Marginal log-linear parameters for graphical Markov models

Robin J. Evans
Department of Statistics
University of Washington
rje42@stat.washington.edu

Thomas S. Richardson
Department of Statistics
University of Washington
tsr@stat.washington.edu

May 31, 2011

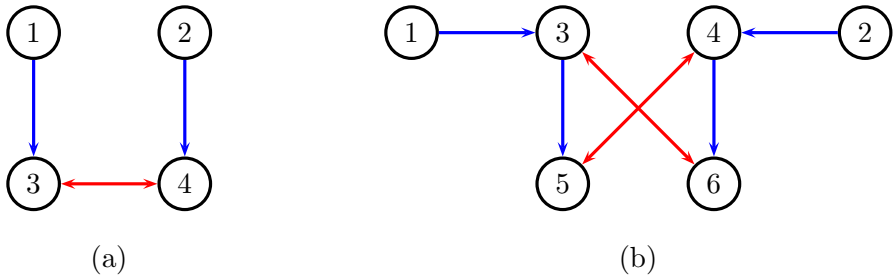
Abstract

The parametrization of multivariate discrete statistical models by marginal log-linear (MLL) parameters provides a great deal of flexibility; in particular, different MLL parametrizations under linear constraints induce various sub-models, including models defined by some collections of conditional independences. Such models are curved exponential families, and therefore have regular asymptotic properties. We introduce a sub-class of MLL models which correspond to Acyclic Directed Mixed Graphs under the usual global Markov property. We characterize for precisely which graphs the resulting parametrization is variation independent, and show how it is both intuitive, and easily adapted to sparse modelling techniques.

1 Introduction

Models defined by conditional independence constraints are central to many methods in multivariate statistics, and in particular to graphical models (Darroch et al., 1980; Whittaker, 1990). In the case of discrete data, *marginal log-linear* (MLL) parameters can be used to parametrize a broad range of models, including some graphical classes and models for conditional independence (Rudas et al., 2010; Forcina et al., 2010). The parameters are defined by considering any sequence, $M_1, M_2, \ldots$, of margins of the distribution which respects inclusion (i.e. M_i precedes M_j if $M_i \subset M_j$), with each such sequence giving rise to a different smooth parametrization of the saturated model. Useful sub-models can be induced by setting some of the parameters to zero, or more generally by restricting attention to a linear subspace of the parameter space.

The amount of flexibility in these models requires some restriction in order to lead to a tractable search space. We describe a sub-class of marginal log-linear models suitable in the context of directed acyclic graphs (DAGs) with hidden variables; these correspond to a class of graphs known as *acyclic directed mixed graphs* (ADMGs). ADMGs contain directed ($\rightarrow$) and bidirected ($\leftrightarrow$) edges, subject to the constraint that there are no directed cycles. Two examples are shown in Figure 1. All the work herein can easily be extended to graphs which also contain an undirected component, provided no undirected edge is adjacent to an arrowhead. This latter case strictly includes all ancestral graphs (Richardson and Spirtes, 2002).



(a) (b)

Figure 1: Two acyclic directed mixed graphs.

Markov properties for ADMGs and a parametrization in the case of discrete random variables were provided by Richardson (2003, 2009). However, that parametrization is subject to variation dependence constraints, in the sense that setting some parameters to particular values may restrict the valid range of other parameters; this makes maximum likelihood fitting, for example, more challenging (Evans and Richardson, 2010). Further, though it parametrizes all distributions satisfying the Markov model, it does not naturally lead to parsimonious submodels.

Taking the graph in Figure 1(a) as an example, the parametrization of Richardson consists (in the binary case) of the probabilities

$$\begin{aligned} &P(X_1 = 0) && P(X_3 = 0 \mid X_1 = x_1) \\ &P(X_2 = 0) && P(X_4 = 0 \mid X_2 = x_2) \\ &P(X_3 = 0, X_4 = 0 \mid X_1 = x_1, X_2 = x_2), \end{aligned}$$

where $x_1, x_2 \in \{0, 1\}$. A disadvantage of this parametrization is that the joint probability of X_3 and X_4 both being zero (conditional on $X_1 = x_1$ and $X_2 = x_2$) is bounded above by the marginal probability of those events. Consequently, from the point of view of parameter interpretation, it makes little sense to consider the joint probabilities in isolation. For example, strong correlation between X_3 and X_4 is present when the joint probability is large relative to the marginals.

If we use the conditional odds ratios

$$\frac{P(X_3 = 0, X_4 = 0 \mid X_1 = x_1, X_2 = x_2) \cdot P(X_3 = 1, X_4 = 1 \mid X_1 = x_1, X_2 = x_2)}{P(X_3 = 1, X_4 = 0 \mid X_1 = x_1, X_2 = x_2) \cdot P(X_3 = 0, X_4 = 1 \mid X_1 = x_1, X_2 = x_2)}$$

instead of the joint probabilities, then the variation dependence disappears. The odds ratio is a measure of correlation without reference to the marginal distributions. This means that if, for example, we wish to define a prior distribution over the univariate probabilities and the odds ratios, we can simply use a product of univariate distributions; similarly, to fit a generalized linear model with the parameters as responses, we only need simple univariate link functions. We will see that this approach to discrete parametrizations can be generalized.

In Section 2 we introduce marginal log-linear parameters and some of their properties. Section 3 gives background theory about ADMGs and the parametrization of Richardson (2009); the application of MLLs to parametrizing these models is presented in Section 4. In Section 5 we relate this to the framework of Bergsma and Rudas (2002), and classify for which models the new parametrization is variation independent. Section 6 gives a slightly different formulation of the parametrization, and demonstrates its advantages for parameter

interpretation. Section 7 discusses approaches to sparse modelling using MLLs, and contains simulated and data-based examples. Longer proofs are in Section 8.

