C_1} \perp \perp Y_{A_2}$. By iteratively applying the same procedure to the pairs of subsets $C_i$ and $A_{i+1} = C_{i+1} \cup \ldots \cup C_r$, for $i = 1, \ldots, (r - 1)$, one obtains $Y_{C_1} \perp \perp \ldots \perp Y_{C_r}$, and therefore that $Y$ obeys the connected set Markov property with respect to $G$. Consider now the case where the distribution of $Y$ obeys the connected set Markov property with respect to $G$ and assume that $A, B$ and $S$ is an arbitrary triple of pairwise disjoint subsets of $V$ such that $S$ separates $A \neq \emptyset$ form $B \neq \emptyset$ in $G$. Then, one can always find two disjoint sets $D_1$ and $D_2$ such that $D_1 \cup D_2 = V \setminus (A \cup B \cup S)$ and $S$ separates $A \cup D_1$ from $B \cup D_2$ in $G$. Hence, the sets $A \cup D_1$ and $B \cup D_2$ are not connected to each other in $G$ and the connected set Markov property implies that $Y_{A \cup D_1} \perp \perp Y_{B \cup D_2}$. In turn, it follows from the properties of conditional independence (see Lauritzen, 1996, section 3.1) that $Y_A \perp \perp Y_B | Y_{D_1 \cup D_2}$, i.e., that $Y_A \perp \perp Y_B | Y_{V \setminus (A \cup B \cup S)}$ as required. □

The connected set Markov property is not complete, and therefore it is less exhaustive than the dual global Markov property. However, it makes it clear that every independence model associated with a bidirected graph is characterized by a collection of marginal independencies. Furthermore, the connected set Markov property is useful in the specification of suitable parameterizations of bidirected graph models. Indeed, unlike the undirected case where for positive distributions one can exploit the equivalence of the three Markov properties, this is not possible in the bidirected case. To see this, we now turn to the analysis of the connections among the four Markov properties for bidirected graphs. It is easy to see that the dual global Markov property implies the dual local Markov property because for every $a \in V$ it holds that
nb(a) separates a from $V \setminus (nb(a) \cup \{a\})$ in $G$. Furthermore, the dual local Markov property implies the dual pairwise Markov property because if $\{a, b\} \notin E^{\leftrightarrow}$ then $b \in V \setminus (nb(a) \cup \{a\})$, and therefore the local independence $Y_a \perp \!\!\!\!\perp Y_{V \setminus (nb(a) \cup \{a\})}$ implies the pairwise independence $Y_a \perp \!\!\!\!\perp Y_b$. We can thus summarize the relationships between the Markov properties for bidirected graphs as follows,

\[
\text{(connected set)} \iff \text{(dual global)} \Rightarrow \text{(dual local)} \Rightarrow \text{(dual pairwise)}.
\]

Unlike the undirected case, the dual pairwise Markov property does not imply the dual global Markov property, even for positive distributions. To show this, Table 5.1 provides an example, taken from Drton and Richardson (2008a), of a positive probability distribution for the binary vector $Y_{\{a,b,c\}}$ such that both $Y_a \perp \!\!\!\!\perp Y_b$ and $Y_a \perp \!\!\!\!\perp Y_c$ but $Y_a \not\perp \!\!\!\!\perp Y_{\{b,c\}}$.

<table>
<thead>
<tr>
<th>$Y_a$</th>
<th>$Y_b$</th>
<th>$Y_c$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
<td>0.02 0.08</td>
</tr>
<tr>
<td>1</td>
<td>0.03 0.12</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>0.05 0.25</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>0.10 0.35</td>
<td></td>
</tr>
</tbody>
</table>

5.3 The Log-mean Linear Parameterization

Marginal independencies between categorical variables cannot be conveniently implemented through the log-linear parameterization because marginal independence constraints do not correspond to linear restrictions in the log-linear parameter space. Nevertheless, a suitable parameterization can be obtained by exploiting the existing duality between the undirected and the bidirected case, which also extends to the parameterization level. The log-linear parameter $\lambda$ is obtained from the log-
linear expansion of the probability parameter $\pi$, and the role it plays in undirected graph models follows from the existing connection between its terms and cross-product ratios of the “conditional” type, $\text{cpr}_{cc}(\cdot)$. The cross-classified Bernoulli distribution belongs to the regular exponential family, and a parameterization alternative to the probability parameter $\pi$ is provided by the mean parameter $\mu$. This is a vector of marginal probabilities and its log-linear expansion yields a parameter $\gamma$ named the log-mean linear parameter of $Y$. Hereafter, we will describe the connection existing between the entries of $\gamma$ and the cross-product ratios of the “marginal” type, $\text{cpr}_{mm}(\cdot)$, used to assess marginal independence relationships.

The Binary Case

For a binary random vector $Y$ with probability parameter $\pi$ the mean parameter, as defined in Section 4.10.4, is given by

$$\mu = (\mu_D)_{D \subseteq V}$$

where

$$\mu_D = p(Y_D = 1) \text{ for every } D \subseteq V,$$

with the convention that $\mu_{\emptyset} = 1$. We recall that by Proposition 4.16 the vector $\mu$ can be computed as $\mu = \mathbb{Z}\pi$ so that $\pi = \mathbb{M}\mu$. As an example, we give below the mean parameter relative to Table 5.1,

<table>
<thead>
<tr>
<th>$\mu_{\emptyset}$</th>
<th>$\mu_a$</th>
<th>$\mu_b$</th>
<th>$\mu_c$</th>
<th>$\mu_{{a,b}}$</th>
<th>$\mu_{{a,c}}$</th>
<th>$\mu_{{b,c}}$</th>
<th>$\mu_{{a,b,c}}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.75</td>
<td>0.60</td>
<td>0.80</td>
<td>0.45</td>
<td>0.60</td>
<td>0.47</td>
<td>0.35</td>
</tr>
</tbody>
</table>

Figure 5.2 Four graphs of marginal independence for $Y_{\{a, b, c\}}$. 

Graphical Models for Categorical Data
The parametric space of $\mu$ is more involved than that of $\pi$. The probabilities in $\pi$ have no restrictions, apart from the sum-to-one constraint. On the other hand, the entries of $\mu$ need to satisfy several constraints. It is straightforward to see that $\mu_{D'} \geq \mu_D$ whenever $D' \subseteq D$ so that, for example, $\mu_a \geq \mu\{a\}$ and $\mu_b \geq \mu\{a,b\}$; but there are also additional, less-immediate, constraints. For example, $\mu_a + \mu_b - \mu\{a,b\}$ is the probability of the event $\{Y_a = 1\} \cup \{Y_b = 1\}$ and this implies that both $\mu_a + \mu_b \geq \mu\{a,b\}$ and $\mu_a + \mu_b \leq 1 + \mu\{a,b\}$.

We now introduce the log-mean linear parameter $\gamma = (\gamma_D)_{D \subseteq V}$ (Roverato et al., 2013) that, in parallel with (4.1), is defined as

$$\gamma = M^T \log \mu. \quad (5.1)$$

It follows from (5.1) and the fact that $\mu = \mathbb{Z}\pi$ that one can compute $\gamma$ from $\pi$ as $\gamma = M^T \log \mathbb{Z}\pi$, and thus $\pi$ from $\gamma$ as $\pi = M \exp(\mathbb{Z}^T \gamma)$. Hence, the functional relationship between $\gamma$ and $\pi$ is a smooth bijection and this implies that $\gamma$ is a valid parameterization of the distribution of $Y$.

The individual entries of $\gamma$ can be written by (3.3) as

$$\gamma_D = \sum_{D' \subseteq D} (-1)^{|D| - |D'|} \log \mu_{D'} \text{ for every } D \subseteq V, \quad (5.2)$$

and this also shows that the log-mean linear parameters satisfy a property called upward compatibility, i.e., every interaction term $\gamma_D$ is computed in the distribution of the relevant margin $Y_D$. It is also useful to compare the theoretical quantities with a more concrete example and to this aim we provide the log-mean linear terms relative to Table 5.1,

<table>
<thead>
<tr>
<th>$\gamma_\emptyset$</th>
<th>$\gamma_a$</th>
<th>$\gamma_b$</th>
<th>$\gamma_c$</th>
<th>$\gamma{a}$</th>
<th>$\gamma{b}$</th>
<th>$\gamma{c}$</th>
<th>$\gamma{a,b}$</th>
<th>$\gamma{a,c}$</th>
<th>$\gamma{b,c}$</th>
<th>$\gamma{a,b,c}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>-0.288</td>
<td>-0.511</td>
<td>-0.223</td>
<td>0</td>
<td>0</td>
<td>-0.021</td>
<td>-0.00712</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

The computation of the log-mean linear parameter is based on a Möbius inversion formula, and this allows us to exploit the tools introduced in Chapter 3. The rest of this section shows the connection between independence relationship and vanishing log-mean linear terms, and it runs in parallel with the theory of the log-linear parameterization given in Section 4.3.

The entry of $\gamma$ indexed by the empty set is always equal to zero,

$$\gamma_\emptyset = \log \mu_\emptyset = \log (1) = 0.$$
The main effect $\gamma_a$, that is the log-mean linear term for the singleton set \{a\}, is computed in the marginal distribution of $Y_a$,

$$\gamma_a = \log \frac{\mu_a}{\mu_\emptyset} = \log \mu_a = \log p(Y_a = 1).$$

The two-way interaction relative to $D = \{a, b\}$ is computed in the marginal distribution of $Y_{\{a,b\}}$, and it is the logarithm of a cross-product ratio,

$$\gamma_{\{a,b\}} = \log \frac{\mu_{\{a,b\}}\mu_\emptyset}{\mu_a \mu_b} = \log \frac{p(Y_a = 1, Y_b = 1)}{p(Y_a = 1)p(Y_b = 1)} = \log \text{cpr}_{mm}(a, a : b, b).$$

It follows from Lemma 2.2 that $\gamma_{\{a,b\}} = 0$ is equivalent to the independence of $Y_a$ and $Y_b$. Similarly to the two-way interaction $\gamma_{\{a,b\}}$, the term $\gamma_{\{a,c\}}$ is

$$\gamma_{\{a,c\}} = \log \frac{\mu_{\{a,c\}}\mu_\emptyset}{\mu_a \mu_c} = \log \text{cpr}_{mm}(a, a : c, c),$$

and therefore $Y_a \independent Y_c$ if and only if $\gamma_{\{a,c\}} = 0$. Consider for example the log-mean linear terms given above, relative to Table 5.1. For this probability distribution $\gamma_{\{a,b\}} = \gamma_{\{a,c\}} = 0$ and therefore it holds both that $Y_a \independent Y_b$ and $Y_a \independent Y_c$.

The three-way interaction for \{a, b, c\} can be decomposed as

$$\gamma_{\{a,b,c\}} = \log \frac{\mu_{\{a,b,c\}}\mu_{\{b,c\}}\mu_{\{a,c\}}\mu_\emptyset}{\mu_{\{a,b\}}\mu_{\{a,c\}}\mu_{\{b,c\}}\mu_\emptyset} = \log \frac{\mu_{\{a,b,c\}}}{\mu_{\{a,b\}}} - \log \frac{\mu_{\{a,c\}}}{\mu_{\{a\}}} = \log \text{cpr}_{mm}(a, a : \{b, c\}, \{b, c\}) - \gamma_{\{a,b\}} - \gamma_{\{a,c\}}.$$

By Lemma 2.2 it holds that $Y_a \independent Y_{\{b,c\}}$ if and only if

$$\text{cpr}_{mm}(a, a : \{b, c\}, \{b, c\}) = \text{cpr}_{mm}(a, a : b, \{b, c\}) = \text{cpr}_{mm}(a, a : c, \{b, c\}) = 1$$

where

$$\text{cpr}_{mm}(a, a : b, \{b, c\}) = \text{cpr}_{mm}(a, a : b, b) = \gamma_{\{a,b\}}$$
and
\[ \text{cpr}_{mm}(a, a : c, \{b, c\}) = \text{cpr}_{mm}(a, a : c, c) = \gamma_{\{a,c\}}. \]

Hence, when \( \gamma_{\{a,b\}} = \gamma_{\{a,c\}} = 0 \), the vanishing of the three-way interaction, \( \gamma_{\{a,b,c\}} = 0 \), is a necessary and sufficient condition for the independence of \( Y_a \) and \( Y_{\{b,c\}} \). In the example of Table 5.1 we have \( \gamma_{\{a,b,c\}} = -0.00712 \) and thus \( Y_a \perp Y_{\{b,c\}} \).

We can now formally state the connection between the vanishing of certain log-mean linear interactions and independence relationships. This theorem is the dual of the equivalent result for the undirected case given in Theorem 4.2. Also in this case the proof is an immediate consequence of the basic lemmas of Chapter 2, connecting cross-product ratios with independence relationships, together with Lemma 3.3 for Möbius inversion; see also Roverato et al. (2013) and La Rocca and Roverato (2017).

**Theorem 5.2**  For a vector \( Y_V \) of binary variables with probability parameter \( \pi > 0 \) let \( \gamma = M^T \log \mu \) where \( \mu = Z \pi \). Then, for a pair of disjoint nonempty subsets \( A \) and \( B \) of \( V \) the following conditions are equivalent:

(i) \( Y_A \perp Y_B \);

(ii) for every \( D \subseteq A \cup B \) such that both \( D \cap A \neq \emptyset \) and \( D \cap B \neq \emptyset \) it holds that
\[ \gamma_D = 0; \]

(iii) for every \( A', B' \subseteq A, B \), such that \( A', B' \neq \emptyset \) it holds that
\[ \log \mu_{A' \cup B'} - \log \mu_{A'} - \log \mu_{B'} + \log \mu_{\emptyset} = 0. \]

**Proof.** The equivalence of (i) and (iii) follows from Lemma 2.2 because the identity (5.3) is equal to \( \log \text{cpr}_{mm}(A', A : B', B) = 0 \). The equivalence of (ii) and (iii) is that stated in Lemma 3.3 when \( \theta = \gamma \) and \( \omega = \log \mu \). \( \square \)

**The Non-binary Case**

The extension of the parameterization \( \mu \) to categorical variables with arbitrary number of levels was developed by Drton (2009). Let \( Y \) be a vector of categorical variables with state space \( I \) and restricted state space \( J \). Drton (2009) introduced the saturated Möbius parameter
\[ \mu_j = (\mu_{jd})_{D \subseteq V} \text{ for all } j \in \mathcal{J}, \]

where
\[ \mu_{jd} = p(Y_D = j_D), \]

and showed that it smoothly parameterizes the distribution of \( Y \). Consequently, the log-mean linear parameter \( \gamma \) of \( Y \) can be defined as the collection of vectors
\[ \gamma_j = (\gamma_{jd})_{D \subseteq V} \text{ for all } j \in \mathcal{J}, \]

where
\[ \gamma_j = M^T \log \mu_j \text{ and } \gamma_{jd} = (\gamma_j)_D. \]

Recall that we adopt the convention that \( j \emptyset = \mathcal{J} \emptyset = \emptyset \). As well as for the probability parameter and for the log-linear parameter also in this case it is possible a more compact representation of the Möbius parameter of \( Y \) as \( \mu = (\mu_{jd})_{D \subseteq V, j_D \in \mathcal{J}_D} \). Likewise, \( \gamma = (\gamma_{jd})_{D \subseteq V, j_D \in \mathcal{J}_D} \).

