August 18, 2011

Parametrizations of Discrete Graphical Models

Robin J. Evans

In presenting this dissertation in partial fulfillment of the requirements for the doctoral degree at the University of Washington, I agree that the Library shall make its copies freely available for inspection. I further agree that extensive copying of the dissertation is allowable only for scholarly purposes, consistent with "fair use" as prescribed in the U.S. Copyright Law. Requests for copying or reproduction of this dissertation may be referred to ProQuest Information and Learning, 300 North Zeeb Road, Ann Arbor, MI 48106-1346, 1-800-521-0600, to whom the author has granted "the right to reproduce and sell (a) copies of the manuscript in microform and/or (b) printed copies of the manuscript made from microform."

Signature _____

Date _____

University of Washington

Abstract

Parametrizations of Discrete Graphical Models

Robin J. Evans

Chair of the Supervisory Committee:

Professor Thomas S. Richardson
Department of Statistics

Graphical models relate graphs to collections of conditional independences among a set of random variables, via Markov properties. In the context of discrete data, we consider a broad class of these models, known as acyclic directed mixed graphs (ADMGs); these contain DAGs and bidirected graphs as special cases.

We present the first fitting algorithm for discrete ADMGs, using an existing parametrization based on conditional probabilities. We present a new parametrization, which we term the ingenuous parametrization, using marginal log-linear parameters. The properties of this parametrization are explored, and in particular we characterize for which models it is variation independent.

The new parametrization is used to produce parsimonious sub-models, and to perform automatic consistent model selection using the adaptive lasso. This is illustrated with data examples and simulations. Finally we consider variation dependence, and show that every discrete ADMG has a smooth variation independent parametrization.

Contents

List of Figures

List of Tables

Preface

This thesis considers a large class of graphical models, known as acyclic directed mixed graphs (ADMGs), and explores their properties in the case of discrete random variables.

Chapter 1 introduces graphical models, and the factorization of Richardson (2009) for discrete distributions on ADMGs. Much of the content of this Introduction is found in Lauritzen (1996)[1], Richardson and Spirtes (2002) and Richardson (2009). Chapter 2 describes a parametrization for binary ADMG models, and gives a method for fitting such models to data via maximum likelihood estimation, as shown in Evans and Richardson (2010).

In Chapter 3 we discuss the marginal log-linear (MLL) parameters of Bergsma and Rudas (2002), and show that they may be used to smoothly parametrize all ADMG models. We also establish the variation independence properties of such parametrizations. Chapter 4 considers the applicability of MLL parameters to finding parsimonious sub-models, and to automatic model selection; these applications are illustrated with simulations. Chapter 5 expands upon the issue of variation independence, and demonstrates how to construct a variation independent parametrization of any ADMG model.

Two indices found at the end of this document should help those readers needing to refer back to definitions and notations quickly. An appendix contains details of how the work in this thesis can be extended from ADMGs to a slightly broader class which allows undirected edges; this class is termed *mixed euphonious graphs*.

Acknowledgements

Firstly I would like to thank my advisor, Thomas Richardson, for so much patient support and encouragement over the last three years. I will be eternally grateful for his role as a mentor, including all the time he spent assisting in the planning of presentations and papers, as well as for financial support through U.S. National Science Foundation grant

[1]However, some of the terminology used differs from Lauritzen's book.

CNS-085523.

To the other members of my doctoral committee, Adrian Dobra, Brian Flaherty, Peter Hoff, Steffen Lauritzen and James Robins, I give thanks for their comments, suggestions, support, and taking time to listen to my examinations or read this thesis.

Others who have contributed with academic assistance, personal encouragement, or both, include Charles Doss, Mathias Drton, Antonio Forcina, Tamás Rudas, Ilya Shpitser, Alex Volfovsky, Jon Wakefield and Jon Wellner.

I also received funding to attend various conferences and workshops over the last three years. Contributors include: the 26th Conference for Uncertainty in Artificial Intelligence; Workshop on Geometric and Algebraic Statistics 3; and the American Institute of Mathematics (workshop *Parameter Identification in Graphical Models*).

Most of all I am indebted to my partner Gao Gao for her part in motivating me even when, as has too often happened, we were thousands of miles apart.

To Margaret Freeman and Peggy Evans, in loving memory.

Chapter 1

Introduction

Graphical models are an intuitive and visual way of encoding a structure of conditional independence relationships among a set of random variables. The nodes of a graph are used to represent the random variables, and the (conditional) independences arise from Markov properties based on the absence of edges between those nodes.

Models based on undirected graphs were pioneered by Darroch et al. (1980), followed later by directed acyclic graph (DAG) models and chain graph models (see, for example, Lauritzen, 1996). Richardson and Spirtes (2002) developed ancestral graph models to create a class of graphs which is closed under conditioning and marginalization. The class of models we work with is the closely related acyclic directed mixed graphs (ADMGs), whose Markov properties were established by Richardson (2003), and which were parametrized in the discrete case by Richardson (2009).

Sections 1.1 and 1.2 contain elementary definitions for graphs and graphical models respectively. Section 1.3 introduces a factorization criterion for discrete ADMGs due to Richardson (2009), and this is further developed in Section 1.4.

1.1 Basic Definitions

Definition 1.1.1. A *mixed graph* $\mathcal{G}$ is a pair (V, E), where V is a set of *vertices* and E represents the *edges*; specifically, E is a function from $V \times V$ to $\mathscr{P}(\mathcal{E})$, where $\mathcal{E} = \{-, \rightarrow, \leftrightarrow\}$, and $\mathscr{P}(A)$ denotes the power set of A.

If $\rightarrow \in E(v, w)$ we write $v \rightarrow w$, and similarly for the other two kinds of edge. We require that there are no loops, i.e. $E(v, v) = \emptyset$, and symmetry for the bidirected and undirected

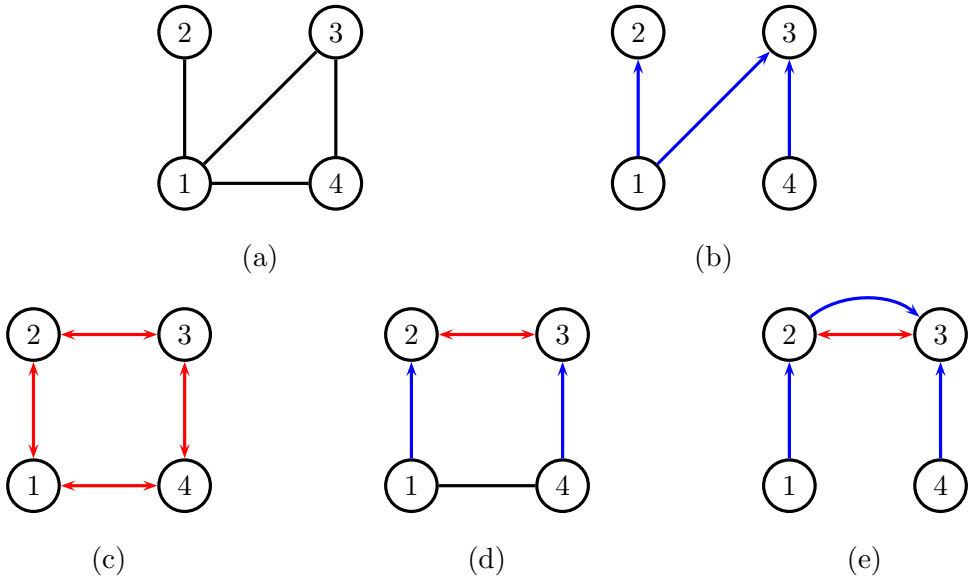


Figure 1.1: (a) An undirected graph; (b) a directed graph; (c) a bidirected graph; (d) a mixed graph; and (e) a directed mixed graph. We use colour only to distinguish between types of edge.

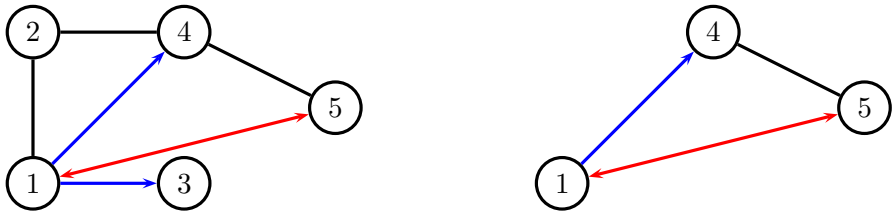
edges:

$$- \in E(v,w) \iff - \in E(w,v), \qquad \leftrightarrow \in E(v,w) \iff \leftrightarrow \in E(w,v).$$

If only undirected edges ('$-$') are present then $\mathcal{G}$ is *undirected*; if only directed edges ($\rightarrow$) are present then it is *directed*; if only bidirected edges ($\leftrightarrow$) are present it is a *bidirected* graph. If there are no undirected edges then $\mathcal{G}$ is a *directed mixed* graph (DMG).

A strength of graphical models comes from their visual nature, and the reader is encouraged to treat the examples in Figure 1.1 as something close to a definition. Note that there cannot be repeated edges of the same type and orientation between two vertices. A further warning for those unfamiliar with mixed graphs is that in spite of the appearance, the bidirected edge $\leftrightarrow$ is *not* equivalent to having both $\leftarrow$ and $\rightarrow$.

Definition 1.1.2. For a mixed graph $\mathcal{G}$ and a subset $A \subseteq V$ of the vertices in $\mathcal{G}$, we define the *induced subgraph*, $\mathcal{G}_A$, to be the graph formed by taking the vertices in A together with all edges whose endpoints are both in A (or equivalently by restricting the domain of the



(a) (b)

Figure 1.2: (a) A graph $\mathcal{G}$, and (b) the induced subgraph $\mathcal{G}_A$ for $A = \{1, 4, 5\}$.

function E to $A \times A$).

We define the *bidirected skeleton*, $\mathcal{G}_{\leftrightarrow}$, of $\mathcal{G}$ to be the graph formed by removing any edges from $\mathcal{G}$ which are not bidirected; similarly $\mathcal{G}_{-}$ is the *undirected skeleton*, formed by removing edges which are not undirected.

An example of a graph and its induced subgraph is shown in Figure 1.2.

Definition 1.1.3. Let $v, w \in V(\mathcal{G})$. If $v \to w$ we say v is a *parent* of w, and w a *child* of v; if $v - w$ or $v \leftrightarrow w$ then v is respectively a *neighbour* or a *spouse* of w. The collections of parents, children, spouses and neighbours of a vertex v in a graph $\mathcal{G}$ are denoted by

$$\mathrm{pa}_{\mathcal{G}}(v) \qquad \mathrm{ch}_{\mathcal{G}}(v) \qquad \mathrm{sp}_{\mathcal{G}}(v) \qquad \mathrm{ne}_{\mathcal{G}}(v)$$

respectively. Further, let

$$\begin{aligned} \mathrm{an}_{\mathcal{G}}(v) &\equiv \{w \mid w \to \cdots \to v \text{ in } \mathcal{G} \text{ or } w = v\}, \\ \mathrm{de}_{\mathcal{G}}(v) &\equiv \{w \mid w \leftarrow \cdots \leftarrow v \text{ in } \mathcal{G} \text{ or } w = v\} \\ \text{and} \quad \mathrm{dis}_{\mathcal{G}}(v) &\equiv \{w \mid w \leftrightarrow \cdots \leftrightarrow v \text{ in } \mathcal{G} \text{ or } w = v\} \end{aligned}$$

be the set of *ancestors*[1], the set of *descendants* and the *district* of v respectively. A district in the graph $\mathcal{G}$ is any set of the form $D = \mathrm{dis}_{\mathcal{G}}(v)$ where $v \in D$.

[1]This definition, though standard, differs from that of Lauritzen (1996), who takes $v \notin \mathrm{an}_{\mathcal{G}}(v)$.

All these definitions are applied disjunctively to sets of vertices so that, for example,

$$\mathrm{pa}_{\mathcal{G}}(A) \equiv \bigcup_{v \in A} \mathrm{pa}_{\mathcal{G}}(v);$$

notice that it is possible for $A \cap \mathrm{pa}_{\mathcal{G}}(A)$ to be non-empty.

On an induced subgraph we will sometimes write $\mathrm{pa}_A(v)$ to denote $\mathrm{pa}_{\mathcal{G}_A}(v)$, and similarly for other definitions; we may omit the subscript entirely when context allows.

Definition 1.1.4. A *path* in a graph $\mathcal{G}$ is a sequence of edges $\epsilon_1, \ldots, \epsilon_k$, such that there is a sequence of distinct vertices $w_1, \ldots, w_{k+1}$, where the endpoints of ϵ_i are w_i and w_{i+1}. We refer to this as a path from w_1 to w_{k+1}. We define paths in terms of edges since there may be more than one edge between two vertices (see Figure 1.1). A path may have length 0, or equivalently consist only of a single vertex. Note that the requirement that vertices are distinct means that paths may not intersect themselves.

A *cycle* is defined similarly to a path, but it must contain at least one edge, and we require $w_1 = w_{k+1}$, the first and last vertices to be the same; otherwise all vertices are distinct. A path or cycle of the form $w_1 \to w_2 \to \cdots \to w_{k+1}$ is said to be *directed*. A graph which contains no directed cycles is said to be *acyclic*.

A path (respectively cycle) containing only bidirected edges is *bidirected*. A path (cycle), possibly containing a mixture of directed and bidirected edges, such that all the directed edges are oriented in the same direction is *semi-directed*. For example, $v_1 \to v_2 \leftrightarrow v_3 \to v_4$ is a semi-directed path from v_1 to v_4. This definition strictly includes all directed and bidirected paths (cycles).

Remark 1.1.5. Some special cases of acyclic graphs are well known. The purely directed case is known as a *directed acyclic graph* (DAG); if a graph containing only directed and bidirected edges is acyclic, it is called an *acyclic directed mixed graph* (ADMG).

Definition 1.1.6. A non-endpoint vertex v on a path π is said to be a *collider* on π if the two edges adjacent to v on π both have arrows pointing towards v. Otherwise v is a non-collider. Thus $\to v \leftarrow$ and $\leftrightarrow v \leftarrow$ are colliders, but $- v \leftarrow$ and $\leftarrow v \leftarrow$ are non-colliders. Note that a vertex is only a (non-)collider relative to a path, and not in an absolute sense.

We now define the most general class of graphs on which most of our results will hold.

Definition 1.1.7. An acyclic mixed graph $\mathcal{G}$ is said to be *euphonious* if for every vertex $v \in \mathcal{G}$, we have $\mathrm{ne}_{\mathcal{G}}(v) \neq \emptyset \Rightarrow \mathrm{pa}_{\mathcal{G}}(v) \cup \mathrm{sp}_{\mathcal{G}}(v) = \emptyset$. In other words, if there is an undirected

edge incident to v, then there must be no arrowheads incident to v. We write MEG for mixed euphonious graph.

Euphonious graphs generalize both the *ancestral graphs* of Richardson and Spirtes (2002) and ADMGs, and hence also DAGs, undirected graphs and purely bidirected graphs. For simplicity, in the rest of this thesis we will only consider the special case of ADMGs, however the work herein can easily be applied to MEGs; Appendix A provides a more detailed explanation of how this is achieved.

Definition 1.1.8. Let $\mathcal{G}$ be an ADMG; for a set $W \subseteq V(\mathcal{G})$, define the *barren subset* of W to be

$$\mathrm{barren}_{\mathcal{G}}(W) \equiv \{v \mid \mathrm{de}_{\mathcal{G}}(v) \cap W = \{v\}\}.$$

If $\mathrm{barren}_{\mathcal{G}}(W) = W$, we say that W *is barren*.

A set A is *ancestral* if $\mathrm{an}_{\mathcal{G}}(A) = A$.

1.2 Graphical Models

For a graph $\mathcal{G}$ with vertex set V, we consider collections of random variables $(X_v)_{v\in V}$ taking values in finite discrete probability spaces $(\mathfrak{X}_v)_{v\in V}$. For $A \subseteq V$ we let $\mathfrak{X}_A \equiv \times_{v\in A}(\mathfrak{X}_v)$, $\mathfrak{X} \equiv \mathfrak{X}_V$ and $X_A \equiv (X_v)_{v\in A}$. We abuse notation in the usual way: v denotes both a vertex and the random variable X_v, likewise A denotes both a set of vertices and the random vector X_A. For fixed elements of $\mathfrak{X}_v$ and $\mathfrak{X}_A$ we write i_v and i_A respectively.

The relationship between a graph $\mathcal{G}$ and random variables X_V is governed by Markov properties.

Definition 1.2.1. A path π in $\mathcal{G}$ between two vertices $v, w \in V(\mathcal{G})$ is said to *m-connect* v and w given a set $C \subseteq V \setminus \{v, w\}$ if both:

(i) no non-collider on π is in C; and

(ii) every collider on π is an ancestor of an element of C.

We say v and w are *m-separated* given C in $\mathcal{G}$ if every path from v to w in $\mathcal{G}$ fails to m-connect them given C. Note that C may be empty.

Sets $A, B \subseteq V$ are said to be m-separated given $C \subseteq V \setminus (A \cup B)$ if every pair $a \in A$ and $b \in B$ are m-separated given C.

The special case of m-separation in purely directed graphs is the better known d-separation (Lauritzen, 1996; Pearl, 1988); for an undirected graph we have the usual separation criterion (Darroch et al., 1980). We next relate m-separation to conditional independence, for which we use the now standard notation of Dawid (1979): for random variables X, Y and Z we denote the statement 'X is independent of Y conditional on Z' by $X \perp\!\!\!\perp Y \mid Z$. If Z is empty we write $X \perp\!\!\!\perp Y$.

Definition 1.2.2. A probability measure P on $\mathfrak{X}$ is said to satisfy the *global Markov property* (GMP) for a mixed graph $\mathcal{G}$, if for all disjoint sets $A, B, C \subseteq V$ with A and B non-empty, A being m-separated from B given C implies that $X_A \perp\!\!\!\perp X_B \mid X_C$ under P.

Definition 1.2.3. Let $\mathcal{G}$ be an ADMG with a vertex v, and an ancestral set A such that $v \in \text{barren}_{\mathcal{G}}(A)$. Define

$$\text{mb}(v, A) = \text{pa}_{\mathcal{G}}\left(\text{dis}_A(v)\right) \cup \left(\text{dis}_A(v) \setminus \{v\}\right)$$

to be the *Markov blanket* for v in the induced subgraph on A.

Let $<$ be a *topological ordering* on the vertices of $\mathcal{G}$, meaning that no vertex appears before any of its ancestors; let $\text{pre}_{\mathcal{G},<}(v)$ be the set of vertices preceding v in the ordering. A probability distribution P is said to satisfy the *ordered local Markov property* for $\mathcal{G}$ with respect to $<$, if for any v and ancestral set A such that $v \in A \subseteq \text{pre}_{\mathcal{G},<}(v)$,

$$v \perp\!\!\!\perp A \setminus \left(\text{mb}(v, A) \cup \{v\}\right) \mid \text{mb}(v, A)$$

with respect to P.

Proposition 1.2.4 (Richardson (2003), Theorem 2)**.** *Let $\mathcal{G}$ be an ADMG, and $<$ a topological ordering of its vertices; further let P be a probability distribution on $\mathfrak{X}_V$. The following are equivalent:*

(i) P obeys the global Markov property with respect to $\mathcal{G}$;

(ii) P obeys the ordered local Markov property with respect to $\mathcal{G}$ and $<$.

In particular note that this result implies that if the ordered local Markov property is satisfied for some topological ordering $<$, then it is satisfied for all topological orderings.

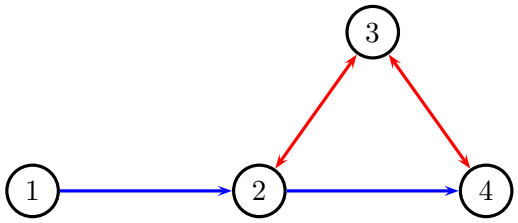


Figure 1.3: An acyclic directed mixed graph, $\mathcal{G}_1$.

1.3 Factorizations

A positive discrete probability distribution P obeys the global Markov property with respect to a DAG if and only if it factorizes as

$$P(X_V = i_V) = \prod_{v \in V} P(X_v = i_v \,|\, X_{\mathrm{pa}(v)} = i_{\mathrm{pa}(v)}),$$

for all $i_V \in \mathfrak{X}_V$ (see, for example, Lauritzen, 1996). Factorizations can also be used to characterize ADMGs, although the criterion is more complicated.

Example 1.3.1. Consider the ADMG in Figure 1.3. A distribution which obeys the global Markov property with respect to this graph satisfies $X_1 \perp\!\!\!\perp X_3$ and $X_1 \perp\!\!\!\perp X_4 \,|\, X_2$. It is not possible to specify a factorization on the joint distribution of X_1, X_2, X_3 and X_4 which implies precisely these two independences. Instead, we require factorizations of certain marginal distributions:

$$\begin{aligned} P(X_1 = i_1,\, X_3 = i_3) &= P(X_1 = i_1) \cdot P(X_3 = i_3), \\ P(X_1 = i_1,\, X_2 = i_2,\, X_4 = i_4) &= P(X_1 = i_1) \cdot P(X_2 = i_2 \,|\, X_1 = i_1) \cdot P(X_4 = i_4 \,|\, X_2 = i_2). \end{aligned}$$

In this section we will see how such marginal factorizations can be used to represent distributions which obey the global Markov property with respect to an ADMG.

Definition 1.3.2. A set of vertices W is *(bidirected-) connected* (in $\mathcal{G}$) if there is a (bidirected) path between every pair of vertices in W, such that every vertex on the path is in W.

We say that a set of vertices W is *(bidirected-) path-connected* in $\mathcal{G}$ if a (bidirected) path exists in $\mathcal{G}$ between each pair of vertices in W (the paths not necessarily being contained within W).

A vertex set $H \subseteq V$ is a *head* if it is barren in $\mathcal{G}$ and is a bidirected-path-connected subset

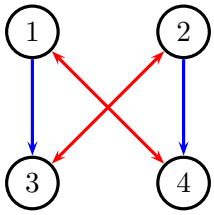


Figure 1.4: An ADMG in which there is no vertex ordering such that all parents of a head precede every vertex in the head.

of $\mathcal{G}_{\mathrm{an}(H)}$. We write $\mathcal{H}(\mathcal{G})$ for the collection of all heads in $\mathcal{G}$.

For any head H, the *tail* of H is the set

$$\mathrm{tail}_{\mathcal{G}}(H) \equiv (\mathrm{dis}_{\mathrm{an}(H)}(H) \setminus H) \cup \mathrm{pa}(\mathrm{dis}_{\mathrm{an}(H)}(H)).$$

We denote the first set in this union by $\text{dis-tail}_{\mathcal{G}}(H)$, and the second by $\text{pa-tail}_{\mathcal{G}}(H)$. These sets need not be disjoint. If the context makes it clear which head we are referring to, we will sometimes denote a tail simply by T.

Example 1.3.3. Note that the tail is a subset of the ancestors of the head. In the special case of a DAG, the heads are all singleton vertices $\{v\}$, and the tails are the sets of parents $\mathrm{pa}_{\mathcal{G}}(v)$. In a purely bidirected graph, the heads are just the connected sets, and the tails are all empty.

Example 1.3.4. The graph $\mathcal{G}_1$ in Figure 1.3 has the following head-tail pairs:

H	$\{1\}$	$\{2\}$	$\{3\}$	$\{2,3\}$	$\{4\}$	$\{3,4\}$
T	$\emptyset$	$\{1\}$	$\emptyset$	$\{1\}$	$\{2\}$	$\{1,2\}$

.

Note that the bidirected-path-connected set $\{2,3,4\}$ is not a head, because it is not barren.

In general, it is not possible to order the vertices in an acyclic directed mixed graph such that, for each head H, all the vertices in $\mathrm{pa}_{\mathcal{G}}(H)$ precede all the vertices in H. A counter example is given in Figure 1.4, which is taken from Richardson (2009). The head $\{1,4\}$ has parent 2, and whilst the head $\{2,3\}$ has parent 1; clearly, whichever way we order these two heads, the condition will be violated.

However, there is a well-defined partial ordering on heads which will be useful to us.

Definition 1.3.5. For two distinct heads H_i and H_j in an ADMG $\mathcal{G}$, say that $H_i \prec H_j$ if $H_i \subseteq \mathrm{an}_{\mathcal{G}}(H_j)$.

Lemma 1.3.6. *The (strict) partial ordering $\prec$ is well-defined.*

Proof. We need to verify that $\prec$ is irreflexive, asymmetric and transitive; irreflexivity is by definition. Asymmetry amounts to $H_i \prec H_j \implies H_j \not\prec H_i$; suppose not for contradiction, so that there exist distinct heads H_i and H_j with $H_i \prec H_j$ and $H_j \prec H_i$. Since H_i and H_j are distinct, there exists a vertex v which is in one of these heads but not the other; assume with out loss of generality that $v \in H_j \setminus H_i$.

Since $H_j \subseteq \mathrm{an}_{\mathcal{G}}(H_i)$, we can find a directed path π_1 from v to some vertex $w \in H_i$; the path is non-empty because $v \notin H_i$. However, since we also have $H_i \subseteq \mathrm{an}_{\mathcal{G}}(H_j)$, we can find a (possibly empty) directed path π_2 from w to some $x \in H_j$. Now, the concatenation of π_1 and π_2 is also a path, because any repeated vertices would imply a directed cycle in the graph. Call this new path π.

But π is a non-empty directed path between two vertices in H_j, which violates the requirement that heads are barren. Hence asymmetry holds.

For transitivity, if $H_i \prec H_j$ and $H_j \prec H_k$, then clearly we can find a directed path from any element $v \in H_i$ to some element of H_k, simply by concatenating paths from $v \in H_i$ to some $w \in H_j$ and from w to H_k. Hence $H_i \subseteq \mathrm{an}_{\mathcal{G}}(H_k)$, and so $H_i \prec H_k$. □

This partial ordering on heads allows us to factorize probabilities for ADMGs into expressions based upon heads and tails.

Definition 1.3.7. We define a function which partitions sets of vertices $W \subseteq V$ by repeatedly removing heads. First, define a function $\Phi_{\mathcal{G}}$ such that $\Phi_{\mathcal{G}}(\emptyset) \equiv \emptyset$ and

$$\Phi_{\mathcal{G}}(W) \equiv \{H \in \mathcal{H}(\mathcal{G}) \cap \mathscr{P}(W) \mid H \text{ maximal head under } \prec \text{ in } W\}$$

for $W \neq \emptyset$; thus $\Phi_{\mathcal{G}}(W)$ returns the heads which are maximal under $\prec$ among those heads which are subsets of W. Then let

$$\begin{aligned}
\psi_{\mathcal{G}}(W) &\equiv W \setminus \bigcup_{H \in \Phi_{\mathcal{G}}(W)} H, \\
\psi_{\mathcal{G}}^{(0)}(W) &\equiv W, \\
\psi_{\mathcal{G}}^{(k)}(W) &\equiv \psi_{\mathcal{G}}(\psi_{\mathcal{G}}^{(k-1)}(W)), \qquad k \in \mathbb{N}.
\end{aligned}$$

Then $\psi_{\mathcal{G}}(W)$ returns the subset of W defined by removing the maximal heads found by $\Phi_{\mathcal{G}}$,

and $\psi_{\mathcal{G}}^{(k)}$ is the function defined by k applications of $\psi_{\mathcal{G}}$. We define the partition

$$[W]_{\mathcal{G}} \equiv \bigcup_{k \geq 0} \Phi_{\mathcal{G}}\left(\psi_{\mathcal{G}}^{(k)}(W)\right).$$

This sequentially removes heads from the set W until no vertices remain.

Proposition 1.3.8. *For any ADMG $\mathcal{G}$ and set $W \subseteq V$, the heads returned by $\Phi_{\mathcal{G}}(W)$ are disjoint. Hence, the function $[\cdot]_{\mathcal{G}}$ partitions sets.*

Proof. Suppose that two heads $H_1, H_2 \subseteq W$ are distinct and $H_1 \cap H_2 \neq \emptyset$. We will show that they cannot both be maximal under $\prec$ in W. Clearly if either $H_1 \prec H_2$ then H_1 is not maximal, and vice versa; assume that $H_1 \not\prec H_2$ and $H_2 \not\prec H_1$.

Let $H = \text{barren}_{\mathcal{G}}(H_1 \cup H_2)$. We first claim that H is a head: clearly it is barren, so we need to prove that it is bidirected-path-connected in $\text{an}_{\mathcal{G}}(H)$. By definition, $\text{an}_{\mathcal{G}}(H) \supseteq H_1 \cup H_2$; we need to find a bidirected path between any distinct $v, w \in H \subseteq H_1 \cup H_2$. If v, w are either both in H_1 or both in H_2, then the existence of such a path follows from the fact that these are heads. If $v \in H_1$ and $w \in H_2$, then construct a bidirected path in $\text{an}_{\mathcal{G}}(H_1)$ to some vertex $x \in H_1 \cap H_2$, and a bidirected path in $\text{an}_{\mathcal{G}}(H_2)$ from x to w; these paths can then be concatenated into a new path meeting the requirements, shortening the resulting sequence of edges if necessary to avoid repetition of vertices. Hence H is a head.

Now H is clearly in W, and also $H_1, H_2 \subseteq \text{an}_{\mathcal{G}}(H)$, so for each $i = 1, 2$, either $H_i \prec H$ or $H_i = H$. Since H_1 and H_2 are distinct, H is not equal to both of them, but then $H_i \prec H$ implies that H_i is not maximal. Thus at least one of H_1 or H_2 is not maximal under $\prec$ in W. □

Remark 1.3.9. The function $\Phi_{\mathcal{G}}$ (and therefore $\psi_{\mathcal{G}}$) is defined incorrectly in Richardson (2009) and Evans and Richardson (2010), but the construction above rectifies this.[2]

It is possible to define an equivalent partition replacing $\Phi_{\mathcal{G}}$ with $\Phi_{\mathcal{G}}^{\triangleleft}$ on maximal heads under a partial ordering $\triangleleft$, where

$$H_i \cup \text{dis-tail}_{\mathcal{G}}(H_i) \subseteq H_j \cup \text{dis-tail}_{\mathcal{G}}(H_j) \quad \Longrightarrow \quad H_i \triangleleft H_j.$$

Heads which are maximal under $\prec$ are also maximal under $\triangleleft$, but the converse is not true, meaning that more heads are removed at each step under $\triangleleft$. However the partition which results is the same and $\prec$ is useful in other contexts, as we will see in Chapter 3.

[2] The two definitions coincide when W is ancestral, but (1.3) does not hold for the incorrect partition in general.

Lemma 1.3.10. *Let A be an ancestral set in $\mathcal{G}$, and let $H \in \Phi_{\mathcal{G}}(A)$ be a head removed from A at the first stage of the partition. If $H \subseteq B \subseteq A$ for some (not necessarily ancestral) set B, we have $H \in \Phi_{\mathcal{G}}(B)$.*

Proof. Let $\mathcal{H}_A$ be the set of heads contained within A. If $H \in \Phi_{\mathcal{G}}(A) \subseteq \mathcal{H}_A$ then H is maximal with respect to $\prec$. It is trivial that $\mathcal{H}_B \subseteq \mathcal{H}_A$, and so H is also maximal in $\mathcal{H}_B$. Thus $H \in \Phi_{\mathcal{G}}(A)$. □

Lemma 1.3.11. *Let $C \in [W]_{\mathcal{G}}$. Then $[W]_{\mathcal{G}} = \{C\} \cup [W \setminus C]_{\mathcal{G}}$.*

Proof. If $C = W$ (including any case where $|W| = 1$) then the result is trivial. We proceed by induction on the size of W.

Since $C \in [W]_{\mathcal{G}}$, if C is not maximal in W with respect to $\prec$, then it is clear that $\Phi_{\mathcal{G}}(W \backslash C) = \Phi_{\mathcal{G}}(W)$. Then by definitions,

$$\begin{aligned}[W]_{\mathcal{G}} &= \Phi_{\mathcal{G}}(W) \cup [\psi_{\mathcal{G}}(W)]_{\mathcal{G}} \\ &= \Phi_{\mathcal{G}}(W \setminus C) \cup [\psi_{\mathcal{G}}(W)]_{\mathcal{G}},\end{aligned}$$

and the problem reduces to showing that $[\psi_{\mathcal{G}}(W)]_{\mathcal{G}} = \{C\} \cup [\psi_{\mathcal{G}}(W) \backslash C]_{\mathcal{G}}$. Thus without loss of generality, assume that $C \in \Phi_{\mathcal{G}}(W)$, since otherwise we can simply repeat the argument.

Clearly $\Phi_{\mathcal{G}}(W \setminus C) \cup \{C\} \supseteq \Phi_{\mathcal{G}}(W)$; if equality holds then we are done. Otherwise let $C_1, \ldots, C_k$ be the sets which are in $\Phi_{\mathcal{G}}(W \setminus C)$ but not $\Phi_{\mathcal{G}}(W)$. These are maximal in $W \setminus C$, and therefore found in $\Phi_{\mathcal{G}}(\psi_{\mathcal{G}}(W))$. Then the problem reduces to showing that

$$[\psi_{\mathcal{G}}(W)]_{\mathcal{G}} = \{C_1, \ldots, C_k\} \cup [\psi_{\mathcal{G}}(W) \setminus (C_1 \cup \cdots \cup C_k)]_{\mathcal{G}},$$

which follows from k applications of the induction hypothesis. □

Lemma 1.3.12. *Let $W = D_1 \cup \cdots \cup D_k$ such that for each pair D_s, D_t, $s \neq t$, there are no bidirected paths from D_s to D_t in $\mathrm{an}_{\mathcal{G}}(W)$. Then*

$$[W]_{\mathcal{G}} = \bigcup_s [D_s]_{\mathcal{G}}.$$

Proof. We prove the result for $k = 2$, from which the general case follows by repeated application. We proceed by induction on the size of W; if D_1 or D_2 is empty, then the result is trivial. Otherwise, note that each head in W is contained within either D_1 or D_2,

because there are no bidirected paths between the two sets in $\mathrm{an}_{\mathcal{G}}(W)$. By definitions

$$[W]_{\mathcal{G}} = \Phi_{\mathcal{G}}(W) \cup [\psi_{\mathcal{G}}(W)]_{\mathcal{G}}\,.$$

Now, $\psi_{\mathcal{G}}(W)$ is strictly smaller than W, and can also be written as $\psi_{\mathcal{G}}(W) = D_1' \cup D_2'$ where $D_t' \subseteq D_t$; then by the induction hypothesis, this gives

$$[W]_{\mathcal{G}} = \Phi_{\mathcal{G}}(W) \cup \left[D_1'\right]_{\mathcal{G}} \cup \left[D_2'\right]_{\mathcal{G}}\,.$$

The heads in $\Phi_{\mathcal{G}}(W)$ are maximal with respect to $\prec$ in W, so they must also be maximal within their respective set D_i; thus by Lemmas 1.3.10 and 1.3.11,

$$[W]_{\mathcal{G}} = [D_1]_{\mathcal{G}} \cup [D_2]_{\mathcal{G}}\,.$$

□

Example 1.3.13. For the graph $\mathcal{G}_1$ in Figure 1.3, we have

H	$\{1\}$	$\{2\}$	$\{3\}$	$\{2,3\}$	$\{4\}$	$\{3,4\}$
$\mathrm{an}_{\mathcal{G}}(H)$	$\{1\}$	$\{1,2\}$	$\{3\}$	$\{1,2,3\}$	$\{1,2,4\}$	$\{1,2,3,4\}$

.