2 Marginal Log-Linear Parameters

We consider collections of random variables $(X_v)_{v\in V}$ with finite index set V, taking values in finite discrete probability spaces $(\mathfrak{X}_v)_{v\in V}$ under a strictly positive probability measure P; without loss of generality, $\mathfrak{X}_v = \{0, 1, \ldots, |\mathfrak{X}_v| - 1\}$. For $A \subseteq V$ we let $\mathfrak{X}_A \equiv \times_{v\in A}(\mathfrak{X}_v)$, $\mathfrak{X} \equiv \mathfrak{X}_V$ and $X_A \equiv (X_v)_{v\in A}$.

For $x \in \mathfrak{X}$, we denote by x_A the sub-vector of x consisting of those indices which belong to variables in A. Further, $\tilde{\mathfrak{X}}$ is the subset of $\mathfrak{X}$ which does not contain the last possible element in any co-ordinate; that is $\tilde{\mathfrak{X}}_v = \{0, 1, \ldots, |\mathfrak{X}_v| - 2\}$, and $\tilde{\mathfrak{X}} = \times_{v\in V}(\tilde{\mathfrak{X}}_v)$.

We use the shorthands $p_A(x_A) \equiv P(X_A = x_A)$ and $p_{A|B}(x_A \,|\, x_B) \equiv P(X_A = x_A \,|\, X_B = x_B)$; for particular instantiations of x we write, for example,

$$\begin{aligned} p_{011} &\equiv P(X_1 = 0, X_2 = 1, X_3 = 1), \\ p_{0\cdot 1} &\equiv \sum_{j\in\mathfrak{X}_2} p_{0j1} \\ &= P(X_1 = 0, X_3 = 1). \end{aligned}$$

Following Bergsma and Rudas (2002), we define a general class of parameters on discrete distributions. The definition relies upon abstract collections of subsets, so it may be helpful to the reader to keep in mind that the sets $M_i \in \mathbb{M}$ are margins, or subsets, of the distribution over V, and each set $\mathbb{L}_i$ is a collection of effects in the margin M_i. A pair (L, M_i) corresponds to a log-linear interaction over the set L, within the margin M_i.

Definition 2.1. 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. Then let

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

be the collection of margins in $\mathbb{P}$. If $\mathbb{M} = \{M_1, \ldots, M_k\}$, write

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

for the set of effects present in the margin M_i. We say that the collection $\mathbb{P}$ is *hierarchical* if the ordering on $\mathbb{M}$ is chosen so 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 saying that each L is associated only with the first margin M 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 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.

Definition 2.2. For $L \subseteq M \subseteq V$ and $x_L \in \mathfrak{X}_L$, let

$$\lambda_L^M(x_L) \equiv \frac{1}{|\mathfrak{X}_M|} \sum_{y_M \in \mathfrak{X}_M} \log p_M(y_M) \prod_{v \in L} \left(|\mathfrak{X}_v| \mathbb{I}_{\{x_v = y_v\}} - 1 \right).$$

This is a *marginal log-linear parameter*. For a collection of ordered pairs of subsets $\mathbb{P}$ (see Definition 2.1), we let

$$\Lambda(\mathbb{P}) \equiv \{\lambda_L^M(x_L) \mid (L, M) \in \mathbb{P},\ x_L \in \mathfrak{X}_L\},$$

be the collection of parameters associated with $\mathbb{P}$.

This definition is equivalent to the recursive one given in Bergsma and Rudas (2002); both expositions are fairly abstract, so we invite the reader to consult the examples below. In particular note that for binary random variables, the product is always ± 1. Any collection $\Lambda(\mathbb{P})$ where $\mathbb{P}$ is hierarchical and complete smoothly parametrizes the saturated model.

There is redundancy in the parameters (see Proposition 2.5) which leads to some flexibility in the definition of a marginal log-linear parameter; ours corresponds to 'effect coding'. We could use instead 'dummy coding', a corner point constraint which sets $\lambda_L^M(x_L)$ to be zero whenever any element of x_L is zero; this is used in Marchetti and Lupparelli (2010). This would not affect any major results.

The case where $L = \emptyset$ will not interest us here; in terms of a contingency table, the value of this marginal log-linear parameter is controlled by other MLL parameters and the sum over all cells, which we assume to be 1. Various examples for these definitions can be found in Bergsma and Rudas (2002).

2.1 Properties and Examples

We will write λ_L^M to mean the collection $\{\lambda_L^M(x_L) \mid x_L \in \mathfrak{X}_L\}$; if we write $\lambda_L^M = 0$, we are setting all the parameters in this collection to 0.

Example 2.3. *Log-linear and multivariate logistic parameters*

The usual log-linear parameters for the saturated model of a discrete distribution over a set of vertices V are $\{\lambda_L^V \mid L \subseteq V\}$. In the more general case of an undirected graph $\mathcal{G}$, the set of discrete distributions P obeying the global Markov property with respect to $\mathcal{G}$ is parametrized by $\{\lambda_L^V \mid L \in \mathcal{C}(\mathcal{G})\}$, where $\mathcal{C}(\mathcal{G})$ is the collection of complete subsets of $\mathcal{G}$.

Up to trivial transformations, the multivariate logistic parameters of Glonek and McCullagh (1995) are $\{\lambda_L^L \mid L \subseteq V\}$.