A formal extension of Theorem 5.2 to the general case of categorical variables with arbitrary number of levels can be stated as follows; see also Roverato (2015) and La Rocca and Roverato (2017).

**Corollary 5.3** Let \( Y_V \) be a vector of categorical variables with probability parameter \( \pi > 0 \), Möbius parameter \( \mu = (\mu_{jd})_{D \subseteq V, j_D \in \mathcal{J}_D} \) and log-mean linear parameter \( \gamma = (\gamma_{jd})_{D \subseteq V, j_D \in \mathcal{J}_D} \). Then, for a pair of disjoint nonempty subsets \( A \) and \( B \) of \( V \) the following conditions are equivalent:

(i) \( Y_A \perp \! \! \! \perp Y_B \);

(ii) for every \( D \subseteq A \cup B \) such that both \( D \cap A \neq \emptyset \) and \( D \cap B \neq \emptyset \) it holds that
\[ \gamma_{jd} = 0, \]

for every \( j \in \mathcal{J} \);

(iii) for every \( A' \subseteq A, B' \subseteq B \), such that \( A', B' \neq \emptyset \) it holds that
\[ \log \mu_{j_{A' \cup B'}} + \log \mu_{j_{A'}} + \log \mu_{j_{B'}} - \log \mu_{\emptyset} = 0, \]

for every \( j \in \mathcal{J} \).
Proof. The equivalence of (i) and (iii) follows from Lemma 2.6 because the identity (5.4) is equal to \( \log \text{cpr}_{mm}(j_{A'}, A : j_{B'}, B) = 0 \). The equivalence of (ii) and (iii) holds for every \( j \in J \) and is that stated in Lemma 3.3 when \( \theta = \gamma_j \) and \( \omega = \log \mu_j \). □

5.4 Log-mean Linear Graphical Models

A graphical model is a family of probability distributions defined through the independence relationships encoded by a graph under a given Markov property. Unlike the undirected case where, for positive probability distributions, the graphical models under the three Markov properties coincide, in models defined by bidirected graphs it is always necessary to specify the Markov property being used. The results given in the previous section show that models for bidirected graphs can be identified by means of vanishing log-mean linear interactions. Specifically, Theorem 5.2 and Corollary 5.3 can be immediately applied to identify the log-mean linear restrictions required to specify a bidirected graph model under either the dual pairwise or the dual local Markov properties. Establishing the connection of the dual global Markov property with the log-mean linear parameter is less straightforward and carefully described below; pairwise independencies are also discussed for a comparison.

The Binary Case

The independence of two binary variables, say \( Y_a \) and \( Y_b \), is equivalent to the vanishing of a single log-mean linear interaction, regardless of the dimension of \( Y_V \). This follows from Theorem 5.2 where it is shown that \( Y_a \) and \( Y_b \) are independent if and only if \( \gamma_{\{a, b\}} = 0 \). It is thus immediate to assess whether a distribution belongs to a bidirected graph model under the dual pairwise Markov property because this happens whenever every two-way log-mean linear interaction associated with a missing edge is equal to zero. More formally, the distribution of \( Y_V \) obeys the dual pairwise Markov property with respect to \( G = (V, E^{\leftrightarrow}) \) if and only if for every pair of distinct vertices \( a, b \in V \) such that \( \{a, b\} \notin E^{\leftrightarrow} \) it holds that \( \gamma_{\{a, b\}} = 0 \).

The log-mean linear expansion of the individual entries of \( \mu \) is

\[
\log \mu_D = \sum_{D' \subseteq D} \gamma_{D'} \text{ for every } D \subseteq V,
\]
so that the expansion of $\log \mu_V$ involves all the entries of $\gamma$. Consider the example where $V = \{a, b, c\}$. The saturated model is associated with the complete graph of Figure 5.2A, and this corresponds to the log-mean linear model with no restrictions, but for $\gamma_{\emptyset} = 0$. Hence we can write

$$\log \mu_{\{a,b,c\}} = \gamma_a + \gamma_b + \gamma_c + \gamma_{\{a,b\}} + \gamma_{\{a,c\}} + \gamma_{\{b,c\}} + \gamma_{\{a,b,c\}}. \quad (5.5)$$

To gain insight into the meaning of the log-mean linear expansion it is useful to notice that the upward compatibility property implies invariance of the log-mean linear terms with respect to marginalization. Thus, marginalization over some variables can be carried out by removing the proper log-mean linear terms. For instance, for the log-mean linear expansion of the probability distribution of $Y_{\{a,b\}}$ it holds that

$$\log \mu_{\{a,b\}} = \gamma_a + \gamma_b + \gamma_{\{a,b\}},$$

where the $\gamma$-terms are the same as those in (5.5). By further removing the terms $\gamma_b$ and $\gamma_{\{a,b\}}$ one obtains $\log \mu_a = \gamma_a$ that is the log-mean linear expansion relative to $Y_a$.

Under the dual pairwise Markov property, the model defined by the bidirected graph in Figure 5.2B is the log-mean linear model satisfying the constraint $\gamma_{\{a,b\}} = 0$,

$$\log \mu_{\{a,b,c\}} = \gamma_a + \gamma_b + \gamma_c + \gamma_{\{a,c\}} + \gamma_{\{b,c\}} + \gamma_{\{a,b,c\}}.$$

Accordingly, the distribution of $Y$ belongs to the (pairwise) bidirected graph model for the graph Figure 5.2C if and only if both $\gamma_{\{a,b\}}$ and $\gamma_{\{a,c\}}$ are constrained to vanish,

$$\log \mu_{\{a,b,c\}} = \gamma_a + \gamma_b + \gamma_c + \gamma_{\{b,c\}} + \gamma_{\{a,b,c\}}.$$

Finally, the distribution of $Y$ is pairwise Markov with respect to the graph in Figure 5.2D if and only if $\gamma_{\{a,b\}} = \gamma_{\{a,c\}} = \gamma_{\{b,c\}} = 0$, so that

$$\log \mu_{\{a,b,c\}} = \gamma_a + \gamma_b + \gamma_c + \gamma_{\{a,b,c\}}.$$

We now turn to the dual global Markov property. Consider the pairwise and the global bidirected graph models for the four graphs in Figure 5.2. For each of the graphs in Figures 5.2A and 5.2B the pairwise and the global Markov properties are equivalent. On the other hand, for the graph in Figures 5.2C the pairwise Markov property is different from the global Markov property, and the same is true for the graph in Figure 5.2D. This can be seen by noticing that in both graphs the subset $\{a, b, c\}$ is disconnected in such a way that the dual global Markov property implies $Y_a \perp\!\!\!\!\!\!\!\!\perp Y_{\{b,c\}}$. However, this independence is not
satisfied by any of the respective pairwise models because the interaction $\gamma_{\{a,b,c\}}$ is unconstrained. This suggests that the disconnected subsets of $V$ can be used to identify the vanishing log-mean linear interactions. The connection between zero restrictions in $\gamma$ and the dual global Markov property can be formally proved by applying Theorem 5.2 to the independencies implied by the connected set Markov property (Roverato et al., 2013, theorem 2).

**Proposition 5.4** Let $Y_V$ be a binary random vector with probability parameter $\pi > 0$ and let $\gamma = M^T \log \mu$ where $\mu = \mathbb{Z} \pi$. The distribution of $Y_V$ obeys the connected set Markov property with respect to a bidirected graph $G$ if and only if $\gamma_D = 0$ for every subset $D \subseteq V$ that is disconnected in $G$.

Consider the graph in Figure 5.2C. The connected subsets of $V$ are $\{a\}$, $\{b\}$, $\{c\}$ and $\{b,c\}$, whereas $\{a,b\}$, $\{a,c\}$ and $\{a,b,c\}$ are disconnected. Accordingly, the distribution of $Y$ is globally Markov with respect to this graph if and only if $\gamma_{\{a,b\}} = \gamma_{\{a,c\}} = \gamma_{\{a,b,c\}} = 0$.

It is interesting to compare the dimensions of models for undirected graphs with those of models for bidirected graphs. Let $G_1 = (V, E^\sim)$ and $G_2 = (V, E^{++})$ be an undirected and a bidirected graph, respectively, with the same skeleton, i.e. with the same missing edges. An instance of this situation is provided by the graphs in Figures 4.1 and 5.1. In the distributions which are globally Markov with respect to $G_1$ the unrestricted log-linear terms are those indexed by the complete subsets of $V$. On the other hand, in the distributions which are dual globally Markov with respect to $G_2$ the unrestricted log-mean linear terms are those indexed by the connected subsets of $V$. Because every complete subset is connected, it follows that the undirected graph model is typically defined by a smaller number of parameters than the relative bidirected graph model. Thus, undirected graph models are typically more parsimonious than bidirected graph models. In turn, the model defined by a bidirected graph under the global Markov property is typically more parsimonious than the model defined by the same graph under the pairwise Markov property.

**The Non-binary Case**

The use of log-mean linear models to parameterize bidirected graph models can be straightforwardly extended to categorical variables with
arbitrary number of levels as described in Roverato (2015). In this case
the log-mean linear expansion of the Möbius parameters is

$$\log \mu_{jD} = \sum_{D' \subseteq D} \gamma_{jD'}$$

for every $D \subseteq V$ and $j \in J$,

and the constraint $\gamma_D = 0$ in the binary case needs to be reformulated as
$\gamma_{jD} = 0$ for every $j \in J$. The connection between vanishing log-linear
parameters and marginal independence follows from Corollary 5.3,
which generalizes Theorem 5.2. This implies that, in the non-binary
case, the distribution of $Y$ obeys the dual pairwise Markov property with
respect to $G = (V, E^{\leftrightarrow})$ if and only if for every pair of distinct vertices
$a, b \in V$ such that $\{a, b\} \notin E^{\leftrightarrow}$ it holds that $\gamma_{j(a,b)} = 0$ for every $j \in J$.
Furthermore, the distribution of $Y$ obeys the connected set Markov
property with respect to a bidirected graph $G$ if and only if for every
subset $D \subseteq V$ that is disconnected in $G$ it holds that $\gamma_{jD} = 0$ for every
$j \in J$ (Roverato, 2015, corollary 1).

### 5.5 Example: Symptoms in Psychiatric Patients

Table 5.2 shows data from Coppen (1966) for a set of four binary
variables concerning symptoms in 362 psychiatric patients. The symptoms are: $Y_1 =$ stability (levels: extroverted, introverted),
$Y_2 =$ validity (levels: psychasthenic, energetic), $Y_3 =$ acute depression
(levels: yes, no) and $Y_4 =$ solidity (levels: hysteric, rigid).

In the graphical model literature, these data have been analyzed in
two different ways. Firstly, Wermuth (1976) applied a conditional
independence approach and selected the undirected graph model with

<table>
<thead>
<tr>
<th>Stability</th>
<th>Depression</th>
<th>Solidity</th>
<th>Hysteric</th>
<th>Rigid</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Validity</td>
<td>Psych.</td>
<td>Energetic</td>
</tr>
<tr>
<td>Extro.</td>
<td>No</td>
<td>12</td>
<td>47</td>
<td>8</td>
</tr>
<tr>
<td></td>
<td>Yes</td>
<td>16</td>
<td>14</td>
<td>22</td>
</tr>
<tr>
<td>Intro.</td>
<td>No</td>
<td>27</td>
<td>46</td>
<td>22</td>
</tr>
<tr>
<td></td>
<td>Yes</td>
<td>32</td>
<td>9</td>
<td>30</td>
</tr>
</tbody>
</table>
Figure 5.3 Graphs for Coppen data: (A) undirected graph, (B) bidirected graph, (C) directed graph including a vertex *U* for an unobserved variable.

graph in Figure 5.3A. Subsequently, Lupparelli *et al.* (2009) considered a marginal independence approach and obtained the bidirected graph model, under the dual global Markov property, for the graph in Figure 5.3B; see also Roverato *et al.* (2013). In applied work, the choice of the approach to follow usually depends on subject matter working hypotheses and background knowledge on the problem. This application to the Coppen data is of special interest because the two selected graphs have the same missing edges. It is thus of interest to compare the selected models with respect to their interpretation and statistical properties. Hereafter, we will denote the undirected and bidirected graphs in Figure 5.3 by $G^\sim = (V, E^\sim)$ and $G^{\leftrightarrow} = (V, E^{\leftrightarrow})$, respectively, where $V = \{1, 2, 3, 4\}$.

The undirected graph $G^\sim$ is decomposable with cliques $\mathcal{C} = \{\{1, 2\}, \{2, 3\}, \{3, 4\}\}$ and separators $\{\{2\}, \{3\}\}$. For decomposable models a closed-form expression for the ML estimates is available, and the application of Proposition 4.12 shows that the ML estimates of the expected cell counts of the model for $G^\sim$ can be computed from the marginal observed counts as follows,
\[ \hat{\mu}(Y = i) = \frac{n(Y_{1,2} = i_{1,2})n(Y_{2,3} = i_{2,3})n(Y_{3,4} = i_{3,4})}{n(Y_2 = i_2)n(Y_3 = i_3)} \]

for all \( i \in \mathcal{I} \).