Then $\Phi_{\mathcal{G}_1}(\{2,3,4\}) = \{\{3,4\}\}$, and $\Phi_{\mathcal{G}_1}(\psi_{\mathcal{G}_1}(\{2,3,4\})) = \Phi_{\mathcal{G}_1}(\{2\}) = \{\{2\}\}$, giving

$$[\{2,3,4\}]_{\mathcal{G}_1} = \{\{3,4\},\{2\}\}.$$

Now we can provide a factorization criterion for acyclic directed mixed graphs.

Theorem 1.3.14 (Richardson (2009), Theorem 4)**.** *Let $\mathcal{G}$ be an ADMG, and P a probability distribution on $\mathfrak{X}_V$. Then P obeys the global Markov property with respect to $\mathcal{G}$ if and only if for every ancestral set A in $\mathcal{G}$,*

$$P\left(X_A = i_A\right) = \prod_{H \in [A]_{\mathcal{G}}} P(X_H = i_H \,|\, X_T = i_T). \tag{1.1}$$

Example 1.3.15. For the graph in Figure 1.4, observe that the global Markov property implies precisely that $X_3 \perp\!\!\!\perp X_4 \,|\, X_{12}$, and $X_1 \perp\!\!\!\perp X_2$. Theorem 1.3.14 gives us

$$P(X_{1234} = i_{1234}) = P(X_{23} = i_{23} \,|\, X_1 = i_1) \cdot P(X_{14} = i_{14} \,|\, X_2 = i_2).$$

Though a strange factorization, it does indeed imply that $X_3 \perp\!\!\!\perp X_4 \,|\, X_{12}$; summing over i_3

and i_4 gives

$$P(X_{12} = i_{12}) = P(X_2 = i_2 \,|\, X_1 = i_1) \cdot P(X_1 = i_1 \,|\, X_2 = i_2),$$

which implies that $X_1 \perp\!\!\!\perp X_2$.

Remark 1.3.16. It follows from the factorizations above that if H is a head, $\text{tail}_{\mathcal{G}}(H)$ is the Markov blanket for H in the set $\text{an}_{\mathcal{G}}(H)$, in the sense that

$$H \perp\!\!\!\perp \text{an}_{\mathcal{G}}(H) \setminus (H \cup \text{tail}_{\mathcal{G}}(H)) \,|\, \text{tail}_{\mathcal{G}}(H). \tag{1.2}$$

1.4 Towards a Parametrization

Theorem 1.4.1. *Let $\mathcal{G}$ be an ADMG, and P a probability distribution on $\{0,1\}^{|V|}$. Then P obeys the global Markov property with respect to $\mathcal{G}$ if and only if for any ancestral set A*

$$P\left(X_A = i_A\right) = \sum_{C : O \subseteq C \subseteq A} (-1)^{|C \setminus O|} \prod_{H \in [C]_{\mathcal{G}}} P(X_H = \mathbf{0} \,|\, X_T = i_T), \tag{1.3}$$

where $O \equiv \{v \in A \,|\, i_v = 0\}$ and the empty product is taken to be 1.

The required result is tricky to prove because the sets C in (1.3) may not be ancestral.

The following result, due to Evans and Richardson (2010), shows that the summation in (1.3) can be factorized into districts.

Lemma 1.4.2. *Suppose $D_1 \cup D_2 \cup \cdots \cup D_k = D$ and that each pair D_s and D_t, $s \neq t$, there are no bidirected paths from D_s to D_t in $\mathcal{G}_{\text{an}(D)}$. Further let $O_s = O \cap D_s$ for each s. Then*

$$\sum_{C : O \subseteq C \subseteq D} (-1)^{|C \setminus O|} \prod_{H \in [C]_{\mathcal{G}}} P(X_H = \mathbf{0} \,|\, X_T = i_T) \tag{1.4}$$

$$= \prod_{s=1}^{k} \sum_{C : O_s \subseteq C \subseteq D_s} (-1)^{|C \setminus O_s|} \prod_{H \in [C]_{\mathcal{G}}} P(X_H = \mathbf{0} \,|\, X_T = i_T). \tag{1.5}$$

Proof. We prove the case $k = 2$, from which the full result follows trivially by induction.

By Lemma 1.3.12, we have

$$\sum_{C : O \subseteq C \subseteq D} (-1)^{|C \setminus O|} \prod_{H \in [C]_{\mathcal{G}}} P(X_H = \mathbf{0} \,|\, X_T = i_T)$$

$$= \sum_{C:O\subseteq C\subseteq V} (-1)^{|C\setminus O|}\left[\prod_{H_1\in[C\cap D_1]_{\mathcal{G}}} P(X_{H_1}=\mathbf{0}\,|\,X_{T_1}=i_{T_1}) \prod_{H_2\in[C\cap D_2]_{\mathcal{G}}} P(X_{H_2}=\mathbf{0}\,|\,X_{T_2}=i_{T_2})\right].$$

Also

$$\{C : O\subseteq C\subseteq V\} = \{C_1\cup C_2 : O_1\subseteq C_1\subseteq D_1,\ O_2\subseteq C_2\subseteq D_2\}$$

and $C\cap D_i = C_i$, so

$$\sum_{C:O\subseteq C\subseteq D} (-1)^{|C\setminus O|} \prod_{H\in[C]_{\mathcal{G}}} P(X_H=\mathbf{0}\,|\,X_T=i_T)$$

$$= \sum_{\substack{C_1:O_1\subseteq C_1\subseteq D_1\\ C_2:O_2\subseteq C_2\subseteq D_2}} (-1)^{|(C_1\cup C_2)\setminus O|}\left[\prod_{H_1\in[C_1]_{\mathcal{G}}} P(X_{H_1}=\mathbf{0}\,|\,X_{T_1}=i_{T_1}) \prod_{H_2\in[C_2]_{\mathcal{G}}} P(X_{H_2}=\mathbf{0}\,|\,X_{T_2}=i_{T_2})\right].$$

Noting that

$$\begin{aligned}|(C_1\cup C_2)\setminus O| &= |(C_1\setminus O_1)\cup(C_2\setminus O_2)|\\ &= |C_1\setminus O_1| + |C_2\setminus O_2|,\end{aligned}$$

gives the result. □

The induction argument we will use in the proof of Theorem 1.4.1 requires the following definition and lemma.

Definition 1.4.3. Let $\mathcal{G}$ be an ADMG, and W be a subset of its vertices. We say W is an *ancestrally closed district* for $\mathcal{G}$ if W is a bidirected-connected and $\mathrm{dis}_{\mathrm{an}(W)}(W) = W$. In other words, it is a district in $\mathrm{an}_{\mathcal{G}}(W)$.

Lemma 1.4.4. *If P obeys the global Markov property with respect to $\mathcal{G}$, then for every ancestrally closed district D in $\mathcal{G}$ and $v\in\mathrm{barren}_{\mathcal{G}}(D)$,*

$$\sum_{i_v}\prod_{H\in[D]_{\mathcal{G}}} P(X_H=i_H\,|\,X_T=i_T) = \prod_{H\in[D\setminus\{v\}]_{\mathcal{G}}} P(X_H=i_H\,|\,X_T=i_T). \tag{1.6}$$

Proof. Suppose that the GMP is satisfied, and let $A=\mathrm{an}_{\mathcal{G}}(D)$. Then

$$\prod_{H'\in[A\setminus D]_{\mathcal{G}}} P(X_{H'}=i_{H'}\,|\,X_{T'}=i_{T'}) \sum_{i_v}\prod_{H\in[D]_{\mathcal{G}}} P(X_H=i_H\,|\,X_T=i_T)$$

$$= \sum_{i_v} \prod_{H \in [D]_{\mathcal{G}}} P(X_H = i_H \,|\, X_T = i_T) \prod_{H' \in [A \setminus D]_{\mathcal{G}}} P(X_{H'} = i_{H'} \,|\, X_{T'} = i_{T'})$$
$$= \sum_{i_v} \prod_{H \in [A]_{\mathcal{G}}} P(X_H = i_H \,|\, X_T = i_T)$$

by Lemma 1.3.11. Then by Theorem 1.3.14,

$$= \sum_{i_v} P(X_A = i_A)$$
$$= P(X_{A \setminus \{v\}} = i_{A \setminus \{v\}}),$$

which, since $A \setminus \{v\}$ is ancestral, is

$$= \prod_{H \in [A \setminus \{v\}]_{\mathcal{G}}} P(X_H = i_H \,|\, X_T = i_T)$$
$$= \prod_{H \in [D \setminus \{v\}]_{\mathcal{G}}} P(X_H = i_H \,|\, X_T = i_T) \prod_{H' \in [A \setminus D]_{\mathcal{G}}} P(X_{H'} = i_{H'} \,|\, X_{T'} = i_{T'}).$$

Then comparing the first and last expressions in this sequence gives (1.6). □

Proof of Theorem 1.4.1. Suppose that P obeys the factorization in (1.1). We will show that for any disjoint union of ancestrally closed districts D,

$$\prod_{H \in [D]_{\mathcal{G}}} P(X_H = i_H \,|\, X_T = i_T) = \sum_{O \subseteq C \subseteq D} (-1)^{|C \setminus O|} \prod_{H \in [C]_{\mathcal{G}}} P(X_H = \mathbf{0} \,|\, X_T = i_T)$$

where $O \equiv \{v \in D \,|\, i_v = 0\}$. Since ancestral sets are also disjoint unions of ancestrally closed districts, this gives the 'only if' part of the statement. We proceed by induction on the size of D and the number of 1s in the vector i_D. If $i_D = \mathbf{0}$ then the result is trivial, since the left and right hand sides are identical; if $|D| = 1$ then this is just a trivial application of the laws of probability.

Suppose that $i_D \neq \mathbf{0}$ and $|D| > 1$, and let $D = D_1 \cup \cdots \cup D_k$ for disjoint ancestrally closed districts $D_1, \ldots, D_k$.

If $i_v = 0$ for all $v \in \text{barren}_{\mathcal{G}}(D_1)$, then there is some head $H \subseteq \text{barren}_{\mathcal{G}}(D_1)$ which, by Lemma 1.3.10 appears in $[C]_{\mathcal{G}}$ for all $O \subseteq C \subseteq D$; this means that we can remove the factor $P(X_H = \mathbf{0} \,|\, X_T = i_T)$ from both sides of the above expression, and the problem is reduced to a strictly smaller disjoint union of ancestrally closed districts, $D \setminus H$.

Otherwise, let $i_v = 1$ for $v \in \text{barren}_{\mathcal{G}}(D_1)$; then

$$\begin{aligned}
&\sum_{C:O_1 \subseteq C \subseteq D_1} (-1)^{|C \setminus O_1|} \prod_{H \in [C]_{\mathcal{G}}} P(X_H = \mathbf{0} \,|\, X_T = i_T) \\
&\quad = \sum_{C:O_1 \subseteq C \subseteq D_1 \setminus \{v\}} (-1)^{|C \setminus O_1|} \prod_{H \in [C]_{\mathcal{G}}} P(X_H = \mathbf{0} \,|\, X_T = i_T) \\
&\qquad - \sum_{C:O_1 \cup \{v\} \subseteq C \subseteq D_1} (-1)^{|C \setminus (O_1 \cup \{v\})|} \prod_{H \in [C]_{\mathcal{G}}} P(X_H = \mathbf{0} \,|\, X_T = i_T) \\
&\quad = \prod_{H \in [D_1 \setminus \{v\}]} P(X_H = i_H \,|\, X_T = i_T) - \prod_{H \in [D_1]} P(X_H = i'_H \,|\, X_T = i_T),
\end{aligned}$$

where $i' = i$ except that $i'_v = 0$; this last expression follows from the induction hypothesis applied to the first term because $|D_1 \setminus \{v\}| < |D_1|$, and the second because i'_{D_1} has strictly fewer 1s than i_{D_1}. By Lemma 1.4.4 this is just

$$= \prod_{H \in [D_1]} P(X_H = i_H \,|\, X_T = i_T).$$

Now, using Lemma 1.4.2,

$$\begin{aligned}
&\sum_{C:O \subseteq C \subseteq D} (-1)^{|C \setminus O|} \prod_{H \in [C]_{\mathcal{G}}} P(X_H = \mathbf{0} \,|\, X_T = i_T) \\
&\quad = \prod_j \sum_{C:O_j \subseteq C \subseteq D_j} (-1)^{|C \setminus O_j|} \prod_{H \in [C]_{\mathcal{G}}} P(X_H = \mathbf{0} \,|\, X_T = i_T),
\end{aligned}$$

which by the above result for D_1 and application of the induction hypothesis to D_s for $s > 1$, is

$$\begin{aligned}
&= \prod_j \prod_{H \in [D_j]} P(X_H = i_H \,|\, X_T = i_T) \\
&= \prod_{H \in [D]} P(X_H = i_H \,|\, X_T = i_T)
\end{aligned}$$

by Lemma 1.3.12.

For the converse result, suppose that P satisfies the conditions given in (1.3); we will show that it also satisfies the ordered local Markov property and therefore the global Markov property for $\mathcal{G}$.

Let A be an ancestral set and $v \in \text{barren}_{\mathcal{G}}(A)$. Suppose further that $A = D_1 \cup \cdots \cup D_k$ for

disjoint ancestrally closed districts $D_1, \ldots, D_k$. By Lemma 1.4.2 we have

$$\begin{aligned} P(X_A = i_A) &= \sum_{O \subseteq C \subseteq A} (-1)^{|C \setminus O|} \prod_{H \in [C]} P(X_H = 0 \,|\, X_T = i_T) \\ &= \prod_{j=1}^{k} \sum_{O_j \subseteq C \subseteq D_j} (-1)^{|C \setminus O_j|} \prod_{H \in [C]} P(X_H = 0 \,|\, X_T = i_T) \\ &= \prod_{j=1}^{k} f_j(i_{D_j}, i_{\mathrm{pa}(D_j)}), \end{aligned}$$

for some functions f_j. Now if $v \in D_l$, then i_v appears only in the function f_l because $v \in \mathrm{barren}_{\mathcal{G}}(A)$. But we have

$$P(X_A = i_A) = f_l(i_v, i_{D_l \setminus \{v\}}, i_{\mathrm{pa}(D_l)}) \prod_{j \neq l} f_j(i_{D_j}, i_{\mathrm{pa}(D_j)}),$$

which shows that

$$v \perp\!\!\!\perp A \setminus (D_l \cup \mathrm{pa}_{\mathcal{G}}(D_l)) \,|\, D_l \cup \mathrm{pa}_{\mathcal{G}}(D_l)$$

Note also that $D_l = \mathrm{dis}_A(v)$, so

$$v \perp\!\!\!\perp A \setminus \mathrm{mb}(v, A) \,|\, \mathrm{mb}(v, A)$$

where $\mathrm{mb}(v, A) = \mathrm{dis}_A(v) \cup \mathrm{pa}_{\mathcal{G}}(\mathrm{dis}_A(v))$, which is just the ordered local Markov property for v and A. □

Chapter 2

Parametrization and Fitting

Suppose that we have independent and identically distributed observations generated from some positive binary probability distribution P. Let $\boldsymbol{p} = (p_i)_{i \in \mathfrak{X}_V}$ denote the vector of probabilities, where $p_i \equiv P(X_V = i)$. We can record data from this distribution as counts $\boldsymbol{n} = (n_i)_{i \in \mathfrak{X}_V}$, where n_i is the number of observations of the response pattern i.

We now consider the problem of fitting models to data generated in this manner, where P is assumed to obey the global Markov property with respect to some ADMG $\mathcal{G}$; this chapter closely follows Evans and Richardson (2010). Although we only consider binary probability distributions for ease of notation and explication, everything which follows is easily extended to general discrete state spaces.

Section 2.1 shows that the conditional probabilities used in (1.3) constitute a smooth parametrization of the associated model, and hence that ADMG models are curved exponential families. Section 2.2 motivates a maximum likelihood fitting algorithm using this parametrization, and formulates maps using matrices. The issue of inequality constraints on the parameters is tackled in Section 2.3. Section 2.4 gives the fitting algorithm, and finally Section 2.5 contains a formula for calculating asymptotic standard errors of maximum likelihood estimates.

2.1 Parametrizations

We denote the k-dimensional strictly positive probability simplex by Δ_k:

$$\Delta_k \equiv \left\{ \boldsymbol{p} \in \mathbb{R}^{k+1} \,\middle|\, \boldsymbol{p} > \mathbf{0}, \sum_{j=1}^{k+1} p_j = 1 \right\}.$$

Definition 2.1.1. Let $\mathcal{M} \subseteq \Delta_k$ be some collection of discrete probability distributions; $\mathcal{M}$ is referred to as a *model*. A *parametrization* is a bijective function $\boldsymbol{\theta} : \mathcal{M} \to \Theta$, for some open set $\Theta \subseteq \mathbb{R}^d$, called the parameter space.

We say that the parametrization $\boldsymbol{\theta}$ is *smooth* if $\boldsymbol{\theta}$ is twice continuously differentiable and the Jacobian $\frac{\partial \boldsymbol{\theta}}{\partial \boldsymbol{p}}$ of $\boldsymbol{\theta}(\boldsymbol{p})$ is of full rank d ($\leq k$) everywhere; this implies, by application of the inverse function theorem, that $\boldsymbol{\theta}^{-1}$ is also twice continuously differentiable. d is the dimension of the model.

If a model $\mathcal{M}$ admits a smooth parametrization, it is called a *curved exponential family* of order d.

We will thus assume that the probability distribution P is strictly positive; that is, $\boldsymbol{p} \in \Delta_k$. The collection of all positive probability measures on $\mathfrak{X}$, which is the model defined by $\mathcal{M} = \Delta_k$, is known as the *saturated model*.

For an ADMG $\mathcal{G}$, the model associated with $\mathcal{G}$, denoted $\mathcal{P}_\mathcal{G} \subseteq \Delta_k$, is the set of positive probability distributions which obey the global Markov property with respect to $\mathcal{G}$. From Proposition 1.2.4 and Theorem 1.3.14 it follows that we could equivalently define $\mathcal{P}_\mathcal{G}$ as the set of distributions which obey the ordered local Markov property with respect to $\mathcal{G}$, or which factorize according to (1.1).

Definition 2.1.2. Let $\mathfrak{X} = \{0,1\}^{|V|}$, so $X \in \mathfrak{X}$ is a binary vector. For $A \subseteq \{1, \ldots, |V|\}$, define

$$q_A = P(X_A = \mathbf{0}).$$

This is the *Möbius parameter* associated with A.

Drton and Richardson (2008a) show that the class of multivariate binary distributions obeying the global Markov property with respect to a bidirected graph $\mathcal{G}$ is smoothly parametrized by

$$\mathscr{Q}(\mathcal{G}) = \{q_A \mid A \text{ a connected subset of } \mathcal{G}\}.$$

Definition 2.1.3. Let $\mathfrak{X} = \{0,1\}^{|V|}$; for $A, B \subseteq \{1, \ldots, |V|\}$ with $A \cap B = \emptyset$, define

$$q_{A|B}^{i_B} = P(X_A = \mathbf{0} \mid X_B = i_B).$$

This is the *generalized Möbius parameter* associated with A, B and i_B.

In subscripts and superscripts, we generally omit braces and $\cup$ symbols for brevity so that, for example, q_{Av} appears in place of $q_{A\cup\{v\}}$. Similarly, we write q_{12} instead of $q_{\{1,2\}}$.

We will show that the generalized Möbius parameters can be used to construct a smooth parametrization of ADMG models. From (1.3) it follows that the collection of generalized Möbius parameters

$$\mathscr{Q}'(\mathcal{G}) \equiv \{q_{H|T}^{i_T} \mid H \in \mathcal{H}(\mathcal{G}), i_T \in \{0,1\}^{|T|}\},$$

is sufficient to fully specify a probability distribution in the model $\mathcal{P}_\mathcal{G}$. The probability distribution can be recovered using the functions

$$p_i(\boldsymbol{q}) = \sum_{C:O\subseteq C\subseteq V} (-1)^{|C\setminus O|} \prod_{H\in[C]_\mathcal{G}} q_{H|T}^{i_T}, \qquad i \in \mathfrak{X}_V.$$

Define $\mathcal{Q}_\mathcal{G}$ to be the collection of vectors $\boldsymbol{q}$ which may be obtained by computing the appropriate conditional probabilities for a distribution $\boldsymbol{p} \in \mathcal{P}_\mathcal{G}$. The following result shows that this set is open, and hence that $\mathcal{Q}_\mathcal{G}$ is of the appropriate dimension to be a parameter space for $\mathcal{P}_\mathcal{G}$.

Theorem 2.1.4. *For an ADMG $\mathcal{G}$, a vector of generalized Möbius parameters $\boldsymbol{q}$ is valid (i.e. $\boldsymbol{q} \in \mathcal{Q}_\mathcal{G}$) if and only if for each $i_V \in \mathfrak{X}_V$ we have*

$$f_{i_V}(\boldsymbol{q}) \equiv \sum_{C\,:\,i_V^{-1}(0)\subseteq C\subseteq V} (-1)^{|C\setminus i_V^{-1}(0)|} \prod_{H\in[C]_\mathcal{G}} q_{H|T}^{i_T} > 0, \tag{2.1}$$

where $i_V^{-1}(0) \equiv \{v \in V \mid i_v = 0\}$.

Remark 2.1.5. The boundary of the space is the set of $\boldsymbol{q}$ for which $f_{i_V}(\boldsymbol{q}) = 0$ for some $i_V \in \mathfrak{X}_V$.

The definition of $f_{i_V}(\boldsymbol{q})$ is just the expression given for $P(X_V = i_V)$ in (1.3) and so the result might at first seem trivial; clearly probabilities must be non-negative. However, it is not immediately obvious that this condition is *sufficient* for validity of the parameters. If we take some $\boldsymbol{q}^\dagger \notin \mathcal{Q}_\mathcal{G}$ and apply to it the non-linear functional form in (1.3) to obtain $\boldsymbol{p}(\boldsymbol{q}^\dagger)$, without this result there is no apparent reason why $\boldsymbol{p}(\boldsymbol{q}^\dagger)$ should not be a valid probability distribution, or indeed a probability distribution in $\mathcal{P}_\mathcal{G}$.

To prove Theorem 2.1.4, we need the following lemma.

Lemma 2.1.6. *Let A be an ancestral set in $\mathcal{G}$, and let $i_A \in \mathfrak{X}_A$. Then*

$$\sum_{j_V : j_A = i_A} f_{j_V}(\boldsymbol{q}) = \sum_{C\,:\, i_A^{-1}(0) \subseteq C \subseteq A} (-1)^{|C \setminus i_A^{-1}(0)|} \prod_{H \in [C]_\mathcal{G}} q_{H|T}^{i_T},$$

where $i_A^{-1}(0) \equiv \{v \in A \,|\, i_v = 0\}$. In particular,

$$\sum_{j_V} f_{j_V}(\boldsymbol{q}) = 1.$$

Proof. If $A = V$ the result is trivial. If not, pick some $v \in \mathrm{barren}_\mathcal{G}(V) \setminus A$; this is possible because if $A \supseteq \mathrm{barren}_\mathcal{G}(V)$ then $A = V$ by ancestrality of A. So

$$\begin{aligned}
\sum_{\substack{j_V \\ j_A = i_A}} f_{j_V}(\boldsymbol{q}) &= \sum_{\substack{j_V \\ j_A = i_A}} \sum_{j_V^{-1}(0) \subseteq C \subseteq V} (-1)^{|C \setminus j_V^{-1}(0)|} \prod_{H \in [C]_\mathcal{G}} q_{H|T}^{j_T} \\
&= \sum_{\substack{j_{V\setminus\{v\}} \\ j_A = i_A}} \sum_{j_v} \sum_{j_V^{-1}(0) \subseteq C \subseteq V} (-1)^{|C \setminus j_V^{-1}(0)|} \prod_{H \in [C]_\mathcal{G}} q_{H|T}^{j_T} \\
&= \sum_{\substack{j_{V\setminus\{v\}} \\ j_A = i_A}} \Bigg(\sum_{j_{V\setminus\{v\}}^{-1}(0) \subseteq C \subseteq V} (-1)^{|C \setminus j_{V\setminus\{v\}}^{-1}(0)|} \prod_{H \in [C]_\mathcal{G}} q_{H|T}^{j_T} \\
&\qquad\qquad + \sum_{j_{V\setminus\{v\}}^{-1}(0) \cup \{v\} \subseteq C \subseteq V} (-1)^{|C \setminus (j_{V\setminus\{v\}}^{-1}(0) \cup \{v\})|} \prod_{H \in [C]_\mathcal{G}} q_{H|T}^{j_T} \Bigg).
\end{aligned}$$

The last equation simply breaks the sum into cases where $j_v = 1$ and $j_v = 0$ respectively, which is possible because v does not appear in any tail sets. The first inner sum in the last expression can be further divided into case where C contains v, and those where it does not, giving

$$\begin{aligned}
\sum_{\substack{j_V \\ j_A = i_A}} f_{j_V}(\boldsymbol{q}) = \sum_{\substack{j_{V\setminus\{v\}} \\ j_A = i_A}} \Bigg(& \sum_{j_{V\setminus\{v\}}^{-1}(0) \subseteq C \subseteq V\setminus\{v\}} (-1)^{|C \setminus j_{V\setminus\{v\}}^{-1}(0)|} \prod_{H \in [C]_\mathcal{G}} q_{H|T}^{j_T} \\
& + \sum_{j_{V\setminus\{v\}}^{-1}(0) \cup \{v\} \subseteq C \subseteq V} (-1)^{|C \setminus j_{V\setminus\{v\}}^{-1}(0)|} \prod_{H \in [C]_\mathcal{G}} q_{H|T}^{j_T} \\
& + \sum_{j_{V\setminus\{v\}}^{-1}(0) \cup \{v\} \subseteq C \subseteq V} (-1)^{|C \setminus (j_{V\setminus\{v\}}^{-1}(0) \cup \{v\})|} \prod_{H \in [C]_\mathcal{G}} q_{H|T}^{j_T} \Bigg).
\end{aligned}$$

The second and third terms differ only by a factor of -1, and so cancel leaving

$$\sum_{\substack{j_V \\ j_A = i_A}} f_{j_V}(\boldsymbol{q}) = \sum_{\substack{j_{V\setminus\{v\}} \\ j_A = i_A}} \left(\sum_{C \,:\, j_{V\setminus\{v\}}^{-1}(0) \subseteq C \subseteq V\setminus\{v\}} (-1)^{|C \setminus j_{V\setminus\{v\}}^{-1}(0)|} \prod_{H \in [C]_{\mathcal{G}}} q_{H|T}^{j_T} \right).$$

Repeating this until no vertices outside A are left gives

$$\sum_{\substack{j_V \\ j_A = i_A}} f_{j_V}(\boldsymbol{q}) = \sum_{j_A^{-1}(0) \subseteq C \subseteq A} (-1)^{|C \setminus j_A^{-1}(0)|} \prod_{H \in [C]_{\mathcal{G}}} q_{H|T}^{j_T}.$$

□

Proof of Theorem 2.1.4. The 'only if' part of the statement follows from the fact that if the parameters are valid, then $f_{i_V}(\boldsymbol{q}) = P(X_V = i_V)$, and is therefore non-negative.

For the converse, suppose that the inequalities hold; we will show that we can retrieve the generalized Möbius parameters simply by calculating the appropriate conditional probabilities. Lemma 2.1.6 ensures that $\sum_{i_V} f_{i_V}(\boldsymbol{q}) = 1$, and that therefore this is a valid probability distribution.

Next, choose some $H^* \in \mathcal{H}(\mathcal{G})$, with $T^* = \mathrm{tail}_{\mathcal{G}}(H^*)$ and $A = \mathrm{an}_{\mathcal{G}}(H^*)$; also set $i_{H^*} = \mathbf{0}$ and pick $i_{T^*} \in \{0,1\}^{|T^*|}$. By Lemma 2.1.6,

$$\sum_{j_V \,:\, j_A = i_A} f_{j_V}(\boldsymbol{q}) = \sum_{j_A^{-1}(0) \subseteq C \subseteq A} (-1)^{|C \setminus j_A^{-1}(0)|} \prod_{H \in [C]_{\mathcal{G}}} q_{H|T}^{j_T}.$$

Now, applying Lemma 1.3.10 and the fact that $H^* \subseteq j_A^{-1}(0)$ means that we can factor out the parameter associated with H^*, giving

$$\begin{aligned} &= q_{H^*|T^*}^{j_{T^*}} \sum_{j_A^{-1}(0) \subseteq C \subseteq A} (-1)^{|C \setminus j_A^{-1}(0)|} \prod_{H \in [C\setminus H^*]_{\mathcal{G}}} q_{H|T}^{j_T} \\ &= q_{H^*|T^*}^{j_{T^*}} \sum_{j_{A\setminus H^*}^{-1}(0) \subseteq C \subseteq A\setminus H^*} (-1)^{|C \setminus j_{A\setminus H^*}^{-1}(0)|} \prod_{H \in [C]_{\mathcal{G}}} q_{H|T}^{j_T}. \end{aligned}$$

But note that $A \setminus H^*$ is also an ancestral set, and thus using Lemma 2.1.6 again,

$$\sum_{j_V \,:\, j_{A\setminus H^*} = i_{A\setminus H^*}} f_{j_V}(\boldsymbol{q}) = \sum_{j_{A\setminus H^*}^{-1}(0) \subseteq C \subseteq A\setminus H^*} (-1)^{|C \setminus j_{A\setminus H^*}^{-1}(0)|} \prod_{H \in [C]_{\mathcal{G}}} q_{H|T}^{j_T}.$$

And thus

$$\frac{\sum_{j_{V\setminus A}} f_{i_V}(\boldsymbol{q})}{\sum_{j_{V\setminus(A\setminus H^*)}} f_{i_V}(\boldsymbol{q})} = q_{H^*|T^*}^{i_{T^*}}.$$

So we can recover the original parameters from the probability distribution $\boldsymbol{f}$ in the manner we would expect; that $\boldsymbol{f}$ satisfies the global Markov property for $\mathcal{G}$ then follows from Theorem 1.4.1. Thus $\boldsymbol{f} \in \mathcal{P}_\mathcal{G}$ and $\boldsymbol{q} = \boldsymbol{q}(\boldsymbol{f}) \in \mathcal{Q}_\mathcal{G}$, so the generalized Möbius parameters are valid. □

This brings us to the main result in this section.

Theorem 2.1.7. *For an ADMG $\mathcal{G}$, the model $\mathcal{P}_\mathcal{G}$ of strictly positive binary probability distributions satisfying the global Markov property with respect to $\mathcal{G}$ is smoothly parametrized by the generalized Möbius parameters*

$$\mathscr{Q}'(\mathcal{G}) \equiv \{q_{H|T}^{i_T} \mid H \in \mathcal{H}(\mathcal{G}), i_T \in \{0,1\}^{|T|}\}.$$

Consequently the model $\mathcal{P}_\mathcal{G}$ is a curved exponential family of dimension $d = |\mathscr{Q}'(\mathcal{G})|$.

Proof. By Theorem 2.1.4, the set $\mathcal{Q}_\mathcal{G} \subseteq \mathbb{R}^d$ is open. The map $\boldsymbol{p}(\boldsymbol{q}) : \mathcal{Q}_\mathcal{G} \to \mathcal{P}_\mathcal{G}$ is multilinear, and therefore infinitely differentiable. Its inverse $\boldsymbol{q} : \mathcal{P}_\mathcal{G} \to \mathcal{Q}_\mathcal{G}$ is also infinitely differentiable.

The composition $\boldsymbol{q} \circ \boldsymbol{p}$ is the identity function on $\mathcal{Q}_\mathcal{G}$, and therefore its Jacobian is the identity matrix I_d. However, the Jacobian of a composition of differentiable functions is the product of the Jacobians, so

$$I_d = \frac{\partial \boldsymbol{q}}{\partial \boldsymbol{p}} \frac{\partial \boldsymbol{p}}{\partial \boldsymbol{q}}.$$

But this implies that each of the Jacobians has full rank d, and therefore the map $\boldsymbol{q}$ is a smooth parametrization of $\mathcal{P}_\mathcal{G}$. □

We note that the parametrization could also be achieved using ordinary Möbius parameters

$$\mathscr{Q}(\mathcal{G}) \equiv \{q_{H\cup A} \mid H \in \mathcal{H}(\mathcal{G}), A \subseteq T\}.$$

To see this, make the inductive hypothesis that the distribution over $\mathrm{an}_\mathcal{G}(H) \setminus H$ has been

parametrized, and note that

$$\begin{aligned} q_{H|T}^{i_T} &= P(X_H = \mathbf{0} \mid X_T = i_T) \\ &= \frac{P(X_H = \mathbf{0}, X_T = i_T)}{P(X_T = i_T)}, \end{aligned}$$

and since

$$P(X_H = \mathbf{0}, X_T = \mathbf{1}) = \sum_{C \subseteq T} (-1)^{|C|} q_{H \cup C},$$

the Möbius parameters $\mathscr{Q}$ suffice.

Some classes of discrete graphical models, such as AMP chain graphs, are not smooth and therefore not curved exponential families (Drton, 2009). It should also be remarked that all discrete ADMG models are everywhere identified on the interior of the simplex, which follows from the fact that the parameters are just conditional probabilities.

2.2 Motivating an Algorithm

Definition 2.2.1. Let θ_i, for $i = 1, \ldots, d$ be a collection of parameters such that θ_i takes values in a set Θ_i. We say that the vector $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_d)^T$ is *variation independent* if $\boldsymbol{\theta}$ can take any value in the set $\Theta_1 \times \cdots \times \Theta_d$.

Clearly the parametrization given above is not variation independent in general because, for example, $q_1 > q_{12}$. Indeed, the generalized Möbius parameters obey a complex pattern of variation dependence, as characterized in Theorem 2.1.4. To ensure that the parameters are valid, we need to verify that they yield valid probabilities. We proceed by reformulating the map between generalized Möbius parameters $\boldsymbol{q} \equiv (q_{H|T}^{i_T})$ and probabilities $\boldsymbol{p}$ with matrices.

Definition 2.2.2. We refer to a product of the form

$$\prod_{H \in [C]_{\mathcal{G}}} q_{H|T}^{i_T}, \tag{2.2}$$

as a *term*. From the fact that $[\cdot]_{\mathcal{G}}$ partitions sets, it is clear that for each $v \in C$, the term has exactly one factor $q_{H|T}^{i_T}$ whose head contains v; hence the expression for p_i is a multi-linear polynomial in the generalized Möbius parameters.

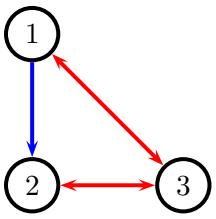


Figure 2.1: An ADMG, $\mathcal{G}_2$, used to illustrate construction of matrices M and P.

We show below that the expression (1.3) can be written in the form

$$\boldsymbol{p}(\boldsymbol{q}) = M \exp(P \log \boldsymbol{q}) \tag{2.3}$$

for matrices M and P; here the operations exp and log are taken pointwise over vectors.