Example 2.4. Let $V = \{1, 2, 3\}$ and assume all random variables are binary. Then

$$\lambda_1^1(0) = \frac{1}{2} \log \frac{p_{0\cdot\cdot}}{p_{1\cdot\cdot}},$$

which, up to a multiplicative constant, is the logit of the probability of the event $\{X_1 = 0\}$. Also,

$$\lambda_1^{12}(0) = \frac{1}{4} \log \frac{p_{00\cdot}\, p_{01\cdot}}{p_{10\cdot}\, p_{11\cdot}} \qquad \text{and} \qquad \lambda_{12}^{12}(0,0) = \frac{1}{4} \log \frac{p_{00\cdot}\, p_{11\cdot}}{p_{10\cdot}\, p_{01\cdot}},$$

the log odds product and log odds ratio between X_1 and X_2 respectively. Here we have written, for example, 12 instead of $\{1,2\}$; similarly, for sets A and B we sometimes write AB for $A \cup B$, and aB for $\{a\} \cup B$.

Proposition 2.5. *For any $v \in L$, and fixed $x_{L\setminus\{v\}}$*

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

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

Remark 2.6. A collection of parameters which avoids the redundancy in Proposition 2.5 is

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

The next result relates the marginal log-linear parameters to conditional independences.

Lemma 2.7 (Rudas et al. (2010), Lemma 1)**.** *For any disjoint sets A, B and C, where C may be empty, $A \perp\!\!\!\perp B \mid C$ if and only if*

$$\lambda_{abD}^{ABC} = 0 \qquad \text{for every } a \in A, \quad b \in B, \quad D \subseteq A \cup B \cup C \setminus \{a, b\}.$$

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

Example 2.8. Take a complete and hierarchical parametrization of 3 variables,

$$\lambda_1^1 \qquad \lambda_2^2 \qquad \lambda_3^3 \qquad \lambda_{12}^{12} \qquad \lambda_{13}^{13} \qquad \lambda_{23}^{123} \qquad \lambda_{123}^{123}.$$

Then we can force $X_1 \perp\!\!\!\perp X_3$ by setting $\lambda_{13}^{13} = 0$. Similarly we can force $X_2 \perp\!\!\!\perp X_3 \mid X_1$ by setting $\lambda_{23}^{123} = \lambda_{123}^{123} = 0$.

Lemma 2.9. *Suppose that $A \perp\!\!\!\perp B \mid C$, and A is non-empty. Then for any $D \subseteq C$,*

$$\lambda_{AD}^{ABC}(x_{AD}) = \lambda_{AD}^{AC}(x_{AD}), \qquad \text{for each } x_{AD} \in \mathfrak{X}_{AD}.$$

3 Acyclic Directed Mixed Graphs

Definition 3.1. A *directed mixed graph* $\mathcal{G}$ consists of a set of vertices V, and both directed ($\rightarrow$) and bidirected ($\leftrightarrow$) edges. Edges of the same type and orientation may not be repeated, but there may be multiple edges of different types between a pair of vertices.

A *path* in $\mathcal{G}$ is a sequence of adjacent edges, without repetition of a vertex; a path may be empty, or equivalently consist of only one vertex. A *directed* path is one in which all the edges are directed ($\rightarrow$) and are oriented in the same direction. A *bidirected path* is a path consisting entirely of bidirected edges. A *semi-directed* path is one in which all directed edges are oriented in the same direction, but which may contain bidirected edges as well; thus in our definition, every bidirected or directed path is also semi-directed.

A *cycle* is a non-empty sequence of adjacent edges, also without repetition of a vertex, except that the last vertex is the same as the first. A directed cycle is non-empty sequence of edges of the form $v \to \cdots \to v$. An *acyclic* directed mixed graph (ADMG) is one which contains no directed cycles.

Definition 3.2. For a graph $\mathcal{G}$ and a subset of the vertices $A \subseteq V$, we denote by $\mathcal{G}_A$ the *induced subgraph* formed by A; that is, the graph containing the vertices A, and the edges in $\mathcal{G}$ whose end points are both in A.

Definition 3.3. Let a, b and d be vertices in a mixed graph $\mathcal{G}$. If $b \to a$ we say that b is a *parent* of a, and a is a *child* of b. The set of vertices which are parents of a is written $\mathrm{pa}_{\mathcal{G}}(a)$, and the set of children of b is $\mathrm{ch}_{\mathcal{G}}(b)$.

If there is a directed path from a to d, or $a = d$, we say that a is an *ancestor* of d, and that d is a *descendant* of a. Sets of ancestors and descendants are denoted $\mathrm{an}_{\mathcal{G}}(d)$ and $\mathrm{de}_{\mathcal{G}}(a)$ respectively. The *district* of a, denoted $\mathrm{dis}_{\mathcal{G}}(a)$, is the set containing a and all vertices which are connected to a by a bidirected path; a district is referred to by some authors as a c-component. A district of $\mathcal{G}$ is a maximal set of vertices all connected by bidirected paths.

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

$$\mathrm{pa}_{\mathcal{G}}(W) \equiv \bigcup_{w \in W} \mathrm{pa}_{\mathcal{G}}(w), \qquad \mathrm{dis}_{\mathcal{G}}(W) \equiv \bigcup_{w \in W} \mathrm{dis}_{\mathcal{G}}(w).$$

A set is of vertices A is *ancestral* if $A = \mathrm{an}_{\mathcal{G}}(A)$; that is, A contains all its own ancestors.

Note that by the definitions of some authors, vertices are not their own ancestors (Lauritzen, 1996). The above notations may be shortened on induced subgraphs so that $\mathrm{pa}_A \equiv \mathrm{pa}_{\mathcal{G}_A}$, and similarly for other definitions. In some cases where the meaning is clear, we will dispense with the subscript altogether.