The dimension of the saturated model is \( 2^{|V|} - 1 = 15 \), whereas the selected undirected graph model has dimension 7, with eight vanishing log-linear interactions,

\[ \log \pi_{\{1,2,3,4\}} = \lambda_0 + \lambda_1 + \lambda_2 + \lambda_3 + \lambda_{\{1,2\}} + \lambda_{\{2,3\}} + \lambda_{\{3,4\}}. \]

This selected model is characterized by the two conditional independencies \( Y_1 \perp Y_{\{3,4\}} | Y_2 \) and \( Y_4 \perp Y_{\{1,2\}} | Y_3 \), and it provides an adequate fit of the data with \( \text{dev}(G^\sim) = 13.9 \) on 8 degrees of freedom (\( p \)-value = 0.08). In order to interpret this model, we first recall that in undirected models every separation in the graph implies a connection between variables that is only indirect. This can be applied to see that, for example, depression (\( Y_3 \)) separates validity (\( Y_2 \)) from solidity (\( Y_4 \)), and therefore the latter two variables are not associated directly, but only in an indirect way, through depression. Furthermore, for the prediction of a variable it is sufficient to have the information provided by its neighbors, in the sense that the remaining variables become irrelevant. Hence, for example, to predict solidity (\( Y_4 \)) the variable depression (\( Y_3 \)) is sufficient because the remaining two variables provide no additional information with respect to that provided by depression. Likewise, stability (\( Y_1 \)) and depression (\( Y_3 \)) are sufficient to predict validity (\( Y_2 \)).

We now turn to the model for \( G^{\leftrightarrow} \) under the dual global Markov property. The graph \( G^{\leftrightarrow} \) has five disconnected sets, namely \( \{1, 3\} \}, \{1, 4\}, \{2, 4\}, \{1, 2, 4\} \) and \( \{1, 3, 4\} \). Accordingly, the bidirected graph model defined by \( G^{\leftrightarrow} \) has dimension 10 with five constrained log-mean linear interactions,

\[ \log \mu_{\{1,2,3,4\}} = \gamma_1 + \gamma_2 + \gamma_3 + \gamma_4 + \gamma_{\{1,2\}} + \gamma_{\{2,3\}} + \gamma_{\{3,4\}} + \gamma_{\{1,2,3\}} + \gamma_{\{2,3,4\}}. \]

As expected, this model is less parsimonious than the model defined by \( G^\sim \). Inferential issues concerning bidirected graph models are deferred to the next chapter. Here we mention that the computation of the ML estimates for this model requires the application of an iterative procedure. Furthermore, this model also provides an adequate fit of the data (deviance equal to 8.6 on 5 degrees of freedom, \( p \)-value = 0.13).
The interpretation of the model implied by $G^{\leftrightarrow}$ is less intuitive than that implied by $G\sim$. The bidirected graph model is characterized by the marginal independencies $Y_1 \perp Y_{\{3,4\}}$ and $Y_4 \perp Y_{\{1,2\}}$; that is, depression ($Y_3$) and solidity ($Y_4$) are jointly independent of stability ($Y_1$) and, moreover, stability ($Y_1$) and validity ($Y_2$) are jointly independent of solidity ($Y_4$). More interestingly, this model may suggest the presence of relevant unobserved variables. In the next chapter we will consider independence models defined by graphs with directed edges. We shall see that the graph in Figure 5.3C represents a multivariate regression model with three covariates, $Y_1$, $Y_4$ and $U$, and two responses, $Y_2$ and $Y_3$. The three covariates are mutually independent, and the responses are conditionally independent given the covariates. The system described by the graph in Figure 5.3C is thus clearly interpretable. When this system is marginalized over $U$, that is when $U$ is unobserved, the independence structure of the four observed variables is that given by $G^{\leftrightarrow}$, and therefore the interpretation of the model for $G^{\leftrightarrow}$ is compatible with that of a multivariate regression model with unobserved variables.

### 5.6 Parsimonious Graphical Modeling

A general difficulty when dealing with categorical variables is that the dimension of the parametric space increases exponentially with the number of variables. Specifically, the distribution of a vector of $|V| = p$ binary variables has $2^p - 1$ parameters, and the number of parameters increases to $(\prod_{v \in V} |\mathcal{I}_v|) - 1$ in the polytomous case. For this reason, it is important to be able to implement additional substantive constraints so as to specify parsimonious submodels. Parsimonious modeling is an important issue in graphical modeling, and more widely in multivariate analysis, but it is a problem of especial relevance in bidirected graph modeling because the number of parameters can be relatively large even for sparse graphs.

For the models considered in this text, one way to achieve parsimony is through the removal of additional interaction terms with respect to those implied by the relevant graphical model. However, it is always desirable that models are interpretable, and thus an important issue concerns the interpretation of the additional constraints. From this perspective, when the interest is in independence relationships, it is somehow natural to investigate the identification of constraints admitting an interpretation of this kind. Relevant work in this area has focused
on the implementation of context-specific independencies, i.e., conditional independencies that are not satisfied for every value of the conditioning variables but that only hold in specific contexts; see Boutilier et al. (1996), Corander (2003), Højsgaard (2004), Nyman et al. (2014), La Rocca and Roverato (2017) and references therein.

Context-specific independencies represent natural additional restrictions in the context of conditional independence models, but they do not suit well models of marginal independence. A natural approach for specifying parsimonious submodels of marginal independence models concerns the implementation of marginal independencies that hold between dichotomized versions of the variables. Indeed, one can notice that collapsing two or more levels of a non-binary variable into a single level can be regarded as a special type of marginalization. Accordingly, independence relationships that appear after some levels of a variable are collapsed into a single level can be regarded as special kinds of marginal independencies. The log-mean linear parametrization satisfies upward compatibility, but for variables with more than two levels it also satisfies a stronger property named dichotomization invariance (Roverato, 2015). This implies that marginal independencies between dichotomized versions of the variables correspond to zero log-mean linear terms. Interestingly, the bidirected graph of the model can be enlarged to accommodate such additional relationships, which can be read from the enlarged graph through the usual Markov properties. This approach is illustrated with an example, and we refer to Roverato (2015) for a formal presentation of this issue.

Table 5.3 gives a $4 \times 3 \times 3 \times 2$ contingency table originally described by Madsen (1976) that refers to a sample of 1681 residents of Copenhagen (see also Agresti, 2013, exercise 8.28). The variables are $Y_1 =$ type of housing (levels: $Ap =$ apartments, $At =$ atrium houses, $Te =$ terraced houses, $To =$ tower blocks), $Y_2 =$ feeling of influence on apartment management (levels: $L_2 =$ low, $M_2 =$ medium, $H_2 =$ high), $Y_3 =$ satisfaction with housing conditions (levels: $L_3 =$ low, $M_3 =$ medium, $H_3 =$ high) and $Y_4 =$ degree of contact with other residents (levels: $L_4 =$ low, $H_4 =$ high).

An exhaustive search within the family of bidirected graph models led to the selection of the graph in Figure 5.4 which encodes the structural relationship $Y_3 \perp Y_4$ (Roverato, 2015). The selected model has deviance 5.13 on 2 degrees of freedom ($p$-value = 0.08). Parsimony is not achieved by this model that has 69 parameters, to be compared
<table>
<thead>
<tr>
<th>Housing Type</th>
<th>Contact: Low</th>
<th>Satisfaction: Medium</th>
<th>Satisfaction: High</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tower blocks</td>
<td>Low: 21</td>
<td>14</td>
<td>37</td>
</tr>
<tr>
<td></td>
<td>Medium: 34</td>
<td>12</td>
<td>40</td>
</tr>
<tr>
<td></td>
<td>High: 10</td>
<td>11</td>
<td>23</td>
</tr>
<tr>
<td>Apartments</td>
<td>Low: 61</td>
<td>28</td>
<td>36</td>
</tr>
<tr>
<td></td>
<td>Medium: 43</td>
<td>17</td>
<td>36</td>
</tr>
<tr>
<td></td>
<td>High: 26</td>
<td>11</td>
<td>5</td>
</tr>
<tr>
<td>Atrium houses</td>
<td>Low: 13</td>
<td>9</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>Medium: 8</td>
<td>8</td>
<td>12</td>
</tr>
<tr>
<td></td>
<td>High: 6</td>
<td>7</td>
<td>9</td>
</tr>
<tr>
<td>Terraced houses</td>
<td>Low: 18</td>
<td>13</td>
<td>23</td>
</tr>
<tr>
<td></td>
<td>Medium: 15</td>
<td>13</td>
<td>31</td>
</tr>
<tr>
<td></td>
<td>High: 7</td>
<td>5</td>
<td>11</td>
</tr>
</tbody>
</table>

Table 5.3: Data from Madsen (1976) on satisfaction with housing.
with the dimension of the saturated model that is equal to 71. Notice also that the analysis of these data in the context of undirected graph models would not give a more parsimonious model because no pairwise conditional independence can be identified on the basis of the asymptotic chi-squared distribution of the deviance.

Let $Y_v$ be a variable with state space $\mathcal{I}_v = \{0, 1, \ldots, d_v\}$ and restricted state space $\mathcal{J}_v = \{1, \ldots, d_v\}$. Every level $i \in \mathcal{I}_v$ of $Y_v$ is associated with a binary variable obtained by collapsing all the levels, with the exception of $i$, into a single level. More formally we let

$$X^i_v = \begin{cases} 
1 & \text{if } Y_v = i \\
0 & \text{otherwise.}
\end{cases}$$

This gives $|\mathcal{I}_v| = d_v + 1$ binary variables, and one can replace $Y_v$ with this collection of binary variables with no loss of information. On the other hand, this set of variables is redundant because any subset of $d_v$ of such binary variables is sufficient to recover $Y_v$. Hence, we replace $Y_v$ with the vector of binary variables indexed by its restricted state space; that is, we let $X_{\mathcal{J}_v} = (X^1_v, \ldots, X^{d_v}_v)$ and call this vector a binary expansion of $Y_v$. The dichotomization invariance property of the log-mean linear parameterization implies that, if the restricted state space used to expand the variables is the same used in the computation of the log-mean linear parameter, then independence relationships between the entries of the binary expansions are equivalent to zero constraints in the log-mean linear parameters. Furthermore, it is possible to represent the independence structure through an expanded bidirected graph that is a bidirected graph where one or more variables are replaced by their binary expansions. This graph provides a complete description of the constraints characterizing the submodel, which are all feasible of an interpretation in terms of marginal independencies. Figure 5.5 shows

**Figure 5.4** Housing data: bidirected graph of the selected model.
two expanded bidirected graphs for the housing data where, to improve readability, the gray color is used to depict the complete subgraphs of the expanded variables.

The restricted state space is obtained by removing one level from $I_v$, and the choice of such baseline level is arbitrary. Likewise, one can replace any variable $Y_v$ with one binary vector arbitrarily chosen among $d_v + 1$ different, but equivalent, binary expansions. One drawback of this approach is therefore that the selected model depends on the chosen baseline level. On the other hand, sometimes there exists a natural way to choose the baseline level. For instance, for the housing data the variable $Y_2$ has three ordinal levels, $L_2 = \text{low}$, $M_2 = \text{medium}$, $H_2 = \text{high}$; and the central level $\text{medium}$ is a natural baseline because it makes sense to collapse it with either the level $\text{low}$ or the level $\text{high}$ leading to the binary expansions $(X^L_2, X^H_2)$. Similarly, a natural binary expansion of $Y_3$ is $(X^L_3, X^H_3)$. The variable $Y_1$ has no natural baseline and in the application presented below the level $\text{tower blocks}$ is set as baseline thereby obtaining the binary expansion $(X^{Ap}_1, X^{At}_1, X^{Te}_1)$. The variable $Y_4$ is binary and, therefore, it cannot be expanded.

Roverato (2015) applied this approach to the housing data and identified the model with graph in Figure 5.5A, which has deviance equal to 34.34 on 23 degrees of freedom ($p$-value = 0.06). This submodel is much more parsimonious than the model in Figure 5.4 with 21 additional vanishing log-mean linear terms that can be interpreted as $X^{Ap}_1 \perp (Y_2, Y_3, Y_4)$ and $X^{At}_1 \perp (X^{L_2}_2, X^{H_3}_3)$. The graph in Figure 5.5A fully identifies the selected model, but has the disadvantage of being dense and thus difficult to read. For this reason, we also give the graph in

![Figure 5.5 Housing data: expanded bidirected graphs.](https://www.cambridge.org/core/core/559BC2/289251CEF3077F20D)
Figure 5.5B, where only variable $Y_1$ is expanded. The latter graph fails to represent the independence $X_1^{A_1} \perp (X_2^{L_2}, X_3^{H_3})$, but may be regarded as a good compromise between readability of the graphical representation and completeness of the information provided by the graph.

A relevant feature of this approach is that the variables to be expanded do not need to be defined a priori, but they can be specified after a model has been selected from data. For instance, the set of variables to expand can be chosen so as to optimize the trade-off between readability of the graph and the need to explicitly represent the learnt independencies involving every single expanded variable, as shown in the application to the housing data. In this sense, it is possible to see the operation of variable expansion in the context of an interactive display, whereby the analyst could select the nodes to be expanded by, for instance, clicking on them.
6

Directed Acyclic and Regression Graph Models

In this chapter we introduce graphs containing directed edges, which are used to deal with asymmetric relationships between variables. We first consider statistical models associated with a family of graphs, called directed acyclic graphs, whose edge set only comprises directed edges. These models have a long history that may be traced back to the pioneering work of Wright (1921) on path analysis, and they are also commonly known in the literature as Bayesian networks (Pearl, 1986). Graphical models based on directed acyclic graphs constitute a very general framework. They can be used to represent several classical statistical formalisms such as linear structural equation models, models for factor analysis and various types of regression and latent variable models. Furthermore, they have proved useful in a wide range of applications including expert systems (Pearl, 1988; Cowell et al., 1999), causal analysis (Spirtes et al., 2000; Lauritzen, 2001; Pearl, 2009) and machine learning (Barber, 2012). Next, we will consider models for graphs, called regression graphs, that comprise the three kinds of edges: undirected, bidirected and directed. In a directed graph arrows point from vertices associated with explanatory variables to vertices associated with response variables. For example, the graph $a \rightarrow b \rightarrow c$ implies a relationship between variables where $Y_a$ is a purely explanatory variable, $Y_b$ is an intermediate variable, being both an explanatory variables of $Y_c$ and a response of $Y_a$, and $Y_c$ is a primary response. Regression graph models extend this idea to a multivariate regression setting (Cox and Wermuth, 1993, 1996; Wermuth and Sadeghi, 2012; Wermuth and Cox, 2015). In this approach, variables are arranged in a sequence of blocks ordered on the basis of subject matter considerations. Every block contains variables which are multiple responses of the previous blocks and explanatory of the following blocks. The first block has no incoming arrows, and therefore it contains purely explanatory variables, also called context variables. Likewise,
the last block contains the primary responses. The variables within each block are considered to be on an equal standing, with the relationships between context variables represented by an undirected graph and those in each block of response variables by a bidirected graph. Graphical models for regression graphs form a flexible class that includes, as special cases, undirected, bidirected and directed acyclic graph models. Furthermore, they have close connections with other relevant families of graphical models such as models for multivariate regression chain graphs (Drton, 2009) and acyclic directed mixed graphs (Richardson, 2003).