We first restrict our attention to ADMGs containing only one district. In this case p_i is simply a sum of terms of the form (2.2) (up to sign), each characterized by C and the tail states i_T. We define a matrix M whose rows correspond to the possible states i, and whose columns correspond to possible terms of the form (2.2). Let M have (j,k)th entry ± 1 if the term associated with column k appears with that coefficient in the expression for the probability associated with row j; otherwise the entry is 0. For example, in the graph $\mathcal{G}_2$ in Figure 2.1,

$$p_{101} = q^{(1)}_{2|1} - q^{(1)}_{2|1}\, q_1 - q^{(1)}_{23|1} + q^{(1)}_{23|1}\, q_1.$$

The row of M associated with the state $(1,0,1)^T$ contains entries

$$\begin{array}{cccccccccccc}
\emptyset & \{1\} & \{2\}_{i_1=0} & \{2\}_{i_1=1} & \{1,2\}_{i_1=0} & \{1,2\}_{i_1=1} & \{3\} & \{1,3\} & \{2,3\}_{i_1=0} & \{2,3\}_{i_1=1} & \{1,2,3\}_{i_1=0} & \{1,2,3\}_{i_1=1} \\
0 & 0 & 0 & +1 & 0 & -1 & 0 & 0 & 0 & -1 & 0 & +1
\end{array}.$$

The full matrix M is

$$\begin{array}{c}
\\ p_{000} \\ p_{100} \\ p_{010} \\ p_{110} \\ p_{001} \\ p_{101} \\ p_{011} \\ p_{111}
\end{array}
\begin{array}{c}
\begin{array}{cccccccccccc}
\emptyset & \{1\} & \{2\}_{i_1=0} & \{2\}_{i_1=1} & \{1,2\}_{i_1=0} & \{1,2\}_{i_1=1} & \{3\} & \{1,3\} & \{2,3\}_{i_1=0} & \{2,3\}_{i_1=1} & \{1,2,3\}_{i_1=0} & \{1,2,3\}_{i_1=1}
\end{array} \\
\begin{pmatrix}
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & +1 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & +1 & 0 & -1 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & +1 & 0 & 0 & -1 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & +1 & -1 & 0 & -1 & 0 & +1 \\
0 & 0 & 0 & 0 & +1 & 0 & 0 & 0 & 0 & 0 & -1 & 0 \\
0 & 0 & 0 & +1 & 0 & -1 & 0 & 0 & 0 & -1 & 0 & +1 \\
0 & +1 & 0 & 0 & -1 & 0 & 0 & -1 & 0 & 0 & +1 & 0 \\
+1 & -1 & 0 & -1 & 0 & +1 & -1 & +1 & 0 & +1 & 0 & -1
\end{pmatrix}
\end{array}.$$

Note that the terms for $\{2\}$ with $i_1 = 0$ and $\{2,3\}$ with $i_1 = 0$ cannot logically occur, so those columns could be removed together with the corresponding rows of P below.

We create a second matrix P which contains a row for each term of the form (2.2), and a column for each element of $\boldsymbol{q}$; it will be used to map generalized Möbius parameters to terms. The (j,k)th entry of P is 1 if the term associated with row j contains the parameter associated with column k as a factor, and 0 otherwise. Thus in $\mathcal{G}_2$, for $C = \{1,2\}$ and $i_1 = 1$, the associated term is $q_{2|1}^{(1)}\, q_1$, and the associated row of P contains the entries

$$\begin{array}{ccccccc} q_1 & q_{2|1}^{(0)} & q_{2|1}^{(1)} & q_3 & q_{13} & q_{23|1}^{(0)} & q_{23|1}^{(1)} \\ \Big(1 & 0 & 1 & 0 & 0 & 0 & 0 \Big), \end{array}$$

where the parameters are shown above their respective columns.

It is clear that the operation $\exp(P \log \boldsymbol{q})$ maps the vector of parameters $\boldsymbol{q}$ to a vector containing all possible terms. The full matrix for $\mathcal{G}_2$ is

$$P = \begin{array}{c} \\ \emptyset \\ \{1\} \\ \{2\},i_1{=}0 \\ \{2\},i_1{=}1 \\ \{1,2\},i_1{=}0 \\ \{1,2\},i_1{=}1 \\ \{3\} \\ \{1,3\} \\ \{2,3\},i_1{=}0 \\ \{2,3\},i_1{=}1 \\ \{1,2,3\},i_1{=}0 \\ \{1,2,3\},i_1{=}1 \end{array} \begin{array}{c} \begin{array}{ccccccc} q_1 & q_{2|1}^{(0)} & q_{2|1}^{(1)} & q_3 & q_{13} & q_{23|1}^{(0)} & q_{23|1}^{(1)} \end{array} \\ \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix} \end{array}.$$

For a graph with multiple districts $D_1, \ldots, D_k$, it is most efficient to construct a pair (M^j, P^j) for each district D_j, so that

$$\boldsymbol{p}(\boldsymbol{q}) = \prod_j M^j \exp(P^j \log \boldsymbol{q}^j)$$

using the result of Lemma 1.4.2; here $\boldsymbol{q}^j$ is a vector of parameters whose heads are in the district D_j.

For binary random variables, the number of columns in M is $\sum_{H\in\mathcal{H}} 2^{|\operatorname{tail}(H)|}$, but there are at most $|\mathcal{H}|$ entries in each row. Hence M grows quickly for with district size, but will also be sparse if tail sets are large relative to the number of heads. Similar comments apply to P.

2.3 Inequality Constraints

Taken together with the result of Theorem 2.1.4, this means that we can check that the parameters $\boldsymbol{q}$ are legitimate by evaluating $\boldsymbol{p}(\boldsymbol{q}) = M\exp(P\log\boldsymbol{q})$, and ensuring that the resulting probabilities are positive. Note that $\sum_i p_i(\boldsymbol{q}) = 1$ follows from Lemma 2.1.6.

Recall that $\mathcal{P}_\mathcal{G}$ is the collection of all strictly positive probabilities distributions $\boldsymbol{p}$ which satisfy the global Markov property for $\mathcal{G}$; recall also that $\mathcal{Q}_\mathcal{G}$ is the image of $\mathcal{P}_\mathcal{G}$ under the map $\boldsymbol{p}^{-1} = \boldsymbol{q}$ which takes probabilities to generalized Möbius parameters.

We approach the fitting by constructing local constraints, considering only the parameters whose heads contain a particular vertex v: $\theta^v \equiv (q^{i_T}_{H|T} \mid v \in H \in \mathcal{H}(\mathcal{G}))$; the rest are held fixed. Since each term in the map $p_i(\boldsymbol{q})$ contains at most one factor with v in its head, $\boldsymbol{p}$ is a linear function of θ^v; i.e. $\boldsymbol{p} = A^v\theta^v - \boldsymbol{b}^v$ for some matrix A^v and vector $\boldsymbol{b}^v$. We need to ensure that $p_i > 0$ for each i, so the constraints amount to

$$A^v\theta^v > \boldsymbol{b}^v, \tag{2.4}$$

where the inequality is understood to act pointwise on vectors. For graphs with multiple districts the value of A^v, where v is in a district D, depends only on the value of parameters whose heads are contained in D.

2.4 Maximum Likelihood Estimation

Proposition 2.4.1. *Suppose that the observed counts n_i are all positive. Then for any ADMG $\mathcal{G}$ a maximum likelihood estimator of $\boldsymbol{q} \in \mathcal{Q}_\mathcal{G}$ exists.*

Proof. Let

$$l_n(\boldsymbol{q}) \equiv \sum_i n_i \log p_i(\boldsymbol{q})$$

be the log-likelihood function with respect to $\boldsymbol{q}$. From Theorem 2.1.4, on the boundary of $\mathcal{Q}_{\mathcal{G}}$ we have $p_i(\boldsymbol{q}) = 0$ for some $i \in \mathfrak{X}$; hence $l_n(\boldsymbol{q}) = -\infty$ on the boundary of $\mathcal{Q}_{\mathcal{G}}$ because $n_i > 0$.

Further, on the interior of $\mathcal{Q}_{\mathcal{G}}$ each p_i is positive, and therefore l_n is finite. Since $\mathcal{Q}_{\mathcal{G}}$ is bounded (generalized Möbius parameters lie between 0 and 1), its closure is compact. l_n is smooth, so it must attain a local maximum somewhere in the closure of $\mathcal{Q}_{\mathcal{G}}$, and since this point is clearly not on the boundary, it must be in the interior. □

Remark 2.4.2. The above result established the existence of a maximum likelihood estimate, but not its uniqueness, which is not guaranteed.

If some counts are not positive, as is often the case, a maximum likelihood estimator will still exist, but may be on the boundary of $\mathcal{Q}_{\mathcal{G}}$. Note also that $q_{H|T}^{(i_T)}$ may be unidentified if the event $X_T = i_T$ has not been observed.

The basis of our algorithm is a block co-ordinate updating scheme with gradient ascent. For simplicity we will assume that all the counts $\boldsymbol{n} = (n_i)_{i \in \{0,1\}^{|V|}}$ are strictly positive, so that by Proposition 2.4.1, the possibility of optima on the boundary need not be taken into account. In the case of zero counts, the partial likelihood function that is considered below is still concave but need no longer be strictly concave.

At each step we will increase the likelihood by updating the parameters whose heads contain a vertex v, considering each vertex in turn. The partial likelihood has the form

$$l(\theta^v) = \sum_i n_i \log p_i^v(\theta^v)$$

where p_i^v are purely linear functions in θ^v. This function is strictly concave in θ^v, and can be maximized using a gradient ascent approach, subject to the linear constraints $A^v\theta^v > \boldsymbol{b}^v$. A feasible starting value is easily found using, for example, full independence.

Algorithm 2.4.3. Cycle through each vertex $v \in V$ performing the following steps:

Step 1. Construct the constraint matrix A^v.

Step 2. Solve the non-linear program

$$\begin{aligned} &\text{maximize} && l(\theta^v) = \sum_i n_i \log p_i^v(\theta^v) \\ &\text{subject to} && A^v\theta^v > \boldsymbol{b}^v. \end{aligned}$$

Stop when a complete cycle of V results in a sufficiently small increase in the likelihood. □

The programme in Step 2 has a unique maximum θ^v, and is easy to solve using a gradient ascent method; a line search using the Armijo rule ensures convergence. See Bertsekas (1999) for examples. The maximum at each step is on the interior of $\mathcal{Q}_\mathcal{G}$, because if $A^v\theta^v = \boldsymbol{b}^v$ the log-likelihood takes the value $-\infty$ (see Proposition 2.4.1). The likelihood is guaranteed not to decrease at each step, and if the algorithm cycles through all vertices v without moving, we are guaranteed to have reached a (possibly local) maximum (see Drton and Eichler, 2006). For graphs with more than one district, we can apply Algorithm 2.4.3 to each district, possibly in parallel.

A 'black box' fitting algorithm could also be used to find ML estimates; however our approach gives more clarity to the parametrization and fitting problem. This approach also proves useful for implementing extensions to these models, such as with generalized Markov properties (Shpitser et al., 2011).

2.5 Standard Errors

Since this is a curved exponential family, asymptotic standard errors can be obtained from the Fisher information matrix, $I(\boldsymbol{q})$ (Johansen, 1979). Let $\boldsymbol{p}^* = \boldsymbol{p}(\boldsymbol{q}^*)$ be the 'true' probability distribution of X_V, where $\boldsymbol{p}^*$ is assumed not to be on the boundary of the simplex, and $\hat{\boldsymbol{p}} = \boldsymbol{p}(\hat{\boldsymbol{q}})$ the maximum likelihood estimate. Define the augmented likelihood l_λ for the sample by

$$l_\lambda(\boldsymbol{p}) = \sum_i n_i \log p_i + \lambda \left(1 - \sum_i p_i\right)$$

and note that $\nabla_{\boldsymbol{q}} l_\lambda = \nabla_{\boldsymbol{q}} l$ since $1 - \sum_i p_i(\boldsymbol{q}) = 0$ for all $\boldsymbol{q}$. We have

$$\nabla_{\boldsymbol{q}} l_\lambda(\boldsymbol{p}(\boldsymbol{q})) = \frac{\partial \boldsymbol{p}}{\partial \boldsymbol{q}} \cdot \nabla_{\boldsymbol{p}} l_\lambda(\boldsymbol{p}),$$

where

$$\frac{\partial l}{\partial p_i} = n_i p_i^{-1} - \lambda.$$

Choosing $\lambda = n$ gives $\mathbb{E}_{\boldsymbol{p}^*}\left[\nabla_{\boldsymbol{p}} l_\lambda(\boldsymbol{p}^*)\right] = \boldsymbol{0}$, and

$$n^{-1}\mathbb{E}_{\boldsymbol{p}^*}\left[\left(\nabla_{\boldsymbol{p}} l_\lambda\right)\left(\nabla_{\boldsymbol{p}} l_\lambda\right)^T\right] = \operatorname{diag} 1/\boldsymbol{p}^* - \boldsymbol{1}\boldsymbol{1}^T,$$

where $(1/\boldsymbol{p})_i = 1/p_i$. Thus

$$\begin{aligned} I(\boldsymbol{q}^*) &= n^{-1}\mathbb{E}_{\boldsymbol{p}(\boldsymbol{q}^*)}\left[(\nabla_{\boldsymbol{q}} l)\,(\nabla_{\boldsymbol{q}} l)^T\right] \\ &= \left(\left.\frac{\partial \boldsymbol{p}}{\partial \boldsymbol{q}}\right|_{\boldsymbol{q}=\boldsymbol{q}^*}\right)(\operatorname{diag} 1/\boldsymbol{p}^* - \mathbf{1}\mathbf{1}^T)\left(\left.\frac{\partial \boldsymbol{p}}{\partial \boldsymbol{q}}\right|_{\boldsymbol{q}=\boldsymbol{q}^*}\right)^T. \end{aligned}$$

Here

$$\frac{\partial \boldsymbol{p}}{\partial \boldsymbol{q}} = M \operatorname{diag}\left[\exp(P \log \boldsymbol{q})\right] P \frac{1}{\boldsymbol{q}},$$

where $\frac{1}{\boldsymbol{q}}$ is a vector with jth element $1/q_j$. Then

$$\sqrt{n}(\hat{\boldsymbol{q}} - \boldsymbol{q}^*) \xrightarrow{\mathcal{D}} \mathrm{N}(\mathbf{0}, I(\boldsymbol{q}^*)^{-1})$$

and we approximate the standard error of q_j by $\sqrt{[I(\hat{\boldsymbol{q}})^{-1}]_{jj}}$.

Chapter 3

Marginal Log-Linear Parameters

One difficulty with generalized Möbius parameters is their variation dependence, which we encountered in the previous chapter. The maximum likelihood fitting algorithm was carefully designed to overcome this problem, but we might ask whether a parametrization of the model exists which is variation independent.

In light of this question we turn our attention to marginal log-linear parameters, which have well understood variation dependence properties. In this chapter we again take $\mathfrak{X}$ to be a finite discrete probability space and P a strictly positive probability distribution over $\mathfrak{X}$.

Section 3.1 introduces marginal log-linear parameters and their properties; we present a parametrization of discrete ADMG models using them in Section 3.2, and show that they represent smooth sub-models of the saturated model in Section 3.3. Variation independence for the new parametrization is discussed in Section 3.4, followed by some alternative formulations in Section 3.5. Section 3.6 contains a simple method for recovering probabilities from marginal log-linear parameters.

3.1 Introduction

Marginal log-linear parameters, introduced by Bergsma and Rudas (2002), are a generalization of ordinary log-linear parameters, being defined with respect to a particular marginal distribution. Our exposition uses abstract collections of sets, so it may be helpful for the reader to keep in mind that the sets $M_i \in \mathbb{M}$ represent margins of a distribution over V, and each set $\mathbb{L}_i$ is a collection of effects in the margin M_i. Further, a pair (L, M) corresponds to a log-linear interaction over the set L, within the margin M.

Definition 3.1.1. For $L \subseteq M \subseteq V$, the pair (L, M) is an ordered pair of subsets of V. Let $\mathbb{P}$ be a collection of such pairs, and

$$\mathbb{M} \equiv \{M \,|\, (L, M) \in \mathbb{P} \text{ for some } L\},$$

be the collection of margins in $\mathbb{P}$. For a particular ordering $M_1, \ldots, M_k$ of the margins in $\mathbb{M}$, write

$$\mathbb{L}_i \equiv \{L \,|\, (L, M_i) \in \mathbb{P}\}.$$

We say that the collection $\mathbb{P}$ is *hierarchical* if there is some ordering on the M_i such that if $i < j$, then $M_j \not\subseteq M_i$ and also $L \in \mathbb{L}_j \Rightarrow L \not\subseteq M_i$; the second condition is equivalent to requiring that each effect L is associated with the first margin in the ordering of which it is a subset. We say the collection is *complete* if every non-empty subset of V is an element of precisely one set $\mathbb{L}_i$.

The case $L = \emptyset$ will not interest us as a parameter; in terms of a contingency table, $\lambda_\emptyset^M$ is determined by other parameters in the same margin and the sum over all cells, which we assume to be 1.

The term 'hierarchical' is used because each log-linear interaction is defined in the first possible margin in an ascending class; 'complete' is used because all interactions are present. Some papers (Rudas et al., 2010; Lupparelli et al., 2009) consider only collections which are complete. Various examples of these definitions can be found in Bergsma and Rudas (2002).

In this chapter, for disjoint sets $A, B \subseteq V$ we denote

$$p_A(i_A) \equiv P(X_A = i_A)$$
$$p_{A|B}(i_A \,|\, i_B) \equiv P(X_A = i_A \,|\, X_B = i_B).$$

For particular instantiations of small numbers of variables we write, for example,

$$p_{010} \equiv P(X_1 = 0,\, X_2 = 1,\, X_3 = 0)$$
$$p_{0\cdot 0} \equiv P(X_1 = 0,\, X_3 = 0).$$

Definition 3.1.2. For $\emptyset \subseteq L \subseteq M \subseteq V$ and $i_L \in \mathfrak{X}_L$, let

$$\nu_L^M(i_L) \equiv \frac{1}{|\mathfrak{X}_{M \setminus L}|} \sum_{\substack{j_M \in \mathfrak{X}_M \\ j_L = i_L}} \log p_M(j_M)$$

and

$$\lambda_L^M(i_L) \equiv \sum_{L' \subseteq L} (-1)^{|L \setminus L'|} \nu_{L'}^M(i_{L'}).$$

We call $\lambda_L^M(i_L)$ a *marginal log-linear parameter*. For a collection of ordered subsets $\mathbb{P}$ (see Definition 3.1.1), we let

$$\Lambda(\mathbb{P}) = \{\lambda_L^M(i_L) \,|\, (L, M) \in \mathbb{P},\ i_L \in \mathfrak{X}_L\}$$

be the collection of marginal log-linear parameters associated with $\mathbb{P}$.

Note that λ_L^M is just a Möbius transformation of ν_L^M, and thus

$$\nu_L^M(i_L) = \sum_{L' \subseteq L} \lambda_{L'}^M(i_{L'}).$$

We use λ_L^M to denote the collection $\{\lambda_L^M(i_L) \,|\, i_L \in \mathfrak{X}_L\}$, and in particular when we write $\lambda_L^M = 0$, we mean that every parameter in the collection is being set to zero. Similarly for distinct margins M and N we write $\lambda_L^M = \lambda_L^N$ to indicate that $\lambda_L^M(i_L) = \lambda_L^N(i_L)$ for every $i_L \in \mathfrak{X}_L$.

Theorem 3.1.3 (Bergsma and Rudas (2002), Theorem 2)**.** *Any complete and hierarchical collection of marginal log-linear parameters is a smooth parametrization of the saturated model, subject to the redundancy shown in Corollary 3.1.6.*

There are clearly myriad choices of margins; an illustration of two well known possibilities is given in the next example.

Example 3.1.4. *Log-linear and multivariate logistic parameters*

The ordinary log-linear parameters for the saturated model for a discrete distribution over a set of vertices V are $\{\lambda_L^V \,|\, \emptyset \neq L \subseteq V\}$. Following Bergsma and Rudas (2002), we denote $\mathbb{P}^{\max} \equiv \{(L, V) \,|\, \emptyset \neq L \subseteq V\}$; note that although $\lambda_\emptyset^V$ is a valid log-linear parameter, it is redundant for multinomial distrbutions. The multivariate logistic parameters of Glonek and McCullagh (1995) correspond to $\mathbb{P}^{\min} \equiv \{(L, L) \,|\, \emptyset \neq L \subseteq V\}$.

More generally, for an undirected graph $\mathcal{G}$, the set of discrete distributions P obeying the global Markov property with respect to $\mathcal{G}$ is parametrized by $\{\lambda_L^V \,|\, L \in \mathcal{C}(\mathcal{G})\}$, where $\mathcal{C}(\mathcal{G})$

is the collection of complete subsets of $\mathcal{G}$. With this in mind, we define

$$\mathbb{P}^{\max}(\mathcal{G}) \equiv \{(L, V) \,|\, \emptyset \neq L \in \mathcal{C}(\mathcal{G})\}$$

for any connected undirected graph $\mathcal{G}$; the case of disconnected graphs is trivial, since the separate components are completely independent, and may be parametrized as such.

We now examine some properties of marginal log-linear parameters.

Proposition 3.1.5. *We have*

$$\lambda_L^M(i_L) = \frac{1}{|\mathfrak{X}_M|} \sum_{j_M \in \mathfrak{X}_M} \log p_M(j_M) \prod_{v \in L} \left(|\mathfrak{X}_v| \mathbb{I}_{\{i_v = j_v\}} - 1 \right). \tag{3.1}$$

Here $\mathbb{I}_A$ is the indicator function of the event A.

Proof. First,

$$\begin{aligned}
\lambda_L^M(i_L) &= \sum_{L' \subseteq L} (-1)^{|L \setminus L'|} \nu_{L'}^M(i_{L'}) \\
&= \sum_{L' \subseteq L} (-1)^{|L \setminus L'|} \frac{1}{|\mathfrak{X}_{M \setminus L'}|} \sum_{\substack{j_M \in \mathfrak{X}_M \\ j_{L'} = i_{L'}}} \log p_M(j_M).
\end{aligned}$$

Now we establish the coefficient of $\log p_M(j_M)$ for some j_M. This probability will appear in terms of the inner sum if and only if L' is a (possibly empty) subset of $A = \{v \in L \,|\, i_v = j_v\}$. Thus, its coefficient is

$$\begin{aligned}
\sum_{L' \subseteq A} (-1)^{|L \setminus L'|} \frac{1}{|\mathfrak{X}_{M \setminus L'}|} &= \frac{1}{|\mathfrak{X}_{M \setminus A}|} (-1)^{|L \setminus A|} \sum_{L' \subseteq A} (-1)^{|A \setminus L'|} \frac{1}{|\mathfrak{X}_{A \setminus L'}|} \\
&= \frac{1}{|\mathfrak{X}_{M \setminus A}|} (-1)^{|L \setminus A|} \sum_{L' \subseteq A} (-1)^{|L'|} \frac{1}{|\mathfrak{X}_{L'}|} \\
&= \frac{1}{|\mathfrak{X}_{M \setminus A}|} (-1)^{|L \setminus A|} \sum_{L' \subseteq A} \prod_{v \in L'} \frac{-1}{|\mathfrak{X}_v|} \\
&= \frac{1}{|\mathfrak{X}_{M \setminus A}|} (-1)^{|L \setminus A|} \prod_{a \in A} \left(1 - \frac{1}{|\mathfrak{X}_a|} \right) \qquad \text{(binomial theorem)} \\
&= \frac{1}{|\mathfrak{X}_{M \setminus L}|} \prod_{v \in L} \left(\mathbb{I}_{\{i_v = j_v\}} - \frac{1}{|\mathfrak{X}_v|} \right) \\
&= \frac{1}{|\mathfrak{X}_M|} \prod_{v \in L} \left(|\mathfrak{X}_v| \mathbb{I}_{\{i_v = j_v\}} - 1 \right),
\end{aligned}$$

where we take the empty product to be 1. Using this in the expression above gives the result. □

The form of $\lambda_L^M(i_L)$ in (3.1) is generally easier to apply than the recursive definition. We obtain two corollaries to this result.

Corollary 3.1.6. *For any $v \in L$, and fixed $i_{L\setminus\{v\}}$*

$$\sum_{i_v \in \mathfrak{X}_v} \lambda_L^M(i_{L\setminus\{v\}}, i_v) = 0.$$

That is, the sum of the parameters across the support of any variable is 0.

Proof. The coefficient of $\log p_M(j_M)$ is

$$\sum_{i_v \in \mathfrak{X}_v} \prod_{w \in L} \left(|\mathfrak{X}_w|^{\mathbb{I}_{\{i_w = j_w\}}} - 1\right) = \prod_{w \in L\setminus\{v\}} \left(|\mathfrak{X}_w|^{\mathbb{I}_{\{i_w = j_w\}}} - 1\right) \sum_{i_v \in \mathfrak{X}_v} \left(|\mathfrak{X}_v|^{\mathbb{I}_{\{i_v = j_v\}}} - 1\right),$$

where the summand takes the value -1 in $|\mathfrak{X}_v| - 1$ of the possible values of i_v, and the value $|\mathfrak{X}_v| - 1$ once. Hence the result. □

The next provides a useful linear transformation of marginal log-linear parameters to a conditional form, which we apply in Theorem 3.2.3 and Section 3.6.

Corollary 3.1.7. *For disjoint sets $L, N \subseteq V$ where L is non-empty, with $M = L \cup N$ and $i_M \in \mathfrak{X}_M$, let*

$$\kappa_{L|N}(i_L \,|\, i_N) \equiv \sum_{L \subseteq A \subseteq M} \lambda_A^M(i_A).$$

Then

$$\begin{aligned}\kappa_{L|N}(i_L \,|\, i_N) &= \frac{1}{|\mathfrak{X}_L|} \sum_{j_L \in \mathfrak{X}_L} \log p_M(j_L,\, i_N) \prod_{v \in L} \left(|\mathfrak{X}_v|^{\mathbb{I}_{\{i_v = j_v\}}} - 1\right) \\ &= \frac{1}{|\mathfrak{X}_L|} \sum_{j_L \in \mathfrak{X}_L} \log p_{L|N}(j_L \,|\, i_N) \prod_{v \in L} \left(|\mathfrak{X}_v|^{\mathbb{I}_{\{i_v = j_v\}}} - 1\right).\end{aligned}$$

Proof.

$$\begin{aligned}
&\kappa_{L|N}(i_L \,|\, i_N) \\
&= \sum_{L\subseteq A\subseteq M} \frac{1}{|\mathfrak{X}_M|} \sum_{j_M\in\mathfrak{X}_M} \log p_M(j_M) \prod_{v\in A} \left(|\mathfrak{X}_v|\mathbb{I}_{\{i_v=j_v\}} - 1\right) \\
&= \frac{1}{|\mathfrak{X}_M|} \sum_{j_M\in\mathfrak{X}_M} \log p_M(j_M) \sum_{L\subseteq A\subseteq M} \prod_{v\in A} \left(|\mathfrak{X}_v|\mathbb{I}_{\{i_v=j_v\}} - 1\right) \\
&= \frac{1}{|\mathfrak{X}_M|} \sum_{j_M\in\mathfrak{X}_M} \log p_M(j_M) \sum_{L\subseteq A\subseteq M} \prod_{v\in L} \left(|\mathfrak{X}_v|\mathbb{I}_{\{i_v=j_v\}} - 1\right) \prod_{v\in A\setminus L} \left(|\mathfrak{X}_v|\mathbb{I}_{\{i_v=j_v\}} - 1\right) \\
&= \frac{1}{|\mathfrak{X}_M|} \sum_{j_M\in\mathfrak{X}_M} \log p_M(j_M) \prod_{v\in L} \left(|\mathfrak{X}_v|\mathbb{I}_{\{i_v=j_v\}} - 1\right) \sum_{B\subseteq N} \prod_{v\in B} \left(|\mathfrak{X}_v|\mathbb{I}_{\{i_v=j_v\}} - 1\right).
\end{aligned}$$

Now, consider the value of the inner sum, for a fixed j_M. In the case that there is some $w \in N$ with $i_w \neq j_w$, then

$$\begin{aligned}
\sum_{B\subseteq N} \prod_{v\in B} \left(|\mathfrak{X}_v|\mathbb{I}_{\{i_v=j_v\}} - 1\right) &= \sum_{B\subseteq N\setminus\{w\}} \left[\prod_{v\in B} \left(|\mathfrak{X}_v|\mathbb{I}_{\{i_v=j_v\}} - 1\right) + \prod_{v\in B\cup\{w\}} \left(|\mathfrak{X}_v|\mathbb{I}_{\{i_v=j_v\}} - 1\right) \right] \\
&= \sum_{B\subseteq N\setminus\{w\}} \left[\prod_{v\in B} \left(|\mathfrak{X}_v|\mathbb{I}_{\{i_v=j_v\}} - 1\right) - \prod_{v\in B} \left(|\mathfrak{X}_v|\mathbb{I}_{\{i_v=j_v\}} - 1\right) \right] \\
&= 0.
\end{aligned}$$

Alternatively, if $i_N = j_N$, then

$$\begin{aligned}
\sum_{B\subseteq N} \prod_{v\in B} \left(|\mathfrak{X}_v|\mathbb{I}_{\{i_v=j_v\}} - 1\right) &= \sum_{B\subseteq N} \prod_{v\in B} \left(|\mathfrak{X}_v| - 1\right) \\
&= |\mathfrak{X}_N|,
\end{aligned}$$

again by the binomial theorem. This last expression is independent of j_M, so

$$\kappa_{L|N}(i_L \,|\, i_N) = \frac{1}{|\mathfrak{X}_L|} \sum_{j_L\in\mathfrak{X}_L} \log p_M(j_L, i_N) \prod_{v\in L} \left(|\mathfrak{X}_v|\mathbb{I}_{\{i_v=j_v\}} - 1\right),$$

since $\mathfrak{X}_M = \mathfrak{X}_L \times \mathfrak{X}_N$. For the second form given, it was noted in the proof of Corollary 3.1.6 that if L is non-empty,

$$\sum_{j_L\in\mathfrak{X}_L} \prod_{v\in L} \left(|\mathfrak{X}_v|\mathbb{I}_{\{i_v=j_v\}} - 1\right) = 0,$$

so

$$\kappa_{L|N}(i_L \,|\, i_N) = \frac{1}{|\mathfrak{X}_L|} \sum_{j_L \in \mathfrak{X}_L} \log p_M(j_L, i_N) \prod_{v \in L} \left(|\mathfrak{X}_v| \mathbb{I}_{\{i_v = j_v\}} - 1 \right) - \frac{1}{|\mathfrak{X}_L|} \sum_{j_L \in \mathfrak{X}_L} \log p_N(i_N) \prod_{v \in L} \left(|\mathfrak{X}_v| \mathbb{I}_{\{i_v = j_v\}} - 1 \right)$$

and bringing both terms into a single sum gives the result. □

Remark 3.1.8. For $v \in V$, define $\tilde{\mathfrak{X}}_v \equiv \{0, 1, \ldots, |\mathfrak{X}_v| - 2\}$, that is $\mathfrak{X}_v$ without its largest value, and $\tilde{\mathfrak{X}}_A$ for $A \subseteq V$ analogously. From Corollary 3.1.6, the collection

$$\{\lambda_L^M(i_L) \,|\, i_L \in \mathfrak{X}_L\}$$

is determined completely by

$$\{\lambda_L^M(i_L) \,|\, i_L \in \tilde{\mathfrak{X}}_L\}.$$

Hence

$$\tilde{\Lambda}(\mathbb{P}) = \{\lambda_L^M(i_L) \,|\, (L, M) \in \mathbb{P},\ i_L \in \tilde{\mathfrak{X}}_L\}$$

fully determines $\Lambda(\mathbb{P})$. Bergsma and Rudas (2002) show that this collection of parameters contains no redundancies.

We often consider the easier binary case, when $\tilde{\mathfrak{X}}_L$ contains just one state, $\mathbf{0}$. Letting $\|i\|$ be the number of 1s in the binary vector i,

$$\lambda_L^M(\mathbf{0}) = \frac{1}{2^{|M|}} \sum_{i_M \in \{0,1\}^{|M|}} (-1)^{\|i_L\|} \log p_M(i_M) \tag{3.2}$$
$$= \frac{1}{2^{|M|}} \log \frac{\prod_{i_M : \|i_L\| \text{ even}} p_M(i_M)}{\prod_{i_M : \|i_L\| \text{ odd}} p_M(i_M)}$$

For example,

$$\lambda_{23}^{123}(\mathbf{0}) = \frac{1}{8} \log \frac{p_{000}\, p_{100}\, p_{011}\, p_{111}}{p_{010}\, p_{001}\, p_{110}\, p_{101}}.$$

Naturally, we are interested in relating these parameters to conditional independences; this is aided by the following result, which is part of Lemma 1 in Rudas et al. (2010), and Equation (6) of Forcina et al. (2010).

Lemma 3.1.9. *For any disjoint sets A, B and C, where C may be empty, $X_A \perp\!\!\!\perp X_B \,|\, X_C$ if and only if*

$$\lambda^{ABC}_{A'B'C'} = 0 \qquad \text{for every} \qquad \emptyset \neq A' \subseteq A, \quad \emptyset \neq B' \subseteq B, \quad C' \subseteq C.$$

The special case $C = \emptyset$ (giving marginal independence) was proved in the context of multivariate logistic parameters by Kauermann (1997).

Example 3.1.10. Suppose that we have a complete and hierarchical parametrization of 3 variables,

$$\lambda^1_1 \qquad \lambda^2_2 \qquad \lambda^3_3 \qquad \lambda^{12}_{12} \qquad \lambda^{13}_{13} \qquad \lambda^{123}_{23} \qquad \lambda^{123}_{123}.$$

We can obtain $X_1 \perp\!\!\!\perp X_3$ by setting $\lambda^{13}_{13} = 0$. Similarly, $X_2 \perp\!\!\!\perp X_3 \,|\, X_1$ if we set $\lambda^{123}_{23} = \lambda^{123}_{123} = 0$.

3.2 Parametrizations of Acyclic Directed Mixed Graphs

We now present a method for parametrizing ADMGs using marginal log-linear parameters. As shown in Section 2.1, the collection of binary probability distributions obeying the global Markov property for an ADMG $\mathcal{G}$ is parametrized by

$$\mathscr{Q}'(\mathcal{G}) \equiv \{q^{i_T}_{H|T} \mid H \in \mathcal{H}(\mathcal{G}),\ i_T \in \mathfrak{X}_T\}.$$

Let

$$\mathbb{M} = \{H \cup T \,|\, H \in \mathcal{H}(\mathcal{G})\}.$$

Further, if $M_i = H_i \cup T_i$ for some head H_i, then let $\mathbb{L}_i = \{A \,|\, H_i \subseteq A \subseteq H_i \cup T_i\}$. We call this collection of $\mathbb{M}$ and the $\mathbb{L}_i$s the *ingenuous* parametrization of $\mathcal{G}$, denoted $\mathbb{P}^{\text{ing}}(\mathcal{G})$.

Example 3.2.1. For the graph in Figure 1.3, recall that the head-tail pairs are

H	$\{1\}$	$\{2\}$	$\{3\}$	$\{2,3\}$	$\{4\}$	$\{3,4\}$
T	$\emptyset$	$\{1\}$	$\emptyset$	$\{1\}$	$\{2\}$	$\{1,2\}$

.

So the ingenuous parametrization consists of

M_i	$\mathbb{L}_i$
$\{1\}$	$\{1\}$
$\{1,2\}$	$\{2\},\{1,2\}$
$\{3\}$	$\{3\}$
$\{1,2,3\}$	$\{2,3\},\{1,2,3\}$
$\{2,4\}$	$\{4\},\{2,4\}$
$\{1,2,3,4\}$	$\{3,4\},\{1,3,4\},\{2,3,4\},\{1,2,3,4\}$.

Note that this gives a hierarchical ordering, and indeed such an ordering can always be obtained, as our next result shows.

Lemma 3.2.2. *For an ADMG $\mathcal{G}$, there is an ordering on the margins M_i of the ingenuous parametrization which is hierarchical.*

Proof. We need to show that some ordering of the margins respects inclusion, and that each effect is associated with the earliest margin of which it is a subset.

Recall that the partial ordering on heads $\prec$ is well-defined (see Lemma 1.3.6). Since there is a one-to-one correspondence between heads H_i and margins M_i, this induces a partial ordering on margins. We claim that a total ordering on the margins is hierarchical if it respects the partial ordering.

Suppose $M_i \subseteq M_j$. This implies that

$$H_i \cup T_i \subseteq H_j \cup T_j \subseteq \mathrm{an}_{\mathcal{G}}(H_j),$$

and so $H_i \subseteq \mathrm{an}_{\mathcal{G}}(H_j)$, giving $H_i \prec H_j$. Hence the ordering respects inclusion.