We use the now standard notation of Dawid (1979), and represent the statement 'X is independent of Y given Z under a probability measure P', for random variables X, Y and Z, by $X \perp\!\!\!\perp Y \mid Z$ $[P]$. If P is unambiguous, this part is dropped, and if Z is empty we write simply $X \perp\!\!\!\perp Y$. Finally, we abuse notation in the usual way: v and X_v are used interchangeably as both a vertex and a random variable; likewise A denotes a vertex set and X_A.

3.1 Global Markov Property

A Markov property relates a graph to the probability distributions it represents.

A non-endpoint vertex c on a path is a *collider* on the path if the edges preceding and succeeding c on the path have an arrowhead at c, for example $\to c \leftarrow$ or $\leftrightarrow c \leftarrow$; otherwise c is a *non-collider*. A path between vertices a and b in a mixed graph is said to be *m-connecting given a set* C if

(i) every non-collider on the path is not in C, and

(ii) every collider on the path is an ancestor of C.

If there is no path m-connecting a and b given C, then a and b are said to be *m-separated given* C. Sets A and B are said to be m-separated given C if every $a \in A$ and every $b \in B$

are m-separated given C. This concept naturally extends the d-separation criterion of Pearl (1988) to graphs with bidirected edges.

A probability measure P on $\mathfrak{X}$ is said to satisfy the *global Markov property* for $\mathcal{G}$ if for arbitrary disjoint sets A, B and C,

$$A \text{ is m-separated from } B \text{ given } C \text{ in } \mathcal{G} \qquad \Longrightarrow \qquad X_A \perp\!\!\!\perp X_B \mid X_C \; [P].$$

3.2 Existing Parametrization of ADMGs

This subsection explains the parameters of Richardson (2009) for multivariate discrete distributions satisfying the global Markov property.

Definition 3.4. Let $\mathcal{G}$ be an ADMG with vertex set V. We say that a collection of vertices $W \subseteq V$ is *barren* if for each $v \in W$, we have $W \cap \mathrm{de}_{\mathcal{G}}(v) = \{v\}$; in other words v has no non-trivial descendants in W.

A *head* is a collection of vertices H which is barren and is $\leftrightarrow$-connected in $\mathcal{G}_{\mathrm{an}(H)}$. We write $\mathcal{H}(\mathcal{G})$ for the collection of heads in $\mathcal{G}$. The *tail* of a head H is the set

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

We typically write T for a tail, provided it is clear which head it belongs to.

Richardson (2009) shows that discrete distributions obeying the global Markov property for an ADMG $\mathcal{G}$ are parametrized by the conditional probabilities

$$\left\{ P(X_H = x_H \mid X_T = x_T) \;\middle|\; H \in \mathcal{H},\; T = \mathrm{tail}_{\mathcal{G}}(H),\; x_H \in \tilde{\mathfrak{X}}_H,\; x_T \in \mathfrak{X}_T \right\}.$$

This is achieved via the factorization

$$P(X_V = x_V) = \prod_{H \in [V]_{\mathcal{G}}} P(X_H = x_H \mid X_T = x_T). \tag{1}$$

The function $[\cdot]_{\mathcal{G}}$ partitions sets of vertices into heads; see Richardson (2009) for details. In the case of a directed acyclic graph (DAG), this corresponds to the probability distribution of each vertex conditional on its parents: $p(x_v \mid x_{\mathrm{pa}(v)})$.

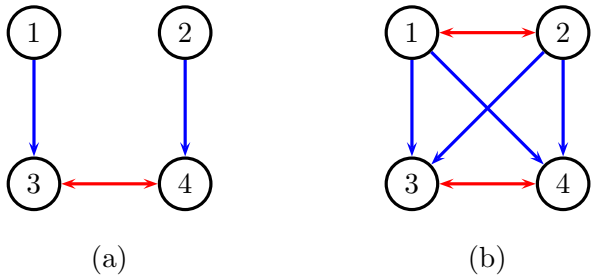
Consider again the ADMG in Figure 1(a); its head-tail pairs (H, T) are $(1, \emptyset)$, $(2, \emptyset)$, $(3, 1)$, $(4, 2)$ and $(34, 12)$. Then the multivariate binary distributions obeying the global Markov property with respect to this graph are parametrized by

$$p_1(0) \qquad p_2(0) \qquad p_{3|1}(0 \mid x_1) \qquad p_{4|2}(0 \mid x_2) \qquad p_{34|12}(0, 0 \mid x_1, x_2),$$

for $x_1, x_2 \in \{0, 1\}$.

3.3 Graphical Completions

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



(a) (b)

Figure 2: (a) An acyclic directed mixed graph, and (b) its balanced completion.

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, the completion scheme defined below preserves heads (see Proposition 3.7).

Definition 3.5. Given an ADMG $\mathcal{G}$, we define $\bar{\mathcal{G}}$, the *balanced completion of* $\mathcal{G}$ to be the graph obtained by adding edges to $\mathcal{G}$ in the following manner: for two vertices i and j with no edge between them, add in the edge $i \to j$ if there is a directed path from i to j, *or* there is a semi-directed path from i to j and no such path from j to i; otherwise consider $j \to i$ in the same manner. If there are semi-directed paths in both directions (but no directed paths), or in neither direction, insert $i \leftrightarrow j$.

It is easy to verify that this notion is well defined; it is also clear that no m-separations hold in the graph $\bar{\mathcal{G}}$, and thus it represents the saturated model. Note that it 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 $\bar{\mathcal{G}}$ is complete. The name 'balanced' is intended to convey that the scheme is a compromise between completions which add only directed edges, and a scheme which adds only bidirected edges.