In the parameterization of regression graph models we will follow the approach of La Rocca and Roverato (2017). Alternative approaches available in the literature are either based on a collection of marginal and conditional probabilities called (generalized) Möbius parameters (Drton, 2009; Evans and Richardson, 2010, 2014) or on the theory of marginal log-linear models (Bergsma and Rudas, 2002; Bartolucci et al., 2007; Evans and Richardson, 2013; Marchetti and Lupparelli, 2011; Rudas et al., 2010).

6.1 Directed Acyclic Graphs

Let $G = (V, E)$ be a graph. A directed edge is an ordered pair $(a, b) \in E$ representing an asymmetric relationship between the two vertices $a, b \in V$. The edge $(a, b)$ is then depicted as an arrow pointing from $a$ to $b$, $a \rightarrow b$, and in this case $a$ is called a parent of $b$ whereas $b$ is a child of $a$. A graph with only directed edges is called directed, and when we want to highlight that an edge set only comprises directed edges we write it as $E^-$. All the graphs we consider are acyclic. Specifically, a directed acyclic graph (DAG) is a directed graph with no directed cycles in the sense that it is not possible to return to any vertex by following the directions of the arrows; see Figure 6.1 for an example of a DAG.

A path is defined for directed graphs like in the undirected case, in the sense that no role is played by the directions of its arrows. Formally, in a directed graph a path with endpoints $a$ and $b$ is a sequence $a = v_0, \ldots, v_k = b$ of distinct vertices such that either $v_{r-1} \rightarrow v_r$ or $v_{r-1} \leftarrow v_r$ for all $r = 1, \ldots, k$. Examples of paths in the graph of Figure 6.1 are $(b, e, g, f)$, $(b, e, c, a, d, f)$ and $(a, d, f, g, h)$. A non-endpoint vertex $z$ of a path is called a collider on the path if the
edges of the path meet head-to-head at $z$, and it is a non-collider otherwise. Hence, the vertex $e$ is a collider on the path $(b, e, c, a, d, f)$, whereas $c$, $a$ and $d$ are non-collider vertices on the same path. Conversely, $e$ is a non-collider vertex on the path $(b, e, g, f)$. When all the edges of the path $a = v_0, \ldots, v_k = b$ are arrows pointing in the direction of $b$, that is $v_{r-1} \rightarrow v_r$ for all $r = 1, \ldots, k$, then the path is said to be directed from $a$ to $b$. Thus, $(a, d, f, g, h)$ is the only directed path among those considered in the example above.

We now introduce some relevant sets of vertices of a DAG $G = (V, E)$. Given a vertex $b \in V$, the set of all $v \in V$ such that $v \rightarrow b$ is the set of parents of $b$, denoted by $\text{pa}_G(b)$. If there is a directed path from $a$ to $b$ then the vertex $b$ is called a descendant of $a$, and the set of descendants of $a$ is denoted by $\text{de}_G(a)$. Finally, the set of non-descendants of a vertex $a$ is defined as $\text{nd}_G(a) = V \setminus (\text{de}_G(a) \cup \{a\})$. For example, in the graph of Figure 6.1, $\text{pa}(g) = \{e, f\}$, $\text{de}(c) = \{e, g, h\}$, whereas $\text{nd}(c) = \{a, b, d, f\}$. Recall that we omit the subscript when it is clear from the context which graph is being considered.

The vertices of a DAG can always be well-ordered so that if two vertices form an edge, the arrow points from the lower to the higher of the two vertices with respect to the ordering. Notice that a DAG may not have a unique well-ordering. For example, two well-orderings of the vertices of the graph in Figure 6.1 are $(a, b, c, d, e, f, g, h)$ and $(b, a, d, f, c, e, g, h)$. If a graph $G$ is well-ordered, the set of predecessors of a vertex $a$, denoted by $\text{pr}_G(a)$, is the set of those vertices that have a lower number than $a$. Hence, $\text{pr}(f) = \{a, b, c, d, e\}$ in the well-ordering $(a, b, c, d, e, f, g, h)$, whereas in the well-ordering $(b, a, d, f, c, e, g, h)$ we have $\text{pr}(f) = \{b, a, d\}$. It follows from the definition of descendant that for any vertex $a \in V$ of a well-ordered DAG it holds that $\text{pr}(a) \cap \text{de}(a) = \emptyset$ and therefore that $\text{pr}(a) \subseteq \text{nd}(a)$.
6.2 Markov Properties for Directed Acyclic Graphs

As for the two types of graphs considered so far, also for DAGs there exist a Markov property named pairwise that concerns independence relationships between pairs of variables, a local Markov property focusing on the relationship between every single variable and all the remaining variables, and a more general global Markov property involving independencies identified through a separation criterion. However, in this case an additional ordered Markov property is available. It applies to well-ordered DAGs, and suits the case when background knowledge on the causal or temporal ordering of the variables allows one to specify the directions of the arrows, and thus an ordering of the vertices.

The Directed Pairwise Markov Property

The probability distribution of a random vector \( Y_V \) is said to obey the directed pairwise Markov property with respect to the DAG \( G = (V, E^-) \) if for every pair \( a, b \in V \) of non-adjacent vertices with \( b \in \text{nd}(a) \) it holds that

\[
Y_a \perp Y_b | Y_{\text{nd}(a) \setminus \{b\}}.
\]

Examples of conditional independencies implied by the pairwise Markov property for the DAG of Figure 6.1 are \( Y_h \perp Y_a | Y_{\{b,c,d,e,f,g\}} \), \( Y_g \perp Y_c | Y_{\{a,b,d,e,f\}} \) and \( Y_a \perp Y_b \). Notice that this property is less interesting than the pairwise properties for undirected and bidirected graphs because the specification of the conditioning set is less natural.

The Directed Local Markov Property

The probability distribution of a random vector \( Y_V \) is said to obey the directed local Markov property with respect to the DAG \( G = (V, E^-) \) if for every \( a \in V \) it holds that

\[
Y_a \perp Y_{\text{nd}(a) \setminus \text{pa}(a)} | Y_{\text{pa}(a)}.
\]

Hence, for the DAG of Figure 6.1 the local Markov property implies, for instance, \( Y_g \perp Y_{\{a,b,c,d\}} | Y_{\{e,f\}} \), \( Y_f \perp Y_{\{a,b,c,e\}} | Y_d \) and \( Y_a \perp Y_b \).
The Ordered Markov Property

The probability distribution of a random vector $Y_V$ is said to obey the ordered Markov property with respect to the well-ordered DAG $G = (V, E^-)$ if for every $a \in V$ it holds that

$$Y_a \perp Y_{\text{pr}(a) \setminus \text{pa}(a)} | Y_{\text{pa}(a)}.$$  

For the DAG in Figure 6.1, the ordered Markov property with respect to the well-order $(a, b, c, d, e, f, g, h)$ implies, for instance, $Y_c \perp Y_{\{a,d\}} | Y_{\{b,c\}}$ and $Y_e \perp Y_b | Y_a$.

The ordered Markov property replaces all non-descendants of $a$, $\text{nd}(a)$, in the local Markov property by its predecessors $\text{pr}(a)$ in some given well-ordering of the vertices. By noticing that $\text{pr}(a) \subseteq \text{nd}(a)$ for every $a \in V$ and for every well-ordering of the vertices, it follows immediately that the local directed Markov property implies the ordered Markov property with respect to any well-ordering of the vertices. In fact, we will see below that the ordered and the local Markov properties are equivalent because it is also true that the former implies the latter. This also shows that, if the distribution of $Y$ is ordered Markov with respect to a given well-ordering, then it is ordered Markov with respect to all well-orderings of the vertices.

The Directed Global Markov Property

The global Markov property for DAGs requires the introduction of a kind of separation that is more involved than in the undirected and bidirected cases.

A path between $a$ and $b$ in a DAG $G = (V, E^-)$ is blocked given a set $S \subseteq V$ if there is an inner vertex $z$ of the path such that

(b1) either $z$ is a non-collider on the path and $z \in S$,

(b2) or $z$ is a collider on the path and neither $z$ nor any descendant of $z$ is in $S$; i.e., $(\text{de}(z) \cup \{z\}) \cap S = \emptyset$.

A path that is not blocked is a connecting path. More specifically, a path between $a$ and $b$ in $G$ is connecting given a set $S \subseteq V$ if the following conditions are both satisfied:

(c1) every non-collider on the path is not contained in $S$;

(c2) every collider on the path is either in $S$ or has a descendant in $S$. 

Clearly, condition (b1) is satisfied for some vertex if and only if (c1) is not true and, likewise, condition (b2) holds true for some vertex if and only if (c2) is false.

We illustrate the notion of blocked and connecting paths by means of some examples. The most basic instance of a path containing inner vertices is called a tripath and has exactly three vertices, two of which are endpoints of the path and one is an inner vertex. Figure 6.2 displays four DAGs each containing a tripath \((a, b, c)\). In the tripaths of Figure 6.2A and Figure 6.2B the inner vertex \(b\) is a non-collider and therefore these two paths are blocked given \(S = \{b\}\), but they are both connecting paths given \(S = \emptyset\). In the DAGs of Figure 6.2C and Figure 6.2D the inner vertex \(b\) is a collider and therefore in both DAGs the path \((a, b, c)\) is blocked given \(S = \emptyset\), but connecting given \(S = \{b\}\). Furthermore, in the DAG of Figure 6.2D the vertex \(d\) is a descendant of the collider \(b\) and therefore the path \((a, b, c)\) is connecting given \(S = \{d\}\), and also given \(S = \{b, d\}\). We now turn to the DAG of Figure 6.1. The path \((b, e, g, f)\) is blocked, for instance, given any of the sets \(S = \{e, g\}\) (the vertex \(e\) satisfies condition (b1)) \(S = \emptyset\) and \(S = \{a, c, d\}\) (in both cases the vertex \(g\) satisfies condition (b2)). On the other hand, the latter is a connecting path given, for instance, each of the sets \(S = \{g\}, S = \{h\}\) and \(S = \{a, c, d, h\}\). Finally, consider the path \((b, e, c, a, d, f)\). Examples of sets blocking this path are \(S = \emptyset\).
(the vertex \( e \) satisfies condition (b2)) and \( S = \{ d, h \} \) (the vertex \( d \) satisfies condition (b1)). Conversely, this is a connecting path given, for instance, \( S = \{ e \} \), \( S = \{ h \} \) and \( S = \{ g, h \} \).

The notion of separation in a DAG \( G = (V, E^-) \) was introduced by Pearl (1988). Two subsets \( A, B \subseteq V \) are \( d \)-separated by \( S \subseteq V \) if all paths from \( A \) to \( B \) are blocked by \( S \). To see an example, in the DAG of Figure 6.1 there are two paths between the sets \( A = \{ b \} \) and \( B = \{ f \} \), specifically \( (b, e, g, f) \) and \( (b, e, c, a, d, f) \). These are both blocked given \( S = \emptyset \) and therefore \( A \) and \( B \) are separated by the empty set. Other instances of separating sets are \( S = \{ e, d, g, f \} \), \( S = \{ a, d, g, f \} \) and \( S = V \setminus \{ b, f \} = \{ a, c, d, e, g, h \} \). In the same DAG, three sets separating \( A = \{ b, e \} \) from \( B = \{ d, f \} \) are \( \{ a \} \), \( \{ c \} \) and \( \{ a, c \} \). Notice that neither \( S = \emptyset \) nor \( S = V \setminus (A \cup B) = \{ a, c, g, h \} \) \( d \)-separate \( \{ b, e \} \) from \( \{ d, f \} \).

The probability distribution of a random vector \( Y_V \) is said to obey the directed global Markov property with respect to the DAG \( G = (V, E^-) \) if for any triple of pairwise disjoint subsets \( A, B, S \subseteq V \) such that \( S \) \( d \)-separates \( A \) from \( B \) in \( G \) it holds that

\[
Y_A \perp \!\!\! \perp Y_B | Y_S.
\]

Hence, if \( Y \) is globally Markov with respect to the DAG of Figure 6.1 then, for instance, we have \( Y_b \perp Y_f \), \( Y_b \perp Y_f | Y_V \setminus \{ b, f \} \) and \( Y_{\{ b, e \}} \perp Y_{\{ d, f \}} | Y_{\{ a, c \}} \).

The criterion of \( d \)-separation might be difficult to verify, and therefore the independence relationships implied by the global Markov property for DAGs are less intuitive than in the undirected and bidirected cases. However, this Markov property is important because Geiger and Pearl (1988) showed that it is complete, so that the \( d \)-separation criterion cannot be improved; see also Meek (1995).

**The Factorization Property**

Assume that the distribution of \( Y_V \) admits a density \( f(\cdot) \) with respect to a product measure. Then the distribution of \( Y_V \) is said to factorize over the DAG \( G = (V, E^-) \) if its density has the form

\[
f(y) = \prod_{v \in V} f(y_v | y_{pa(v)}),
\]  

(6.1)

where \( f(\cdot | \cdot) \) are conditional densities.
In the directed case the relationships among alternative Markov properties, and between Markov properties and the factorization property, are simpler than in the undirected and bidirected cases. It can be shown (see, e.g. Cowell et al., 1999, theorem 5.14) that if the distribution of $Y$ admits a density, then the factorization property, the directed local, ordered and global Markov properties are all equivalent, i.e.,

$$(\text{factorization}) \iff (\text{global}) \iff (\text{local}) \iff (\text{ordered}).$$

In fact, Lauritzen et al. (1990, propositions 4 and 5) showed that the ordered, local and global Markov properties are equivalent even without assuming the existence of a density. The directed pairwise Markov property is implied by each of the four properties above, but the converse is not true in general. A sufficient condition for the equivalence of the pairwise Markov property with the other four properties is that the density of $Y$ is positive (see Lauritzen, 1996, section 3.2.2).