Now suppose that $A \in \mathbb{L}_i$ and $A \subseteq M_j$. The first condition implies that $H_i \subseteq A$, and the second that $A \subseteq H_j \cup T_j \subseteq \mathrm{an}_{\mathcal{G}}(H_j)$, which together imply $H_i \subseteq \mathrm{an}_{\mathcal{G}}(H_j)$, and so again $H_i \prec H_j$. Hence under the ordering, all effects are subsets of the earliest margin of which they are a subset. □

One can also show that if the total ordering on margins does not respect the partial ordering $\prec$, then the ordering cannot be hierarchical, but this is not necessary for the result.

The next two results demonstrate that the ingenuous parameters for an ADMG provide a smooth parametrization of the associated model.

Theorem 3.2.3. *The collection of parameters*

$$\{\lambda_A^M(i_A) \,|\, L \subseteq A \subseteq M, i_M \in \tilde{\mathfrak{X}}_M\},$$

together with the $(|L|-1)$*-dimensional marginal distributions of* X_L *conditional on* $X_{M\setminus L}$*, smoothly parametrizes the saturated distribution of* X_L *conditional on* $X_{M\setminus L}$*.*

Proof. First we show that we can construct all the local log $|L|$-way interaction parameters ($L \neq \emptyset$) as a smooth function of the given MLL parameters. Let $N \equiv M \setminus L$, and recall from Corollary 3.1.7 that the parameters $\kappa_{L|N}$ are just a linear transformation of λ_A^M for $L \subseteq A \subseteq M$.

Pick some $i_L \in \tilde{\mathfrak{X}}_L$ and $i_N \in \mathfrak{X}_N$; for $A \subseteq L$, let $\mathbf{1}_A$ denote a vector of length $|L|$ with a 1 in position t if $t \in A$, and 0 otherwise. Consider

$$\begin{aligned}
&\sum_{A\subseteq L} (-1)^{|L\setminus A|} \kappa_{L|N}(i_L + \mathbf{1}_A \,|\, i_N) \\
&= \frac{1}{|\mathfrak{X}_L|} \sum_{j_L \in \mathfrak{X}_L} \log p_M(j_L, i_N) \sum_{A\subseteq L} (-1)^{|L\setminus A|} \prod_{v\in L} \left(|\mathfrak{X}_v|^{\mathbb{I}_{\{i_v + \mathbb{I}_{\{v\in A\}} = j_v\}}} - 1 \right).
\end{aligned}$$

The coefficient of $\log p_M(j_M)$ for some j_M depends upon the inner sum in this expression. If for some $w \in L$, $j_w \notin \{i_w, i_w + 1\}$, its value is

$$\begin{aligned}
&\sum_{A\subseteq L} (-1)^{|L\setminus A|} \prod_{v\in L} \left(|\mathfrak{X}_v|^{\mathbb{I}_{\{i_v + \mathbb{I}_{\{v\in A\}} = j_v\}}} - 1 \right) \\
&\quad = \sum_{A\subseteq L\setminus\{w\}} (-1)^{|L\setminus A|} \left[\prod_{v\in L} \left(|\mathfrak{X}_v|^{\mathbb{I}_{\{i_v + \mathbb{I}_{\{v\in A\}} = j_v\}}} - 1 \right) - \prod_{v\in L} \left(|\mathfrak{X}_v|^{\mathbb{I}_{\{i_v + \mathbb{I}_{\{v\in A\cup\{w\}\}} = j_v\}}} - 1 \right) \right] \\
&\quad = 0,
\end{aligned}$$

because the value of the outer indicator function is 0 in both terms when $v = w$, and the two indicators are clearly equal for all other v.

Otherwise, if $j_w \in \{i_w, i_w + 1\}$ for all $w \in L$, then

$$B(A) \equiv \{v \in L \,|\, i_v + \mathbb{I}_{\{v\in A\}} = j_v\}$$

defines a one-to-one map from $\mathscr{P}(L)$ to itself. Hence we can rewrite:

$$\sum_{A\subseteq L} (-1)^{|L\setminus A|} \prod_{v\in L} \left(|\mathfrak{X}_v|^{\mathbb{I}_{\{i_v + \mathbb{I}_{\{v\in A\}} = j_v\}}} - 1 \right)$$

$$
\begin{aligned}
&= (-1)^{\|i_L - j_L\|} \sum_{B \subseteq L} (-1)^{|L \setminus B|} \prod_{v \in L} \left(|\mathfrak{X}_v| \mathbb{I}_{\{v \in B\}} - 1 \right) \\
&= (-1)^{\|i_L - j_L\|} \sum_{B \subseteq L} \prod_{v \in B} \left(|\mathfrak{X}_v| \mathbb{I}_{\{v \in B\}} - 1 \right) \\
&= (-1)^{\|i_L - j_L\|} \prod_{v \in L} |\mathfrak{X}_v| \\
&= (-1)^{\|i_L - j_L\|} |\mathfrak{X}_L|,
\end{aligned}
$$

where $\|i_L - j_L\|$ is just the number of entries in which i_L and j_L differ. Then since $\|i_L - j_L\| = |A|$,

$$
\begin{aligned}
\sum_{A \subseteq L} (-1)^{|L \setminus A|} \kappa_{L|N}(i_L + \mathbf{1}_A \,|\, i_N) &= \sum_{A \subseteq L} (-1)^{|A|} \log p_M(i_L + \mathbf{1}_A,\, i_N) \\
&= \sum_{A \subseteq L} (-1)^{|A|} \left\{ \log p_{L|N}(i_L + \mathbf{1}_A \,|\, i_N) + \log p_N(i_N) \right\}
\end{aligned}
$$

and because L is non-empty, it has as many even subsets as odd subsets, so

$$
= \sum_{A \subseteq L} (-1)^{|A|} \log p_{L|N}(i_L + \mathbf{1}_A \,|\, i_N),
$$

which is the (conditional) local log $|L|$-way interaction.

The collection of all the (conditional) local log $|L|$-way interactions together with the (conditional) $(|L| - 1)$-dimensional marginal distributions smoothly parametrizes the $|L|$-way table (Csiszár, 1975). □

Corollary 3.2.4. *The ingenuous parametrization $\tilde{\Lambda}(\mathbb{P}^{\text{ing}}(\mathcal{G}))$ of an ADMG $\mathcal{G}$ smoothly parametrizes precisely those distributions P obeying the global Markov property with respect to $\mathcal{G}$.*

Proof. We proceed by induction. For the base case, we know that the distribution of a singleton head $\{h\}$ with empty tail is parametrized by the logits

$$
\lambda_h^h(i+1) - \lambda_h^h(i) = \frac{1}{|\mathfrak{X}_h|} \log \frac{P(X_h = i+1)}{P(X_h = i)},
$$

for each $i \in \tilde{\mathfrak{X}}_h$.

We use the partial ordering $\prec$ on heads from Definition 1.3.5: $H_i \prec H_j$ if $H_i \neq H_j$ and $H_i \subset \text{an}_{\mathcal{G}}(H_j)$.

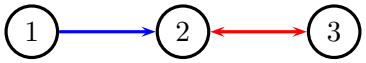


Figure 3.1: A small graph used to illustrate the ingenuous parametrization.

Suppose that we wish to find the distribution of a head H conditional on its tail T. Assume for induction that we have the distribution of all heads H' which precede H under $\prec$, conditional on their respective tails; we claim this is sufficient to give the $(|H|-1)$-dimensional marginal distributions of H conditional on T.

Let $v \in H$, and let $L = H \setminus \{v\}$ be a $(|H|-1)$-dimensional marginal of interest. The set $A = \mathrm{an}_{\mathcal{G}}(H) \setminus \{v\}$ is ancestral, since v cannot have (non-trivial) descendants in $\mathrm{an}_{\mathcal{G}}(H)$; in particular $L \cup T \subseteq A$. By Theorem 1.3.14,

$$P(X_A) = \prod_{H' \in [A]_{\mathcal{G}}} P(X_{H'} \,|\, X_{\mathrm{tail}\, H'}).$$

But all the probabilities given here are smoothly parametrized by the induction hypothesis, and the distribution of L conditional on T is given by the joint distribution of A.

The ingenuous parametrization, by definition, contains $\lambda_A^{H \cup T}$ for $H \subseteq A \subseteq H \cup T$, and thus the result follows from Theorem 3.2.3 above. □

Example 3.2.5. Consider the graph in Figure 3.1. The head-tail pairs are $(\{1\}, \emptyset)$, $(\{2\}, \{1\})$, $(\{3\}, \emptyset)$ and $(\{2,3\}, \{1\})$, giving the parameterization

$$\lambda_1^1, \qquad \lambda_2^{12}, \qquad \lambda_{12}^{12}, \qquad \lambda_3^3, \qquad \lambda_{23}^{123}, \qquad \lambda_{123}^{123},$$

up to the redundancy in Corollary 3.1.6. In the binary case,

$$\lambda_1^1(0) = \frac{1}{2} \log \frac{p_{0\cdot\cdot}}{p_{1\cdot\cdot}} = \log \frac{q_1}{1-q_1}, \qquad \lambda_3^3(0) = \frac{1}{2} \log \frac{p_{\cdot\cdot 0}}{p_{\cdot\cdot 1}} = \log \frac{q_3}{1-q_3},$$

from which the generalized Möbius parameters q_1 and q_3 can be recovered. Also,

$$\lambda_2^{12}(0) = \frac{1}{4} \log \frac{p_{00\cdot} p_{10\cdot}}{p_{01\cdot} p_{11\cdot}}, \qquad \lambda_{12}^{12}(0,0) = \frac{1}{4} \log \frac{p_{00\cdot} p_{11\cdot}}{p_{01\cdot} p_{10\cdot}},$$

so

$$\kappa_{2|1}(0\,|\,0) = \lambda_2^{12}(0) + \lambda_{12}^{12}(0,0) = \frac{1}{2} \log \frac{p_{00\cdot}}{p_{01\cdot}} = \frac{1}{2} \log \frac{q_{2|1}^{(0)}}{1 - q_{2|1}^{(0)}},$$

$$\kappa_{2|1}(0\,|\,1) = \lambda_2^{12}(0) - \lambda_{12}^{12}(0,0) = \frac{1}{2}\log\frac{p_{10\cdot}}{p_{11\cdot}} = \frac{1}{2}\log\frac{q_{2|1}^{(1)}}{1-q_{2|1}^{(1)}},$$

which gives us simple equations for $q_{2|1}^{(0)}$ and $q_{2|1}^{(1)}$. Lastly,

$$\begin{aligned}
\kappa_{23|1}(0,0\,|\,0) &= \lambda_{23}^{123}(0,0) + \lambda_{123}^{123}(0,0,0) \\
&= \frac{1}{4}\log\frac{p_{000}\,p_{011}}{p_{001}\,p_{010}} \\
&= \frac{1}{4}\log\frac{P(X_2=0,X_3=0\,|\,X_1=0)\cdot P(X_2=1,X_3=1\,|\,X_1=0)}{P(X_2=0,X_3=1\,|\,X_1=0)\cdot P(X_2=1,X_3=0\,|\,X_1=0)} \\
&= \frac{1}{4}\log\frac{q_{23|1}^{(0)}\cdot(1-q_{2|1}^{(0)}-q_3+q_{23|1}^{(0)})}{(q_{2|1}^{(0)}-q_{23|1}^{(0)})\cdot(q_3-q_{23|1}^{(0)})}
\end{aligned}$$

which can be rearranged to give a quadratic equation for $q_{23|1}^{(0)}$. We use $\kappa_{23|1}(0,0\,|\,1) = \lambda_{23}^{123}(0,0) - \lambda_{123}^{123}(0,0,0)$ to obtain $q_{23|1}^{(1)}$ similarly.

The following lemma is useful in the context of demonstrating that the ingenuous parametrization arises from sub-models of the ingenuous parametrization of complete graphs.

Lemma 3.2.6. *Suppose that $X_A \perp\!\!\!\perp X_B \mid X_C$, and A is non-empty. For any $D \subseteq C$,*

$$\lambda_{AD}^{ABC} = \lambda_{AD}^{AC}.$$

Proof. We have

$$\begin{aligned}
\lambda_{AD}^{ABC}(i_{AD}) &= \sum_{L'\subseteq A\cup D}(-1)^{|(A\cup D)\setminus L'|}\frac{1}{|\mathfrak{X}_{(A\cup B\cup C)\setminus L'}|}\sum_{\substack{j_{ABC}\in\mathfrak{X}_{ABC}\\ j_{L'}=i_{L'}}}\log p_{ABC}(j_{ABC}) \\
&= \sum_{L'\subseteq A\cup D}\frac{(-1)^{|(A\cup D)\setminus L'|}}{|\mathfrak{X}_{(A\cup B\cup C)\setminus L'}|}\sum_{\substack{j_{ABC}\in\mathfrak{X}_{ABC}\\ j_{L'}=i_{L'}}}\left(\log p_{AC}(j_{AC}) + \log p_{B|C}(j_B\,|\,j_C)\right). \quad (3.3)
\end{aligned}$$

Let

$$c(L') \equiv \frac{(-1)^{|(A\cup D)\setminus L'|}}{|\mathfrak{X}_{(A\cup B\cup C)\setminus L'}|}.$$

The sum over the terms involving $\log p_{AC}(j_{AC})$ is

$$\begin{aligned}
&\sum_{L' \subseteq A \cup D} c(L') |\mathfrak{X}_B| \sum_{\substack{j_{AC} \in \mathfrak{X}_{AC} \\ j_{L'} = i_{L'}}} \left(\frac{1}{|\mathfrak{X}_B|} \sum_{j_B \in \mathfrak{X}_B} \log p_{AC}(j_{AC}) \right) \\
&\quad = \sum_{L' \subseteq A \cup D} (-1)^{|(A \cup D) \setminus L'|} \frac{1}{|\mathfrak{X}_{(A \cup C) \setminus L'}|} \sum_{\substack{j_{AC} \in \mathfrak{X}_{AC} \\ j_{L'} = i_{L'}}} \log p_{AC}(j_{AC}) \\
&\quad = \lambda^{AC}_{AD}(i_{AD}).
\end{aligned}$$

Taking the remaining terms in (3.3) (i.e. those involving $\log p_{B|C}(j_B \,|\, j_C)$) we pick an arbitrary element $a \in A$ and separate the outer sum into cases where $a \in L'$ and $a \notin L'$:

$$\begin{aligned}
&\sum_{L' \subseteq A \cup D \setminus \{a\}} c(L') \left(\sum_{\substack{j_{ABC} \in \mathfrak{X}_{ABC} \\ j_{L'} = i_{L'}}} \log p_{B|C}(j_B \,|\, j_C) - |\mathfrak{X}_a| \sum_{\substack{j_{ABC} \in \mathfrak{X}_{ABC} \\ j_{L'} = i_{L'}}} \mathbb{I}_{\{j_a = i_a\}} \log p_{B|C}(j_B \,|\, j_C) \right) \\
&\quad = \sum_{L' \subseteq A \cup D \setminus \{a\}} c(L') \sum_{\substack{j_{ABC} \in \mathfrak{X}_{ABC} \\ j_{L'} = i_{L'}}} \left(\log p_{B|C}(j_B \,|\, j_C) - |\mathfrak{X}_a| \mathbb{I}_{\{j_a = i_a\}} \log p_{B|C}(j_B \,|\, j_C) \right) \\
&\quad = \sum_{L' \subseteq A \cup D \setminus \{a\}} c(L') \sum_{\substack{j_{ABC \setminus \{a\}} \in \mathfrak{X}_{ABC \setminus \{a\}} \\ j_{L'} = i_{L'}}} \log p_{B|C}(j_B \,|\, j_C) \underbrace{\sum_{j_a \in \mathfrak{X}_a} \left(1 - |\mathfrak{X}_a| \mathbb{I}_{\{j_a = i_a\}} \right)}_{=0} \\
&\quad = 0.
\end{aligned}$$

□

Remark 3.2.7. Any marginal log-linear parametrization $\boldsymbol{\eta}$ can be written in generalized log-linear form

$$\boldsymbol{\eta} = C \log(M \boldsymbol{p}),$$

where $\boldsymbol{p}$ is a vector of joint probabilities, and log is understood to act pointwise on vectors. Here the matrix M sums probabilities into appropriate marginal distributions, and these are multiplied (added on the log-scale) by the matrix of contrasts C. This form has been studied by various authors; see Lang (1996) for an overview.

3.3 Graphical Models as Sub-models

The ingenuous parametrization we have described relies upon two components: the values of the parameters themselves, and the constraints imposed by the structure of the graph. Other authors (Lupparelli et al., 2009; Rudas et al., 2010) first take a complete and hierarchical parametrization, and set some of the parameters to zero to impose the relevant conditional independences. This approach has the advantage that all the models thereby created are smooth sub-models of the saturated model. In their Theorem 5, Bergsma and Rudas (2002) show that this leads to a curved exponential family whose dimension is equal to the number of non-zero parameters; their results about variation independence (their Theorem 4) also depend upon having a complete parametrization.

It is not immediately clear that the ingenuous parametrization is of this nature. Consider again the graph in Figure 3.1, and its associated ingenuous parametrization $\mathbb{P}^{\text{ing}}$:

M_i	$\mathbb{L}_i$
$\{1\}$	$\{1\}$
$\{2\}$	$\{2\}, \{1,2\}$
$\{3\}$	$\{3\}$
$\{1,2,3\}$	$\{2,3\}, \{1,2,3\}$.

The effect $\{1,3\}$ is not represented, and so the parametrization is not complete. Bergsma and Rudas (2002) provide a method for completing parametrizations in a manner which preserves the property of hierarchy: one simply inserts the missing effects into the first margin of which they are a subset, according to some hierarchical ordering of margins. We call this *greedy completion*. In our case this would mean adding in the pair $(\{1,3\}, \{1,2,3\})$, giving a complete and hierarchical parametrization; call it $\mathbb{P}^{\text{ing}\,*}$.

The value of λ_{13}^{123} is just a function of joint probabilities, and so its value is completely determined by the values of the other parameters and the global Markov property associated with the graph. However, there is no simple expression for λ_{13}^{123} as a function of these parameters. This means that even though variation independence holds for the parameters $\mathbb{P}^{\text{ing}\,*}$ in the saturated model (more details in Section 3.4), there is no guarantee that this is true of the parameters $\tilde{\Lambda}(\mathbb{P}^{\text{ing}})$ in the sub-model defined by the graph.

Instead of $(\{1,3\}, \{1,2,3\})$, we could chose to include $(\{1,3\}, \{1,3\})$, and obtain a hierarchical and complete parameterization $\mathbb{P}^{\text{ing}\,\dagger}$. Since $X_1 \perp\!\!\!\perp X_3$ according to the graph, we have $\lambda_{13}^{13} = 0$ under this model. In fact, a probability distribution P satisfies the global Markov property with respect to this graph if and only if $\lambda_{13}^{13} = 0$, so the ingenuous parametrization does represent a (smooth) bijection between $\mathbb{R}^6$ and the model space.

This motivates the idea of a *completion*; a completion $\mathbb{P}^*$ of a parametrization $\mathbb{P}$ is any collection such that $\mathbb{P}^* \supseteq \mathbb{P}$ and $\mathbb{P}^*$ is complete. Of course, there may be many different completions of a given parametrization, and as we have just observed, some completions are more useful than others. We say that a completion is *sound* relative to a model $\mathcal{M}$ if $\mathbb{P}^*$ is hierarchical, and all the parameters in $\mathbb{P}^* \setminus \mathbb{P}$ are identically zero under $\mathcal{M}$.

We remark that a completion which preserves hierarchy, as we would surely wish any sensible scheme to do, can be put in the form of a greedy completion if we allow the addition of margins which have no associated effects. For example, adding the 'empty' margin $\{1,3\}$ into the example above between $\{3\}$ and $\{1,2,3\}$, gives

M_i	$\mathbb{L}_i$
$\{1\}$	$\{1\}$
$\{2\}$	$\{2\}, \{1,2\}$
$\{3\}$	$\{3\}$
$\{1,3\}$	
$\{1,2,3\}$	$\{2,3\}, \{1,2,3\}$.

Applying greedy completion now adds the effect $\{1,3\}$ into the first margin of which it is a subset, i.e. $\{1,3\}$.

3.3.1 Graphical Completion

Given a discrete model defined by a set of conditional independence constraints, it is natural to consider it as a sub-model of the saturated model, which contains all positive probability distributions. In a setting where the model is graphical, it becomes equally natural to think of the graph as a subgraph of a complete graph, by which we mean a graph containing at least one edge between any pair of vertices. We can achieve this by inserting edges between each pair of vertices which lack one, but this leaves a choice of edge type and orientation. These choices may affect how much of the structure and spirit of the original graph is retained; in particular, we require that any completion scheme preserves heads (see Proposition **??**).

Definition 3.3.1. Given an ADMG $\mathcal{G}$ and a complete supergraph $\bar{\mathcal{G}}$, we say that $\bar{\mathcal{G}}$ is a *head-preserving completion* of $\mathcal{G}$ if $\mathcal{H}(\mathcal{G}) \subseteq \mathcal{H}(\bar{\mathcal{G}})$.

It is easy to verify that such a completion always exists: we can simply add bidirected edges between all pairs of vertices which lack one. It is also clear that no m-separations hold in a complete graph $\bar{\mathcal{G}}$, and thus a completion represents the saturated model. Note that it

(a) (b)

Figure 3.2: (a) An acyclic directed mixed graph, and (b) a head-preserving completion.

is not necessary for every pair of vertices to be joined by an edge in order for a graph to represent the saturated model, however we do require this property of our completions.

Example 3.3.2. Figure 3.2(a) shows an ADMG, together with a head-preserving completion (b).

3.3.2 Completion of the Ingenuous Parametrization

Using a head-preserving completion we can represent the ingenuous parametrization as a sub-model of the saturated model.

Lemma 3.3.3. *Let $\mathcal{G}$ be an ADMG and $\bar{\mathcal{G}}$ any head-preserving completion of $\mathcal{G}$. Under the ingenuous parametrization, the model for $\mathcal{G}$ is a linear subspace of that for $\bar{\mathcal{G}}$. Under the sub-model $\mathcal{G}$, the non-zero parameters in the two parametrizations are equal.*

Proof. Let (H, T) be a head-tail pair in $\bar{\mathcal{G}}$. There are three possibilities for how this pair relates to $\mathcal{G}$: if (H, T) is also a head-tail pair in $\mathcal{G}$, there is no work to be done; otherwise either (i) H is not a head in $\mathcal{G}$, or (ii) H is a head in $\mathcal{G}$ but T is not its tail.

If (i) holds, we claim that under $\mathcal{G}$, $\lambda_A^{HT} = 0$ for all $H \subseteq A \subseteq H \cup T$. To see this, first note that H is a barren set in $\bar{\mathcal{G}}$, and since it is maximally connected, this means that all elements are joined by bidirected edges. H must also be barren in $\mathcal{G}$, and since it is not a head in $\mathcal{G}$ this means that $H = K \cup L$ for disjoint non-empty sets K and L with no edges directly between them. But this implies that K and L are m-separated conditional on T, and thus $X_K \perp\!\!\!\perp X_L \,|\, X_T$ under the Markov property for $\mathcal{G}$. Then, by Lemma 3.1.9, these parameters are all identically zero under $\mathcal{G}$.

(ii) implies that H is a head in both $\mathcal{G}$ and $\bar{\mathcal{G}}$, but that $T \equiv \mathrm{tail}_{\bar{\mathcal{G}}}(H) \supset \mathrm{tail}_{\mathcal{G}}(H) \equiv T'$. We claim that $\lambda_A^{HT} = 0$ for all $H \subseteq A \subseteq H \cup T$ such that $A \cap (T \setminus T') \neq \emptyset$; this follows from the fact that T' is the Markov blanket for H in $\mathrm{an}_{\mathcal{G}}(H)$, and Lemma 3.1.9 (see Definition 1.2.3 for the meaning of 'Markov blanket').

We have shown that all parameters corresponding to effects not found in $\mathbb{P}^{\text{ing}}(\mathcal{G})$ are identically zero under $\mathcal{G}$. The vanishing of these parameters defines the correct sub-model, but note that some of the remaining margins are not the same in $\mathbb{P}^{\text{ing}}(\bar{\mathcal{G}})$ as in $\mathbb{P}^{\text{ing}}(\mathcal{G})$. These remaining cases are again from (ii), but where $H \subseteq A \subseteq H \cup T'$; in this case $\lambda_A^{HT} = \lambda_A^{HT'}$ under $\mathcal{G}$, again because T' is the Markov blanket for H in $\text{an}_{\mathcal{G}}(H)$, and by application of Lemma 3.2.6. Thus the parameters are equal under the constraints imposed by the parameters set to zero above. □

This allows us to apply Theorem 5 of Bergsma and Rudas (2002) to ADMG models, and thus shows that each model corresponds to a curved exponential family of distributions with dimension equal to its ingenuous parameter count. It follows from this that $\mathbb{P}^{\text{ing}}(\mathcal{G})$ is a smooth parametrization of distributions satisfying the Markov property for $\mathcal{G}$; however, the direct proof is instructive.

Example 3.3.4. Consider again the ADMG $\mathcal{G}$ in Figure 3.2(a) with the head-preserving completion $\bar{\mathcal{G}}$ in (b). The ingenuous parametrization for $\bar{\mathcal{G}}$ is

M	$\mathbb{L}$
{1}	{1}
{2}	{2}
{1,2}	{1,2}
{1,2,3}	{3}, {1,3}, {2,3}, {1,2,3}
{1,2,4}	{4}, {1,4}, {2,4}, {1,2,4}
{1,2,3,4}	{3,4}, {1,3,4}, {2,3,4}, {1,2,3,4}.

The sub-model $\mathcal{G}$ corresponds to setting

$$\lambda_{12}^{12} = \lambda_{23}^{123} = \lambda_{123}^{123} = \lambda_{14}^{124} = \lambda_{124}^{124} = 0,$$

under which conditions the following equalities hold:

$$\lambda_3^{123} = \lambda_3^{13} \qquad \lambda_{13}^{123} = \lambda_{13}^{13} \qquad \lambda_4^{124} = \lambda_4^{24} \qquad \lambda_{24}^{124} = \lambda_{24}^{24}.$$

Removing the parameters which are identical to zero, and renaming the four others according to these last equations returns us to the ingenuous parametrization of $\mathcal{G}$:

M	$\mathbb{L}$
{1}	{1}
{2}	{2}
{1,3}	{3}, {1,3}
{2,4}	{4}, {2,4}
{1,2,3,4}	{3,4}, {1,3,4}, {2,3,4}, {1,2,3,4}.

Remark 3.3.5. If $\mathcal{G}$ is a purely bidirected graph, the ingenuous parametrization consists of sets of the form (C, C), where C is connected in $\mathcal{G}$. The complete bidirected graph is a head-preserving completion; this adds the sets (D, D) to the parametrization, where D is disconnected in $\mathcal{G}$. We have $\lambda_D^D = 0$ for a disconnected set, because any two disconnected components will be marginally independent.

Remark 3.3.6. In the case where $\mathcal{G}$ is a DAG, the ingenuous parametrization consists of sets of the form $(\{v\} \cup Q, \{v\} \cup P)$, where $P = \mathrm{pa}_{\mathcal{G}}(v)$, and $Q \subseteq P$. The ingenuous parametrization of $\bar{\mathcal{G}}$ contains sets of the form $(H \cup Q, H \cup A)$, where H is any set which is barren in $\mathcal{G}$, $A = \mathrm{an}_{\mathcal{G}}(H)$ and $Q \subseteq A$. If $H = \{v\}$ and $Q \subseteq P$, then Lemma 3.2.6 and the Markov property for $\mathcal{G}$ gives $\lambda_{HQ}^{HA} = \lambda_{vQ}^{vP}$.

3.4 Ordered Decomposability and Variation Independence

Bergsma and Rudas (2002) characterize precisely which hierarchical and complete parametrizations of the saturated model are variation independent, through an extended notion of decomposability of sets. The main result in this section characterizes precisely when the ingenuous parametrization is variation independent.

Definition 3.4.1. Two sets M_1 and M_2 are *incomparable*, if $M_1 \not\subseteq M_2$ and $M_2 \not\subseteq M_1$.

A collection $\mathbb{M}$ of incomparable subsets of V is *decomposable* if it has at most two elements, or there is an ordering $M_1, \ldots, M_k$ on the elements of $\mathbb{M}$ wherein for each $i = 3, \ldots, k$, there exists $j_i < i$ such that

$$\left(\bigcup_{l=1}^{i-1} M_l \right) \cap M_i = M_{j_i} \cap M_i.$$

This is also known as the *running intersection property*.

A collection $\mathbb{M}$ of (possibly comparable) subsets is *ordered decomposable* if it has at most

two elements, or there is an ordering $M_1, \ldots, M_k$ such that

$$M_i \not\subseteq M_j \quad \text{for } i > j,$$

and for each $i = 3, \ldots, k$, the inclusion maximal elements of $\{M_1, \ldots, M_i\}$ form a decomposable collection. We say that a collection $\mathbb{P}$ of parameters is ordered decomposable if there is an ordering on the margins $\mathbb{M}$ which is both hierarchical and ordered decomposable.

Decomposability here should not be confused with the notion of decomposability of an undirected graph. The ingenuous parametrization of an undirected graph (i.e. the ordinary log-linear parameters, $\mathbb{P}^{\max}$) is *always* ordered decomposable, even if the graph is not, because we only use the single margin V (see Appendix A).

Bergsma and Rudas show that if $\mathbb{P}$ is hierarchical and complete, the parameters $\tilde{\Lambda}(\mathbb{P})$ are variation independent if and only if $\mathbb{P}$ is ordered decomposable. However, they give no general results for variation independence in the case of an incomplete parametrization; the following observation follows trivially from their work, and gives a sufficient but not necessary criterion.

Remark 3.4.2. Let $\mathcal{M} \subset \Delta_{|\mathfrak{X}_V|-1}$ be a model. Suppose that $\mathcal{M}$ is of the form

$$\mathcal{M} = \{\boldsymbol{p} \,|\, \lambda_L^M(\boldsymbol{p}) = 0 \text{ for all } (L, M) \in \mathbb{P}^0 \subset \mathbb{P}^*\}$$

for some complete and hierarchical parametrization $\mathbb{P}^*$ and subset $\mathbb{P}^0$. Then $\mathcal{M}$ is smoothly parametrized by $\mathbb{P} = \mathbb{P}^* \setminus \mathbb{P}^0$, and if $\mathbb{P}^*$ is ordered decomposable then this parametrization is variation independent.

The following graphical lemma is used in the proof of the main result in this section, Theorem 3.4.5.

Lemma 3.4.3. *An ADMG contains at least one head of size three or more if and only if it contains two heads of the form $\{v_1, v_2\}$ and $\{v_2, v_3\}$, where $\{v_1, v_2, v_3\}$ is barren.*

Remark 3.4.4. Note that an ADMG may contain a head H of size three, such that no subset of H of size two is a head. Figure 3.3 contains an example: the set $H = \{1, 3, 5\}$ is barren and bidirected connected in $\mathrm{an}_{\mathcal{G}}(H)$; however no subset of size two shares these properties. The lemma states that there is *some* head of size three which has two subsets of size two which are also heads; in this case $\{2, 3, 4\}$ is such a set.

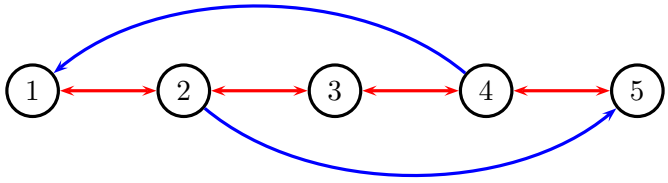


Figure 3.3: An ADMG with a head $H = \{1, 3, 5\}$ of size three, such that no subset of size two of H is a head.

Proof. First let $\mathcal{G}$ be an ADMG with a head H of size ≥ 3, and suppose that the implication fails to hold. Pick 3 vertices $\{w_1, w_2, w_3\}$ in H. By the definition of a head, we can pick a bidirected path π, through $\text{an}_{\mathcal{G}}(H)$, from w_1 to w_2; assume that π contains no other element of H, otherwise shorten the path and redefine w_1 or w_2. Then create a similar path ρ from w_2 to w_3; again assume that ρ contains no other element of H, else shorten the path and redefine w_3. If w_1 lies on ρ, we can swap w_1 and w_2 to ensure that neither π nor ρ pass through elements of H other than at their endpoints.

According to our assumption that the result is false, at least one of $\{w_1, w_2\}$ or $\{w_2, w_3\}$ is not a head; assume the former without loss of generality. This implies that π passes through at least one vertex v which is not an ancestor of $\{w_1, w_2\}$. If there is more than one such vertex, choose one which has no distinct descendants on the path π. By the construction of π we have $v \in \text{an}_{\mathcal{G}}(H) \setminus H$ and $\text{an}_{\mathcal{G}}(v) \cap \pi = \{v\}$.

Let W be the set of vertices in π, and $H^* \equiv \text{barren}_{\mathcal{G}}(W)$. Since W is $\leftrightarrow$-path-connected, H^* is a head, and $\{w_1, w_2, v\} \subseteq H^*$. Thus we have created a head distinct from H, of size at least 3, and $H^* \prec H$.

The assumption we have made implies that we can repeat this process indefinitely, but since we have a finite set of heads and $\prec$ is a well defined strict partial ordering on heads, this is a contradiction. Thus we must eventually reach a set of size three with the desired properties.

For the converse, note that since $\{v_1, v_2\}$ is bidirected-path-connected in $\text{an}_{\mathcal{G}}(\{v_1, v_2\})$, then clearly it is in $\text{an}_{\mathcal{G}}(\{v_1, v_2, v_3\})$ as well. The same holds for $\{v_2, v_3\}$, and hence $\{v_1, v_2, v_3\}$ is a head. □

We remark that the converse of this result is clearly also true.

Theorem 3.4.5. *The ingenuous parametrization of the distributions satisfying the global Markov property for an ADMG $\mathcal{G}$ is variation independent if and only if $\mathcal{G}$ contains no heads of size ≥ 3.*

Proof. ($\Leftarrow$). Suppose that $\mathcal{G}$ contains no heads of size ≥ 3, and let $1, \ldots, n$ be a topological ordering on the vertices of $\mathcal{G}$. We will construct a complete, hierarchical and variation independent parametrization, and show that it is equivalent to the ingenuous parametrization.

Let $\mathbb{M}_i \subseteq \mathbb{M}$ be the collection of margins which involve only the vertices in $[i] = \{1, \ldots, i\}$. Assume for induction, that $\mathbb{M}_{i-1}$ includes the set $[i-1]$, is hierarchical and complete up to this point, and that it satisfies the ordered decomposability criterion. The base case for $i = 1$ is trivial.

Now, let the heads involving i contained within $[i]$ be $H_0 = \{i\}, H_1 = \{j_1, i\}, \ldots, H_k = \{j_k, i\}$, where $j_1 < \ldots < j_k < i$. Call the associated tails $T_0, \ldots, T_k$. We have

$$\text{barren}_{\mathcal{G}}\left(\text{dis}_{\mathcal{G}}(i)\right) = \{j_k, i\},$$

since $\text{barren}_{\mathcal{G}}\left(\text{dis}_{\mathcal{G}}(i)\right)$ is a head, and cannot have size ≥ 3. Since i has no proper descendants in $[i]$, then $H_k \cup T_k \setminus \{i\} = \text{mb}(i, [i])$, where $\text{mb}(v, A)$ is the Markov blanket of v in the ancestral set A.