Example 3.6. Figure 2(a) shows an ADMG, together with its balanced completion (b).

Proposition 3.7. $\mathcal{H}(\mathcal{G}) \subseteq \mathcal{H}(\bar{\mathcal{G}})$*; in other words, heads are preserved by balanced completion.*

Of course tails will, in general, be greatly expanded by the process.

4 Ingenuous Parametrization

In this section we use the marginal log-linear parameters defined in Section 2 to parametrize the distributions discussed in Section 3.

Definition 4.1. Consider an ADMG $\mathcal{G}$ with head-tail pairs (H_i, T_i) over some index i, and let $M_i = H_i \cup T_i$. Further, let $\mathbb{L}_i = \{A \mid H_i \subseteq A \subseteq H_i \cup T_i\}$. This collection of margins and associated effects is the *ingenuous* parametrization of $\mathcal{G}$, denoted $\mathbb{P}^{\text{ing}}(\mathcal{G})$.

Example 4.2. We return again to the ADMG in Figure 1(a); the head-tail pairs are $(1, \emptyset)$, $(2, \emptyset)$, $(3, 1)$, $(4, 2)$ and $(34, 12)$, meaning that the ingenuous parametrization is given by the

following margins and effects:

M	$\mathbb{L}$
1	1
2	2
13	3, 13
24	4, 24
1234	34, 134, 234, 1234.

In order to use most of the results of Bergsma and Rudas (2002), we need to confirm that the definition above corresponds to a hierarchical parametrization, which is shown by the following result.

Lemma 4.3. *For any ADMG $\mathcal{G}$, there is an ordering on the sets M_i of the ingenuous parametrization $\mathbb{P}^{\text{ing}}(\mathcal{G})$ which is hierarchical.*

Proof. Firstly, we show that for distinct heads H_i and H_j, the collections $\mathbb{L}_i$ and $\mathbb{L}_j$ are disjoint. To see this, assume for contradiction that there exists A such that $H_i \subseteq A \subseteq H_i \cup T_i$ and $H_j \subseteq A \subseteq H_j \cup T_j$. Then since $H_i \neq H_j$, assume without loss of generality that there exists $v \in H_i \cap H_j^c \subseteq A$.

Then $v \in H_j \cup T_j$ implies that $v \in T_j$, and thus there is a directed path from v to some $w \in H_j$. Now, $w \notin H_i$, since $v, w \in H_i$ would imply that H_i is not barren. But then if $w \in H_j \cap H_i^c$, then by the same argument as above we can find a directed path from w to some $x \in H_i$. Then $v \to \cdots \to w \to \cdots \to x$ is a directed path between elements of H_i, which is a contradiction. Thus $\mathbb{L}_i$ and $\mathbb{L}_j$ are disjoint.

Now, choose the labelling of heads such that $i < j$ if $H_i \neq H_j$ and $H_i \subset \text{an}_{\mathcal{G}}(H_j)$. This is a well defined partial ordering, since if it were not then $\mathcal{G}$ would contain a directed cycle. Any total ordering which respects this partial ordering is hierarchical, because any set $A \in \mathbb{L}_i$ is a subset of the ancestors of H_i. □

We proceed to show that the ingenuous parameters produce the set of distributions corresponding to the global Markov property.

Lemma 4.4. *Given sets M and $L \subseteq M$, the collection of MLL parameters*

$$\{\lambda_A^M(x_A) \mid L \subseteq A \subseteq M, x_M \in \tilde{\mathfrak{X}}_M\},$$

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

A proof is given in Section 8. Note that in the case of binary random variables, the collection consists of just one parameter.

Theorem 4.5. *The ingenuous parametrization $\tilde{\Lambda}(\mathbb{P}^{\text{ing}}(\mathcal{G}))$ of an ADMG $\mathcal{G}$ parametrizes precisely those distributions P obeying the global Markov property with respect to $\mathcal{G}$.*

Proof. We proceed by induction. Use the same partial ordering on heads from the proof of Lemma 4.3: that is, $H_i \prec H_j$ if $H_i \neq H_j$ and $H_i \subset \mathrm{an}_{\mathcal{G}}(H_j)$. For the base case, we know that singleton heads $\{h\}$ with empty tails are parametrized by the logits λ_h^h.

Now, suppose that we wish to find the distribution of a head H conditional on its tail T. Assume that we have the distribution of all heads H' which precede H, 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 $L \cup T \subseteq A$. Theorem 4 of Richardson (2009) states that the factorization in equation (1) holds for any ancestral set, so

$$p_A(x_A) = \prod_{\substack{H' \in [A]_{\mathcal{G}} \\ T' = \mathrm{tail}(H)}} p_{H'|T'}(x_{H'} \mid x_{T'}).$$

But all the probabilities in the product are known by our induction hypothesis, and the marginal distribution of L conditional on T is given by the 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 Lemma 4.4 above. □

Example 4.6. Returning to our running example, the graph in Figure 1(a) has the ingenuous parametrization

$$\begin{array}{cccccc} \lambda_1^1 & \lambda_2^2 & \lambda_3^{13} & \lambda_{13}^{13} & \lambda_4^{24} & \lambda_{24}^{24} \\ \lambda_{34}^{1234} & \lambda_{134}^{1234} & \lambda_{234}^{1234} & \lambda_{1234}^{1234}. & & \end{array}$$