**Markov Equivalence**

It is possible that two different graphs, under the respective Markov properties, encode the same collection of independence relationships. In this case the two graphs define the same statistical model and are said to be *Markov equivalent*. For example, the collection of independence relationships implied by the DAG in Figure 6.2A comprises the unique independence $Y_a \perp Y_c | Y_b$. Also the DAG in Figure 6.2B only implies the relationship $Y_a \perp Y_c | Y_b$ and therefore the two DAGs are Markov-equivalent. Furthermore, these two DAGs are both Markov-equivalent to the undirected graph $a \rightarrow b \rightarrow c$. The DAG in Figure 6.2A encodes the independence $Y_a \perp Y_c$ and it is Markov-equivalent to the bidirected graph $a \leftrightarrow b \leftrightarrow c$. Markov equivalence is an important question in graphical modelling with relevant implications in many domains such as structure learning (Drton and Maathuis, 2017) and causal discovery (Spirtes et al., 2000; Pearl, 2009). We do not provide additional details on this subject, and the interested reader is referred to Frydenberg (1990), Andersson et al. (1997), Pearl and Verma (1990), Volf and Studený (1999), Chickering (2002), Roverato (2005), Roverato and Studený (2006), Roverato and La Rocca (2006), Drton and Richardson (2008B), Ali et al. (2009), Wermuth and Sadeghi (2012) and references therein for a comprehensive presentation of the theory of Markov equivalence and related problems.
6.3 Regression Graphs

Up to now we have considered graphs consisting of a single type of edge. Here we introduce mixed graphs whose edge set may contain undirected, bidirected and directed edges. Our main focus is on the class of *regression graphs* which belong to the wider family of *chain graphs*. In a chain graph the vertex set can be partitioned into an ordered sequence of pairwise disjoint blocks. Directed edges are not allowed within blocks, and all the edges joining vertices belonging to different blocks are arrows pointing from the lower to the higher of the two blocks with respect to the ordering. The connected components of the blocks are called *chain components*, and we denote the set of chain components of a chain graph by $\mathcal{K} = \{K_1, \ldots, K_s\}$. Notice that the partition of the vertex set into blocks may not be unique, but the set of chain components is unique and it does not depend on the block partition. A chain graph where all chain components are singletons is a DAG.

A regression graph is a chain graph $G = (V, E)$ whose vertex set can be partitioned into an ordered sequence of blocks $(T_0, T_1, \ldots, T_k)$ such that the subgraph $G_{T_0}$ induced by the first block $T_0$ is undirected, whereas the subgraphs induced by the remaining blocks, $G_{T_r}$ for $r = 1, \ldots, k$, are all bidirected. The graph in Figure 6.3 is a regression graph with blocks $T_0 = \{a, b, c, d, e\}$, $T_1 = \{f, g, h\}$ and $T_2 = \{l, m, q, r\}$, whereas the set of chain components is $\mathcal{K} = \{\{a, b, c\}, \{d, e\}, \{f, g, h\}, \{l, m, q, r\}\}$.

The notion of parent, given for DAGs, can also be applied to chain graphs. If $G = (V, E)$ is a chain graph then the parents of the vertex

![Example of regression graph with blocks ($T_0$, $T_1$, $T_2$).](https://www.cambridge.org/core/product/F9D4523D3559BC2BA9251CEF3077F20D)
$a \in V$, $\text{pa}_G(a)$, is the set of vertices $v \in V$ such that $G$ contains the edge $v \rightarrow a$. Furthermore, the set of parents of a subset $A \subseteq V$ that is defined as $\text{pa}_G(A) = \bigcup_{a \in A} \text{pa}(a) \setminus A$. Hence, in the graph of Figure 6.3 it holds, for example, that $\text{pa}(l) = \{c, h\}$, $\text{pa}(q) = \emptyset$ and $\text{pa}\{l, q\} = \{c, h\}$. Furthermore, $\text{pa}(T_2) = \{c, f, h\}$, $\text{pa}(T_1) = \{e\}$ and $\text{pa}(T_0) = \emptyset$. For a regression graph with block ordering $(T_0, T_1, \ldots, T_k)$, we define the sets of predecessors of a bidirected block $T_r$ as $\text{pr}_G(T_r) = T_0 \cup \ldots \cup T_{r-1}$ for $r = 1, \ldots, k$. Thus, in the graph of Figure 6.3 we have $\text{pr}(T_2) = \{a, b, c, d, e, f, g, h\}$ and $\text{pr}(T_1) = \{a, b, c, d, e\}$.

### 6.4 Markov Properties for Regression Graphs

The Markov property for regression graphs considered here unifies the global Markov properties for undirected graphs, bidirected graphs and DAGs. Consider a random vector $Y_V$ and a regression graph $G = (V, E)$ with ordered sequence of blocks $(T_0, T_1, \ldots, T_k)$. The distribution of $Y_V$ obeys the ordered regression graph Markov property with respect to $G = (V, E)$ when

(RG.d) $Y_A \perp Y_{\text{pr}(T_r) \setminus \text{pa}(A)} | Y_{\text{pa}(A)}$ for all $A \subseteq T_r$ and every $r = 1, \ldots, k$;

(RG.b) the distribution of $Y_{T_r} | Y_{\text{pr}(T_r)}$ obeys the dual global Markov property with respect to the induced subgraph $G_{T_r}$, for every $r = 1, \ldots, k$;

(RG.u) the distribution of $Y_{T_0}$ obeys the global Markov property with respect to the induced subgraph $G_{T_0}$.

Conditions (RG.b) and (RG.u) concern the independence relationships implied by missing bidirected and undirected edges, respectively. Condition (RG.d) is well-defined when both $A \neq \emptyset$ and $\text{pr}(T_r) \setminus \text{pa}(A) \neq \emptyset$, and it gives the independence relationships implied by missing arrows. When all blocks are singletons, (RG.d) simplifies to the ordered directed Markov property for DAGs.

This formulation of the Markov property for regression graphs was first given in Marchetti and Lupparelli (2011), where it is shown that models defined by the above conditions do not depend on the chosen block ordering of variables; see also Drton (2009). The ordered Markov property is the only Markov property that we introduce for regression graphs, and in the following we will shortly say that a distribution is Markov with respect to a regression graph. The reader interested in the alternative Markov properties for regression graphs can find the theory...
relative to the pairwise Markov property in Sadeghi and Wermuth (2016). A global Markov property for regression graphs can be obtained by exploiting the Markov properties defined for other classes of models for mixed graphs, which have close connections with regression graphs. Several classes of graphical models for mixed graphs are available; see Sadeghi and Lauritzen (2014) for an overview. A relevant subfamily of models for mixed graphs is known in the literature as multivariate regression chain graphs (MRCGs), but it was also called chain graphs of type IV by Drton (2009). MRCGs belong to the wider class of acyclic directed mixed graphs (ADMGs), and there exists a global Markov property for ADMG models. This is based on the notion of \( m \)-separation, that is an extension of the \( d \)-separation criterion (Richardson and Spirtes, 2002; Richardson, 2003; Sadeghi and Lauritzen, 2014). An MRCG is a regression graph without the undirected block, i.e., with \( T_0 = \emptyset \), and it can be shown that in this case the ordered Markov property for regression graphs is equivalent to the global Markov property for MRCGs (Drton, 2009; Marchetti and Lupparelli, 2011). This equivalence can be extended to the case where \( T_0 \neq \emptyset \) by considering the global Markov property for the undirected component and the ADMG Markov property for the distribution of \( Y_V \) conditional on the undirected component.

Assume that the distribution of \( Y_V \) admits a density \( f(\cdot) \) with respect to a product measure. If \( Y \) is Markov with respect to a regression graph \( G = (V, E) \) then the density factorizes as,

\[
f(y) = \prod_{K \in \mathcal{K}} f(y_K | y_{\text{pa}(K)});
\]

see Lauritzen (1996, section 3.2.3) and Drton (2009). This factorization is essentially identical to the factorization (6.1) for DAGs. However, in general the factorization (6.2) is not equivalent to the Markov property for regression graphs because it typically implies only a subset of the independence relationships implied by the Markov property.

### 6.5 On the Interpretation of Models defined by Regression Graphs

The family of chain graphs first introduced in the graphical model literature has two types of edges, directed and undirected (Lauritzen and Wermuth, 1989). Accordingly, the classical Markov property for
chain graphs is a natural generalization of both the Markov property for undirected graphs and DAGs (Frydenberg, 1990). One motivation for extending DAGs to graphs including undirected edges was to allow the modelling of “symmetric associations” and “simultaneous responses” (Lauritzen and Richardson, 2002). More recently, chain graph models have been extended to include a third type of edge, i.e., bidirected arrows, to express a second type of symmetric relationship between variables, and today four different Markov properties are available for chain graphs (Frydenberg, 1990; Wermuth and Lauritzen, 1990; Andersson et al., 2001; Drton, 2009).

We focus on regression graphs because, whenever it is reasonable to assume the existence of a partial ordering of the variables, these models play a central role among the existing families of chain graph models; see Wermuth and Sadeghi (2012) and Wermuth and Cox (2015) for further discussion on this issue. As far as interpretation is of concern, in regression graph models the distribution of every block conditioned on the previous blocks is a multivariate regression and, as such, it focuses on the marginal distributions of response variables (see, e.g. Whittaker, 1990, section 10.5). To fix ideas, consider the model implied by the graph in Figure 6.3. In the multivariate regression with response $Y_T$, one can restrict the attention to specific subsets of responses, for example $Y_l$ and $Y_{\{m,q\}}$, and apply (RG.d) to read from the graph that $Y_l \independent Y_{\{a,b,e,d,f,g\}}|Y_{\{c,h\}}$ and that $Y_{\{m,q\}} \independent Y_{T_0 \cup \{g\}}|Y_{\{f,h\}}$. Likewise, in the regression with response $Y_{T_1}$ it holds for example that $Y_{\{g,h\}} \independent Y_{\{a,b,c,d\}}|Y_e$.

Regression graph models also model the possible presence of residual associations among the responses. This is done through a bidirected graph and, in this regard, it is worth mentioning that bidirected edges are feasible of an interpretation in terms of latent variables; see, e.g., Richardson and Spirtes (2002) and Evans (2016). For example, the DAG model with graph

$$a \leftarrow u_1 \rightarrow b \leftarrow u_2 \rightarrow c$$

implies the marginal independence of $Y_a$ and $Y_c$ so that, if the variables for the vertices $u_1$ and $u_2$ are unobserved, then the independence structure of the observed variables is provided by the bidirected graph

$$a \leftrightarrow b \leftrightarrow c.$$
Another example is given in the application of Section 5.5 where the DAG in Figure 5.3C implies, for the four observed variables, the independence structure encoded by the regression graph that is also Markov equivalent to the bidirected graph of Figure 5.3B.

We now turn to the first block $T_0$ of context variables. For background variables, which are purely explanatory, no incoming arrow is assumed, and thus the association structure may be suitably represented by means of an undirected graph that provides insight into the conditional independence relationship holding among these variables.

### 6.6 The Log-hybrid Linear Parameterization

As a first step towards the parameterization of regression graph models, we consider the case where $V$ is partitioned into two subsets, which we also call blocks, $T$ and $U$. We write this partition as an ordered pair $(U, T)$ to underline the different role played by $U$, that indexes a vector of explanatory variables $Y_U$, with respect to $T$, that indexes a vector of response variables $Y_T$. Relative to this partition, we describe a parameterization that unifies the log-linear and the log-mean linear approaches, in the sense that independence relationships among explanatory variables correspond to the vanishing of interactions associated with cross-product ratios of the “conditional” type, $\text{cpr}_{cc}()$, whereas independence relationships among response variables correspond to the vanishing of interactions associated with cross-product ratios of the “marginal” type, $\text{cpr}_{mm}()$. The novel feature of this parameterization concerns independence relationships between variables belonging to different blocks, which correspond to the vanishing of interactions associated with cross-product ratios of the “marginal-conditional” type, $\text{cpr}_{mc}()$.

In the following, we first introduce a generalization of the zeta matrix, and then we show how this can be used in the computation of the parameters of interest and in the derivation of their properties.
As usually done in this text, the binary case is considered in detail whereas for the general case of polytomous variables only a list of technical results is provided.

**The Matrices** $\mathbb{I}_U \otimes \mathbb{Z}_T$ and $\mathbb{I}_U \otimes \mathbb{M}_T$

In order to facilitate the computation of some relevant quantities, it is useful to extend the zeta and the Möbius matrices so as to explicitly take into account the asymmetric block partitioning of the vertices. This is done by associating an identity matrix with the block $U$. Initially, we consider singleton sets, and denote by $\mathbb{I}_v$ the identity matrix associated with $v \in V$, that is

$$
\begin{pmatrix}
\emptyset & \{v\} \\
\{v\} & \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}
\end{pmatrix} = \mathbb{I}_v,
$$

where the matrix is displayed with its row and column indexes. Next, we define the identity matrix associated with a subset $A \subseteq V$ as

$$
\mathbb{I}_A = \bigotimes_{v \in A} \mathbb{I}_v,
$$

with the convention that $\mathbb{I}_\emptyset = 1$. The rows and columns of $\mathbb{I}_A$ are indexed by the subsets of $A$, so that the entry of $\mathbb{I}_A$ indexed by the pair $D,E \subseteq A$ is $(\mathbb{I}_A)_{D,E} = 1(D = E)$; it is also a well-known fact that $\mathbb{I}_A^{-1} = \mathbb{I}_A$. The set $U$ can thus be associated with $\mathbb{I}_U$, the set $T$ with $\mathbb{Z}_T$, and the pair $(U, T)$ with the Kronecker product of $\mathbb{I}_U$ and $\mathbb{Z}_T$; i.e., $\mathbb{I}_U \otimes \mathbb{Z}_T$. By applying the properties of the Kronecker product, one easily obtains that $(\mathbb{I}_U \otimes \mathbb{Z}_T)^{-1} = \mathbb{I}_U \otimes \mathbb{M}_T$ and, furthermore, Proposition 3.1 can be extended to give an explicit formulation of the entries of $\mathbb{I}_U \otimes \mathbb{Z}_T$ and $\mathbb{I}_U \otimes \mathbb{M}_T$.

**Proposition 6.1** Given a partition $(U, T)$ of $V$, let $\mathbb{I}_U$ be the identity matrix associated with $U$, and let $\mathbb{Z}_T$ and $\mathbb{M}_T$ be the zeta and the Möbius matrix, respectively, associated with $T$. Then, for every pair of subsets $D, E \subseteq V$

$$(\mathbb{I}_U \otimes \mathbb{Z}_T)_{D,E} = 1(U \cap E \subseteq D \subseteq E),$$

whereas

$$(\mathbb{I}_U \otimes \mathbb{M}_T)_{D,E} = (-1)^{|E \setminus D|} 1(U \cap E \subseteq D \subseteq E).$$

\[\square\]
Proof. We first show that the pair of subsets $D, E \subseteq V$ satisfies the condition (i) \((U \cap D = U \cap E) \land (T \cap D \subseteq T \cap E)\) if and only if it satisfies (ii) \((U \cap E \subseteq D \subseteq E)\). Condition (i) implies (ii) because the identity $U \cap E = U \cap D$ implies $U \cap E \subseteq D$, whereas $U \cap D = U \cap E$ together with $T \cap D \subseteq T \cap E$ implies \((U \cap D) \cup (T \cap D) \subseteq (U \cap E) \cup (T \cap E)\) and therefore that $D \subseteq E$. To show the reverse implication, i.e., that (ii) implies (i), we notice that $D \subseteq E$ implies $T \cap D \subseteq T \cap E$ whereas $D \subseteq E$ implies $U \cap D \subseteq U \cap E$ and $U \cap E \subseteq D$ implies $U \cap E \subseteq U \cap D$ so that $U \cap D = U \cap E$.