Now, since the ordering is topological, $A_k \equiv [i]$ is an ancestral set, and the ordered local Markov property shows that

$$i \perp\!\!\!\perp A_k \setminus (\text{mb}(i, A_k) \cup \{i\}) \mid \text{mb}(i, A_k),$$

so

$$i \perp\!\!\!\perp A_k \setminus (H_k \cup T_k) \mid H_k \cup T_k \setminus \{i\}.$$

Then

$$\begin{aligned} \lambda_C^{A_k} &= \lambda_C^{T_k} && \text{for any } H_k \subseteq C \subseteq H_k \cup T_k \\ \lambda_C^{A_k} &= 0 && \text{for any } \{i\} \subset C \not\subseteq H_k \cup T_k, \end{aligned}$$

where first equality follows from the independence and Lemma 3.2.6, and the second from the above independence and Lemma 3.1.9.

Note that these conditions include every set C which contains both i and any descendant of j_k, since no descendant of j_k is in $H_k \cup T_k$. Thus we have created parameters for every subset of A_k which contains some descendant of j_k, and shown that the non-zero parameters are equivalent to the ingenuous parameters.

Now let $A_{k-1} = A_k \setminus \text{de}_{\mathcal{G}}(j_k)$. The set A_{k-1} is ancestral and contains i, so applying the

ordered local Markov property again gives

$$\begin{aligned} \lambda_C^{A_{k-1}} &= \lambda_C^{T_{k-1}} && \text{for any } H_{k-1} \subseteq C \subseteq H_{k-1} \cup T_{k-1} \\ \lambda_C^{A_{k-1}} &= 0 && \text{for any } \{i\} \subset C \nsubseteq H_{k-1} \cup T_{k-1}. \end{aligned}$$

Continuing this approach gives a parameter, possibly fixed to zero, for every subset of $[i]$ containing some descendant of any of $j_1, \ldots, j_k$. Lastly let $A_0 = A_1 \setminus \mathrm{de}_{\mathcal{G}}(j_1)$.

$$\begin{aligned} \lambda_C^{A_0} &= \lambda_C^{T_0} && \text{for any } \{i\} \subseteq C \subseteq \{i\} \cup T_0 \\ \lambda_C^{A_0} &= 0 && \text{for any } \{i\} \subset C \nsubseteq \{i\} \cup T_0. \end{aligned}$$

The margins we have added are $A_0 \subset \cdots \subset A_k$, and since they all contain $\{i\}$, they are not a subset of any existing margin. Further, each set C we associate with A_l contains a vertex which is not in A_{l-1}. Thus our new parametrization is complete and hierarchical. Setting $\mathbb{M}_i = \mathbb{M}_{i-1} \cup \{A_0, \ldots, A_k\}$, the new maximal subsets created are all of the form $[i-1] \cup A_l$; thus $\mathbb{M}_i$ is clearly also ordered decomposable.

($\Rightarrow$). Our construction will assume the random variables are binary; the general case is a trivial but tedious extension. Suppose that $\mathcal{G}$ has a head of size ≥ 3, and assume for contradiction that its ingenuous parametrization is variation independent. By Lemma 3.4.3, there exist two heads $H_1 = \{v_1, v_2\}$ and $H_2 = \{v_2, v_3\}$ such that $\{v_1, v_2, v_3\}$ is barren. Let $H_3 \equiv \{v_3, v_1\}$ noting that this set may or may not be a head.

Also let $T_i = \mathrm{tail}_{\mathcal{G}}(H_i)$, where if H_3 is not a head, T_3 is taken to be the tail of H_3 *if there were* a bidirected arrow between v_1 and v_3. Further let $A = \mathrm{an}_{\mathcal{G}}(\{v_1, v_2, v_3\}) \setminus \{v_1, v_2, v_3\}$.

Now choose $\lambda_{C_i}^{B_i} = 0$, where $B_i = \{v_i\} \cup \mathrm{tail}_{\mathcal{G}}(v_i)$ and $\{v_i\} \subseteq C_i \subseteq B_i$; this sets each v_i to be uniform on $\{0, 1\}$ for any instantiation of its tail.

Similarly, by choosing $\lambda_{C_1}^{H_1 \cup T_1}(0)$ to be large and positive for each $H_1 \subseteq C_1 \subseteq H_1 \cup T_1$, we can force v_1 and v_2 to be arbitrarily highly correlated conditional on T_1, and therefore conditional on A. We can do the same for v_2 and v_3, so for any $0 < \epsilon < \frac{1}{2}$:

		v_1 = 0	v_1 = 1
v_2	0	$\frac{1}{2} - \epsilon$	ϵ
	1	ϵ	$\frac{1}{2} - \epsilon$

		v_2 = 0	v_2 = 1
v_3	0	$\frac{1}{2} - \epsilon$	ϵ
	1	ϵ	$\frac{1}{2} - \epsilon$

,

where these tables are understood to show the two-way marginal distributions conditional on any instantiation i_A of A.

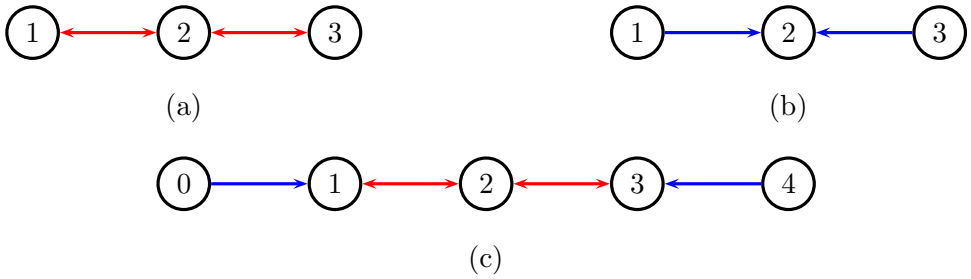


Figure 3.4: (a) A graph with a variation dependent ingenuous parametrization; (b) a Markov equivalent graph with a variation independent parametrization; (c) a graph with no known variation independent MLL parametrization.

But now either $\lambda_{C_3}^{H_3 \cup T_3} = 0$ by design (because H_3 is not a head, and v_1 and v_3 are independent conditional on their 'tail'), or we can choose this to be the case by the assumption of variation independence. This implies that v_1 and v_3 are independent conditional on A so, for example,

$$\begin{aligned}
\frac{1}{4} &= P(v_1 = 1 \,|\, A = i_A) \cdot P(v_3 = 0 \,|\, A = i_A) \\
&= P(v_1 = 1, v_3 = 0 \,|\, A = i_A) \\
&= P(v_1 = 1, v_2 = 0, v_3 = 0 \,|\, A = i_A) + P(v_1 = 1, v_2 = 1, v_3 = 0 \,|\, A = i_A) \\
&< P(v_1 = 1, v_2 = 0 \,|\, A = i_A) + P(v_2 = 1, v_3 = 0 \,|\, A = i_A) \\
&= 2\epsilon,
\end{aligned}$$

which is a contradiction if $\epsilon < \frac{1}{8}$. Thus the parameters are variation dependent. □

Remark 3.4.6. The simplest ADMG for which the ingenuous parametrization is variation dependent is the bidirected 3-chain, shown in Figure 3.4(a); see the next example for details. However, since there are no colliders at 1 or 3, we could change to the Markov equivalent graph in 3.4(b), whose ingenuous parametrization is variation independent.

On the other hand, the structure in Figure 3.4(c) shows why heads of size 3 are a problem. The graph is Markov equivalent to a bidirected 5-chain, which has no known variation independent parametrization in the MLL framework, and it seems unlikely that one exists. In Chapter 5 we present the first variation independent parametrization of this model using a slight extension to MLL parameters.

The bidirected 4-cycle has a variation independent MLL parametrization, but the ingenuous parametrization fails to find it: see Example 3.5.1.

In general, the ingenuous parametrization's variation independence properties are best when one chooses a graph from the Markov equivalence class created by the model which has the smallest possible maximum head size. This can be achieved by finding a Markov equivalent graph which contains the fewest possible arrowheads; see, for example, Ali et al. (2005) and Drton and Richardson (2008b) for approaches to this.

Example 3.4.7. The ingenuous parametrization avoids the trivial variation dependence which arises from the fact that marginal probabilities are always greater than joint probabilities, but it still suffers from variation dependence due to marginal distributions which are not *strongly compatible* with each other, in the terminology of Bergsma and Rudas (2002). As illustrated in the proof of Theorem 3.4.5, it is possible to choose marginal distributions which are weakly compatible, in the sense that their marginal distributions agree, but not strongly compatible. Suppose we have a probability distribution on 3 binary variables according to the graph in Figure 3.4(a), so that $1 \perp\!\!\!\perp 3$. Using the ingenuous parametrization of this graph, select $\lambda_1^1 = \lambda_2^2 = \lambda_3^3 = 0$ and $\lambda_{12}^{12}(0,0) = \lambda_{23}^{23}(0,0) = \log 2$. Then the two-way marginal distributions of $\{1,2\}$ and $\{2,3\}$ are

X_1 \ X_2	0	1
0	$\frac{2}{5}$	$\frac{1}{10}$
1	$\frac{1}{10}$	$\frac{2}{5}$

X_2 \ X_3	0	1
0	$\frac{2}{5}$	$\frac{1}{10}$
1	$\frac{1}{10}$	$\frac{2}{5}$

.

Each of these two-way marginal distributions is, on its own, compatible with the model; in fact we may choose any real values of λ_1^1, λ_2^2, λ_{12}^{12} and λ_3^3 that we wish, but having done so, we are no longer necessarily free to choose any value of λ_{23}^{23}. These issues are discussed in more detail in Section 5.2.

3.5 Alternative Parametrizations

The ingenuous parametrization is not the only way to parametrize probability distributions over graphical models using marginal log-linear parameters, and we briefly mention two different schemes here.

Example 3.5.1. The bidirected 4-cycle (Figure 3.5) contains the head $\{1,2,3,4\}$ of size 4, and thus its ingenuous parametrization is variation dependent. However it *can* be given an ordered decomposable (and so variation independent) parametrization in the framework of marginal log-linear parameters.

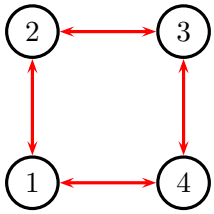


Figure 3.5: A bidirected 4-cycle.

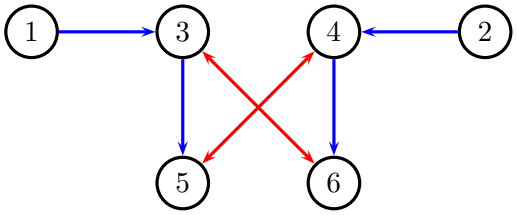


Figure 3.6: An acyclic directed mixed graph not equivalent to any type IV chain graph.

The 4-cycle corresponds to precisely the model with $X_1 \perp\!\!\!\perp X_3$ and $X_2 \perp\!\!\!\perp X_4$. Set $\mathbb{M} = \{\{1,3\},\{2,4\},\{1,2,3,4\}\}$, with

$$\begin{aligned}\mathbb{L}_1 &= \{\{1\},\{3\},\{1,3\}\}\\ \mathbb{L}_2 &= \{\{2\},\{4\},\{2,4\}\}\\ \mathbb{L}_3 &= \mathscr{P}(\{1,2,3,4\}) \setminus (\mathbb{L}_1 \cup \mathbb{L}_2);\end{aligned}$$

here $\mathscr{P}(A)$ denotes the power set of A. This gives a hierarchical, complete and ordered decomposable parametrization, and thus the parameters are variation independent. But the 4-cycle corresponds exactly to setting $\lambda^{13}_{13} = \lambda^{24}_{24} = 0$, and thus the remaining parameters are still variation independent.

The above is an example of the approach taken by Lupparelli (2006) and Lupparelli et al. (2009), who parametrize distributions over bidirected graphs using disconnected sets. Their scheme gives variation independent parametrizations for some models which the ingenuous parametrization does not, including the above example; however, it is not clear how to generalize their approach to ADMGs. In the bidirected case, this approach gives a variation independent parametrization only for dense graphs where the number of constraints is small. In particular, the disconnected sets cannot overlap.

Example 3.5.2. Rudas et al. (2010) and Marchetti and Lupparelli (2010) both parametrize chain graph models of multivariate regression type, also known as type IV chain graph models, using marginal log-linear parameters. Type IV chain graph models are a special

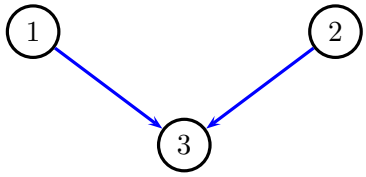


Figure 3.7: A directed acyclic graph.

case of ADMG models, in the sense that by replacing the undirected edges in a type IV chain graph with bidirected edges, the global Markov property on the resulting ADMG is equivalent to the Markov property for the chain graph (see Drton, 2009). The graphs in Figure 3.2(a) and (b) are examples of Type IV models. However, there are models in the class of ADMGs which do not correspond to any chain graph, such as the one in Figure 3.6.

The parametrization of Rudas et al. (2010) uses different choices of margins to the ingenuous parametrization, though their parameters can be shown to be equal to the ingenuous parameters under the appropriate Markov property, using Lemma 3.2.6. It follows that the variation dependence properties of that parametrization are identical to those of the ingenuous parametrization (see next section). Forcina et al. (2010) provide an algorithm which gives a range of 'admissible' margins in which collections of conditional independence constraints may be defined.

Marchetti and Lupparelli (2010) also parametrize type IV chain graph models in a similar manner to Rudas et al. (2010), but using multivariate logistic contrasts.

Example 3.5.3. The work of Rudas et al. (2006) inspires an alternative parametrization for DAGs which is always ordered decomposable. First, label the vertices according to a topological ordering $1, \ldots, n$. Set $M_i = \{1, \ldots, i\}$ for $i = 1, \ldots, n$, and let

$$\mathbb{L}_i = \{A \,|\, \{i\} \subseteq A \subseteq \{i\} \cup \mathrm{pa}_{\mathcal{G}}(i)\}.$$

We call this the *RBN parametrization*, after the authors. It should be noted that Rudas et al. only use a subset of the effects above, but the definition above follows naturally from their work.

For the DAG in Figure 3.7 we could choose the topological ordering 1, 2, 3; the parameters associated with the model would be

$$\lambda_1^1 \qquad \lambda_2^{12} \qquad \lambda_3^{123} \qquad \lambda_{13}^{123} \qquad \lambda_{23}^{123} \qquad \lambda_{123}^{123}.$$

The only difference from the ingenuous parametrization for this graph is that λ_2^2 is replaced with λ_2^{12}.

The RBN parametrization has two nice features: firstly it is clearly ordered decomposable, and secondly if we complete the parametrization greedily by throwing missing sets into the earliest margin they can go in, this is equivalent to adding arrows between every pair of vertices according to the topological ordering. One disadvantage is that it introduces parameters which are apparently harder to interpret; for example, λ_{13}^{123} has replaced λ_{13}^{13}. It seems odd to frame the relationship between a vertex and its parents through all its predecessors in an arbitrary topological ordering.

However, it follows from Lemma 3.2.6 that the RBN parameters are equal to the ingenuous parameters under the model, and specifically in the example given that $\lambda_{13}^{13} = \lambda_{13}^{123}$. To see this, recall that in a DAG, a vertex v is independent of its non-descendants given its parents, and note that all vertices preceding v in the topological ordering are its non-descendants. Writing an RBN parameter as λ_{vQ}^{vPN}, where P are the parents of v, N are the other vertices which precede v in the topological ordering, and $Q \subseteq P$, we have $v \perp\!\!\!\perp N \,|\, P$; hence by Lemma 3.2.6

$$\lambda_{vQ}^{vPN} = \lambda_{vQ}^{vP}.$$

But the collection $\{\lambda_{vQ}^{vP} \,|\, Q \subseteq P\}$ is just the set of ingenuous parameters for the head $\{v\}$.

3.6 Probability Calculations

A disadvantage of marginal log-linear parameters is that it is not necessarily easy to recover raw probabilities from them, which is necessary to evaluate the likelihood; in general one can use the Iterative Proportional Fitting (IPF) algorithm to recover a joint distribution from information about marginal distributions (Csiszár, 1975). However because this approach is iterative, if we use it at every step of, for example, an MCMC procedure in order to evaluate the likelihood, it may be very computationally expensive.

Further problems arise when optimizing with respect to MLL parameters if they are variation dependent. 'Black box' methods such as Newton-Raphson for maximum likelihood estimation can leave the space of valid parameters.

In the binary case, we present a method for taking a set of valid ingenuous parameters, and retrieving the joint probability distribution from them; alternatively, if the parameters are invalid, the method will tell us so. This approach is closely related to that of Qaqish and Ivanova (2006), who work with the multivariate logistic parameters. Recall that $q_{H|T}^{(i_T)} \equiv P(X_H = 0 \,|\, X_T = i_T)$; we use the following lemma.

Lemma 3.6.1. *Let P be a strictly positive binary probability distribution which obeys the global Markov property with respect to an ADMG $\mathcal{G}$, and let H a head in $\mathcal{G}$. Then for any $i_H \in \mathfrak{X}_H$ and $i_T \in \mathfrak{X}_T$,*

$$p_{H|T}(i_H \,|\, i_T) = (-1)^{\|i_H\|} \left(q_{H|T}^{(i_T)} - K(i_{HT}) \right),$$

where K is an infinitely differentiable function of $q_{H'|T'}^{(i_{T'})}$ for heads $H' \prec H$. Here $\|i_A\|$ is the number of 1s in the binary vector i_H.

Proof. Let $A = \mathrm{an}_{\mathcal{G}}(H)$; clearly H is the (unique) maximal head in A under $\prec$. By (1.2) we have

$$\begin{aligned} p_{H|T}(i_H \,|\, i_T) &= p_{H|A\setminus H}(i_H \,|\, i_{A\setminus H}) \\ &= \frac{p_A(i_A)}{p_{A\setminus H}(i_A)}, \end{aligned}$$

where $i_{A\setminus(H\cup T)}$ is chosen arbitrarily from $\mathfrak{X}_{A\setminus(H\cup T)}$ if necessary. Then Theorem 1.4.1 gives

$$= \frac{\displaystyle\sum_{O\subseteq C\subseteq A} (-1)^{|C\setminus O|} \prod_{H'\in[A]} q_{H'|T'}^{(i_{T'})}}{\displaystyle\sum_{O\setminus H\subseteq C\subseteq A\setminus H} (-1)^{|C\setminus(O\setminus H)|} \prod_{H'\in[A\setminus H]} q_{H'|T'}^{(i_{T'})}},$$

where $O = \{v \in A \,|\, i_v = 0\}$. Now, clearly $q_{H|T}^{(i_T)}$ only appears in terms for which $C \supseteq H$, so we can write

$$p_{H|T}(i_H \,|\, i_T) = \frac{\displaystyle\sum_{O\cup H\subseteq C\subseteq A} (-1)^{|C\setminus O|} \prod_{H'\in[A]} q_{H'|T'}^{(i_{T'})} + D(i_{HT})}{\displaystyle\sum_{O\setminus H\subseteq C\subseteq A\setminus H} (-1)^{|C\setminus(O\setminus H)|} \prod_{H'\in[A\setminus H]} q_{H'|T'}^{(i_{T'})}},$$

where D is a multi-linear and therefore infinitely differentiable function of generalized Möbius parameters $q_{H'|T'}^{(i_{T'})}$ for $H' \prec H$. But since H is maximal in A under $\prec$, we have

$$\begin{aligned} \sum_{O\cup H\subseteq C\subseteq A} (-1)^{|C\setminus O|} \prod_{H'\in[A]} q_{H'|T'}^{(i_{T'})} &= q_{H|T}^{(i_T)} \sum_{O\cup H\subseteq C\subseteq A\setminus H} (-1)^{|C\setminus O|} \prod_{H'\in[A]} q_{H'|T'}^{(i_{T'})} \\ &= (-1)^{|H\setminus O|} q_{H|T}^{(i_T)} \sum_{O\cup H\subseteq C\subseteq A\setminus H} (-1)^{|C\setminus(O\setminus H)|} \prod_{H'\in[A]} q_{H'|T'}^{(i_{T'})}. \end{aligned}$$

Then

$$p_{H|T}(i_H \,|\, i_T) = (-1)^{\|i_H\|} \left(q^{(i_{T'})}_{H'|T'} + \frac{D(i_{HT})}{p_{A\setminus H}(i_{A\setminus H})} \right),$$

where $p_{A\setminus H}(i_{A\setminus H})$ is a strictly positive multi-linear function of parameters involving heads $H' \prec H$. Setting $K(i_{HT}) = -D(i_{HT})/p_A(i_A)$ gives the result. □

We now proceed to show how generalized Möbius parameters may be recovered from the ingenuous parametrization.

It is clear that for a singleton head $\{h\}$ with empty tail,

$$\lambda^h_h(0) = \frac{1}{2} \log \frac{q_h}{1 - q_h}$$

$$\text{so} \qquad q_h = \frac{\exp(2\lambda^h_h(0))}{1 + \exp(2\lambda^h_h(0))}.$$

Now, for a general head H, let $i_H = \mathbf{0}$ and $i_T \in \mathfrak{X}_T$, and suppose that $q^{(i_{T'})}_{H'|T'}$ is known for all heads H' preceding H in the partial ordering $\prec$. Then

$$\kappa_{H|T}(i_H \,|\, i_T) = \frac{1}{|\mathfrak{X}_H|} \sum_{j_H \in \mathfrak{X}_H} (-1)^{\|j_H\|} \log p_{H|T}(j_H \,|\, i_T),$$

where $\|j_H\|$ denotes the number of 1s in j_H. By Lemma 3.6.1, we have

$$p_{H|T}(j_H \,|\, i_T) = (-1)^{\|j_H\|} \left(q^{(i_T)}_{H|T} - K(j_H, i_T) \right),$$

where each $K(j_H, i_T)$ is composed entirely of parameters already known by the induction hypothesis. Thus

$$\kappa_{H|T}(i_H \,|\, i_T) = \frac{1}{|\mathfrak{X}_H|} \sum_{j_H \in \mathfrak{X}_H} (-1)^{\|j_H\|} \log \left\{ (-1)^{\|j_H\|} \left(q^{i_T}_{H|T} - K(j_H, i_T) \right) \right\}$$

$$\exp\left\{ |\mathfrak{X}_H| \kappa_{H|T}(i_H \,|\, i_T) \right\} = \frac{\prod_{\|j_H\| \text{ even}} \left\{ q^{i_T}_{H|T} - K(j_H, i_T) \right\}}{\prod_{\|j_H\| \text{ odd}} \left\{ K(j_H, i_T) - q^{i_T}_{H|T} \right\}}. \tag{3.4}$$

Now, it is clear that any valid solution must cause all the expressions in braces to be positive, since they are all probabilities. This is true precisely when

$$\max_{\|j_H\| \text{ even}} \{K(j_H, i_T)\} < q^{i_T}_{H|T} < \min_{\|j_H\| \text{ odd}} \{K(j_H, i_T)\}.$$

Furthermore, it is easy to see that a unique solution to the above equation exists precisely when

$$\max_{\|j_H\| \text{ even}} \{K(j_H, i_T)\} < \min_{\|j_H\| \text{ odd}} \{K(j_H, i_T)\},$$

since the right hand side of (3.4) varies monotonically from 0 to ∞ as $q_{H|T}^{(i_T)}$ varies in this range. An equation of this form can be solved numerically using an elementary method such as interval bisection which, whilst still iterative, is extremely easy for a computer to perform.

Example 3.6.2. Extending this method to the non-binary case is not easy. Consider a graph consisting of two nodes, each taking three states, with a bidirected edge between them. To find $P(X_1 = 0)$, note that

$$\begin{aligned} \lambda_1^1(0) &= \frac{1}{3} \log \frac{p_{0\cdot}^2}{p_{1\cdot}\, p_{2\cdot}} \\ \lambda_1^1(0) - \lambda_1^1(1) &= \log \frac{p_{0\cdot}}{p_{1\cdot}} \\ p_{1\cdot} &= p_{0\cdot} \exp\left(\lambda_1^1(1) - \lambda_1^1(0)\right). \end{aligned}$$

Similarly

$$p_{2\cdot} = p_{0\cdot} \exp\left(-\lambda_1^1(1) - 2\lambda_1^1(0)\right).$$

So

$$\begin{aligned} 1 &= p_{0\cdot} + p_{1\cdot} + p_{2\cdot} \\ &= p_{0\cdot} \left(1 + e^{\lambda_1^1(1) - \lambda_1^1(0)} + e^{-\lambda_1^1(1) - 2\lambda_1^1(0)}\right) \end{aligned}$$

which gives us $P(X_1 = 0)$, and $P(X_1 = 1)$ and $P(X_1 = 2)$ follow. For the joint parameters, we obtain the equations

$$\begin{aligned} \lambda_{12}^{12}(0,0) &= \frac{1}{9} \log \frac{p_{00}\, p_{11}}{p_{01}\, p_{10}}, \\ \lambda_{12}^{12}(1,0) &= \frac{1}{9} \log \frac{p_{10}\, p_{21}}{p_{11}\, p_{20}} = \frac{1}{9} \log \frac{p_{10}\, (p_{\cdot 1} - p_{01} - p_{11})}{p_{11}\, (p_{\cdot 0} - p_{00} - p_{10})}, \\ \lambda_{12}^{12}(0,1) &= \frac{1}{9} \log \frac{p_{01}\, p_{12}}{p_{02}\, p_{11}} = \frac{1}{9} \log \frac{p_{01}\, (p_{1\cdot} - p_{10} - p_{11})}{(p_{0\cdot} - p_{01} - p_{11})\, p_{11}}, \\ \lambda_{12}^{12}(1,1) &= \frac{1}{9} \log \frac{p_{11}\, p_{22}}{p_{21}\, p_{12}} = \frac{1}{9} \log \frac{p_{11}\, (1 - p_{0\cdot} - p_{1\cdot} - p_{\cdot 0} - p_{\cdot 1} + p_{00} + p_{10} + p_{01} + p_{11})}{(p_{\cdot 1} - p_{01} - p_{11})\, (p_{1\cdot} - p_{10} - p_{11})}. \end{aligned}$$

We have 4 simultaneous polynomials in 4 unknowns, p_{00}, p_{10}, p_{01} and p_{11}. These can be solved analytically using Gröbner basis methods, but this is extremely computationally demanding in general. No analogue to the approach used in the binary case appears possible.

Remark 3.6.3. Note that if a random variable takes precisely 2^k states, we can split the node into k binary nodes, all fully connected by directed edges, and each with the same incident edges as the original node. So if every vertex takes precisely 2^k states, possibly with different ks, we can still use the above method. This includes a case of considerable interest to geneticists, since there are 4 base pairs in DNA.

Bartolucci et al. (2007) give an outline of an algorithm for solving general marginal log-linear parametrizations, using a modification of the approach of Aitchison and Silvey (1958). An implementation of this was used to perform the simulations and calculations in Chapter 4, and it is likely to be more useful in practice than the approach outlined above.

Chapter 4

Parsimonious Modelling with Marginal Log-Linear Parameters

In this Chapter we consider the application of marginal log-linear parameters to sparse modelling and model selection.

Section 4.1 motivates this problem and illustrates the difficulties in defining sub-models using the generalized Möbius parameters. In Section 4.2 we use marginal log-linear parameters to create parsimonious sub-models of ADMG models, and provide simulations and applications to real data. Section 4.3 introduces the adaptive lasso, and shows how it can be used to perform consistent automatic model selection for marginal log-linear models. Finally, Section 4.4 contains simulations to demonstrate the performance of the adaptive lasso.

4.1 Motivation

Example 4.1.1. Consider the bidirected k-chain shown in Figure 4.1(a), and the model defined by it for binary random variables. As noted in Section 2.1, it contains precisely one parameter for each connected set in the graph, so in this case there are $\binom{k+1}{2} = O(k^2)$ parameters.

One possible reason for using a model with bidirected structure is as a proxy for a directed acyclic graph model where some of the variables are latent, or unmeasured. Consider the DAG in Figure 4.1(b), where the variables $h_1, \ldots, h_{k-1}$ are latent. Regardless of what assumptions we make (or do not make) assumptions about the state space or distribution of the latent variables, other than that the joint distribution obeys the global Markov property

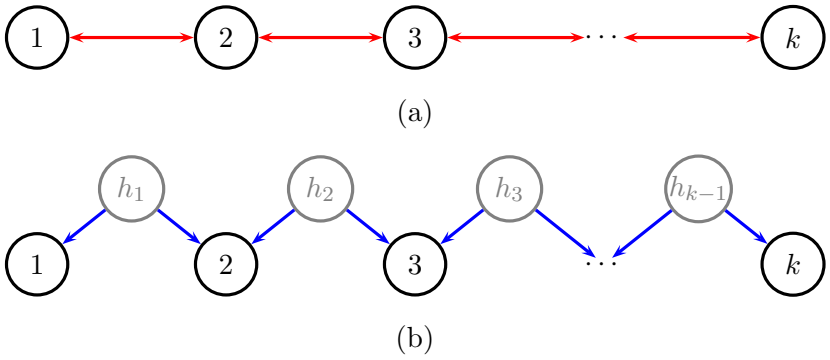


Figure 4.1: (a) A bidirected k-chain and (b) a DAG with latent variables $(h_1, \ldots, h_{k-1})$ generating the same conditional independence structure.

with respect to the DAG, the observed distribution over the remaining k variables satisfies the conditional independence structure given by the chain in Figure 4.1(a).

Now, if the number of states taken by each of the latent variables is fixed at say, m, then the number of parameters in the fully observed DAG model ($k \geq 2$) is $k(m-1)+2m+(k-2)m^2$, which is $O(k)$. The discrepancy between the number of parameters in the two models is quadratic, which suggests that the bidirected chain may have more parameters than are necessary in practice for explaining variation in data generated under the model in Figure 4.1(a).

Unfortunately, the generalized Möbius parameters present no obvious method for reducing the parameter count in a way which preserves the conditional independence structure of the model. In contrast, there are well established methods for sparse modelling with other classes of graphical models. For an undirected graph with discrete random variables, restricting to one parameter for each vertex and each edge leads the auto-logistic model of Besag (1974), also known in the binary case as a Boltzmann Machine (Ackley et al., 1985). Rudas et al. (2006) use marginal log-linear parameters to provide a sparse parametrization of a DAG model, again restricting to one parameter for each vertex and edge. We will see that the ingenuous parametrization allows us to produce analogous sub-models.

4.2 Parsimonious Modelling

Example 4.2.1. Consider again the models in Figure 4.1, for the case $k = 6$; suppose that we take distributions from the DAG model in (b), where each latent variable has $m = 3$ states, and each observed variable is binary.

We generated 2,500 probability distributions satisfying the DAG model in (b) as follows: the distribution of each latent state was taken to be uniform, so the probability of achieving each state was always one third; conditional on any instantiation of its latent parents, the probability of each observed variable being equal to 0 was chosen using an independent uniform random variable on $(0, 1)$. Now, for each of these 2,500 distributions, we drew a single random dataset of size 10,000. Each of our 2,500 datasets was generated from a distribution which satisfies the global Markov property of the ADMG in Figure 4.1(a), and we fitted this model to each dataset.

The histograms in Figure 4.2 show the increase in deviance, compared with the full model in Figure 4.1(a), caused by fixing higher order interaction parameters to be zero. The left plot contains the case where we set the 5- and 6-way interaction parameters to zero, a total of three parameters; a χ^2_3-density is shown for comparison. We see no significant difference between the histogram and the density, which is what we would expect if the three parameters have no explanatory effect at all. The centre plot shows the increase when the 4-, 5- and 6-way parameters are all set to zero; here the tail of the distribution in the histogram is slightly heavier than in the density, which suggests that these 4-way parameters have a measurable effect on the likelihood. Removing the 3-way interactions in addition to the higher order parameters results in a dramatic increase in the deviance, as seen in the very heavy tail of the histogram on the third plot.

If, instead of drawing the conditional probabilities uniformly, we obtain them from a Beta$(2, 2)$ distribution, the effect of the higher order parameters becomes even smaller. The histograms in Figure 4.3 show the associated increases in deviance; note that in the centre plot we cannot observe any significant increase in deviance from removing the 4-, 5- and 6-way interactions. The Beta$(2, 2)$ distribution has less probability density close to 0 and 1 than a uniform distribution, so the joint distributions obtained in this way are less likely to contain small cell probabilities.

Remark 4.2.2. Each of the parameters set to zero for fitting is almost surely non-zero, under the method being used to generate the 'true' distributions. The actual values obtained from our approach are presumably small, which accounts for the likelihood ratio test appearing to have very low power, even at a sample size of 10,000. As we saw above, the exact way in which we generate the probabilities affects how important the higher order interaction parameters appear.

Since each interaction is almost surely non-zero, increasing the sample size will eventually result in significant increases in deviance; this can be observed by performing the same simulations with datasets of size 100,000. That some interactions can appear unimportant

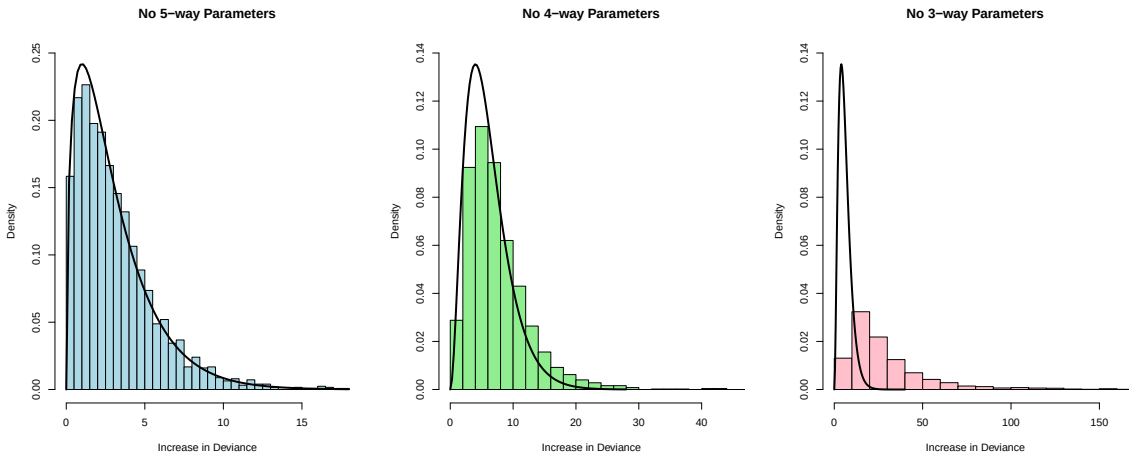


Figure 4.2: Increase in deviance compared with full bidirected chain over 2,500 simulations where probabilities are generated from Uniform$(0, 1)$ distribution. The left plot shows increase caused by setting 5- and 6-way interaction parameters to zero; the centre plot setting 4-, 5- and 6-way interaction parameters to zero; and the right plot 3-, 4-, 5- and 6-way parameters to zero.