Assume the random variables are binary and recall that the parametrization in Richardson (2009) used the probabilities

$$p_1(0) \qquad p_2(0) \qquad p_{3|1}(0 \mid x_1) \qquad p_{4|2}(0 \mid x_2) \qquad p_{34|12}(0,0 \mid x_1, x_2),$$

for $x_1, x_2 \in \{0,1\}$; thus if we can recover these probabilities, we know that we can recover the whole distribution. First,

$$\lambda_1^1(0) = \frac{1}{2}\log\frac{p_1(0)}{p_1(1)} = \frac{1}{2}\log\frac{p_1(0)}{1-p_1(0)} \qquad \lambda_2^2(0) = \frac{1}{2}\log\frac{p_2(0)}{p_2(1)} = \frac{1}{2}\log\frac{p_2(0)}{1-p_2(0)}$$

can be solved for the marginal distributions of 1 and 2. Also

$$\lambda_3^{13}(0) + \lambda_{13}^{13}(x_1, 0) = \frac{1}{2}\log\frac{p_{13}(x_1,0)}{p_{13}(x_1,1)} = \frac{1}{2}\log\frac{p_{3|1}(0 \mid x_1)}{1-p_{3|1}(0 \mid x_1)},$$

which is easily solved for the distribution of 3 conditional upon 1. The joint distribution of 2 and 4 follows similarly. Lastly, for each $x_1, x_2 \in \{0,1\}$,

$$\begin{aligned} \lambda_{34}^{1234}(0,0) + \lambda_{134}^{1234}(x_1,0,0) + \lambda_{234}^{1234}(x_2,0,0) + \lambda_{1234}^{1234}(x_1,x_2,0,0) \\ = \frac{1}{16}\log\frac{p_{34|12}(0,0 \mid x_1,x_2)\cdot p_{34|12}(1,1 \mid x_1,x_2)}{p_{34|12}(1,0 \mid x_1,x_2)\cdot p_{34|12}(0,1 \mid x_1,x_2)}. \end{aligned}$$

Note that each of the probabilities in this last expression can be written in terms of things already known and $p_{34|12}(0,0 \mid x_1, x_2)$; for example,

$$p_{34|12}(1,1 \mid x_1, x_2) = 1 - p_{3|1}(0 \mid x_1) - p_{4|2}(0 \mid x_2) + p_{34|12}(0,0 \mid x_1, x_2).$$

Then we can rearrange to give a quadratic equation for $p_{34|12}(0,0 \mid x_1, x_2)$, which has exactly one valid solution.

4.1 Completion

We have demonstrated that the ingenuous parametrization is hierarchical, but it is clearly not complete. To apply many of the results in Bergsma and Rudas (2002) concerning marginal log-linear parameters, we require completeness; in particular we wish to represent the ingenuous parametrization as a sub-model of the saturated model.

Lemma 4.7. *The ingenuous parametrization of an ADMG $\mathcal{G}$ is a linear subspace of the ingenuous parametrization of its balanced completion $\bar{\mathcal{G}}$, possibly after relabelling some margins.*

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 $\mathcal{G}$, then 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, then 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 must be 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 connecting 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 2.7, these parameters are all identically zero under $\mathcal{G}$.

(ii) implies that H is head in both $\mathcal{G}$ and $\bar{\mathcal{G}}$, but $T \equiv \text{tail}_{\bar{\mathcal{G}}} H \supset \text{tail}_{\mathcal{G}} H \equiv T'$. Then 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 $\text{an}_{\mathcal{G}}(H)$, and Lemma 2.7.

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 $\mathcal{G}$, which follows from the fact that T' is the Markov blanket for H in $\text{an}_{\mathcal{G}}(H)$, and Lemma 2.9. Thus we are simply relabelling some of the parameters.

Setting some of the parameters to zero amounts to a set of linear constraints on the parameters. □

This result allows us to apply Bergsma and Rudas' results 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. We could also have used these methods to prove 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 4.8. Consider again the ADMG $\mathcal{G}$ in Figure 2(a) with its balanced completion $\bar{\mathcal{G}}$ in (b). The ingenuous parametrization for $\bar{\mathcal{G}}$ is

M	$\mathbb{L}$
1	1
2	2
12	12
123	3, 13, 23, 123
124	4, 14, 24, 124
1234	34, 134, 234, 1234.

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 zero parameters and renaming the four others according to these last equations returns us to the ingenuous parametrization of $\mathcal{G}$.

Remark 4.9. Rudas et al. (2010) 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 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 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 that in Figure 1(b).

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 parameters considered here under the appropriate Markov property, using Lemma 2.9. Thus the variation dependence properties of that parametrization are identical to those of the ingenuous parametrization (see next section).

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

5 Ordered Decomposability and Variation Independence

Definition 5.1. Let θ_i, for $i = 1, \ldots, k$ be a collection of parameters such that θ_i takes any value in the set Θ_i. We say that the vector $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_k)$ is *variation independent* if $\boldsymbol{\theta}$ can take any value in the set $\Theta_1 \times \cdots \times \Theta_k$.

We now seek to categorize which ingenuous parametrizations are variation independent. Bergsma and Rudas (2002) characterize precisely which hierarchical and complete parametrizations are variation independent, using a notion they call ordered decomposability.

Definition 5.2. A collection of sets $\mathbb{M} = \{M_1, \ldots, M_k\}$ is *incomparable* if $M_i \not\subseteq M_j$ for every $i \neq j$.