This result can be used to compute the entries of $I_U \otimes \mathbb{Z}_T$ because
\[
(I_U \otimes \mathbb{Z}_T)_{D,E} = (I_U)_{U \cap D, U \cap E} \times (\mathbb{Z}_T)_{T \cap D, T \cap E}
\]
\[
= 1(U \cap D = U \cap E) \times 1(T \cap D \subseteq T \cap E)
\]
\[
= 1(U \cap E \subseteq D \subseteq E)
\]
where the entries of $\mathbb{Z}_T$ are given in Proposition 3.1. Similarly,
\[
(I_U \otimes \mathbb{M}_T)_{D,E} = 1(U \cap D = U \cap E) \times (-1)^{|(T \cap E) \setminus (T \cap D)|} 1(T \cap D \subseteq T \cap E)
\]
\[
= (-1)^{|E \setminus D|} 1(U \cap E \subseteq D \subseteq E),
\]
where \((T \cap E) \setminus (T \cap D) = E \setminus D\) because, as shown above, when $U \cap E \subseteq D \subseteq E$ it holds that $U \cap D = U \cap E$. \qed

The Binary Case

Consider a vector $Y$ of binary random variables with probability parameter $\pi$. We have seen that undirected graph models are suitably parameterized by the log-linear expansion $\lambda = \mathbb{M}^T \log \pi$ of $\pi$, whereas for bidirected graph models it is more convenient to employ the log-linear expansion $\gamma = \mathbb{M}^T \log \mu$ of the mean parameter $\mu = \mathbb{Z} \pi$. Every entry $\pi_D = p(Y_D = 1, Y_{V \setminus D} = 0)$ of $\pi$ is a probability involving all the variables in $Y_V$, whereas every entry $\mu_D = p(Y_D = 1)$ of $\mu$ is a probability from the marginal distribution of $Y_D$. We now introduce a further parameterization that combines the features of $\pi$ and $\mu$. This is associated with the partition \((U, T)\), and consists of probabilities involving all the variables in $Y_U$ but only certain subvectors of $Y_T$. Formally, the hybrid probability-mean parameter of $Y_V$, or hybrid parameter for short, with respect to the partition \((U, T)\) is the vector $\pi^{(U,T)}(D) = (\pi_D^{(U,T)})_{D \subseteq V}$ with entries
The hybrid parameter $\pi^{(U,T)}$ includes, as special cases, both the probability and the mean parameters; specifically, $\pi = \pi^{(V,\emptyset)}$ and $\mu = \pi^{(\emptyset,V)}$. Furthermore, $\pi^{(U,T)}$ can be computed from $\pi$ as $\pi^{(U,T)} = (\mathbb{I}_U \otimes \mathbb{Z}_T) \pi$, that for $U = \emptyset$ simplifies to $\mu = \mathbb{Z}_V \pi$.

**Proposition 6.2** Let $Y_V$ be a binary random vector with probability parameter $\pi$. The hybrid parameter $\pi^{(U,T)}$ of $Y_V$, with respect to the partition $(U, T)$ of $V$, can be computed as $\pi^{(U,T)} = (\mathbb{I}_U \otimes \mathbb{Z}_T) \pi$ so that $\pi = (\mathbb{I}_U \otimes \mathbb{M}_T) \pi^{(U,T)}$.

**Proof.** We have to show that for every $D \subseteq V$ it holds that

$$\pi^{(U,T)}_D = \sum_{E \subseteq V} (\mathbb{I}_U \otimes \mathbb{Z}_T)_{D,E} \pi_E,$$

and the result follows from Proposition 6.1 because

$$\sum_{E \subseteq V} (\mathbb{I}_U \otimes \mathbb{Z}_T)_{D,E} \pi_E = \sum_{E \subseteq V} 1(U \cap E \subseteq D \subseteq E) \pi_E$$

$$= \sum_{F \subseteq T \setminus D} \pi_{D \cup F}$$

$$= \sum_{F \subseteq T \setminus D} p(Y_D = 1, Y_F = 1, Y_{(T \setminus D) \setminus F} = 0, Y_{U \setminus D} = 0)$$

$$= p(Y_D = 1, Y_{U \setminus D} = 0)$$

$$= \pi^{(U,T)}_D. \quad \square$$

The parameters $\lambda$ and $\gamma$ are obtained from the log-linear expansion of $\pi$ and $\mu$, respectively, and it is therefore natural to consider the log-linear expansion of $\pi^{(U,T)}$ so as to obtain the log-hybrid linear parameter of $Y$ (La Rocca and Roverato, 2017). This is denoted by $\varphi^{(U,T)}$, or simply by $\varphi$ when it is clear which partition of $V$ is being used, and is defined as

$$\varphi^{(U,T)} = \mathbb{M}^T \log \pi^{(U,T)}. \quad (6.4)$$

It follows from Proposition 6.2 that $\varphi = \mathbb{M}^T \log \{(\mathbb{I}_U \times \mathbb{Z}_T) \pi\}$. This shows that $\lambda = \varphi^{(V,\emptyset)}$ whereas $\gamma = \varphi^{(\emptyset,V)}$ and, furthermore, that $\pi$ can...
be obtained from $\varphi$ as $\pi = (\mathbb{I}_U \times M_T)\exp(Z^T \varphi)$ so that there is a one-to-one correspondence between $\pi$ and $\varphi$. More precisely, the latter is a smooth bijection, and this implies that $\varphi$ is a valid parameterization of the distribution of $Y_V$. The individual entries of $\varphi$ can be computed as

$$\varphi_D = \sum_{D' \subseteq D} (-1)^{|D|\setminus|D'|} \log \pi_{D'}^{(U,T)} \quad \text{for every } D \subseteq V, \quad (6.5)$$

and, conversely,

$$\log \pi_D^{(U,T)} = \sum_{D' \subseteq D} \varphi_{D'} \quad \text{for every } D \subseteq V. \quad (6.6)$$

The importance of the log-hybrid linear parameterization stems from the fact that any independence relationship implied by the regression graph Markov property is equivalent to the vanishing of certain $\varphi$-terms. To illustrate the connection between zero entries of $\varphi$ and independence relationships, we distinguish among three different cases. Firstly, we consider independence relationships between variables belonging to different blocks, and the rules to deal with this situation are stated in Theorem 6.3. Secondly, we focus on independencies holding among response variable and provide the relevant rules in Theorem 6.5. Finally, we consider independencies holding among context variables with the rules pertaining to this case given in Theorem 6.7.

In the Markov property for regression graphs, the independencies involving variables belonging to different blocks are those implied by missing arrows. The simplest case of this situation is when $|U| = 1$ and $|D| = 2$, say $U = \{b\}$ and $D = \{a, b\}$ with $a \in T$, so that $\pi_D^{(U,T)} = p(Y_a = 1, Y_b = 1)$ and (6.6) becomes

$$\log p(Y_a = 1, Y_b = 1) = \varphi \emptyset + \varphi_a + \varphi_b + \varphi_{\{a,b\}}.$$

It is easy to see from (6.5) that $\varphi \emptyset = \log p(Y_b = 0)$, whereas

$$\varphi_a = \log p(Y_a = 1 | Y_b = 0) \quad \text{and} \quad \varphi_b = \log \frac{p(Y_b = 1)}{p(Y_b = 0)}.$$

However, the main interest if for the two-way interaction $\varphi_{\{a,b\}}$ that coincides with the logarithm of the cross product ratio in (2.19),

$$\varphi_{\{a,b\}} = \log \frac{p(Y_a = 1, Y_b = 1)p(Y_b = 0)}{p(Y_a = 1, Y_b = 0)p(Y_b = 1)} = \log \text{cpr}_{mc}(a, a : b, b),$$
and it follows from Lemma 2.2 that $Y_a \perp Y_b$ if and only if $\varphi_{(a,b)} = 0$.

It is also worthwhile noticing that the interaction $\varphi_{(a,b)}$ can also be interpreted as the log-relative risk of the event $\{Y_a = 1\}$ with respect to the two groups identified by $\{Y_b = 1\}$ and $\{Y_b = 0\}$,

$$\varphi_{(a,b)} = \log \frac{p(Y_a = 1 | Y_b = 1)}{p(Y_a = 1 | Y_b = 0)}.$$

The connection between $\varphi$-terms and relative risks is discussed, with more generality, in Lupparelli and Roverato (2017).

Assume now that $Y_U$ contains an arbitrary number of explanatory variables. Condition (RG.d) concerns independence relationships between variables $Y_A$, with $A \subseteq T$, and variables $Y_B$, with $B \subseteq U$, given the remaining variables in $Y_U$; that is $Y_A \perp Y_B | Y_{U \setminus B}$. Consider, for example, the case where $A = \{a\}$ and $B = \{b\}$. The interaction $\varphi_{(a,b)}$ is equal to the logarithm of a conditional cross-product ratio,

$$\varphi_{(a,b)} = \log \frac{p(Y_a = 1, Y_b = 1, Y_{U \setminus \{b\}} = 0) p(Y_b = 0, Y_{U \setminus \{b\}} = 0)}{p(Y_a = 1, Y_b = 0, Y_{U \setminus \{b\}} = 0) p(Y_b = 1, Y_{U \setminus \{b\}} = 0)},$$

$$\varphi_{(a,b)} = \log \frac{p(Y_a = 1, Y_b = 1 | Y_{U \setminus \{b\}} = 0) p(Y_b = 0 | Y_{U \setminus \{b\}} = 0)}{p(Y_a = 1, Y_b = 0 | Y_{U \setminus \{b\}} = 0) p(Y_b = 1 | Y_{U \setminus \{b\}} = 0)}$$

$$= \log \text{cpr}_{mc}(a, a : b, b | \emptyset, U \setminus \{b\}).$$

Hence, by Corollary 2.4, the relationship $Y_a \perp Y_b | Y_{U \setminus \{b\}}$ implies that $\varphi_{(a,b)} = 0$. However, the vanishing of $\varphi_{(a,b)}$ is not sufficient for the reverse implication to hold true. Indeed, the independence $Y_a \perp Y_b | Y_{U \setminus \{b\}}$ is equivalent to the vanishing of the interactions $\varphi_{(a,b) \cup C}$ for all $C \subseteq U \setminus \{b\}$. This is formally stated in the theorem given below, which generally deals with independence relationships between response variables and explanatory variables given the remaining explanatory variables. This result is an immediate consequence of the properties of cross product ratios and Möbius inversion and, in the context of regression graphs, it plays the same role as Theorem 4.2 in the context of undirected graphs and Theorem 5.2 in the context of bidirected graphs; see also La Rocca and Roverato (2017).

**Theorem 6.3** For a vector $Y_V$ of binary variables with probability parameter $\pi > 0$, let $\varphi = \mathbb{M}^T \log \pi^{(U,T)}$, where $\pi^{(U,T)} = (\mathbb{I}_U \otimes \mathbb{Z}_T) \pi$ and $V = U \cup T$ is a partition of $V$. Then, for a pair of
nonempty subsets $A \subseteq T$ and $B \subseteq U$ the following conditions are equivalent:

(i) $Y_A \perp Y_B | Y_{U \setminus B}$;

(ii) for every $D \subseteq A \cup U$ such that both $D \cap A \neq \emptyset$ and $D \cap B \neq \emptyset$ it holds that

$$
\varphi_D = 0;
$$

(iii) for every $A' \subseteq A$, $B' \subseteq B$, and $C' \subseteq U \setminus B$, such that $A', B' \neq \emptyset$ it holds that

$$
\log \pi^{(U,T)}_{A' \cup B' \cup C'} - \log \pi^{(U,T)}_{B' \cup C'} - \log \pi^{(U,T)}_{A' \cup C'} + \log \pi^{(U,T)}_{C'} = 0.
$$

(6.7)

Proof. The equivalence of (i) and (iii) follows from Corollary 2.4 because the identity (6.7) is equal to $\log \operatorname{cpr}_{mc}(A', A : B' \cap B | C', C) = 0$ with $C = U \setminus B$. The equivalence of (ii) and (iii) is that stated in Lemma 3.3 when $\theta = \varphi$ and $\omega = \log \pi^{(U,T)}$.

Assume that the distribution of $Y$ is Markov with respect to the graph in Figure 6.4. Hence, (RG.d) implies, for example, that $Y_e \perp Y_{\{a,b\}} | Y_{\{c,d\}}$ and by Theorem 6.3 this relationship is satisfied if and only if $\varphi_D = 0$ for all $D \subseteq \{a, b, c, d, e\}$ such that both $e \in D$ and $D \cap \{a, b\} \neq \emptyset$. The same graph also implies $Y_{\{g,h\}} \perp Y_{\{a,b,c,d\}}$ and this independence holds if and only if $\varphi_D = 0$ for all $D \subseteq \{a, b, c, d, g, h\}$ such that both $D \cap \{a, b, c, d\} \neq \emptyset$ and $D \cap \{g, h\} \neq \emptyset$.

We now turn to the analysis of the independence relationships between two subsets of response variables, which are considered conditionally on the explanatory variables. Hence, for a subset $U' \subseteq U$ we consider the event $\{Y_{U'} = 1, Y_{U \setminus U'} = 0\}$ and focus on the conditional

Figure 6.4 Regression graph with an undirected block $U = \{a, b, c, d\}$ and a bidirected block $T = \{e, f, g, h\}$.
distribution of $Y_T|\{Y_{U'} = 1, Y_{U\backslash U'} = 0\}$. Then, we denote the probability parameter of this distribution by $\pi_T^{U'} = (\pi_D^{T|U'})_{D \subseteq T}$ where, for $D \subseteq T$,

$$\pi_D^{T|U'} = p(Y_D = 1, Y_{T\backslash D} = 0|Y_{U'} = 1, Y_{U\backslash U'} = 0).$$

In the Markov property for regression graphs, independencies among response variables are implied by missing bidirected arrows, and therefore we compute the log-mean linear parameters

$$\gamma_T^{U'} = M_T^T \log(Z_T \pi_T^{U'}) \text{ for every } U' \subseteq U. \quad (6.8)$$

The parameter $\gamma_T^{U'}$ can be used to establish “marginal” independence relationships in the distribution of $Y_T|\{Y_{U'} = 1, Y_{U\backslash U'} = 0\}$. More precisely, Theorem 5.2 implies that,

- for a pair of disjoint nonempty subsets $A$ and $B$ of $T$ it holds that
  $$Y_A \perp Y_B|\{Y_{U'} = 1, Y_{U\backslash U'} = 0\}$$
- if and only if $\gamma_D^{T|U'} = 0$ for every $D \subseteq A \cup B$ such that $D \cap A \neq \emptyset$ and $D \cap B \neq \emptyset$.