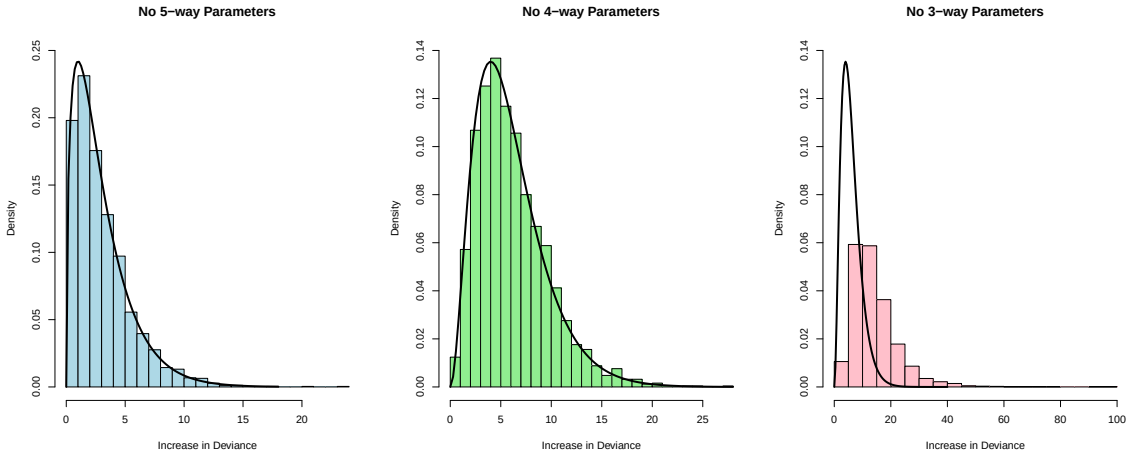


Figure 4.3: Increase in deviance compared with full bidirected chain over 2,500 simulations where probabilities are generated from Beta$(2, 2)$ distribution. The left plot shows increase caused by setting 5- and 6-way interaction parameters to zero; the centre plot setting 4-, 5- and 6-way interaction parameters to zero; and the right plot 3-, 4-, 5- and 6-way parameters to zero.

even with samples as large as 10,000 suggests that fixing these parameters to zero is a reasonable approach in some instances. We make no claim that either method used for generating distributions provides a realistic model of nature, but the next example shows that higher order interactions may indeed not be particularly useful in real datasets with finite samples.

Example 4.2.3. Drton and Richardson (2008a) examine responses to seven questions relating to trust and social institutions, taken from the US General Social Survey between 1975 and 1994. Briefly, the seven questions were:

Trust. Generally speaking, would you say that most people can be trusted or that you can't be too careful in life?

Helpful. Would you say that most of the time people try to be helpful, or that they are mostly just looking out for themselves?

MemUn. Are you a member of a labour union?

MemCh. Are you a member of a church?

ConLegis. Do you have confidence in congress?

ConClerg. Do you have confidence in organized religion?

ConBus. Do you have confidence in major companies?

In that paper, the model given by the graph in Figure 4.4 is shown to adequately explain the data, having a deviance of 32.67 on 26 degrees of freedom, when compared with the saturated model. The authors also provide an undirected graphical model which has one more edge than the graph in Figure 4.4, and yet 62 fewer parameters. It too gives a good fit to the data, having a deviance of 87.62 on 88 degrees of freedom.

For practical and theoretical reasons, the bidirected model may be preferred to the undirected one, even though the latter appears to be much more parsimonious. One might consider the responses to a questionnaire to be jointly affected by unmeasured characteristics of the respondent, such as her political beliefs. Such a system would give rise to an observed independence structure which can be represented by a bidirected graph, but not necessarily by an undirected one.

The greater parsimony of the undirected model, when defined purely by conditional independences, is due to its hierarchical nature: removing an edge between two vertices a and b corresponds to requiring that $\lambda_A^V = 0$ for every effect A containing both a and b. Removing

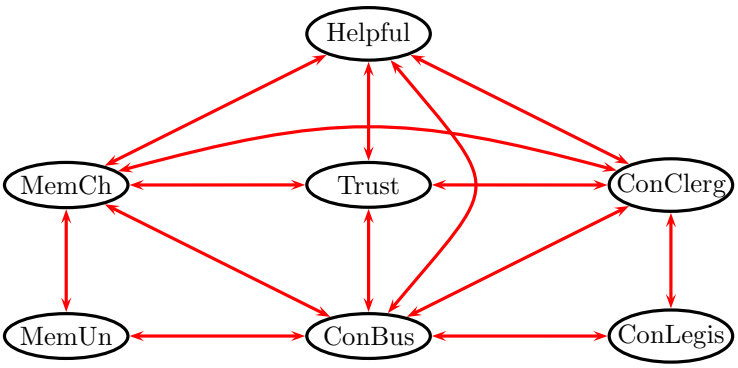


Figure 4.4: Markov model for Trust data given in Drton and Richardson (2008a).

that edge in a bidirected model may merely mean setting $\lambda^{ab}_{ab} = 0$, depending upon the other edges present; there is no cascading effect upwards to higher order parameters. Using the ingenuous parametrization, however, it is easy to also constrain these higher order terms to be zero.

Starting with the model in Figure 4.4 and fixing the 4-, 5-, 6- and 7-way interaction terms to be zero increases the deviance to 84.18 on 81 degrees of freedom; none of the 4-way interaction parameters was found to be significant on its own in a likelihood-ratio test against this sub-model. Furthermore, removing 21 of the remaining 25 three-way interaction terms increases the deviance to 111.48 on 102 degrees of freedom; using an asymptotic χ^2 approximation gives a p-value of 0.755, so this model is not contradicted by the data. Since some expected cell counts are small, the asymptotic approximation should be treated with caution. The only parameters retained are the one-dimensional marginal probabilities, the two-way interactions corresponding to edges, and the following three-way interactions:

MemUn, ConClerg, ConBus	Helpful, MemUn, MemCh
Trust, ConLegis, ConBus	MemCh, ConClerg, ConBus.

This model retains the marginal independence structure of Drton and Richardson's model, but provides a good fit with only 25 parameters, rather than the original 101.

Note that if we start with a different but Markov equivalent graph, we get an alternative ingenuous parametrization and a new set of possible sub-models.

4.3 Automatic Model Selection

The previous section demonstrates that the ingenuous parametrization can be used to generate sub-models of ordinary ADMG models, simply by setting a subset of the parameters to zero; this is known as *subset selection*. However, the approach to subset selection illustrated in Example 4.2.3 is somewhat ad hoc. In this section we develop a principled and automatic method.

Perhaps the most celebrated current approach to simultaneous parameter estimation and subset selection is the *lasso*, due to Tibshirani (1996). Originally developed in the context of linear regression where the number of parameters exceeds the number of observations, the lasso estimate minimizes the penalized log-likelihood function

$$\phi_n(\theta_1, \ldots, \theta_d) = -l_n(\theta_1, \ldots, \theta_d) + \nu_n \sum_{i=1}^{d} |\theta_i|, \qquad \nu_n > 0,$$

where l_n is the ordinary log-likelihood on n observations, and ν_n is a tuning parameter. The L_1-penalty shrinks the parameters towards zero, and is able to produce estimates which are exactly zero, because of the cusp in absolute value function.

Various authors (see, for example, Fan and Li, 2001; Zou, 2006) argue that a good parameter estimate $\boldsymbol{\theta}^n$ should have so-called *oracle* properties:

- $\boldsymbol{\theta}^n$ identifies the correct subset model $A \equiv \{j \,|\, \theta_j \neq 0\}$; and
- the non-zero parameters $\boldsymbol{\theta}^n_A$ have the optimal estimation rate $\sqrt{n}(\boldsymbol{\theta}^n_A - \boldsymbol{\theta}_A) \xrightarrow{\mathcal{D}} \mathrm{N}(0, \Sigma)$ where Σ is the covariance matrix we would expect if we knew A in advance.

However, based on results relating to linear regression, Fan and Li (2001) conjecture that the ordinary lasso is not oracle: essentially, if we have $\nu_n = O(\sqrt{n})$ then the penalty grows too slowly, and the correct subset is not chosen; and if $\nu_n/\sqrt{n} \to \infty$ then the resulting bias on the non-zero parameters is too large.

Zou (2006) developed the *adaptive lasso*, which addresses these deficiencies by weighting the penalties for each parameter. The adaptive lasso estimate minimizes

$$\phi_n(\theta_1, \ldots, \theta_k) = -l_n(\theta_1, \ldots, \theta_k) + \nu_n \sum_{i=1}^{k} w_i |\theta_i|, \qquad \nu_n > 0,$$

where $w_i = |\hat{\theta}^n_i|^{-\gamma}$, for $\gamma > 0$; here $\hat{\theta}^n_i$ is typically the ordinary MLE, but other consistent estimators can be used. The weights penalize small parameters more, enabling the estimates

to be exactly zero, whilst giving relatively small penalties to larger parameters, and thus avoiding the bias of the ordinary lasso. Zou shows that, in the case of linear regression, the adaptive lasso does possess the oracle properties. We show the same for marginal log-linear models.

Application of the lasso to graphical models on Gaussian random variables is performed by Meinshausen and Bühlmann (2004, 2006); ordinary log-linear parameters are considered by Nardi and Rinaldo (2007), who make use of the grouped lasso to ensure that only hierarchical models are selected.

4.3.1 Some Analytical Requirements

Suppose that $\boldsymbol{\eta}(\boldsymbol{p}) = C\log(M\boldsymbol{p})$ is a complete and hierarchical parametrization of the saturated model (see Remark 3.2.7). The function $\boldsymbol{\eta}(\boldsymbol{p})$ is clearly infinitely differentiable for $\boldsymbol{p} > \mathbf{0}$, and Bergsma and Rudas (2002), Theorem 2 shows that the transformation is smooth in the sense of Definition 2.1.1, meaning that $\boldsymbol{\eta}$ is a homeomorphism and the Jacobian $\frac{\partial \boldsymbol{\eta}}{\partial \boldsymbol{p}}$ has full rank for all $\boldsymbol{p} > \mathbf{0}$.

Take a fixed point $\boldsymbol{\eta}^* = \boldsymbol{\eta}(\boldsymbol{p}^*)$ where $\boldsymbol{p}^* > \mathbf{0}$. By the inverse function theorem, there exist open sets $U \ni \boldsymbol{p}^*$ and $V \ni \boldsymbol{\eta}^*$ such that $\boldsymbol{\eta} : U \to V$ is invertible with infinitely differentiable inverse $\boldsymbol{p}(\boldsymbol{\eta})$ (see, for example, Kass and Vos, 1997, pages 300–302). Since $\boldsymbol{p}$ is infinitely differentiable at $\boldsymbol{\eta}^*$, we can find an open set $W \subseteq U$ containing $\boldsymbol{p}^*$, such that all partial derivatives up to third order are bounded in W (this can be seen from the fact that these partial derivatives are continuous, for example).

From this, we can write the log-likelihood function as

$$\begin{aligned} l_n(\boldsymbol{\eta}) &= \sum_{i \in \mathfrak{X}} n_i \log p_i(\boldsymbol{\eta}) \\ &= l_n(\boldsymbol{\eta}^*) + (\boldsymbol{\eta} - \boldsymbol{\eta}^*)\dot{l}_n(\boldsymbol{\eta}^*) + \frac{1}{2}(\boldsymbol{\eta} - \boldsymbol{\eta}^*)^T \ddot{l}_n(\boldsymbol{\eta}^*)(\boldsymbol{\eta} - \boldsymbol{\eta}^*) + r(\boldsymbol{\eta} - \boldsymbol{\eta}^*) \end{aligned}$$

using Taylor's theorem. Further, $r(\cdot - \boldsymbol{\eta}^*)$ is bounded in a neighbourhood of $\boldsymbol{\eta}^*$.

It is easy to show that the score function has zero expectation at the truth and finite variance: $\mathbb{E}_{\boldsymbol{\eta}} \dot{l}_1(\boldsymbol{\eta}) = \mathbf{0}$ and $\mathbb{E}_{\boldsymbol{\eta}} \dot{l}_1(\boldsymbol{\eta})^2 < \infty$. Also, the Fisher Information matrix $I(\boldsymbol{\eta}) = -\mathbb{E}_{\boldsymbol{\eta}} \ddot{l}_1(\boldsymbol{\eta})$ is positive definite.

In the case of an unconstrained multinomial model, it is elementary to prove that the unique MLE is $\hat{\boldsymbol{p}}^n = \boldsymbol{n}/n$, so by the Central Limit Theorem,

$$\sqrt{n}(\hat{\boldsymbol{p}}^n - \boldsymbol{p}) \xrightarrow{\mathcal{D}} \mathrm{N}(\mathbf{0}, \operatorname{diag} \boldsymbol{p} - \boldsymbol{p}\boldsymbol{p}^T),$$

and the MLE is a $\sqrt{n}$-consistent estimator.

Proposition 4.3.1. *Suppose that we have a complete and hierarchical parametrization* $\boldsymbol{\eta} = C\log(M\boldsymbol{p})$*, and that* $\eta_j = 0$*. The ordinary MLE in the saturated model satisfies*

$$\sqrt{n}(\hat{\boldsymbol{\eta}}^n - \boldsymbol{\eta}^*) \xrightarrow{\mathcal{D}} \mathrm{N}(\mathbf{0}, I(\boldsymbol{\eta}^*)^{-1}),$$

so

$$P(\hat{\eta}_j^n = 0) \to 0$$

as $n \to \infty$*.*

Proof. As observed above, the MLE is $\hat{\boldsymbol{p}}^n = \boldsymbol{n}/n$ and $\sqrt{n}(\hat{\boldsymbol{p}}^n - \boldsymbol{p}) \xrightarrow{\mathcal{D}} \mathrm{N}(\mathbf{0}, \Sigma_p)$ for a non-negative definite matrix Σ_p with kernel of dimension 1. Since $\boldsymbol{\eta}(\boldsymbol{p})$ is a smooth map whose Jacobian has full rank everywhere, application of the delta method gives $\sqrt{n}(\hat{\boldsymbol{\eta}}^n - \boldsymbol{\eta}) \xrightarrow{\mathcal{D}} \mathrm{N}(\mathbf{0}, \Sigma_\eta)$, where $\Sigma_\eta = \left(\frac{\partial \boldsymbol{p}}{\partial \boldsymbol{\eta}}\right)^T \Sigma_p \left(\frac{\partial \boldsymbol{p}}{\partial \boldsymbol{\eta}}\right)$ is positive definite; this distribution is non-degenerate, so

$$\sqrt{n}\hat{\eta}_j \xrightarrow{\mathcal{D}} \mathrm{N}(\mathbf{0}, \sigma_j^2),$$

for some $\sigma_j^2 > 0$. But for every $\delta > 0$ we have

$$P(\hat{\eta}_j = 0) \leq P(\sqrt{n}|\hat{\eta}_j| < \delta) \to P(\sigma_j Z < \delta)$$

where $Z \sim \mathrm{N}(0, 1)$, and this last probability can be chosen arbitrarily small by varying δ. □

This result demonstrates that whilst the MLE is consistent and asymptotically normal, it never gives parameter estimates which are identically zero.

4.3.2 Oracle Properties of the Adaptive Lasso

The first result establishes the consistency of the adaptive lasso.

Lemma 4.3.2. *For* $\gamma > 0$*, let* $\nu_n/\sqrt{n} \to 0$ *and* $n^{\frac{\gamma-1}{2}}\nu_n \to \infty$*. With probability 1, for some*

$N \geq 1$ there exists a sequence $(\boldsymbol{\eta}^n)_{n=N}^{\infty}$ of local minima of ϕ_n such that

$$\sqrt{n}(\boldsymbol{\eta}^n - \boldsymbol{\eta}^*) = O(1).$$

In other words, the estimator $\boldsymbol{\eta}^n$ is $\sqrt{n}$-consistent.

Proof. We consider a $\sqrt{n}$-neighbourhood around the truth, $\boldsymbol{\eta}^*$. For some $K > 0$ let $N_K = \{\boldsymbol{u} \,|\, \|\boldsymbol{u}\| = K\sqrt{n}\}$, and for $\boldsymbol{a} \in N_K$ let

$$\begin{aligned}
\Psi_n(\boldsymbol{a}) &\equiv \left\{ \phi_n(\boldsymbol{\eta}^*) - \phi_n\left(\boldsymbol{\eta}^* + \frac{\boldsymbol{a}}{\sqrt{n}}\right) \right\} \\
&= -l_n(\boldsymbol{\eta}^*) + l_n\left(\boldsymbol{\eta}^* + \frac{\boldsymbol{a}}{\sqrt{n}}\right) + \nu_n \sum_j |\hat{\eta}_j|^{-\gamma} \left(|\eta_j^*| - \left|\eta_j^* + \frac{a_j}{\sqrt{n}}\right| \right) \\
&= \boldsymbol{a}^T \frac{\dot{l}_n(\boldsymbol{\eta}^*)}{\sqrt{n}} + \frac{1}{2n}\boldsymbol{a}^T \ddot{l}_n(\boldsymbol{\eta}^*)\boldsymbol{a} + \nu_n \sum_j |\hat{\eta}_j|^{-\gamma} \left(|\eta_j^*| - \left|\eta_j^* + \frac{a_j}{\sqrt{n}}\right| \right) + o_p(1).
\end{aligned}$$

For $j \in A$ we have

$$\begin{aligned}
\nu_n |\hat{\eta}_j|^{-\gamma} \left(|\eta_j^*| - \left|\eta_j^* + \frac{a_j}{\sqrt{n}}\right| \right) &\leq \nu_n |\hat{\eta}_j|^{-\gamma} \left| \frac{a_j}{\sqrt{n}} \right| \\
&\leq \frac{\nu_n}{\sqrt{n}} |\hat{\eta}_j|^{-\gamma} K \\
&\xrightarrow{p} 0,
\end{aligned}$$

since $|\hat{\eta}_j|^{-\gamma} \xrightarrow{p} |\eta_j^*|^{-\gamma}$ by consistency of the MLE, and $\nu_n/\sqrt{n} \to 0$. For $j \notin A$ on the other hand $\eta_j^* = 0$, and therefore the contribution to the sum is always negative. Thus

$$\begin{aligned}
\Psi(\boldsymbol{a}) &\leq \boldsymbol{a}^T \frac{\dot{l}_n(\boldsymbol{\eta}^*)}{\sqrt{n}} + \frac{1}{2n}\boldsymbol{a}^T \ddot{l}(\boldsymbol{\eta}^*)\boldsymbol{a} + \frac{\nu_n}{\sqrt{n}} \sum_{j \in A} |\hat{\eta}_j|^{-\gamma} K + o_p(1) \\
&\leq K \left\| \frac{\dot{l}_n(\boldsymbol{\eta}^*)}{\sqrt{n}} \right\| - \frac{1}{2}\lambda_{\min} K^2 + \frac{\nu_n}{\sqrt{n}} \sum_{j \in A} |\hat{\eta}_j|^{-\gamma} K + o_p(1)
\end{aligned}$$

using fact that $n^{-1}\ddot{l}_n(\boldsymbol{\eta}^*) \xrightarrow{p} -I(\boldsymbol{\eta}^*)$ where $I(\boldsymbol{\eta}^*)$ is positive definite, with minimal eigenvalue $\lambda_{\min} > 0$. So

$$\Psi(\boldsymbol{a}) \leq K \left\| \frac{\dot{l}_n(\boldsymbol{\eta}^*)}{\sqrt{n}} \right\| - \frac{1}{2}\lambda_{\min} K^2 + o_p(1).$$

Then since $n^{-1/2}\dot{l}_n(\boldsymbol{\eta}^*) = O_p(1)$ by the Central Limit Theorem, if K is chosen suitably

large we can ensure that the right hand side is negative for all $\boldsymbol{a} \in N_K$ with arbitrarily large probability.

Then since Ψ_n is continuous, is zero at $\boldsymbol{a} = \mathbf{0}$ and negative at any point on the boundary of N_K, it has a minimum in N_K. Equivalently there must be a minimum of ϕ_n within the $\sqrt{n}$-neighbourhood of ϕ_n. □

Theorem 4.3.3. *Suppose we have count data generated from a complete and hierarchical marginal log-linear model on $\mathfrak{X}_V$ with true parameter $\boldsymbol{\eta}^*$, and let $A^* \equiv \{j \,|\, \eta_j^* \neq 0\}$ be the set of non-zero elements of $\boldsymbol{\eta}^*$. Define $A^n \equiv \{j \,|\, \eta_j^n \neq 0\}$, the set of parameters estimated to be non-zero.*

For some $\gamma > 0$, let $\boldsymbol{\eta}^n$ be the adaptive lasso estimator with penalty ν_n, where $\nu_n/\sqrt{n} \to 0$ and $n^{\frac{\gamma-1}{2}}\nu_n \to \infty$.

The adaptive lasso estimates the model consistently, i.e. $P(A^n = A^*) \to 1$ as $n \to \infty$; and the non-zero parameters are estimated efficiently:

$$\sqrt{n}(\boldsymbol{\eta}_A^n - \boldsymbol{\eta}_A^*) \xrightarrow{\mathcal{D}} \mathrm{N}\left(0, I(\boldsymbol{\eta}^*)_{AA}^{-1}\right),$$

where $I(\boldsymbol{\eta}^*)_{AA}^{-1} = (I(\boldsymbol{\eta}^*)_{AA})^{-1}$.

Remark 4.3.4. Note that the asymptotic variance for the non-zero parameters $\boldsymbol{\eta}_A$ is the optimal Cramér-Rao bound when the values of the parameters in $\bar{A} \equiv V \backslash A$ are known. This variance is smaller in general than that of the ordinary maximum likelihood estimator $\hat{\boldsymbol{\eta}}_n$, which has asymptotic variance $\left(I^{-1}\right)_{AA} = (I_{AA})^{-1} + I^{A\bar{A}}(I^{\bar{A}\bar{A}})^{-1}I^{\bar{A}A}$, where $I^{\bar{A}A} \equiv (I^{-1})_{\bar{A}A}$ and so on.

Letting $\nu_n = Cn^r$, the range for r which satisfies the conditions of the theorem is $\frac{1-\gamma}{2} < r < \frac{1}{2}$. In practice, we choose ν_n using cross validation, and r determines how penalty chosen for the training dataset scales up to the full dataset. In the case of the ordinary lasso ($\gamma = 0$), it is not possible to satisfy the rate requirements on ν_n. Based on related results, Fan and Li (2001) conjecture that the ordinary lasso does not have oracle properties for linear regression.

The result of Theorem 4.3.3 was originally given in the case of linear regression models by Zou (2006), and bits of the proof below are similar to Theorem 2 of that paper. However we avoid having to call upon the theory of epiconvergence by using the consistency result in Lemma 4.3.2. An oracle result for finite regression mixture models was proved by Städler et al. (2010).

Proof. We know that with probability 1 there exists a sequence of adaptive lasso estimates $\boldsymbol{\eta}^n$ which consistently estimates the true value $\boldsymbol{\eta}^*$. This means that if $\eta_j^* \neq 0$, then almost surely $\eta_j^n \neq 0$ eventually for sufficiently large n. We first show that if $\eta_j^* = 0$, then η_j^n will be exactly zero eventually.

Consider the event $\{\eta_j^n \neq 0\}$. Since $\phi_n(\boldsymbol{\eta})$ (treated as a function of η_j with other components held fixed) is differentiable at η_j^n, we have

$$\begin{aligned}
0 = \left.\frac{\partial \phi_n}{\partial \eta_j}\right|_{\eta_j^n} &= -\dot{l}_n(\boldsymbol{\eta}^n) + \nu_n w_j \operatorname{sign}(\eta_j^n) \\
&= -\dot{l}_n(\boldsymbol{\eta}^n) + \nu_n |\hat{\eta}_j^n|^{-\gamma} \operatorname{sign}(\eta_j^n) \\
0 &= -\frac{1}{\sqrt{n}}\dot{l}_n(\boldsymbol{\eta}^*) - \sqrt{n}(\boldsymbol{\eta}^n - \boldsymbol{\eta}^*)\frac{1}{n}\ddot{l}_n(\boldsymbol{\eta}^*) + o_p(1) + n^{-\frac{1}{2}}\nu_n |\hat{\eta}_j^n|^{-\gamma} \operatorname{sign}(\eta_j^n) \\
&= -\frac{1}{\sqrt{n}}\dot{l}_n(\boldsymbol{\eta}^*) - \sqrt{n}(\boldsymbol{\eta}^n - \boldsymbol{\eta}^*)\frac{1}{n}\ddot{l}_n(\boldsymbol{\eta}^*) + n^{\frac{\gamma-1}{2}}\nu_n |\sqrt{n}\hat{\eta}_j^n|^{-\gamma} \operatorname{sign}(\eta_j^n) + o_p(1).
\end{aligned}$$

Now, since $\sqrt{n}(\boldsymbol{\eta}^n - \boldsymbol{\eta}^*) = O_p(1)$, and $n^{-1/2}\dot{l}_n(\boldsymbol{\eta}^*) \xrightarrow{\mathcal{D}} \mathrm{N}(\mathbf{0}, I(\boldsymbol{\eta}^*))$, and $n^{-1}\ddot{l}_n(\boldsymbol{\eta}^*) \xrightarrow{p} I(\boldsymbol{\eta}^*)$, the first two terms converge in distribution by Slutsky's Theorem, and $n^{-1/2}\dot{l}_n(\boldsymbol{\eta}^n) = O_p(1)$. For the remaining term,

$$\pm n^{\frac{\gamma-1}{2}}\nu_n |\sqrt{n}\hat{\eta}_j^n|^{-\gamma},$$

we have $n^{\frac{\gamma-1}{2}}\nu_n \to \infty$ and $\sqrt{n}\hat{\eta}_j^n = O_p(1)$, so

$$P\left(n^{\frac{\gamma-1}{2}}\nu_n |\sqrt{n}\hat{\eta}_j^n|^{-\gamma} \geq C\right) \to 1,$$

for any $C > 0$. Then

$$P(\eta_j^n \neq 0) \leq P\left(\frac{\dot{l}_n(\boldsymbol{\eta}^n)}{\sqrt{n}} = \pm n^{\frac{\gamma-1}{2}}\nu_n |\sqrt{n}\hat{\eta}_j^n|^{-\gamma}\right) \to 0,$$

and thus $P(A^n = A^*) \to 1$ as $n \to \infty$.

Now, for sufficiently large n, we know that $\boldsymbol{\eta}_{\bar{A}}^n = \mathbf{0}$, and thus the minimization problem reduces to

$$\phi_n^A(\boldsymbol{\eta}_A) = -l_n(\boldsymbol{\eta}_A, \mathbf{0}) + \nu_n \sum_{j \in A} w_j |\eta_j|.$$

Further, $\phi_n^A(\boldsymbol{\eta}_A)$ is differentiable in an open set containing $\boldsymbol{\eta}_A^*$, because all of the components of the vector $\boldsymbol{\eta}_A^*$ are non-zero. Thus, by the Lemma 4.3.2, with probability tending to 1

there exists $\boldsymbol{\eta}_A^n$, in a $\sqrt{n}$-neighbourhood of $\boldsymbol{\eta}_A^*$, such that $\dot{\phi}_n(\boldsymbol{\eta}_A^n) = \mathbf{0}$. Thus,

$$\begin{aligned}\mathbf{0} = \frac{1}{\sqrt{n}}\dot{\phi}_n(\boldsymbol{\eta}_A^n) &= -\frac{1}{\sqrt{n}}\dot{l}_n(\boldsymbol{\eta}_A^n) + \frac{\nu_n}{\sqrt{n}}\sum_{j\in A}|\hat{\eta}_j^n|^{-\gamma}\operatorname{sign}(\eta_j^n) \\ &= -\frac{1}{\sqrt{n}}\dot{l}_n(\boldsymbol{\eta}_A^*) + \frac{1}{\sqrt{n}}\left(\int_0^1 \ddot{l}_n(\boldsymbol{\eta}_A^* + t(\boldsymbol{\eta}_A^n - \boldsymbol{\eta}_A^*))\,dt\right)(\boldsymbol{\eta}_A^n - \boldsymbol{\eta}_A^*) \\ &\quad + \frac{\nu_n}{\sqrt{n}}\sum_{j\in A}|\hat{\eta}_j^n|^{-\gamma}\operatorname{sign}(\eta_j^n),\end{aligned}$$

using the multivariate Mean Value Theorem (see, for example, Ferguson, 1996). Here the derivatives $\dot{l}_n$ and $\ddot{l}_n$ should be understood as being with respect to $\boldsymbol{\eta}_A$ rather than $\boldsymbol{\eta}$. Then

$$\mathbf{0} = -\frac{1}{\sqrt{n}}\dot{l}_n(\boldsymbol{\eta}_A^*) + B_n\sqrt{n}(\boldsymbol{\eta}_A^n - \boldsymbol{\eta}_A^*) + \frac{\nu_n}{\sqrt{n}}\sum_{j\in A}|\hat{\eta}_j^n|^{-\gamma}\operatorname{sign}(\eta_j^n),$$

where

$$B_n \equiv \int_0^1 \left(\frac{1}{n}\ddot{l}_n(\boldsymbol{\eta}_A^* + t(\boldsymbol{\eta}_A^n - \boldsymbol{\eta}_A^*))\right)dt.$$

By standard asymptotic results,

$$-\frac{1}{\sqrt{n}}\dot{l}_n(\boldsymbol{\eta}_A^*) \xrightarrow{\mathcal{D}} \mathrm{N}(\mathbf{0}, I(\boldsymbol{\eta}^*)_{AA})$$

and by arguments found in Ferguson (1996, Chapter 18), $B_n \xrightarrow{p} -I(\boldsymbol{\eta}^*)_{AA}$, which is a positive definite matrix; by continuity of matrix inverses, B_n^{-1} will exist eventually. Since $\nu_n/\sqrt{n} \to 0$, the last term tends in probability to 0. Since

$$\sqrt{n}(\boldsymbol{\eta}_A^n - \boldsymbol{\eta}_A^*) = B_n^{-1}\left(\frac{1}{\sqrt{n}}\dot{l}_n(\boldsymbol{\eta}_A^*) - \frac{\nu_n}{\sqrt{n}}\sum_{j\in A}|\hat{\eta}_j^n|^{-\gamma}\operatorname{sign}(\eta_j^n)\right),$$

by Slutsky's Theorem we obtain

$$\sqrt{n}(\boldsymbol{\eta}_A^n - \boldsymbol{\eta}_A^*) \xrightarrow{\mathcal{D}} \mathrm{N}(\mathbf{0}, I(\boldsymbol{\eta}^*)_{AA}^{-1}).$$

□

4.3.3 Scaling

The definition of a marginal log-linear parameter $\lambda_L^M(i_L)$ includes an essentially arbitrary multiplicative constant $c_M \equiv |\mathfrak{X}_M|^{-1}$ (see Proposition 3.1.5). Thus far, we have never been required to compare MLL parameters defined within different margins, and the value of this constant seems irrelevant so long as c_M does not depend upon L or i_L. However, the lasso presents a situation in which the relative penalization of the different parameters does depend upon their scale, even for parameters defined within different margins.

A particular advantage of the adaptive lasso with the choice $\gamma = 1$ is that the penalties are scale invariant. Other values of $\gamma > 0$ will still provide consistent model selection and efficient parameter estimates as indicated by Theorem 4.3.3, but for finite samples the model selected and the parameter estimates may vary with the arbitrary choice of scale, which seems undesirable.

The choice $c_M = |\mathfrak{X}_M|^{-1}$ is due to Bergsma and Rudas (2002), and is a legacy of ordinary log-linear parameters (Tamás Rudas, personal communication). Suppose that we generate the probabilities $(p_i)_{i \in \mathfrak{X}}$ as independent positive random variables from some distribution with mean μ and variance $\sigma^2 < \infty$. Note that if we rescale the probabilities to ensure that they sum to one, this does not affect the value of any MLL parameter, as the rescaling constant will be cancelled out. The variance of the resulting MLL parameters is given by

$$\operatorname{Var} \lambda_L^M(i_L) = c_M^2 |\mathfrak{X}_M| \operatorname{Var} \left(\log p_M(\mathbf{0}) \right).$$

Here $p_M(\mathbf{0})$ is a sum of $|\mathfrak{X}_{V \setminus M}|$ independent random variables with mean μ and variance σ^2. If M is large (i.e. $V \setminus M$ is small), the resulting random variable $\log p_M(\mathbf{0})$ will be relatively unstable, since the probability may be close to 0, suggesting that c_M should compensate for the increased variance which will result.

$$\begin{aligned} \log p_M(\mathbf{0}) &= -\log \left(\mu |\mathfrak{X}_{V \setminus M}| \right) + \log \left(1 + |\mathfrak{X}_{V \setminus M}|^{-1} \mu^{-1} p_M(\mathbf{0}) - 1 \right) \\ &= -\log \left(\mu |\mathfrak{X}_{V \setminus M}| \right) + \frac{p_M(\mathbf{0})}{|\mathfrak{X}_{V \setminus M}| \mu} - 1 + O\left(\left(\frac{p_M(\mathbf{0})}{|\mathfrak{X}_{V \setminus M}| \mu} - 1 \right)^2 \right) \end{aligned}$$

so

$$\operatorname{Var} \left(\log p_M(\mathbf{0}) \right) = \operatorname{Var} \left(\frac{p_M(\mathbf{0})}{|\mathfrak{X}_{V \setminus M}| \mu} - 1 \right) + O\left(\operatorname{Var} \left(\frac{p_M(\mathbf{0})}{|\mathfrak{X}_{V \setminus M}| \mu} - 1 \right)^2 \right).$$

Applying the central limit theorem

$$|\mathfrak{X}_{V\setminus M}|^{1/2}\left(\frac{p_M(\mathbf{0})}{|\mathfrak{X}_{V\setminus M}|\mu}-1\right)\xrightarrow{\mathcal{D}}\mathrm{N}\left(0,\mu^{-2}\sigma^2\right),$$

so

$$\mathrm{Var}\left(\log p_M(\mathbf{0})\right)=|\mathfrak{X}_{V\setminus M}|^{-1}\frac{\sigma^2}{\mu^2}+O\left(\frac{\sigma^4}{\mu^4|\mathfrak{X}_{V\setminus M}|^2}\right).$$

This gives

$$\mathrm{Var}\,\lambda_L^M(i_L)=c_M^2\frac{|\mathfrak{X}_M|^2}{|\mathfrak{X}_V|}\frac{\sigma^2}{\mu}+O\left(\frac{\sigma^4|\mathfrak{X}_M|}{\mu^4|\mathfrak{X}_{V\setminus M}|^2}\right),$$

which suggests that $c_M=|\mathfrak{X}_M|^{-1}$ is indeed the appropriate choice in order to keep parameters on the same scale. The quality of the approximation depends upon how close $p_M(\mathbf{0})$ is to its mean $|\mathfrak{X}_{V\setminus M}|\mu$, or equivalently the size of σ/μ. If each probability is a Gamma$(k,1)$ random variable up to rescaling, which corresponds to probabilities being generated as a Dirichlet$(k,\ldots,k)$, then the approximation improves as $k\to\infty$.

Having established variances for the marginal log-linear parameters under particular mechanisms for generating probabilities, a natural next question is how those parameters are correlated. The next result gives mild conditions for zero correlation between MLL parameters.