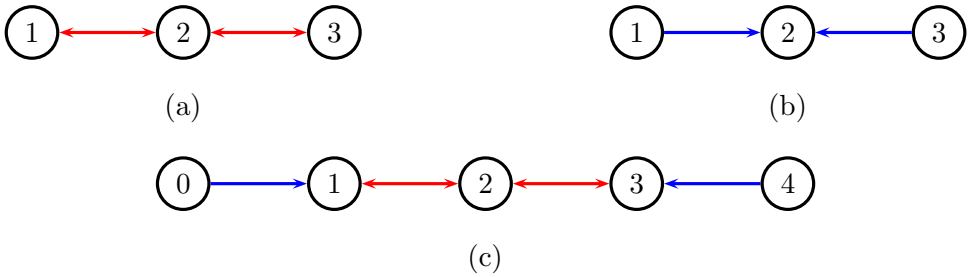


Figure 3: (a) a graph with a variation dependent ingenuous parametrization; (b) a Markov equivalent graph to (a) with a variation independent MLL parametrization; (c) a graph with no variation independent MLL parametrization.

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 $M_i \nsubseteq M_j$ 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.

The following example is found in Bergsma and Rudas (2002).

Example 5.3. Let $\mathbb{M} = \{12, 13, 23, 123\}$. In order to have a hierarchical ordering of these margins it is clear that the set 123 must come last, but there is no way to order the collection of inclusion maximal margins $\{12, 13, 23\}$ such that it has the running intersection property. Thus $\mathbb{M}$ is not ordered decomposable.

The next result links variation independence to ordered decomposability.

Theorem 5.4 (Bergsma and Rudas (2002), Theorem 4)**.** *Let $\mathbb{P}$ be a parametrization which is hierarchical and complete. Then the parameters are variation independent if and only if $\mathbb{P}$ is ordered decomposable.*

For the ingenuous parametrization this has the following consequence.

Theorem 5.5. *The ingenuous parametrization for an ADMG $\mathcal{G}$ is variation independent if and only if $\mathcal{G}$ contains no heads of size greater than or equal to 3.*

The bidirected 3-chain shown in Figure 3(a) has the head 123, and therefore its ingenuous parametrization is variation dependent. This can easily be seen directly: in the binary case,

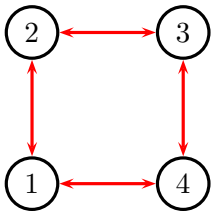


Figure 4: A bidirected 4-cycle.

for example, if the parameters $\lambda_{12}^{12}(0)$ and $\lambda_{23}^{23}(0)$ are chosen to be very large, it induces very high correlation between the variables X_1 and X_2, and between X_2 and X_3 respectively. If these correlations are chosen to be too high, then it is impossible for X_1 and X_3 to be independent, which is implied by the graph.

Observe that we could use the Markov equivalent graph in Figure 3(b), which has no heads of size 3, and thus obtain a variation independent parametrization of the same model. However, if we add incident arrows as shown in Figure 3(c), we obtain a graph where such a trick is not possible. In fact, this third graph is Markov equivalent to the bidirected 5-chain, which has no variation independent parametrization in the Bergsma and Rudas framework.

In general, it would be useful for those concerned about variation dependence to choose a graph from the Markov equivalence class created by their model which has the smallest possible maximum head size. This could be achieved by reducing the number of bidirected edges in the graph, where possible; see, for example, Ali et al. (2005) and Drton and Richardson (2008b) for approaches to this.

Example 5.6. The bidirected 4-cycle, shown in Figure 4, contains a head of size 4, and so its ingenuous parametrization is variation dependent. However, the model can be given an ordered decomposable and thus variation independent parametrization in the framework of marginal log-linear parameters. The 4-cycle is precisely the model with $X_1 \perp\!\!\!\perp X_3$ and $X_2 \perp\!\!\!\perp X_4$. Set $\mathbb{M} = \{13, 24, 1234\}$, with

$$\begin{aligned}\mathbb{L}_1 &= \{1, 3, 13\}\\ \mathbb{L}_2 &= \{2, 4, 24\}\\ \mathbb{L}_3 &= \mathscr{P}(\{1, 2, 3, 4\}) \setminus (\mathbb{L}_1 \cup \mathbb{L}_2);\end{aligned}$$

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. The 4-cycle corresponds exactly to setting $\lambda_{13}^{13} = \lambda_{24}^{24} = 0$, and thus the remaining parameters must still be variation independent under this constraint.

This method of parametrization by considering disconnected sets is discussed in detail by Lupparelli et al. (2009). It produces a variation independent parametrization for graphs where the disconnected sets do not overlap, and may well be preferable to the ingenuous parametrization in these cases. In sparser graphs however, it does not seem as useful; the bidirected 5-cycle, for example, has no variation independent MLL parametrization.

6 Equivalent Parametrizations

It turns out that a simple, invertible, linear transformation of the ingenuous parameters yields equivalent parameters which may be easier to interpret, and perhaps preferable for applied users of statistics. For example, it can lead to rather abstruse higher order interaction parameters being replaced with conditional logits and log odds ratios.

Proposition 6.1. *For disjoint sets $L, N \subseteq V$ with $M = L \cup N$ and $x_M \in \mathfrak{X}_M$, let*

$$\kappa_{L|N}(x_L \,|\, x_N) \equiv \sum_{L \subseteq A \subseteq M} \lambda_A^M(x_A).$$

Then

$$\kappa_{L|N}(x_L \,|\, x_N) = \frac{1}{|\mathfrak{X}_L|} \sum_{\substack{y_M \in \mathfrak{X}_M \\ y_{M\setminus L} = x_{M\setminus L}}} \log p(y_M) \prod_{v \in L} \left(|\mathfrak{X}_v| \mathbb{I}_{\{x_v = y_v\}} - 1 \right).$$

The proof of this result is in Section 8. Note that the summation used to define κ leaves the indices in N fixed; thus we can think of $\kappa_{L|N}(x_L \,|\, x_N)$ as the logarithm of the $|L|$-way interaction parameter for x_L, under the conditional distribution where $X_N = x_N$. For $|L| = 1$ we obtain something equivalent to a conditional probability; for $|L| = 2$ we obtain a conditional log-odds ratio, and $|L| = 3$ gives conditional log three-way interactions; for $|L| \geq 4$ interpretation becomes difficult.