Furthermore, $Y_A \perp Y_B|Y_U$ if and only if $Y_A \perp Y_B|\{Y_{U'} = 1, Y_{U\backslash U'} = 0\}$ for every $U' \subseteq U$, and this leads to the following equivalence:

- for a pair of disjoint nonempty subsets $A$ and $B$ of $T$ it holds that
  $$Y_A \perp Y_B|Y_U$$
- if and only if $\gamma_D^{T|U'} = 0$ for every $D \subseteq A \cup B$ such that $D \cap A \neq \emptyset$ and $D \cap B \neq \emptyset$ and for every $U' \subseteq U$.

A relevant fact concerning the log-hybrid linear parameter $\varphi$ is that it is related with the log-mean linear parameters in (6.8), through Möbius inversion, in such a way that it can be equivalent used to establish the independence relationships implied by condition (RG.b). The relevant Möbius inversion formula follows from the application of Lemma 3.4.

**Proposition 6.4** In the setting of Theorem 6.3, for $U' \subseteq U$ let

$$\gamma_T^{U'} = (\gamma_D^{T|U'})_{D \subseteq T}$$

be the log-mean linear parameter of the distribution of $Y_T|\{Y_U = 1, Y_{U\backslash U'} = 0\}$. Then, for every nonempty subset $D \subseteq T$ it holds that
\[ \gamma_{D}^{T|U'} = \sum_{U^* \subseteq U'} \varphi_{D \cup U^*}. \quad (6.9) \]

**Proof.** For every \( U' \subseteq U \) and nonempty \( D \subseteq V \) we can apply the definition of log-mean linear parameter in (5.2) to compute

\[
\gamma_{D}^{T|U'} = \sum_{D' \subseteq D} (-1)^{|D'|} \log p(Y_{D'} = 1|Y_{U'} = 1, Y_{U \setminus U'} = 0) \quad (6.10)
\]

\[
= \sum_{D' \subseteq D} (-1)^{|D'|} \log p(Y_{D'} = 1, Y_{U'} = 1, Y_{U \setminus U'} = 0) \quad (6.11)
\]

\[
= \sum_{D' \subseteq D} (-1)^{|D'|} \log \pi_{U' \cup D'}^{(U,T)} \quad (6.12)
\]

\[
= \varphi_{U'|D'}^{(U,T)} \quad (6.13)
\]

Because we have assumed \( D \neq \emptyset \), then the sum in (6.10) has the same number of positive and negative terms. As a consequence, the sum does not change if, before taking the logarithm, every probability is multiplied by a positive constant, and we go from (6.10) to (6.11) by choosing such constant equal to \( p(Y_{U'} = 1, Y_{U \setminus U'} = 0) \). Equation (6.12) is obtained by applying the definition of \( \pi_{U' \cup D'}^{(U,T)} \) in (6.3) and, finally, (6.13) follows from the definition of \( \varphi_{U'|D'}^{(U,T)} \) in (3.7). Hence, the required result can be obtained by applying Lemma 3.4 to (6.13). \( \square \)

The result of Proposition 6.4 can be applied to state the connection existing between vanishing log-hybrid linear parameters and marginal independencies in the distribution of \( Y_{T|Y_{U}} \).

**Theorem 6.5** In the setting of Theorem 6.3, for a pair of nonempty disjoint subsets \( A, B \subseteq T \) the following conditions are equivalent:

(i) \( Y_{A} \independent Y_{B}|Y_{U} \);

(ii) it holds that \( \varphi_{D \cup U'}^{(U,T)} = 0 \) for every \( D \subseteq A \cup B \) such that \( D \cap A \neq \emptyset \) and \( D \cap B \neq \emptyset \) and for every \( U' \subseteq U \).

**Proof.** We first show that for every nonempty \( D \subseteq V \) the conditions (a) \( \varphi_{D \cup U'} = 0 \) for every \( U' \subseteq U \), is equivalent to (b) \( \gamma_{D}^{T|U'} = 0 \) for every \( U' \subseteq U \).
Point (a) implies (b) by (6.9) whereas (b) implies (a) because, by Möbius inversion, (6.9) gives
\[ \varphi_{D \cup U'} = \sum_{U'' \subseteq U'} (-1)^{|U'| - |U''|} \gamma_D^{U'} \]
for every \( U' \subseteq U \). Hence, condition (ii) is satisfied if and only if \( \gamma_D^{U'} = 0 \) for every \( D \subseteq A \cup B \) such that \( D \cap A \neq \emptyset \) and \( D \cap B \neq \emptyset \) and for every \( U' \subseteq U \), and this is equivalent to (i) by Theorem 5.2. \( \square \)

To see an example, assume that the distribution of \( Y \) is Markov with respect to the graph in Figure 6.4. Hence, condition (RG.b) implies \( Y_e \perp Y_h | Y_{\{a,b,c,d\}} \), and it follows from Theorem 6.5 that this independence relationship is satisfied if and only if \( \varphi_{\{e,h\} \cup U'} = 0 \) for every \( U' \subseteq U \).

Finally, we consider independence relationships among context variables which, in the Markov property for regression graphs, are implied by missing undirected edges. Accordingly, the interest here lies in the conditional independence relationships in the distribution of \( Y_U \) and therefore in the log-linear parameter \( \lambda^U = M_U T \log \pi_U \). However, \( \varphi \) can be used in place of \( \lambda^U \) because it holds that \( \lambda^U = (\varphi_D)_{D \subseteq U} \).

**Proposition 6.6** In the setting of Theorem 6.3, it holds that
\[ \varphi_D = \lambda_D^U \]
for every \( U' \subseteq U \), where \( \lambda^U \) is the log-linear parameter of \( Y_U \).

**Proof.** This follows from (4.2) and (6.5) by noticing that for every \( D \subseteq U \)
\[ \pi_D^{(U,T)} = p(Y_D = 1, Y_{U \setminus D} = 0) = \pi_D^U \]
where \( \pi^U \) is the probability parameter of \( Y_U \). \( \square \)

Hence, the connection between vanishing log-hybrid linear parameters and conditional independencies in the distribution of \( Y_U \) is as follows.

**Theorem 6.7** In the setting of Theorem 6.3, for a pair of nonempty disjoint subsets \( A, B \subseteq U \) the following conditions are equivalent:

1. \( Y_A \perp Y_B | Y_{U \setminus (A \cup B)} \);
2. it holds that \( \varphi_D^{(U,T)} = 0 \) for every \( D \subseteq U \) such that \( D \cap A \neq \emptyset \) and \( D \cap B \neq \emptyset \).
Proof. This is an immediate consequence of Proposition 6.6 and Theorem 4.2.

For example, in the regression graph of Figure 6.4 condition (RG.c) implies $Y_d \perp Y_{\{a,c\}} | Y_b$, and it follows from Theorem 6.7 that this independence relationship is satisfied if and only if $\varphi_D = 0$ for every $D \subseteq U$ such that $d \in D$ and $D \cap \{a,c\} \neq \emptyset$.

The Non-binary Case

We now consider the case of categorical variables with an arbitrary number of levels. The material described here may be useful in the practical implementation of procedures based on the log-hybrid linear parameterization, and for this reason it is presented in a concise way as a list of technical results. More concretely, we provide a formal extension of the hybrid parameter and of the log-hybrid linear parameter of $Y$ and, furthermore, of Theorems 6.3, 6.5 and 6.7.

Let $Y$ be a vector of categorical variables with probability parameter $\pi$, and let $\mathcal{I}$ and $\mathcal{J}$ be the state space and restricted state space of $Y$, respectively. The hybrid parameter of $Y$ relative to the partition $(U, T)$ can be defined as the collection of vectors

$$\pi_j^{(U,T)} = (\pi_D^{(U,T)})_{D \subseteq V} \text{ for all } j \in \mathcal{J},$$

where

$$\pi_D^{(U,T)} = p(Y_D = j_D, Y_{U \setminus D} = 0).$$

The hybrid parameter of $Y$ can be compactly written as $\pi^{(U,T)} = (\pi_D^{(U,T)})_{D \subseteq V, j_D \in \mathcal{J}_D}$, and it can be shown that there is a one-to-one correspondence between $\pi$ and $\pi^{(U,T)}$. Hence, the log-hybrid linear parameter of $Y$ is the collection of vectors

$$\varphi_j = (\varphi_{j_D})_{D \subseteq V} \text{ for all } j \in \mathcal{J},$$

where

$$\varphi_j = \mathbb{M}_j^T \log \pi_j^{(U,T)} \text{ and } \varphi_D = (\varphi_j)_D,$$

and also in this case we use the compact form $\varphi = (\varphi_{j_D})_{D \subseteq V, j_D \in \mathcal{J}_D}$. 
A formal extension of Theorem 6.3 to the general case of categorical variables with arbitrary number of levels is as follows.

**Corollary 6.8** Let $Y_V$ be a vector of categorical variables with probability parameter $\pi > 0$, hybrid parameter $\pi_{(U,T)}(D \subseteq V, j_D \in J_D)$ and log-hybrid linear parameter $\varphi = (\varphi_{j_D})_{D \subseteq V, j_D \in J_D}$ where $V = U \cup T$ is a partition of $V$. Then, for a pair of nonempty subsets $A \subseteq T$ and $B \subseteq U$ the following conditions are equivalent:

(i) $Y_A \independent Y_B | Y_{U \setminus B}$;

(ii) for every $D \subseteq A \cup U$ such that both $D \cap A \neq \emptyset$ and $D \cap B \neq \emptyset$ it holds that

$$\varphi_{j_D} = 0,$$

for every $j \in J$;

(iii) for every $A' \subseteq A$, $B' \subseteq B$, and $C' \subseteq U \setminus B$, such that $A', B' \neq \emptyset$ it holds that

$$\log \pi_{j_A' \cup B' \cup C'}^{(U,T)} - \log \pi_{j_B' \cup C'}^{(U,T)} - \log \pi_{j_A' \cup B' \cup C'}^{(U,T)} + \log \pi_{j_C'^{(U,T)}} = 0, \quad (6.14)$$

for every $j \in J$.

**Proof.** The equivalence of (i) and (iii) follows from Corollary 2.8 because the identity (6.14) is equal to $\log \text{cpr}_{mc}(j_{A'}, A : j_{B'}, B | j_{C'}, C) = 0$ with $C = U \setminus B$. The equivalence of (ii) and (iii) holds for every $j \in J$ and is that stated in Lemma 3.3 when $\theta = \varphi_j$ and $\omega = \log \pi_j^{(U,T)}$. □

The following corollary is a general version of Theorem 6.5.

**Corollary 6.9** In the setting of Corollary 6.8, for a pair of nonempty disjoint subsets $A, B \subseteq T$ the following conditions are equivalent:

(i) $Y_A \independent Y_B | Y_U$;

(ii) it holds that $\varphi_{j_{D \cup U'}} = 0$ for every $D \subseteq A \cup B$ such that $D \cap A \neq \emptyset$ and $D \cap B \neq \emptyset$ and for every $U' \subseteq U$ and $j \in J$.

**Proof.** If, for $U' \subseteq U$ and $j_U' \in J_U'$, we denote by $\gamma^{Tj_U'} = (\gamma_{j_D})_{D \subseteq T, j_D \in J_D}$ the log-mean linear parameter of the conditional distribution of $Y_T | \{Y_U' = j_U', Y_{U' \setminus U'} = 0\}$, then following the same line used in the proof of Proposition 6.4 it can be shown that for every nonempty subset $D \subseteq T$ it holds that
\[ \gamma_{jD}^{U'} = \sum_{U'' \subset U'} \varphi_{jD \cup U''}. \]

Next, in a similar fashion used to prove Theorem 6.5 it can be shown that (ii) is equivalent to the statement

• it holds that \( \gamma_{jD}^{U'} = 0 \) for every \( D \subseteq A \cup B \) such that \( D \cap A \neq \emptyset \) and \( D \cap B \neq \emptyset \) and for every \( U' \subseteq U \) and \( j \in J \).

Hence, the result follows from Corollary 5.3. \( \square \)

Finally, we provide the extension of Theorem 6.7.

**Corollary 6.10** In the setting of Corollary 6.8, for a pair of nonempty disjoint subsets \( A, B \subseteq U \) the following conditions are equivalent:

(i) \( Y_A \perp Y_B | Y_{U \setminus (A \cup B)}; \)

(ii) it holds that \( \varphi_{jD}^{(U,T)} = 0 \) for every \( D \subseteq U \) such that \( D \cap A \neq \emptyset \) and \( D \cap B \neq \emptyset \) and for every \( j \in J \).

**Proof.** Let \( \pi^U = (\pi^{U|D}_{jD})_{D \subseteq T, jD \in J_D} \) be the probability parameter of \( Y_U \). It follows from the fact that \( \pi^{(U,T)}_{jD} = \pi^{U}_{jD} \) for every \( D \subseteq U \) and \( j \in J \), that \( \varphi_{jD} = \lambda_{jD} \) for every \( D \subseteq U \) and \( j \in J \). Hence, the desired result can be obtained from the application of Corollary 4.3. \( \square \)

### 6.7 Log-hybrid Linear Graphical Models

Similarly to the other graphical models considered so far, a regression graph model for \( Y \) can be defined as the family of probability distributions that are Markov with respect to a given regression graph \( G = (V, E) \). The theory presented in the previous section shows that a regression graph model corresponds to a linear subspace of the log-hybrid linear parameter space and, more specifically, it is characterized by the vanishing of certain \( \varphi \)-terms. Here, we look more closely at the connection between vanishing log-hybrid linear interactions and structure of the graph. Along the lines of Section 6.6, we start by considering the simplified setting of binary variables, and assume that the regression graph \( G = (V, E) \) is basic, in the sense that it has two blocks, \( (U, T) \). The graph of Figure 6.4 is an example of a basic regression graph.