Proposition 4.3.5. *Suppose that $(p_i)_{i\in\mathfrak{X}}$ are generated in such a way that they are exchangeable. For any margins $M,N\subseteq V$, any effects $L\subseteq M$ and $K\subseteq N$ with $K\neq L$, and any $i_L\in\mathfrak{X}_L$ and $j_K\in\mathfrak{X}_K$,*

$$\mathrm{Cov}\left(\lambda_L^M(i_L),\,\lambda_K^N(j_K)\right)=0.$$

Proof. First note that

$$\begin{aligned}\mathbb{E}\lambda_L^M(i_L)&=\sum_{j_M\in\mathfrak{X}}\mathbb{E}\left[\log p_M(j_M)\right]\prod_{v\in L}(|\mathfrak{X}_v|\mathbb{I}_{\{i_v=j_v\}}-1)\\&=\mathbb{E}\left[\log p_M(j_M)\right]\sum_{j_M\in\mathfrak{X}}\prod_{v\in L}(|\mathfrak{X}_v|\mathbb{I}_{\{i_v=j_v\}}-1)\\&=0,\end{aligned}$$

because $\mathbb{E} \log p_M(j_M)$ is constant in j_M by exchangeability. Thus

$$\operatorname{Cov}\left(\lambda_L^M(i_L),\, \lambda_K^N(j_K)\right) = \mathbb{E}\left[\lambda_L^M(i_L) \cdot \lambda_K^N(j_K)\right].$$

Assume without loss of generality that $L \setminus K \neq \emptyset$, and choose $v \in L \setminus K$. Now,

$$\mathbb{E}\left[\lambda_L^M(i_{L\setminus\{v\}},\, i'_v) \cdot \lambda_K^N(j_K)\right] = \mathbb{E}\left[\lambda_L^M(i_{L\setminus\{v\}},\, i_v) \cdot \lambda_K^N(j_K)\right]$$

by exchangeability, because replacing i_v with i'_v in $\lambda_K^N(j_K)$ changes nothing. Thus

$$\begin{aligned}
|\mathfrak{X}_v| \mathbb{E}\left[\lambda_L^M(i_{L\setminus\{v\}},\, i_v) \cdot \lambda_K^N(j_K)\right] &= \sum_{i'_v \in \mathfrak{X}_v} \mathbb{E}\left[\lambda_L^M(i_{L\setminus\{v\}},\, i'_v) \cdot \lambda_K^N(j_K)\right] \\
&= \mathbb{E}\left[\lambda_K^N(j_K) \sum_{i'_v \in \mathfrak{X}_v} \lambda_L^M(i_{L\setminus\{v\}},\, i'_v)\right] \\
&= \mathbb{E}0
\end{aligned}$$

by Corollary 3.1.6, giving the result. □

This result is interesting, because it suggests that we might approximate a uniform distribution over the probability simplex with independent normal distributions over the MLL parameters.

4.4 Simulated Examples

To test the adaptive lasso procedure outlined above, we used a series of simulations. Each simulation was structured as follows: we selected a distribution uniformly at random on the probability simplex; we fixed the values of λ_C^C for C connected in the bidirected 4-chain (Figure 4.1(a) with $k = 4$), and set $\lambda_D^D = 0$ for other (disconnected) sets D. We then generated a sample of size n from this new distribution, which obeys the global Markov property for the bidirected 4-chain.

Given the data set, we applied the adaptive lasso to the ingenuous parametrization for the complete bidirected graph on four variables, and tried to recover the correct model. We did not apply the L_1-penalty to one-way marginal parameters, so there are $2^{11} = 2048$ possible sub-models to select from. Selection of the penalty ν was by 10-fold cross-validation using a Kullback-Leibler loss function on the $\boldsymbol{p}$ scale, under the assumption that $\nu_n = Cn^r$ for some chosen $r > 0$. The minimization steps were performed using an algorithm based on the methods of Aitchison and Silvey (1958), which we do not describe here.

For sample sizes from 1,000 to 300,000, penalty rates $r \in \left\{\frac{1}{4}, \frac{1}{3}, \frac{1}{2}, \frac{2}{3}\right\}$, and weights $\gamma \in \left\{0, \frac{1}{2}, 1\right\}$, we applied the method to $N = 250$ different distributions generated in the way described above. Note that the same 250 datasets were used for each r and γ, although for different sample sizes the experiments were started from scratch. The proportion of times the correct model was recovered is given in Table 4.1.

Consistent model selection is seen clearly for $\gamma = \frac{1}{2}$ and $\gamma = 1$, but seems to fail for the ordinary lasso ($\gamma = 0$). It is unclear precisely why this occurs, since heuristically we might expect consistency for $r > \frac{1}{2}$; small simulation studies based on fixing $\nu_n = Cn^r$ for some C, and increasing the sample size within a single distribution (rather than estimating C by cross-validation) suggest that the ordinary lasso *is* consistent for $r = \frac{2}{3}$. The failure may be due to a problem with cross-validation for selecting C.

Table 4.2 shows the decrease in the root mean squared error (RMSE) for estimation of $\boldsymbol{\eta}^*$, as $n \to \infty$. There is a slight increase in RMSE for larger values of r, which we expect because larger penalty functions will slightly increase the estimation bias. The RMSE of the maximum likelihood estimator is given for comparison, and is very similar to the various lasso approaches. Overall, the particular choices of r and $\gamma > 0$ appear not to have much effect, so we suggest the scale invariant $\gamma = 1$ together with any r which satisfies the conditions on ν_n in Theorem 4.3.3.

n	$\gamma = 0$				$\gamma = \frac{1}{2}$				$\gamma = 1$			
	$r = \frac{1}{4}$	$\frac{1}{3}$	$\frac{1}{2}$	$\frac{2}{3}$	$\frac{1}{4}$	$\frac{1}{3}$	$\frac{1}{2}$	$\frac{2}{3}$	$\frac{1}{4}$	$\frac{1}{3}$	$\frac{1}{2}$	$\frac{2}{3}$
1,000	0.028	0.020	0.052	0.040	0.112	0.136	0.080	0.076	0.100	0.148	0.148	0.104
3,000	0.036	0.024	0.060	0.044	0.116	0.132	0.180	0.228	0.292	0.212	0.272	0.240
10,000	0.020	0.016	0.036	0.064	0.248	0.320	0.388	0.360	0.400	0.456	0.432	0.464
30,000	0.016	0.008	0.012	0.048	0.360	0.432	0.468	0.560	0.596	0.584	0.588	0.636
100,000	0.016	0.020	0.020	0.032	0.512	0.532	0.604	0.608	0.724	0.728	0.740	0.776
300,000	0.016	0.016	0.012	0.024	0.632	0.684	0.668	0.732	0.848	0.836	0.852	0.844

Table 4.1: Proportion of times correct model is recovered by the adaptive lasso (from 250 simulations) for varying sample sizes n, weights γ, and rates on the penalty function, r.

n	MLE	$\gamma = 0$				$\gamma = \frac{1}{2}$				$\gamma = 1$			
		$r = \frac{1}{4}$	$\frac{1}{3}$	$\frac{1}{2}$	$\frac{2}{3}$	$\frac{1}{4}$	$\frac{1}{3}$	$\frac{1}{2}$	$\frac{2}{3}$	$\frac{1}{4}$	$\frac{1}{3}$	$\frac{1}{2}$	$\frac{2}{3}$
1,000	0.134	0.134	0.136	0.140	0.148	0.135	0.137	0.141	0.148	0.138	0.140	0.143	0.150
3,000	0.080	0.080	0.080	0.082	0.087	0.077	0.078	0.080	0.084	0.077	0.078	0.080	0.083
10,000	0.045	0.046	0.046	0.047	0.049	0.043	0.044	0.045	0.047	0.042	0.042	0.043	0.045
30,000	0.026	0.028	0.028	0.029	0.030	0.027	0.027	0.028	0.029	0.026	0.026	0.026	0.027
100,000	0.016	0.017	0.017	0.017	0.018	0.016	0.016	0.017	0.017	0.015	0.015	0.015	0.015
300,000	0.009	0.011	0.011	0.012	0.012	0.011	0.011	0.011	0.012	0.009	0.009	0.009	0.010

Table 4.2: Root mean squared error for estimation of $\boldsymbol{\eta}^*$ by the adaptive lasso (from 250 simulations) for varying sample sizes n, weights γ, and rates on the penalty function, r. MLE error is given for comparison.

Chapter 5

Variation Independence

Results in Chapter 3 summarize when a complete and hierarchical marginal log-linear parametrization is variation independent (VI), as well as which ADMGs lead to a VI ingenuous parametrization. Some models, such as the bidirected 4-chain, were shown to have a variation dependent ingenuous parametrization, even though a different MLL parametrization of the same model exists which is variation independent.

In this chapter we explore the possibility of constructing VI parametrizations of general ADMG models. In Section 5.1 we recall the characterization the variation dependence of generalized Möbius parameters established in Chapter 2, and introduce a notation for variation independence. In Section 5.2 the characterization of variation dependence for generalized Möbius parameters is used with Fourier-Motzkin elimination to produce VI parametrizations of some models. Section 5.3 presents the first VI parametrization of the bidirected 5-chain, and in Section 5.4 we see the difficulty in applying the same methods to the bidirected 5-cycle. Section 5.5 gives a partially constructive proof that any ADMG model admits a VI parametrization.

5.1 Variation Independence as a Graphoid

Let $\mathcal{G}$ be an ADMG with vertex set V, and let $\mathcal{P}_\mathcal{G} \subseteq \Delta_{2^{|V|}-1}$ be the space of probability distributions on binary random variables which satisfy the global Markov property with respect to $\mathcal{G}$. Recall that there is a smooth parameter function $\boldsymbol{q} : \mathcal{P}_\mathcal{G} \to \mathcal{Q}_\mathcal{G}$ which maps probabilities $\boldsymbol{p} \in \mathcal{P}_\mathcal{G}$ to their corresponding generalized Möbius parameters $\boldsymbol{q}(\boldsymbol{p})$.

Recall the following result, proved in Chapter 2.

Theorem 2.1.4. *For an ADMG $\mathcal{G}$, a vector of generalized Möbius parameters $\boldsymbol{q}$ is valid (i.e. $\boldsymbol{q} \in \mathcal{Q}_{\mathcal{G}}$) if and only if for each $i_V \in \mathfrak{X}_V$ we have*

$$f_{i_V}(\boldsymbol{q}) \equiv \sum_{C\,:\,i_V^{-1}(0) \subseteq C \subseteq V} (-1)^{|C \setminus i_V^{-1}(0)|} \prod_{H \in [C]_{\mathcal{G}}} q_{H|T}^{i_T} > 0,$$

where $i_V^{-1}(0) \equiv \{v \in V \,|\, i_v = 0\}$.

We briefly introduce some useful notation relating to variation dependence; this can be found, for example, in Dawid (2001).

Definition 5.1.1. Let $\Theta \subseteq \mathbb{R}^k$, and let $\boldsymbol{\theta}$ be some function with domain $\mathcal{X}$, taking values in Θ. For $A \subseteq \{1, \ldots, k\}$, we denote by $\boldsymbol{\theta}_A$ the sub-vector of $\boldsymbol{\theta}$ from co-ordinates in A. Let

$$R(\boldsymbol{\theta}_A) \equiv \{\boldsymbol{\theta}_A(x) \,|\, x \in \mathcal{X}\},$$

be the *range* of θ_A. Also, for $B \subseteq \{1, \ldots, k\} \setminus A$, let

$$R(\boldsymbol{\theta}_A \,|\, \boldsymbol{\theta}_B = \boldsymbol{y}) \equiv \{\boldsymbol{\theta}_A(x) \,|\, x \in \mathcal{X} \text{ and } \boldsymbol{\theta}_B(x) = \boldsymbol{y}\},$$

be the conditional range of $\boldsymbol{\theta}_A$ given $\boldsymbol{\theta}_B = \boldsymbol{y}$.

We will only consider a very specific situation in which $\boldsymbol{\theta} : \Theta \to \Theta$ is the identity function, and the (conditional) range operator gives us (conditional) projections of the set Θ. From hereon we abbreviate $R(\boldsymbol{\theta}_A \,|\, \boldsymbol{\theta}_B = \boldsymbol{y})$ to $R(\boldsymbol{\theta}_A \,|\, \boldsymbol{\theta}_B)$.

We can now reformulate Definition 2.2.1 as follows: a parameter vector $\boldsymbol{\theta}$ taking values in some set Θ is variation independent if

$$\Theta = R(\boldsymbol{\theta}) = R(\theta_1) \times \cdots \times R(\theta_k),$$

and we write $\perp_{var} \{\theta_1, \ldots, \theta_k\}$. Conditional variation independence of $\boldsymbol{\theta}$ and $\boldsymbol{\eta}$ given $\boldsymbol{\zeta}$ is defined as

$$R(\boldsymbol{\theta}, \boldsymbol{\eta} \,|\, \boldsymbol{\zeta}) = R(\boldsymbol{\theta} \,|\, \boldsymbol{\zeta}) \times R(\boldsymbol{\eta} \,|\, \boldsymbol{\zeta}),$$

and denoted by $\boldsymbol{\theta} \perp_{var} \boldsymbol{\eta} \,|\, \boldsymbol{\zeta}$.

Conditional variation independence $\langle \cdot \perp_{var} \cdot \,|\, \cdot \rangle$ is a graphoid, and so obeys some of the same properties as ordinary conditional independence; see, for example, Dawid and Lauritzen (1993); Dawid (2001).

5.2 Fourier-Motzkin Elimination

We now seek to use the characterization of the variation dependence in generalized Möbius parameters from Theorem 2.1.4 to construct variation independent parametrizations for ADMG models. Firstly we introduce a method for eliminating variables from a system of linear inequalities, which may also be thought of as projection onto a subspace.

Suppose a set $\mathcal{X} \subseteq \mathbb{R}^p$ is described by a set of linear inequalities

$$\mathcal{X} \equiv \{\boldsymbol{x} \mid f_i(\boldsymbol{x}) > 0, i = 1, \ldots, I\},$$

where:

$$f_i(\boldsymbol{x}) = a_{i0} + \sum_{j=1}^{p} a_{ij} x_j, \qquad i = 1, \ldots, I.$$

We can easily rewrite the subset of these inequalities which involve x_p in the form:

$$\begin{aligned} x_p - u_i(x_1, \ldots, x_{p-1}) &< 0, & i &= 1, \ldots, I_u \\ x_p - l_j(x_1, \ldots, x_{p-1}) &> 0, & j &= 1, \ldots, I_l, \end{aligned}$$

where u_i and l_j are also linear, and trivially

$$\max_i l_i(x_1, \ldots, x_{p-1}) < x_p < \min_i u_i(x_1, \ldots, x_{p-1}).$$

This provides an explicit range of values for x_p which, given the values of $x_1, \ldots, x_{p-1}$, will mean that $(x_1, \ldots, x_p) \in \mathcal{X}$. Note that the range will be empty if $\min_i u_i < \max_i l_i$.

Now suppose we choose values of $x_1, \ldots, x_p$ sequentially, in order to obtain some $\boldsymbol{x} \in \mathcal{X}$; which values can we select for the first $p-1$ variables, in order to guarantee that there is some value of x_p left to choose? We must ensure precisely that $\min_i u_i > \max_i l_i$, or equivalently that

$$l_j(x_1, \ldots, x_{p-1}) < u_i(x_1, \ldots, x_{p-1}) \qquad i = 1, \ldots, I_u, \quad j = 1, \ldots, I_l.$$

This creates a new collection of $I_l \times I_u$ linear inequalities which, together with any of the original inequalities which did not involve x_p, defines the projection of $\mathcal{X}$ onto its first $p-1$ co-ordinates. This procedure is known as *Fourier-Motzkin elimination*, having been first described by Fourier (1824), and independently rediscovered by Dines (1919) and Motzkin (1936).

Note that the elimination of a variable in this way may substantially increase the number of inequalities in the system; in the worst case, I may become $\frac{1}{4}I^2$. After k eliminations, this leaves up to $4(\frac{1}{4}I)^{2^k}$ inequalities, which is doubly-exponential in k. This is in contrast the elimination of variables from a system of linear equations, wherein the number of equations is reduced at each step.

Example 5.2.1. Let $\mathcal{G}$ be the complete bidirected graph on 3 vertices. The inequalities (2.1) are linear in q_1, q_2, q_{12}, q_3, q_{13}, q_{23}, and q_{123}, where we use this ordering on the parameters; the bounds may be written as

$$\max\left\{\begin{array}{c} 0, \\ q_{12}+q_{23}-q_2, \\ q_{12}+q_{13}-q_1, \\ q_{13}+q_{23}-q_3 \end{array}\right\} < q_{123} < \min\left\{\begin{array}{c} q_{12}, \\ q_{23}, \\ q_{13}, \\ 1-q_1-q_2-q_3+q_{12}+q_{23}+q_{13} \end{array}\right\}. \tag{5.1}$$

These were dubbed *Fréchet bounds* by Dobra and Fienberg (2000), being an extension of bounds arising from cumulative distribution functions.

In order for us to be able to select a value for q_{123} which satisfies the inequalities, every expression on the right hand side must be larger than every expression on the left. This induces 16 further inequalities, obtained by Fourier-Motzkin elimination of the parameter q_{123}:

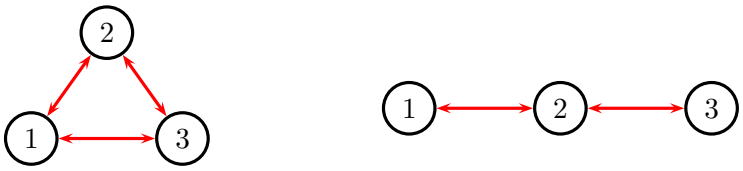
$$\max\left\{\begin{array}{c} 0, \\ q_2+q_3-1, \\ q_{13}+q_{12}-q_1, \\ -1+q_1+q_2+q_3-q_{12}-q_{13} \end{array}\right\} < q_{23} < \min\left\{\begin{array}{c} q_2, \\ q_3, \\ q_{13}+q_2-q_{12}, \\ q_{12}+q_3-q_{13} \end{array}\right\}, \tag{5.2}$$

$$\max\left\{\begin{array}{c} 0, \\ q_1+q_3-1 \end{array}\right\} < q_{13} < \min\left\{\begin{array}{c} q_1, \\ q_3 \end{array}\right\}, \tag{5.3}$$

$$\max\left\{\begin{array}{c} 0, \\ q_1+q_2-1 \end{array}\right\} < q_{12} < \min\left\{\begin{array}{c} q_1, \\ q_2 \end{array}\right\}. \tag{5.4}$$

The inequalities in (5.3) and (5.4) correspond only to the Fréchet bounds for q_{13} and q_{12} respectively, but (5.2) includes four additional constraints.

We may repeat the procedure and eliminate q_{23}, then q_{13}, and so on. However, the only bounds we obtain in addition to those already mentioned are $0 < q_i < 1$ for $i = 1, 2, 3$.



(a) (b)

Figure 5.1: (a) The complete bidirected graph on 3 variables; (b) the bidirected 3-chain.

Once we have a collection of explicit constraints on each parameter conditional on the parameters which precede it in the ordering, it becomes relatively easy to construct a variation independent parametrization. In particular, if the bounds on the kth parameter are of the form

$$\max_i l_i(x_1, \ldots, x_{k-1}) < x_k < \min_i u_i(x_1, \ldots, x_{k-1})$$

then define the kth variation independent 'version' of that parameter by

$$\psi_k = \log \frac{\prod_i \{x_k - l_i(x_1, \ldots, x_{k-1})\}}{\prod_i \{u_i(x_1, \ldots, x_{k-1}) - x_k\}}.$$

This creates a *logarithmic barrier* for x_k with each of its bounds. The value of ψ_k is strictly increasing in x_k, it approaches $+\infty$ as $x_k \to \min_i u_i$, and $-\infty$ as $x_k \to \max_i l_i$. Thus the range of ψ_k is always $\mathbb{R}$, regardless of the value of the parameters $x_1, \ldots, x_{k-1}$, provided that they are valid.

The logarithmic barrier is much used in numerical optimization to penalize a function when close to the boundary of a feasible space (Nocedal and Wright, 1999); the technique of applying it to create variation independent parametrizations is found in Richardson et al. (2010).

Example 5.2.2. Continuing Example 5.2.1, we can construct a variation independent parametrization of the saturated model over three binary random variables (see Figure 5.1(a)) using the logarithmic barrier. In particular, let

$$\begin{aligned}
\psi_i &= \log \frac{q_i}{1 - q_i}, \quad i = 1, 2, 3, \\
\psi_{12} &= \log \frac{q_{12}(1 - q_1 - q_2 + q_{12})}{(q_1 - q_{12})(q_2 - q_{12})}, \qquad \psi_{13} = \log \frac{q_{13}(1 - q_1 - q_3 + q_{13})}{(q_1 - q_{13})(q_3 - q_{13})}, \\
\psi_{23} &= \log \frac{q_{23}(1 - q_2 - q_3 + q_{23})}{(q_2 - q_{23})(q_3 - q_{23})} \\
&\quad + \log \frac{(q_1 - q_{13} - q_{12} + q_{23})(1 - q_1 - q_2 - q_3 + q_{12} + q_{13} + q_{23})}{(q_{13} + q_2 - q_{12} - q_{23})(q_{12} + q_3 - q_{13} - q_{23})},
\end{aligned}$$

$$\psi_{123} = \log \frac{q_{123}(q_1 - q_{12} - q_{13} + q_{123})}{(q_{12} - q_{123})(q_{13} - q_{123})} + \log \frac{(q_2 - q_{12} - q_{23} + q_{123})(q_3 - q_{23} - q_{13} + q_{123})}{(q_{23} - q_{123})(1 - q_1 - q_2 - q_3 + q_{12} + q_{13} + q_{23} - q_{123})}.$$

Then ψ_1, ψ_2, ψ_{12}, ψ_3, ψ_{13}, ψ_{23} and ψ_{123} are a variation independent parametrization of the saturated model on three binary random variables. Note that up to multiplicative constants and with the exception of ψ_{23}, these parameters are identical to the ingenuous parameters for the graph in Figure 5.1(a), which we know to be variation dependent by Theorem 3.4.5 and the fact that the graph has a head of size 3. The second term in ψ_{23} corresponds to the four bounds in (5.2) other than the Fréchet bounds.

Observe that we can rewrite our parametrization in terms of probabilities

$$\begin{aligned}
\psi_1 &= \log \frac{p_{0\cdot\cdot}}{p_{1\cdot\cdot}}, \qquad \psi_2 = \log \frac{p_{\cdot 0\cdot}}{p_{\cdot 1\cdot}}, \qquad \psi_3 = \log \frac{p_{\cdot\cdot 0}}{p_{\cdot\cdot 1}} \\
\psi_{12} &= \log \frac{p_{00\cdot}\, p_{11\cdot}}{p_{01\cdot}\, p_{10\cdot}}, \qquad \psi_{13} = \log \frac{p_{0\cdot 0}\, p_{1\cdot 1}}{p_{0\cdot 1}\, p_{1\cdot 0}}, \\
\psi_{23} &= \log \frac{p_{\cdot 00}\, p_{\cdot 11}}{p_{\cdot 01}\, p_{\cdot 10}} + \log \frac{(p_{100} + p_{011})(p_{000} + p_{111})}{(p_{010} + p_{101})\,(p_{001} + p_{110})}, \\
\psi_{123} &= \log \frac{p_{000}\, p_{110}\, p_{101}\, p_{011}}{p_{100}\, p_{010}\, p_{001}\, p_{111}}.
\end{aligned}$$

This formulation makes it clear why it is necessary that the extra bounds given in (5.2) hold; the rather obtuse requirement that

$$q_{13} + q_2 - q_{12} - q_{23} > 0,$$

for example, is actually equivalent to $p_{010} + p_{101} > 0$, which must be the case since the probabilities are positive. This constraint is violated in Example 4 of Bergsma and Rudas (2002).

Remark 5.2.3. The parameter ψ_{23} is, unlike q_{23} and λ_{23}^{23}, not a function only of the marginal distribution over $\{2, 3\}$, and thus we are in some sense going against the spirit of the bidirected graph. This is an inevitable consequence of the requirement of variation independence: if we are allowed to parametrize each two-way margin independently, we cannot guarantee that there will exist any joint distribution with those margins; see Example 3.4.7.

5.2.1 Application to Non-Linear Systems

In general the expressions given in the inequalities (2.1) are not linear but multi-linear, and thus we cannot hope to have Fourier-Motzkin elimination provide us with variation independent parametrizations for all ADMG models. However, as the following example illustrates, in some cases it is still very useful.

Example 5.2.4. Consider the bidirected 3-chain, shown in Figure 5.1(b). The generalized Möbius parameters for this model are the same as those of the complete bidirected graph on 3 variables, with the exception that q_{13} is not required, and is assumed equal to $q_1 q_3$. This means that the functions $f_{i_V}(\boldsymbol{q})$ are not linear in q_1 and q_3.

However, the inequalities (2.1) *are* linear in q_{123}, in the sense that q_{123} is never multiplied by any of the other variables, so we can still proceed as in Example 5.2.1 to eliminate q_{123}. This gives the same set of inequalities, but q_{13} has been replaced by $q_1 q_3$:

$$\max\left\{\begin{array}{c} 0, \\ q_2+q_3-1, \\ q_1q_3+q_{12}-q_1, \\ -1+q_1+q_2+q_3-q_{12}-q_1q_3 \end{array}\right\} < q_{23} < \min\left\{\begin{array}{c} q_2, \\ q_3, \\ q_1q_3+q_2-q_{12}, \\ q_{12}+q_3-q_1q_3 \end{array}\right\},$$

$$\max\left\{\begin{array}{c} 0, \\ q_1+q_3-1 \end{array}\right\} < q_1q_3 < \min\left\{\begin{array}{c} q_1, \\ q_3 \end{array}\right\},$$

$$\max\left\{\begin{array}{c} 0, \\ q_1+q_2-1 \end{array}\right\} < q_{12} < \min\left\{\begin{array}{c} q_1, \\ q_2 \end{array}\right\}.$$

Note that the second set of inequalities could be replaced by the equivalent requirements that $0 < q_1 < 1$ and $0 < q_3 < 1$.

We can repeat the trick, and eliminate q_{23}, q_{12} and then q_2, all of which always appear without being multiplied by any other parameter. From this we can deduce that a variation independent parametrization of the bidirected 3-chain is given by ψ_1, ψ_2, ψ_{12}, ψ_3, ψ_{23} and ψ_{123}, as defined in Example 5.2.2.

There are two other ways in which we could have seen that this gives a variation independent parametrization. Firstly, the bidirected 3-chain in Figure 5.1(b) is a sub-model of the complete bidirected graph in 5.1(a), corresponding to marginal independence of 1 and 3. This marginal independence can be obtained by setting $\psi_{13} = 0$, and so a variation independent parametrization of the bidirected 3-chain is given by the remaining 6 parameters.

Alternatively, note that ψ_1, ψ_2 and ψ_{12} give a variation independent parametrization for the joint distribution of 1 and 2, and that ψ_3 parametrizes the marginal distribution of 3. It follows immediately that these four parameters are *jointly* variation independent, because given any value of them, we can choose the distribution where $1, 2 \perp\!\!\!\perp 3$ and which has the appropriate marginal distributions. This is certainly contained within the model where $1 \perp\!\!\!\perp 3$, and therefore it is impossible to select values of ψ_1, ψ_2, ψ_{12} and ψ_3 in such a way that we are outside the model.

In Dawid's notation, we have $R(\psi_1, \psi_2, \psi_{12}, \psi_3) = R(\psi_1, \psi_2, \psi_{12})R(\psi_3)$.

The following lemma formalizes the argument at the end of the preceding example.

Lemma 5.2.5. *Let $\mathcal{M}$ be a model defined by (conditional) independences over random variables indexed by V, and parametrized by the vector $(\boldsymbol{\theta}, \boldsymbol{\eta}, \boldsymbol{\zeta})$. Suppose further that the respective marginal models induced by $\mathcal{M}$ upon non-empty disjoint subsets of the random variables $A, B \subset V$ are parametrized precisely by $\boldsymbol{\eta}$ and $\boldsymbol{\zeta}$ respectively. Then $\boldsymbol{\eta}$ is variation independent of $\boldsymbol{\zeta}$, or $R(\boldsymbol{\eta}, \boldsymbol{\zeta}) = R(\boldsymbol{\eta})R(\boldsymbol{\zeta})$.*

Proof. Let $C = \{c_1, \ldots, c_k\} \equiv V \setminus (A \cup B)$. Given some choice of $\boldsymbol{\eta}$ and $\boldsymbol{\zeta}$, consider the probability distribution $\boldsymbol{p}$ which has $\perp\!\!\!\perp \{A, B, c_1, \ldots, c_k\}$ and has the marginal distributions over A and B respectively implied by $\boldsymbol{\eta}$ and $\boldsymbol{\zeta}$.

$\boldsymbol{p} \in \mathcal{M}$, because $\boldsymbol{p}$ obeys every possible (conditional) independence over V other than those within A and B which are not implied by $\boldsymbol{\eta}$ and $\boldsymbol{\zeta}$, and so there exists (a unique) $\boldsymbol{\theta}$ such that $(\boldsymbol{\theta}, \boldsymbol{\eta}, \boldsymbol{\zeta})$ represents the probability distribution $\boldsymbol{p}$. □

Lemma 5.2.6. *Let $\mathcal{G}$ be an ADMG and A be an ancestral set in $\mathcal{G}$; further, let $\mathcal{P}_{\mathcal{G}} \subseteq \Delta_{2^k-1}$ be the model defined by probability distributions which obey the global Markov property with respect to $\mathcal{G}$. Then restriction of $\mathcal{P}_{\mathcal{G}}$ to the variables in A (i.e. by marginalizing out $V \setminus A$ in the probability distributions in $\mathcal{P}_{\mathcal{G}}$) is $\mathcal{P}_{\mathcal{G}_A}$.*

Proof. First assume that $A \equiv V \setminus \{v\}$ for some vertex $v \in \text{barren}_{\mathcal{G}}(V)$. If the result holds for this set A, then by repeatedly removing barren vertices we can see that it holds for any ancestral set. Since v is barren, it is a collider on all paths which pass through it.

We claim that the m-separations concerning the vertices in A in the two graphs $\mathcal{G}$ and $\mathcal{G}_A$ are identical. Suppose that B and C are m-separated in $\mathcal{G}_A$ conditional on D, where $B, C, D \subseteq A$. Then the same m-separation holds in $\mathcal{G}$, because any additional paths between B and C must pass through v, and v is a collider on such paths which is not conditioned

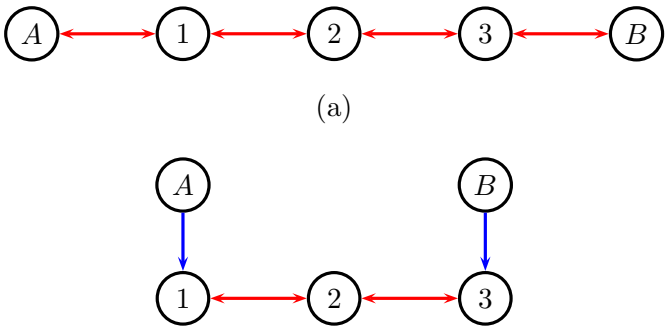


Figure 5.2: (a) The bidirected 5-chain; (b) a graph which is Markov equivalent to a bidirected 5-chain.

upon.

Conversely, assume B and C are m-separated in $\mathcal{G}$ conditional on D; because we are only removing the vertex $v \in \text{barren}_{\mathcal{G}}(V)$ which is not conditioned upon and has no descendants, all paths which are blocked in $\mathcal{G}$ must also be blocked in $\mathcal{G}_A$. Thus the m-separation also holds in $\mathcal{G}_A$.

Now suppose $\boldsymbol{p} \in \mathcal{P}_{\mathcal{G}}$; since the m-separations in the graph $\mathcal{G}_A$ are obeyed by $\boldsymbol{p}$, then the marginal distribution of $\boldsymbol{p}$ over A is in $\mathcal{P}_{\mathcal{G}_A}$.

On the other hand, if $\boldsymbol{p}_A \in \mathcal{P}_{\mathcal{G}_A}$, then the distribution over V which has marginal distribution $\boldsymbol{p}_A$ over A, marginal distribution for v as uniform, and $v \perp\!\!\!\perp A$ clearly satisfies all the conditional independences implied by the Markov property of $\mathcal{G}$; thus there is some $\boldsymbol{p} \in \mathcal{P}_{\mathcal{G}}$ such that marginalizing over $V \setminus A$ gives $\boldsymbol{p}_A$. □

5.3 Variation Independent Parametrization of the Bidirected 5-Chain

The bidirected 5-chain shown in Figure 5.2(a) is defined by the marginal independences

$$A \perp\!\!\!\perp 2, 3, B, \qquad A, 1 \perp\!\!\!\perp 3, B, \qquad A, 1, 2 \perp\!\!\!\perp B.$$

As mentioned in Chapter 3, there is no known variation independent parametrization for the discrete case of this model, and Drton and Richardson (2008a) even conjectured that no such parametrization exists. In this section we show that this conjecture is false, and construct a variation independent parametrization.

For convenience, we work with the Markov equivalent graph shown in Figure 5.2(b). First note that the induced subgraph over $\{A, 1, 2\}$ has a maximal head of size 2, and thus its ingenuous parametrization is variation independent: λ_A^A, λ_1^{A1}, λ_{A1}^{A1}, λ_2^2, λ_{12}^{A12}, λ_{A12}^{A12}. Similarly the induced subgraph over $\{B, 3\}$ has variation independent parametrization λ_B^B, λ_3^{B3}, λ_{B3}^{B3}. Further, by Lemma 5.2.6, these sets of parameters are mutually variation independent:

$$\perp_{var} \left\{\lambda_A^A,\ \lambda_1^{A1},\ \lambda_{A1}^{A1},\ \lambda_2^2,\ \lambda_{12}^{A12},\ \lambda_{A12}^{A12},\ \ \lambda_B^B,\ \lambda_3^{B3},\ \lambda_{B3}^{B3}\right\}.$$

The parametrization can be completed with the six parameters $q_{23|B}^b$ and $q_{123|AB}^{ab}$ for $a, b \in \{0, 1\}$, but this leads to variation dependence. Conditional on having chosen valid values for the first 11 parameters, the remaining four parameters ($q_{123|AB}^{ab}$) satisfy the inequalities

$$\begin{aligned}\max\left\{\begin{array}{c} 0, \\ q_{12|A}^a + q_{23|B}^b - q_2, \\ q_{12|A}^a + q_{1|A}^a q_{3|B}^b - q_{1|A}^a, \\ q_{1|A}^a q_{3|B}^b + q_{23|B}^b - q_{3|B}^b \end{array}\right\} &< q_{123|AB}^{ab} \\ &< \min\left\{\begin{array}{c} q_{12|A}^a, \\ q_{23|B}^b, \\ q_{1|A}^a q_{3|B}^b, \\ 1 - q_{1|A}^a - q_2 - q_{3|B}^b + q_{12|A}^a + q_{23|B}^b + q_{1|A}^a q_{3|B}^b \end{array}\right\},\end{aligned}$$

for $a, b \in \{0, 1\}$; variation independent parameters equivalent to $q_{123|AB}^{ab}$ given by the logarithmic barrier procedure above are just $\kappa_{123|AB}(0, 0, 0 \,|\, a, b)$ for $a, b \in \{0, 1\}$, up to a multiplicative constant. We can also use the equivalent parameters λ_{123}^{AB123}, λ_{A123}^{AB123}, λ_{B123}^{AB123} and λ_{AB123}^{AB123}, which are just an invertible linear combination of those κ-parameters.

After eliminating the parameters $q_{123|AB}^{ab}$ by Fourier-Motzkin, we obtain the following bounds

on the two parameters $q^b_{23|B}$:

$$\max\left\{\begin{array}{c} 0, \\ q_2 + q^b_{3|B} - 1, \\ q^0_{1|A} q^b_{3|B} + q^0_{12|A} - q^0_{1|A}, \\ q^0_{1|A} + q_2 + q^b_{3|B} - q^0_{12|A} - q^0_{1|A} q^b_{3|B} - 1, \\ q^1_{1|A} q^b_{3|B} + q^1_{12|A} - q^1_{1|A}, \\ q^1_{1|A} + q_2 + q^b_{3|B} - q^1_{12|A} - q^1_{1|A} q^b_{3|B} - 1 \end{array}\right\} < q^b_{23|B}$$

$$< \min\left\{\begin{array}{c} q_2, \\ q^b_{3|B}, \\ q^0_{1|A} q^b_{3|B} + q_2 - q^0_{12|A}, \\ q^0_{12|A} + q^b_{3|B} - q^0_{1|A} q^b_{3|B}, \\ q^1_{1|A} q^b_{3|B} + q_2 - q^1_{12|A}, \\ q^1_{12|A} + q^b_{3|B} - q^1_{1|A} q^b_{3|B} \end{array}\right\}$$

for $b \in \{0,1\}$, which using the logarithmic barrier gives the variation independent parameters

$$\psi^b_{23|B} \equiv \log \frac{p_{\cdot 00|b}\, p_{\cdot 11|b}}{p_{\cdot 10|b}\, p_{\cdot 01|b}} + \sum_{a=0}^{1} \log \frac{(p_{000|ab} + p_{111|ab})\,(p_{100|ab} + p_{011|ab})}{(p_{010|ab} + p_{101|ab})\,(p_{001|ab} + p_{110|ab})} \tag{5.5}$$

for $b \in \{0,1\}$, where $p_{000|ab}$ is understood to mean $P(X_1 = 0, X_2 = 0, X_3 = 0 \,|\, X_A = a, X_B = b)$, and so forth.