The definition of the Markov property for regression graphs in Section 6.4 is formulated by means of three statements, (RG.d),
(RG.b) and (RG.u), each concerning the independence relationships implied by the removal of one of the three types of edges. Likewise, there are three distinct rules for identifying the set of vanishing \( \varphi \)-terms from the structure of the graph. If \( G = (V, E) \) is a basic regression graph with blocks \((U, T)\) and \( Y_V \) is a binary random vector with log-hybrid linear parameter \( \varphi \) then the distribution of \( Y \) satisfies

(a) condition (RG.d) with respect to \( G \) if and only if
\[ [H.d] \quad \varphi_D = 0 \text{ for every } D \subseteq V \text{ such that } D \cap T \neq \emptyset \text{ and } D \cap U \nsubseteq \text{pa}(D \cap T); \]

(b) condition (RG.b) with respect to \( G \) if and only if
\[ [H.b] \quad \varphi_D = 0 \text{ for every } D \subseteq V \text{ such that } D \cap T \text{ is disconnected in } G_T; \]

(c) condition (RG.u) with respect to \( G \) if and only if
\[ [H.u] \quad \varphi_D = 0 \text{ for every } D \subseteq U \text{ such that } D \text{ is not complete in } G_U. \]

The statement (b) follows from Theorem 6.5. It refers to the connected set Markov property for bidirected graphs, and when \( U = \emptyset \) it is equivalent to Proposition 5.4. The statement (c) follows from Theorem 6.7 but, in fact, in a regression graph model the block \( U \) can be dealt with marginally as an undirected graph model for the distribution of \( Y_U \) and graph \( G_U \). Recall that, as shown in Proposition 6.6, the \( \varphi \)-terms indexed by the subsets of \( U \) coincide with the log-linear interactions of \( Y_U \). Statement (a) follows from Theorem 6.3, but its derivation is more involved than (b) and (c), and, for this reason, a formal proof is provided.

**Proposition 6.11** Let \( G = (V, E) \) be a basic regression graph with blocks \((U, T)\), and let \( \varphi \) be the log-hybrid linear parameter of a binary random vector \( Y_V \). The distribution of \( Y_V \) satisfies condition (RG.d) with respect to \( G \) if and only if \( \varphi_D = 0 \) for every \( D \subseteq V \) such that both \( D \cap T \neq \emptyset \) and \( D \cap U \nsubseteq \text{pa}(D \cap T) \).

**Proof.** Assume that the distribution of \( Y_V \) satisfies condition (RG.d). If \( D \subseteq V \) is such that \( D \cap U \nsubseteq \text{pa}(D \cap T) \) then \( U \setminus \text{pa}(D \cap T) \neq \emptyset \), and because it also holds that \( D \cap T \neq \emptyset \) then (RG.d) implies that \( Y_D \mid \text{pa}(D \cap T) \perp Y_{U \setminus \text{pa}(D \cap T)} \). Hence, it follows by point (ii) of Theorem 6.3 that \( \varphi_D = 0 \) because \( D \cap T \neq \emptyset \) whereas \( D \cap U \nsubseteq \text{pa}(D \cap T) \) also implies that \( D \cap \{ U \setminus \text{pa}(D \cap T) \} \neq \emptyset \). Conversely, assume that \( \varphi_D = 0 \) for every \( D \subseteq V \) such that \( D \cap T \neq \emptyset \) and \( D \cap U \nsubseteq \text{pa}(D \cap T) \). We have to show that \( Y_A \perp Y_B \mid Y_{U \setminus B} \) for all
non-empty \( A \subseteq T \) such that \( B = U \setminus \text{pa}(A) \) is also non-empty. Hence, by Theorem 6.3, we have to show that \( \varphi_D = 0 \) for every \( D \subseteq A \cup U \) such that both \( D \cap A \neq \emptyset \) and \( D \cap B \neq \emptyset \). Now, \( D \subseteq A \cup U \) implies \( D \cap T = D \cap A \neq \emptyset \), while \( D \cap B \neq \emptyset \) implies \( D \cap U \not\subseteq \text{pa}(A) \); because \( \text{pa}(D \cap A) \subseteq \text{pa}(A) \) it follows that \( D \cap U \not\subseteq \text{pa}(D \cap A) = \text{pa}(D \cap T) \). Therefore, we have \( \varphi_D = 0 \). □

Clearly, the probability distribution of a binary random vector \( Y \) is Markov with respect to a basic regression graph \( G \) if and only if the zero structure of its log-hybrid linear parameter \( \varphi \) satisfies all the conditions [H.d], [H.b] and [H.u]. Conversely, in a basic regression graph model, the only \( \varphi \)-terms which are not constrained to vanish are those indexed by subsets \( D \subseteq V \) such that either \( D \subseteq U \) and \( D \) is complete or \( D \cap T \neq \emptyset \) and both \( D \cap T \) is connected and \( D \cap U \subseteq \text{pa}(D \cap T) \).

To fix ideas, assume that the distribution of \( Y \) is Markov with respect to the regression graph in Figure 6.4 and consider, for example, the set \( D = \{ b, d, f, h \} \). It follows from [H.d] that the corresponding \( \varphi \)-term vanishes, \( \varphi_{\{b,d,f,h\}} = 0 \). Indeed, every subset \( D \subseteq V \), with \( D \cap T \neq \emptyset \), identifies two subsets of \( U \); namely, \( \text{pa}(D \cap T) \) and \( D \cap U \). If \( D \cap U \subseteq \text{pa}(D \cap T) \) then [H.d] imposes no constraints on \( \varphi_D \); otherwise \( \varphi_D \) vanishes. The set \( D = \{ b, d, f, h \} \) can be split into \( D \cap T = \{ f, h \} \) and \( D \cap U = \{ b, d \} \), and because \( \{ b, d \} \not\subseteq \text{pa}(\{ f, h \}) = \{ d \} \) it follows that \( \varphi_{\{b,d,f,h\}} = 0 \). As a second example let \( D = \{ c, d, e, g, h \} \). In this case we find \( D \cap T = \{ e, g, h \} \) and \( D \cap U = \{ c, d \} \), so that \( \{ c, d \} \subseteq \text{pa}(\{ e, g, h \}) = \{ c, d \} \), and therefore \( \varphi_{\{c,d,e,g,h\}} \) is not restricted by [H.d]. The same interaction is not restricted by [H.b] because \( \{ c, d, e, g, h \} \cap T = \{ e, g, h \} \) is a connected set. Because \( \{ c, d, e, g, h \} \not\subseteq U \) then \( \varphi_{\{c,d,e,g,h\}} \) is not restricted by [H.u] either, and we can conclude that \( \varphi_{\{c,d,e,g,h\}} \) is an unrestricted interaction of the model. Consider now the set \( D = \{ d, e, h \} \). In this case, \( D \cap T = \{ e, h \} \) and \( D \cap U = \{ d \} \), so that \( \{ d \} \subseteq \text{pa}(\{ e, h \}) = \{ c, d \} \). Hence, \( \varphi_{\{d,e,h\}} \) is not restricted by condition [H.d]; however it is restricted by [H.b] because \( \{ d, e, h \} \cap T = \{ e, h \} \) is disconnected. More generally, because \( \{ e, h \} \) is disconnected in \( G_T \) it follows by [H.b] that \( \varphi_{\{e,h\} \cup U'} = 0 \) for every \( U' \subseteq U \). Finally, the interaction \( \varphi_{\{a,c,d\}} \) is not restricted by either [H.d] or [H.b] but it is constrained to vanish by [H.u] because \( \{ a, c, d \} \) is a subset of \( U \) and it is not complete.

We now turn to regression graphs with an arbitrary number of blocks. Let \( G = (V, E) \) be an arbitrary regression graph, and let \( (T_0, T_1, \ldots, T_k) \)
be the block ordering of its vertices. The distribution of $Y$ is characterized by the collection of conditional distributions $Y_{T_r} | Y_{pr(T_r)}$, for all $r = 1, \ldots, k$, together with the marginal distribution of $Y_{T_0}$. Accordingly, the $\varphi$-terms for the distribution of $Y$ can be obtained by noticing that, for every $r = 1, \ldots, k$, the quantity $\varphi^{(pr(T_r), T_r)}$ parameterizes the distribution of $Y_{pr(T_r) \cup T_r}$ in such a way that

$$(\varphi^{(pr(T_r), T_r)}_{D \subseteq pr(T_r)}) = \lambda_{pr(T_r)}$$

parameterizes the distribution of $Y_{pr(T_r)}$, while $$(\varphi^{(pr(T_r), T_r)}_{D \nsubseteq pr(T_r)})$$

parameterizes the distribution of $Y_{T_r} | Y_{pr(T_r)}$. Hence, if we write $\varphi^{(T_r|pr(T_r))} = (\varphi^{(pr(T_r), T_r)}_{D \nsubseteq pr(T_r)})$ then the log-hybrid linear parameter of the distribution of $Y'$ are given by

$$\varphi(T_0, T_1), \varphi(T_2|pr(T_2)), \ldots, \varphi(T_k|pr(T_k)).$$

The vanishing $\varphi$-terms can be identified by applying the rules [H.d] and [H.b] iteratively for every $r = 1, \ldots, k$ by setting $U = pr(T_r)$, $T = T_r$ and $V = pr(T_r) \cup T_r$. The rule [H.u] needs to be considered once with $U = T_0$. For example, in the regression graph of Figure 6.3, it holds that $pr(T_2) = T_0 \cup T_1$ and, for example, the [H.d] applied to the partition $(T_0 \cup T_1, T_2)$ implies that $\varphi_{\{l, h, c, b\}} = 0$. On the other hand, the term $\varphi_{\{l, h, c\}}$ is unconstrained because $\{l, h, c\} \cap pr(T_2) = \{h, c\} \subseteq pa(l)$ and, furthermore, $\{l, h, c\} \cap T_2 = \{l\}$ is connected.

A regression graph with singleton chain components is a DAG. In this case one can assign every vertex to a different block and then order the blocks according to any well-ordering of the vertices. In this way, the rules [H.b] and [H.u] become irrelevant, and it follows from [H.d] that the only unconstrained interactions are the terms $\varphi_{\{v\} \cup D}$ for all $v \in V$ and $D \subseteq pa(v)$.

Finally, conditions [H.d], [H.b] and [H.u] can be trivially extended to deal with non-binary variables by requiring $\varphi_{j_D} = 0$ for every $j_D \in J_D$, in place of $\varphi_D = 0$.

### 6.8 Inference in Regression Graph Models

We conclude this chapter by dealing with inferential issues. Assume that $Y^{(1)} = y^{(1)}, \ldots, Y^{(n)} = y^{(n)}$ is a random sample of $n$ independent and identically distributed observations from a distribution with probability parameter $\pi$. Then, the sampling distribution of counts follows a multinomial distribution with probability mass function given in
(4.12). The log-likelihood is therefore that given in (4.13), and it is equal to

\[ l(\pi) = \sum_{i \in I} n(Y = i) \log p(Y = i). \]

Let \( G \) be a regression graph. If the distribution of \( Y \) belongs to the regression graph model \( M(G) \) then it follows from (6.2) that the log-likelihood function can be written as

\[ l(\pi) = \sum_{K \in \mathcal{K}} l_K(\pi), \quad (6.15) \]

where \( l_K(\pi) \), for \( K \in \mathcal{K} \), are the component log-likelihood functions

\[ l_K(\pi) = \sum_{i_K \in I_K} \sum_{i_{\text{pa}(K)} \in I_{\text{pa}(K)}} n(Y_{K \cup \text{pa}(K)} = i_K \cup \text{pa}(K)) \log p(Y_K = i_K | Y_{\text{pa}(K)} = i_{\text{pa}(K)}). \]

The decomposition of the likelihood in (6.15) can be exploited to simplify the computation of the ML estimates. Indeed, it can be shown (see, e.g. Drton, 2009) that \( l(\pi) \) can be maximized in \( M(G) \) by maximizing every \( l_K(\pi) \) separately, and then by combining the optima according to (6.2).

The actual maximization of a component log-likelihood function is a problem in constrained optimization. For categorical data, a general iterative algorithm for constrained likelihood maximization was provided by Lang (1996). This algorithm can be applied to all models considered in this text; see also Evans and Forcina (2013) for a discussion on this issue. Furthermore, dedicated algorithms are available for the maximization of the component functions \( l_K(\pi) \) in (6.15). In the case where the relevant chain component induces an undirected graph, the maximization of the component log-likelihood function can be carried out by applying the classical IPF algorithm. There also exists a dual of the IPF algorithm, named the iterative conditional fitting (ICF) algorithm, that can be applied to maximize the components \( l_K(\pi) \) corresponding to the bidirected chain components. This algorithm was developed by Drton and Richardson (2008b) to fit discrete bidirected graph models and then extended by Drton (2008) to fit the bidirected components of regression graph models.
It is possible that the ML estimates of certain probabilities are simple empirical proportions and can therefore be computed directly. The following result is proved in Drton (2009, proposition 12).

**Proposition 6.12** If the chain component $K$ of the regression graph $G$ is a complete set and $\text{pa}(K) = \emptyset$, then the ML estimator of the marginal probability $p(Y_K = i_K)$ is the empirical proportion

$$\hat{p}(Y_K = i_K) = \frac{n(Y_K = i_K)}{n}.$$ 

Furthermore, if the chain component $K$ is a singleton, then the ML estimator of the conditional probability $p(Y_K = i_K | Y_{\text{pa}(K)} = i_{\text{pa}(K)})$ is the empirical proportion

$$\hat{p}(Y_K = i_K | Y_{\text{pa}(K)} = i_{\text{pa}(K)}) = \frac{n(Y_K = i_K, Y_{\text{pa}(K)} = i_{\text{pa}(K)})}{n(Y_{\text{pa}(K)} = i_{\text{pa}(K)})}. \quad (6.16)$$

All the chain component of a DAG are singletons and therefore (6.16) implies that in DAG models the ML estimates are all empirical proportions. Indeed, equation (6.16) extends to regression graph models a classical result of DAG models; see, e.g., Lauritzen (1996, theorem 4.36). DAGs constitute a distinguished family of regression graphs and a comprehensive description of likelihood inference in these models can be found, for example, in Lauritzen (1996, section 4.5.1).

Finally, we remark that the standard asymptotic theory is valid for regression graph models so that, for instance, the asymptotic distribution of the ML estimator is normal and likelihood ratios are asymptotically chi-squared distributed. This follows from the fact that, as shown in Drton (2009), regression graphs yield models for categorical data that are curved exponential families.