Thus a variation independent parametrization for the bidirected 5-chain is given by

$$\begin{array}{c} \lambda^A_A,\ \lambda^{A1}_1,\ \lambda^{A1}_{A1},\ \lambda^2_2,\ \lambda^{A12}_{12},\ \lambda^{A12}_{A12}, \qquad \lambda^B_B,\ \lambda^{B3}_3,\ \lambda^{B3}_{B3}, \\ \psi^0_{23|B},\ \psi^1_{23|B}, \qquad \lambda^{AB123}_{123},\ \lambda^{AB123}_{A123},\ \lambda^{AB123}_{B123},\ \lambda^{AB123}_{AB123}. \end{array}$$

Note that the only parameters which differ from the ingenuous parametrization of the graph in Figure 5.2(b) are $\psi^0_{23|B}$ and $\psi^1_{23|B}$; in fact, the first term in (5.5) is just a multiple of $\kappa_{23|B}(0,0\,|\,b)$, but the two additional terms are required to ensure that positivity of all conditional probabilities can be satisfied.

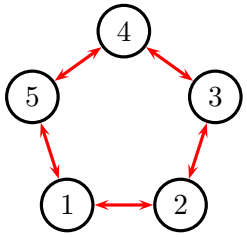
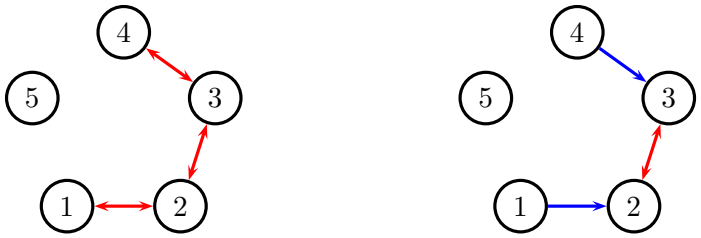


Figure 5.3: The bidirected 5-cycle.



(a) (b)

Figure 5.4: Two Markov equivalent representations of the induced sub-models for the bidirected 5-cycle over $\{1,2,3,4\}$ and $\{5\}$.

5.4 The Bidirected 5-Cycle

The bidirected 5-cycle, shown in Figure 5.3, is parametrized by the following, variation dependent, Möbius parameters:

q_1	q_2	q_3	q_4		q_5
q_{12}	q_{23}	q_{34}		q_{45}	q_{15}
q_{123}	q_{234}		q_{345}	q_{145}	q_{125}
q_{1234}		q_{2345}	q_{1345}	q_{1245}	q_{1235}
			q_{12345}.		

The subgraph induced over $\{1,2,3,4\}$ is the bidirected 4-cycle, shown in Figure 5.4(a), which is Markov equivalent to the ADMG shown in 5.4(b). By Lemma 5.2.5 we have

$$\{q_A \,|\, A \subseteq \{1,2,3,4\}\} \perp_{var} q_5,$$

and using the ingenuous parameters for subgraph over $\{1,2,3,4\}$ in Figure 5.4(b), we may replace $\underline{\mathscr{Q}} \equiv \{q_A \,|\, A \subseteq \{1,2,3,4\}\} \cup \{q_5\}$ with the parameters

λ_1^1	λ_2^{12}	λ_3^{34}	λ_4^4	λ_5^5
λ_{12}^{12}	λ_{23}^{1234}	λ_{34}^{34}		

λ_{123}^{1234} λ_{234}^{1234}
λ_{1234}^{1234},

which are jointly variation independent.

On the other hand, the parameters

$$\overline{\mathscr{Q}} \equiv \{q_{345}, q_{145}, q_{125}, q_{2345}, q_{1345}, q_{1245}, q_{1235}, q_{12345}\}$$

all appear in the expressions for $f_{iv}(\boldsymbol{q})$ without being multiplied by any other parameter, and thus we can use the Fourier-Motzkin approach to find variation independent versions of these, conditional upon the values of the other parameters.

This leaves just two parameters q_{15} and q_{45}, which require special treatment. Using the package `rcdd` of the statistical software `R`, Fourier-Motzkin elimination can be performed automatically (see Fukuda, 2000; Geyer and Meeden, 2008); after elimination of the parameters in $\overline{\mathscr{Q}}$, the variation dependence of the remaining 13 parameters is defined by 392 inequalities. Of these, 24 inequalities do not involve q_{15} or q_{45}, and are therefore satisfied by the parameters $\underline{\mathscr{Q}}$, which we choose to be valid. This leaves us with 368 inequalities of the form

$$a_i q_{15} + b_i q_{45} + c_i > 0, \quad i = 1, \ldots, 368,$$

where a_i, b_i and c_i are infinitely differentiable (indeed linear) functions of the first 11 parameters $\underline{\mathscr{Q}}$. This collection of half spaces defines a non-empty convex polygon. There are 72 distinct pairs of functions (a_i, b_i), so the polygon may have at most 72 sides. Simulations suggest that in practice it has far fewer.

To complete the smooth and variation independent parametrization of the 5-cycle, we therefore need to provide a smooth bijective map between the interior of this polygon and the interior of a rectangle, or $\mathbb{R}^2$.

Lemma 5.4.1. *Let $a_1, \ldots, a_k \in \mathbb{R}^n$ and $b_1, \ldots, b_k \in \mathbb{R}$ be constant and chosen such that*

$$C \equiv \left\{x \in \mathbb{R}^n \,\middle|\, a_i^T x + b_i > 0,\ i = 1, \ldots, k\right\}$$

is non-empty and bounded (i.e. it is an n-dimensional polytope). Then the function

$$f(x) \equiv \prod_{i=1}^{k} (a_i^T x + b_i)$$

is unimodal on C.

Proof. First, note that f is an infinitely differentiable function of x, it is zero on the boundary of C, and that is strictly positive on the interior of C; thus f has at least one local maximum in C. Now suppose for contradiction that x_1 and x_2 are two distinct modes in C, and consider the unique line l which passes through x_1 and x_2. This consists of points of the form $x_1 + z(x_2 - x_1)$ for some scalar z. Let

$$f_l(z) \equiv f\left(x_1 + z(x_2 - x_1)\right) = \prod_{i=1}^{k} \left\{a_i^T \left(x_1 + z(x_2 - x_1)\right) + b_i\right\}$$

which is just a product of linear expressions for z, and let $C_l = \{z \,|\, x_1 + z(x_2 - x_1) \in C\}$. Then f_l is a kth degree polynomial in z, which has its k (possibly repeated) roots at

$$z = \frac{a_i^T x_1 + b_i}{a_i^T (x_1 - x_2)}, \quad \text{for } i = 1, \ldots, k.$$

In any kth degree polynomial with k (possibly repeated) real roots, there can be at most one turning point strictly between two adjacent roots. Since none of the roots of f_l are on the interior of C_l, there is at most one turning point of f_l on C_l; however, this is a contradiction, since both x_1 and x_2 were assumed to be maxima of f_l. □

Lemma 5.4.2. *Consider the situation described in the previous lemma, except now let $a_1, \ldots, a_k$ and $b_1, \ldots, b_k$ be polynomial functions of some vector variable y, such that $C(y)$ is non-empty and bounded for every value of y. Then the mode $m(y)$ of f_y, is an infinitely differentiable function of y.*

Proof. We can write

$$f_y(x) = g(x, y)$$

where g is a polynomial function. The mode function is defined implicitly by

$$h(m(y), y) = 0, \tag{5.6}$$

where $h = \frac{\partial g}{\partial x}$. Now, h is a C^∞ function, and by Lemma 5.4.1 there is a solution to (5.6) in $C(y_0)$ for each fixed y_0. Further, the Jacobian matrix $\frac{\partial h}{\partial x}$ is invertible, indeed negative definite, at the maximum of the polynomial function $f_y(x)$. Thus, by the implicit function theorem, the function $m(y)$ is itself a C^∞ function on a neighbourhood of y_0. □

With f being the unimodal function on the polytope for q_{15} and q_{45} defined in Lemma 5.4.1, we can define a radius function

$$r(q_{15}, q_{45}) = \frac{f(m) - f(q_{15}, q_{45})}{f(q_{15}, q_{45})},$$

which maps each line segment from the mode m to the boundary of C onto $[0, \infty)$. Together with the angle ϕ of the line segment relative to some fixed direction, we can map $C(y)$ onto $\mathbb{R}^2$ using polar co-ordinates. Composing this map with reversion to Cartesian co-ordinates for $\mathbb{R}^2$ provides smooth variation independent parameters equivalent to q_{15} and q_{45}.

5.5 The General Case

In this final section we generalize the approach taken for the bidirected 5-cycle to any ADMG model.

Theorem 5.5.1. *Every ADMG model admits a smooth variation independent parametrization.*

Proof. Let $\mathcal{G}$ be an ADMG with vertex set V; we proceed by induction on the number of vertices. If $|V| < 2$ then the ingenuous parametrization is smooth and variation independent; otherwise let $|V| = n$ and assume that the result holds for ADMGs with fewer than n vertices.

Let $v \in \text{barren}_{\mathcal{G}}(V)$; the sub-model induced over $V \setminus \{v\}$ is just the model associated with the subgraph $\mathcal{G}_{V\setminus\{v\}}$. The generalized Möbius parameters can be divided into two:

$$\mathscr{Q}'(\mathcal{G}) = \mathscr{Q}'(\mathcal{G}_{V\setminus\{v\}}) \cup \mathscr{Q}'_v(\mathcal{G})$$

where $\mathscr{Q}'_v(\mathcal{G}) \equiv \{q^{i_T}_{H|T} \,|\, v \in H \in \mathcal{H}(\mathcal{G})\}$. We denote a vector of these parameters by θ^v, and the vector of parameters in $\mathscr{Q}'(\mathcal{G}_{V\setminus\{v\}})$ by η^{-v}. Note that v does not appear in any tail sets, because it has no descendants. By the induction hypothesis, we have a smooth variation independent version of $\mathscr{Q}'(\mathcal{G}_{V\setminus\{v\}})$. To parametrize the complete model, we need a smooth variation independent version of $\mathscr{Q}'_v(\mathcal{G})$, conditional on the earlier values.

Since any two parameters in $\mathscr{Q}'_v(\mathcal{G})$ both contain v in their heads, they are never multiplied together in a term of the form (2.2). This means that, conditional on the parameters in $\mathscr{Q}'(\mathcal{G}_{V\setminus\{v\}})$, the expressions (2.1) are linear in θ^v. Thus, the inequalities define a series of half spaces, whose intersection we denote by Θ_v; this intersection is non-empty by Lemma 5.2.5.

It is also bounded, because each parameter is a conditional probability, and is therefore between 0 and 1. Thus Θ_v is a non-empty convex polytope, of the form

$$\left\{\theta^v \,\middle|\, a_i^T(\eta^{-v})\theta^v + b_i(\eta^{-v}) > 0,\ i = 1, \ldots, k\right\},$$

where $a_1, \ldots, a_k$ and $b_1, \ldots, b_k$ are smooth functions of η^{-v}.

Lemma 5.4.1 states that the function

$$f(\theta^v) \equiv \prod_{i=1}^{k} \left\{a_i^T(\eta^{-v})\theta^v + b_i(\eta^{-v})\right\}$$

is smooth, positive and unimodal on Θ^v, and zero on its boundary. By Lemma 5.4.2, the mode $\theta^{v*} = m(\eta^{-v})$, and therefore the value of f evaluated at that mode, are smooth functions of η^{-v}. Then the function

$$r(\theta^v) \equiv \frac{f(\theta^{v*}) - f(\theta^v)}{f(\theta^v)}$$

is also smooth on the interior of Θ_v; further $r(\theta^{v*}) = 0$, and r strictly increases as θ^v moves in a straight line away from θ^{v*}, approaching ∞ at the boundary of Θ_v.

We can give a smooth bijective map η^v of this polytope onto $\mathbb{R}^p$ simply by using hyperspherical polar co-ordinates: $r(\theta^v)$ is the radius of θ^v under the mapping, and since Θ_v is convex we can simply preserve the angle of any point θ^v relative to θ^{v*}, with respect to some chosen axes. The set Θ_v under this mapping provides a variation independent parametrization of θ^v. □

We have shown that, in the model defined by a graph $\mathcal{G}$, conditional on the joint distribution of $\{v_1, \ldots, v_{k-1}\}$, the valid space of generalized Möbius parameters which involve v_k forms a convex polytope. It should be remarked that, *a priori*, there is no reason why this should hold, and indeed it does not hold for arbitrary collections of generalized Möbius parameters.

Remark 5.5.2. Clearly we can translate our new parametrization in such a way that any interior point of Θ_v is mapped to the origin. In particular, we can choose it so that $\eta^v = 0$ represents complete independence of v from the vertices $V \setminus \{v\}$. We may also rotate the axes in any way, so that we could fix the vanishing of any component of the vector η^v to imply some context specific conditional independence:

$$q_{H|T}^{i_T} = P(X_v = 0 \,|\, X_T = i_T) \cdot P(X_{H\setminus\{v\}} = 0 \,|\, X_T = i_T).$$

In the case of the bidirected 5-cycle, for example, one could ensure that $\eta_1 = 0$ if and only if $q_{15} = q_1 q_5$, and $\eta_2 = 0$ if and only if $q_{45} = q_4 q_5$.

Index of Notation

Index of Concepts

Bibliography

D. H. Ackley, G. E. Hinton, and T. J. Sejnowski. A learning algorithm for Boltzmann machines. *Cognitive Science*, 9(1):147–169, 1985.

J. Aitchison and S. D. Silvey. Maximum-likelihood estimation of parameters subject to restraints. *The Annals of Mathematical Statistics*, 29(3):813–828, 1958.

R. A. Ali, T. S. Richardson, P. Spirtes, and J. Zhang. Towards characterizing Markov equivalence classes for directed acyclic graphs with latent variables. In *Proceedings of the 21st Conference on Uncertainty in Artificial Intelligence (UAI-05)*, pages 10–17, 2005.

F. Bartolucci, R. Colombi, and A. Forcina. An extended class of marginal link functions for modelling contingency tables by equality and inequality constraints. *Statistica Sinica*, 17:691–711, 2007.

W. P. Bergsma and T. Rudas. Marginal models for categorical data. *Annals of Statistics*, 30(1):140–159, 2002.

D. P. Bertsekas. *Nonlinear Programming*. Athena Scientific, second edition, 1999.

J. Besag. Spatial interaction and the statistical analysis of lattice systems. *Journal of the Royal Statistical Society, Series B*, 36(2):192–236, 1974.

I. Csiszár. i-divergence geometry of probability distributions and minimization problems. *Annals of Probability*, 3(1):146–158, 1975.

G. B. Dantzig and B. C. Eaves. Fourier-Motzkin elimination and its dual. *Journal of Combinatorial Theory, Series A*, 14(3):288–297, 1973.

J. N. Darroch, S. L. Lauritzen, and T. P. Speed. Markov fields and log-linear models for contingency tables. *Annals of Statistics*, 8:522–539, 1980.

A. P. Dawid. Separoids: A mathematical framework for conditional independence and irrelevance. *Annals of Mathematics and Artificial Intelligence*, 32(1):335–372, 2001.

A. P. Dawid. Conditional independence in statistical theory (with discussion). *Journal of the Royal Statistical Society, Series B*, 41:1–31, 1979.

A. P. Dawid and S. L. Lauritzen. Hyper Markov laws in the statistical analysis of decomposable graphical models. *The Annals of Statistics*, 21(3):1272–1317, 1993.

L. L. Dines. Systems of linear inequalities. *Annals of Mathematics*, 20(3):191–199, 1919.

A. Dobra and S. E. Fienberg. Bounds for cell entries in contingency tables given marginal totals and decomposable graphs. *Proceedings of the National Academy of Sciences of the United States of America*, 97(22):11885–11892, 2000.

M. Drton. Iterative conditional fitting for discrete chain graph models. In *Proceedings in Computational Statistics*, pages 93–104, 2008.

M. Drton. Discrete chain graph models. *Bernoulli*, 15(3):736–753, 2009.

M. Drton and M. Eichler. Maximum likelihood estimation in Gaussian chain graph models under the alternative Markov property. *Scandinavian Journal of Statistics*, 33(2):247–257, 2006.

M. Drton and T. S. Richardson. Binary models for marginal independence. *Journal of the Royal Statistical Society, Series B*, 70(2):287–309, 2008a.

M. Drton and T. S. Richardson. Graphical methods for efficient likelihood inference in Gaussian covariance models. *Journal of Machine Learning Research*, 9:893–914, 2008b.

M. Drton, R. Foygel, and S. Sullivant. Global identifiability of linear structural equation models. *Annals of Statistics*, 39(2):865–886, 2011.

R. J. Evans and T. S. Richardson. Maximum likelihood fitting of acyclic directed mixed graphs to binary data. In *Proceedings of the 26th conference on Uncertainty in Artificial Intelligence (UAI-08)*, pages 177–184, 2010.

J. Fan and R. Li. Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, 96(456):1348–1360, 2001.

T. S. Ferguson. *A course in large sample theory*, volume 38. Chapman & Hall/CRC, 1996.

C. A. Floudas and P. M. Pardalos. *Encyclopedia of optimization*. Kluwer Academic, 2001.

A. Forcina, M. Lupparelli, and G. M. Marchetti. Marginal parameterizations of discrete models defined by a set of conditional independencies. *Journal of Multivariate Analysis*, 2010.

L. B. J. Fourier. Reported in: Analyse des travaux de l'academie royale des sciences, pendant l'anne 1824. In *History de l'Academie Royale de Sciences de l'Institute de France*, volume 7, pages xlvii–lv, 1824.

K. Fukuda. *CDD program package*, 2000. Computer program library, see `http://www.ifor.math.ethz.ch/~fukuda/cdd_home/index.html`.

D. Geiger, D. Heckerman, H. King, and C. Meek. Stratified exponential families: graphical models and model selection. *Annals of Statistics*, 29(2):505–529, 2001.

C. J. Geyer. On the asymptotics of constrained m-estimation. *Annals of Statistics*, 22(4): 1993–2010, 1994.

C. J. Geyer. On the asymptotics of convex stochastic optimization. Unpublished manuscript, 1996.

C. J. Geyer and G. D. Meeden. *R package* `rcdd`, 2008. Version 1.1, `http://www.stat.umn.edu/geyer/rcdd`.

G. F. V. Glonek and P. McCullagh. Multivariate logistic models. Technical Report 94-31, School of Information Science and Technology, Flinders University of South Australia, Adelaide, 1994.

G. F. V. Glonek and P. McCullagh. Multivariate logistic models. *Journal of the Royal Statistical Society, Series B*, 57(3):533–546, 1995.

J. C. Huang and B. C. Frey. Cumulative distribution networks and the derivative-sum-product algorithm. In *Proceedings of the 24th conference on Uncertainty in Artificial Intelligence*, pages 290–297, 2008.

S. Johansen. *Introduction to the theory of regular exponential families*. Institute of Mathematical Statistics, University of Copenhagen, Copenhagen, 1979.

R. E. Kass and P. W. Vos. *Geometrical foundations of asymptotic inference*. Wiley New York, 1997.

G. Kauermann. A note on multivariate logistic models for contingency tables. *Australian Journal of Statistics*, 39(3):261–276, 1997.

K. Knight and W. Fu. Asymptotics for lasso-type estimators. *Annals of Statistics*, 28(5): 1356–1378, 2000.

T. W. Körner. *A companion to analysis: a second first and first second course in analysis*. American Mathematical Society, 2004.

S. G. Krantz and H. R. Parks. *The implicit function theorem*. Birkhäuser, 2003.

J. B. Lang. Maximum likelihood methods for a generalized class of log-linear models. *Annals of Statistics*, pages 726–752, 1996.

S. L. Lauritzen. Lectures on contingency tables. Unpublished lecture notes, electronic edition, 2002.

S. L. Lauritzen. *Graphical Models*. Clarendon Press, 1996.

S. I. Lee, V. Ganapathi, and D. Koller. Efficient structure learning of Markov networks using L_1-regularization. In *Advances in Neural Information Processing Systems*, pages 817–824, 2007.

E. L. Lehmann and G. Casella. *Theory of point estimation*. Springer Verlag, second edition, 1998.

M. Lupparelli. *Graphical models of marginal independence for categorical variables*. PhD thesis, University of Florence, 2006.

M. Lupparelli, G. M. Marchetti, and W. P. Bergsma. Parameterizations and fitting of bi-directed graph models to categorical data. *Scandinavian Journal of Statistics*, 36(3): 559–576, 2009.

G. M. Marchetti and M. Lupparelli. Chain graph models of multivariate regression type for categorical data. arXiv:0906.2098, 2010.

N. Meinshausen and P. Bühlmann. Consistent neighborhood selection for sparse high-dimensional graphs with the lasso. *Statistics Surveys*, 2:61–93, 2004.

N. Meinshausen and P. Bühlmann. High-dimensional graphs and variable selection with the lasso. *Annals of Statistics*, 34(3):1436–1462, 2006.

T. Motzkin. *Beitrağe zur Theorie der linearen Ungleichunge*. PhD thesis, University of Basel, 1936.

Y. Nardi and A. Rinaldo. The log-linear group lasso estimator and its asymptotic properties. arXiv:0709.3526, 2007.

J. Nocedal and S. J. Wright. *Numerical optimization*. Springer, 1999.

J. Pearl. *Probabilistic Reasoning in Intelligent Systems*. Morgan Kaufmann, 1988.

B. F. Qaqish and A. Ivanova. Multivariate logistic models. *Biometrika*, 93(4):1011–1017, 2006.

R Development Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2011. URL `http://www.R-project.org`.

T. S. Richardson. Markov properties for acyclic directed mixed graphs. *Scandinavian Journal of Statistics*, 30(1):145–157, 2003.

T. S. Richardson. A factorization criterion for acyclic directed mixed graphs. In *Proceedings of the 25th conference on Uncertainty in Artificial Intelligence*, pages 462–470, 2009.

T. S. Richardson and P. Spirtes. Ancestral graph Markov models. *Annals of Statistics*, 30 (4):962–1030, 2002.

T. S. Richardson, R. J. Evans, and J. M. Robins. Transparent parametrizations of models for potential outcomes. In *Bayesian Statistics 9*, pages 569–610. Oxford University Press, 2010.

T. Rudas, W. P. Bergsma, and R. Németh. Parameterization and estimation of path models for categorical data. In *Proceedings in Computational Statistics, 17th Symposium*, pages 383–394. Physica-Verlag HD, 2006.

T. Rudas, W. P. Bergsma, and R. Németh. Marginal log-linear parameterization of conditional independence models. *Biometrika*, 97(4):1006–1012, 2010.

I. Shpitser, T. S. Richardson, J. M. Robins, and R. J. Evans. Parameter and structure learning in mixed graph models of post-truncation independence. Draft, 2011.

N. Städler, P. Bühlmann, and S. van de Geer. l_1-penalization for mixture regression models. *TEST*, 19(2):209–256, 2010.

R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Series B*, 58(1):267–288, 1996.

P. Tseng. Convergence of a block coordinate descent method for nondifferentiable minimization. *Journal of optimization theory and applications*, 109(3):475–494, 2001.

S. Van De Geer, P. Buhlmann, and S. Zhou. The adaptive and the thresholded lasso for potentially misspecified models. *Arxiv preprint arXiv:1001.5176*, 2010.

A. W. van der Vaart and J. A. Wellner. *Weak convergence and empirical processes*. Springer Verlag, first edition, 1996.

J. Whittaker. *Graphical models in applied multivariate statistics*. Wiley, 1990.

H. Zou. The adaptive lasso and its oracle properties. *Journal of the American Statistical Association*, 101(476):1418–1429, 2006.

Appendix A

Extensions to Euphonious Graphs

This appendix provides some guidance concerning how the work is this paper may be extended from Acyclic Directed Mixed Graphs (ADMGs) to Mixed Euphonious Graphs (MEGs), which contain undirected edges.

A.1 Basic Definitions

Recall that $\mathcal{G}_-$ is the undirected skeleton of $\mathcal{G}$. The first definition gives relational terms for vertices joined by undirected edges (see also Definition 1.1.3).

Definition A.1.1. Let $\mathcal{G}$ be a mixed graph. If two vertices v and w of $\mathcal{G}$ are joined by an undirected edge $v - w$, we say that w is a *neighbour* of v. The set of neighbours of v is denoted by $\mathrm{ne}_{\mathcal{G}}(v)$, and

$$\mathrm{nhd}_{\mathcal{G}}(v) \equiv \{w \mid w - \cdots - v \text{ in } \mathcal{G} \text{ or } w = v\}$$

is the *neighbourhood* of v. A neighbourhood of the graph $\mathcal{G}$ is a maximal connected set in $\mathcal{G}_-$.

Definition A.1.2. In a similar spirit to the set of ancestors, we define the *anterior* of a vertex v to be the set of vertices w such that there is a path from w to v whose edges are all either undirected, or directed and point towards v. We denote this set as $\mathrm{ant}_{\mathcal{G}}(v)$, and note that $v \in \mathrm{ant}_{\mathcal{G}}(v)$. This is applied disjunctively as in Definition 1.1.3.

An *anterior set*, A, is one such that $\mathrm{ant}_{\mathcal{G}}(A) = A$.

Thus, for example, in the graph in Figure A.2, the anterior of the vertex 5 is the set $\{1, 2, 3, 5\}$.

Definition 1.1.7. *A mixed graph $\mathcal{G}$ is said to be* euphonious *if it is acyclic and for every vertex $v \in \mathcal{G}$ we have $\mathrm{ne}_{\mathcal{G}}(v) \neq \emptyset \Rightarrow \mathrm{pa}_{\mathcal{G}}(v) \cup \mathrm{sp}_{\mathcal{G}}(v) = \emptyset$. In other words, if there is an undirected edge incident to v, then there must be no arrowheads incident to v. We write MEG for mixed euphonious graph.*

As a consequence of this definition we can define the *undirected component* of a euphonious graph to be the set of points with no incident arrowheads, i.e.

$$\mathrm{un}_{\mathcal{G}} \equiv \{v \mid \mathrm{pa}_{\mathcal{G}}(v) \cup \mathrm{sp}_{\mathcal{G}}(v) = \emptyset\}.$$

Then the subgraph induced by $\mathrm{un}_{\mathcal{G}}$ contains all undirected edges in $\mathcal{G}$. The *directed component* consists of all other vertices:

$$\mathrm{dir}_{\mathcal{G}} \equiv V \setminus \mathrm{un}_{\mathcal{G}}\,.$$

The definitions of m-separation and the global Markov property in Section 1.2 apply equally to ADMGs and euphonious graphs.

Let $\mathcal{G}$ be a mixed euphonious graph. A collection of vertices $A \subseteq \mathrm{un}_{\mathcal{G}}$ is *complete* if the induced subgraph $\mathcal{G}_A$ has no missing edges. The collection of all such sets in $\mathrm{un}_{\mathcal{G}}$ is denoted $\mathcal{C}(\mathcal{G})$. A set which is maximal in $\mathcal{C}(\mathcal{G})$ with respect to inclusion is called a *clique*; the collection of cliques in $\mathcal{G}$ is denoted $\overline{\mathcal{C}}(\mathcal{G})$.

It is well known that a strictly positive multivariate binary distribution P obeying the global Markov property with respect to an undirected graph $\mathcal{G}$ is parametrized by

$$\mathscr{Q}(\mathcal{G}) = \{q_C \mid C \in \mathcal{C}(\mathcal{G})\},$$

(Lauritzen, 1996). Note that the number of parameters in a bidirected graph is potentially much larger than for an undirected graph with the same skeleton.

The partition of $\mathcal{G}$ into an undirected and a directed component means that the definition of a head must be slightly modified, so as to ensure that it is contained in $\mathrm{dir}_{\mathcal{G}}$. The definition of a tail set is unaltered.

Definition A.1.3. Let $\mathcal{G}$ be a MEG; a subset of vertices $H \subseteq \mathrm{dir}_{\mathcal{G}}$ is a *head* if it is barren in $\mathcal{G}$ and is a $\leftrightarrow$-path-connected subset of $\mathcal{G}_{\mathrm{an}(H)}$.

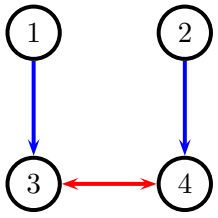


Figure A.1: An acyclic directed mixed graph.

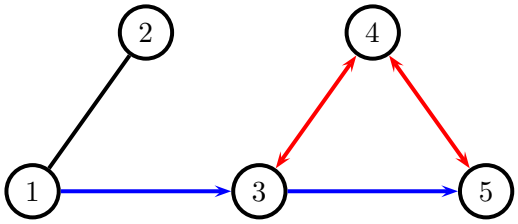


Figure A.2: A mixed euphonious graph $\mathcal{G}_3$.

Recall that for an ADMG $\mathcal{G}$, the set of distributions obeying the global Markov property with respect to $\mathcal{G}$ is parametrized by

$$\mathscr{Q}'(\mathcal{G}) \equiv \{q_{H|T}^{i_T} \mid H \in \mathcal{H}(\mathcal{G}), i_T \in \{0,1\}^{|T|}\}.$$

To extend to MEGs we simply utilize the cut in the parameter space between $\text{un}_{\mathcal{G}}$ and $\text{dir}_{\mathcal{G}}$. We parametrize $\text{un}_{\mathcal{G}}$ as an undirected graph using the Möbius parameters of its complete subsets, and then $\text{dir}_{\mathcal{G}}$ conditionally on $\text{un}_{\mathcal{G}}$. The parameters are

$$\mathscr{Q}'(\mathcal{G}) \equiv \{q_{H|T}^{i_T} \mid H \in \mathcal{H}(\mathcal{G}), i_T \in \{0,1\}^{|T|}\} \cup \{q_C \mid C \in \mathcal{C}(\mathcal{G})\}.$$

Remark A.1.4. Consider the ADMG in Figure A.1. Under the original definition, $\{1\}$ and $\{2\}$ are considered heads, each with an empty tail, leading to the generalized Möbius parameters q_1 and q_2 being included in the parametrization. However, when the graph is considered as a MEG, $\{1,2\} \subseteq \text{un}_{\mathcal{G}}$, because no arrowheads are incident to either vertex.

This does not affect the parametrization, however, because $\{1\}$ and $\{2\}$ are each complete subsets in $\text{un}_{\mathcal{G}}$, and so the parameters q_1 and q_2 are included. The change of definition of a head is therefore merely a technicality.

Example A.1.5. The MEG $\mathcal{G}_3$ in Figure A.2 has the complete subsets $\{1\}$, $\{2\}$ and $\{1,2\}$, and the following head-tail pairs:

H	$\{3\}$	$\{4\}$	$\{3,4\}$	$\{5\}$	$\{4,5\}$
T	$\{1\}$	$\emptyset$	$\{1\}$	$\{3\}$	$\{1,3\}$

.

The only $\leftrightarrow$-path-connected set in $\mathrm{dir}_{\mathcal{G}}$ which is not a head is $\{3,4,5\}$, which is not barren because 3 is a parent of 5.

The parametrization from (1.3) is then

$$P\left(X_V = i_V\right) = P\left(X_{\mathrm{un}_{\mathcal{G}}} = i_{\mathrm{un}_{\mathcal{G}}}\right) \cdot \sum_{C : O \subseteq C \subseteq \mathrm{dir}_{\mathcal{G}}} (-1)^{|C \setminus O|} \prod_{H \in [C]_{\mathcal{G}}} q_{H|T}^{i_T},$$

where $O = \{v \in \mathrm{dir}_{\mathcal{G}} \mid i_v = 0\}$.

The undirected part of a MEG may be handled using standard techniques (Lauritzen, 1996), leaving the directed part of the graph to be fitted using the methods introduced in Chapter 2.

A.2 Marginal Log-Linear Parameters

As noted in Chapter 3, an undirected graph $\mathcal{G}$ can be parametrized with ordinary log-linear parameters

$$\{\lambda_L^V \mid L \in \mathcal{C}(\mathcal{G})\}.$$

To extend the ingenuous parametrization to a MEG we use the margins N_i, where N_i are the neighbourhoods of $\mathrm{un}_{\mathcal{G}}$, and the effects

$$\mathbb{L}_i = \{L \subseteq N_i \mid L \in \mathcal{C}(\mathcal{G})\}.$$

We keep the usual margins in $\mathrm{dir}_{\mathcal{G}}$ of the form $M_i = H_i \cup T_i$ for heads H_i. If $M_1, \ldots, M_k$ is a hierarchical ordering on these margins, then any ordering in which the neighbourhoods N_j are all placed before the M_i is also hierarchical.

Example A.2.1. For the graph in Figure A.2, recall that we have one neighbourhood $\{1,2\}$, and the head-tail pairs

H	$\{3\}$	$\{4\}$	$\{3,4\}$	$\{5\}$	$\{4,5\}$
T	$\{1\}$	$\emptyset$	$\{1\}$	$\{3\}$	$\{1,3\}$

.

Thus the ingenuous parametrization consists of

M_i	$\mathbb{L}_i$
$\{1,2\}$	$\{1\},\{2\},\{1,2\}$
$\{1,3\}$	$\{3\},\{1,3\}$
$\{4\}$	$\{4\}$
$\{1,3,4\}$	$\{3,4\},\{1,3,4\}$
$\{3,5\}$	$\{5\},\{3,5\}$
$\{1,3,4,5\}$	$\{4,5\},\{3,4,5\},\{1,4,5\},\{1,3,4,5\}$.

A.2.1 Completions

Any form of graphical completion for a MEG must proceed as follows: first complete $\mathrm{un}_{\mathcal{G}}$ so that there is an undirected edge between every pair of vertices. Then for every pair of vertices $v \in \mathrm{un}_{\mathcal{G}}$ and $w \in \mathrm{dir}_{\mathcal{G}}$, add in the edge $v \to w$ if it is not already present. The directed part of the graph can be completed according to, some head-preserving completion.

The ingenuous parametrization of a MEG is soundly completed by the ingenuous parametrization for the completion $\bar{\mathcal{G}}$. To see this, note that for the undirected component, the new parameters are $\lambda_L^{\mathrm{un}_{\mathcal{G}}}$; if L is complete in $\mathcal{G}$, then this parameter is equal to $\lambda_L^{N_i}$, where N_i is the neighbourhood of L. If L is not complete in $\mathcal{G}$, then under the Markov property we have $\lambda_L^{\mathrm{un}_{\mathcal{G}}} = 0$.

An extension of Lemma 3.3.3 is easily proved by noting that $\mathrm{tail}_{\mathcal{G}}(H)$ of a head H is the Markov blanket for H in $\mathrm{ant}_{\mathcal{G}}(H)$.

Theorems 3.4.5 and 5.5.1 apply to MEGs just as to ADMGs, because the parametrization of a undirected graph by ordinary log-linear parameters is always variation independent (Bergsma and Rudas, 2002).

Vita

Robin James Evans was born in Chester, UK, where he lived for most of his life. He read mathematics at the University of Cambridge, earning a Bachelor of Arts degree in 2006, and Master of Mathematics in 2007. His masters thesis (officially 'Part III Essay') was supervised by Richard Samworth. He taught at Westminster School in London before joining the Statistics Department at the University of Washington in September 2008. In August 2011 he became a Doctor of Philosophy.