For example, letting $V = \{1, 2, 3\}$ with binary random variables,

$$\begin{aligned} \kappa_{1|2}(0 \,|\, 0) &= \frac{1}{2} \log \frac{p_{1|2}(0 \,|\, 0)}{p_{1|2}(1 \,|\, 0)} \\ &= \frac{1}{2} \log \frac{P(X_1 = 0 \mid X_2 = 0)}{P(X_1 = 1 \mid X_2 = 0)}, \end{aligned}$$

which is, up to the multiplicative constant, the logit of the event $\{X_1 = 0\}$, conditional on $\{X_2 = 0\}$. Similarly,

$$\kappa_{12|3}(0, 0 \,|\, x_3) = \frac{1}{4} \log \frac{p_{12|3}(0, 0 \,|\, x_3)\; p_{12|3}(1, 1 \,|\, x_3)}{p_{12|3}(1, 0 \,|\, x_3)\; p_{12|3}(0, 1 \,|\, x_3)},$$

which is the log odds ratio of X_1 and X_2 conditional on $\{X_3 = x_3\}$.

These parameters are also used by Marchetti and Lupparelli (2010) (see their Lemma 2); other more general parameters were used by Bartolucci et al. (2007).

Definition 6.2. For an ADMG $\mathcal{G}$, define

$$K(\mathcal{G}) \equiv \{\kappa_{H|T}(x_H \,|\, x_T) \mid H \in \mathcal{H},\; x_{HT} \in \mathfrak{X}_{HT}\}.$$

Further, let

$$\tilde{K}(\mathcal{G}) \equiv \{\kappa_{H|T}(x_H \,|\, x_T) \mid H \in \mathcal{H},\; x_{HT} \in \tilde{\mathfrak{X}}_H \times \mathfrak{X}_T\}.$$

It is not difficult to see that the linear transformation from λ's to κ's is invertible, and that therefore the two parametrizations $\tilde{K}(\mathcal{G})$ and $\tilde{\Lambda}(\mathbb{P}^{\text{ing}}(\mathcal{G}))$ are completely equivalent. In particular, because the range of each individual parameter in either parametrization is the whole real line $\mathbb{R}$, then $\tilde{K}(\mathcal{G})$ is variation independent if and only if $\tilde{\Lambda}(\mathbb{P}^{\text{ing}}(\mathcal{G}))$ is variation independent.

Example 6.3. The ingenuous and κ-parametrizations for binary random variables obeying the Markov properties of the graph in Figure 1(a) are presented in the table below.

H	Ingenuous	κ-parametrization
1	λ_1^1	$\kappa_1(\cdot)$
2	λ_2^2	$\kappa_2(\cdot)$
3	λ_3^{13}, λ_{13}^{13}	$\kappa_{3\|1}(\cdot \,\|\, 0)$, $\kappa_{3\|1}(\cdot \,\|\, 1)$
4	λ_4^{24}, λ_{24}^{24}	$\kappa_{4\|2}(\cdot \,\|\, 0)$, $\kappa_{4\|2}(\cdot \,\|\, 1)$
34	λ_{34}^{1234}, λ_{134}^{1234}, λ_{234}^{1234}, λ_{1234}^{1234}	$\kappa_{34\|12}(\cdot \,\|\, 0,0)$, $\kappa_{34\|12}(\cdot \,\|\, 1,0)$, $\kappa_{34\|12}(\cdot \,\|\, 0,1)$, $\kappa_{34\|12}(\cdot \,\|\, 1,1)$

Note, for example, that the ingenuous parameters associated with the head 34 include the 4-way interaction parameter λ_{1234}^{1234}; the meaning of this parameter may not, on its own, mean much to an applied user of statistics who fits the model in Figure 1(a) to some data. On the other hand, in the κ-formulation, the parameters for 34 are all conditional odds-ratios, as discussed in the introduction. This is simply a measure of the correlation between X_3 and X_4 conditional on particular values of X_1 and X_2, which is easier to interpret.

Remark 6.4. We have presented three parametrizations of the model associated with an ADMG $\mathcal{G}$: the generalized Möbius parametrization of Richardson (2009), the ingenuous parametrization, and the 'κ-parametrization'. A notable feature of all three is that parameters may be organized according to the unique head they are associated with.

Consider again the partial ordering on heads used in the proof of Theorem 4.5: $H_i \prec H_j$ if $H_i \subseteq \text{an}_{\mathcal{G}}(H_j)$, and $H_i \neq H_j$. Now let $\mathcal{Q}_k$ be the set of Richardson's conditional probabilities associated with H_k and all the heads preceding it; let $\mathcal{L}_k$ and $\mathcal{K}_k$ be defined analogously for the ingenuous and κ-parametrizations respectively.

It is not hard to see that $\mathcal{Q}_k$, $\mathcal{L}_k$ and $\mathcal{K}_k$ are all equivalent for $k = 1, 2, \ldots$, in the sense that there are smooth bijective maps between these collections of parameters. Further, for any $x_{T_k} \in \mathfrak{X}_{T_k}$, the set $\mathcal{K}_{k-1} \cup \{\kappa_{H_k|T_k}(0 \,|\, x_{T_k})\}$ is equivalent to $\mathcal{Q}_{k-1} \cup \{p_{H_k|T_k}(0 \,|\, x_{T_k})\}$, though there is in general no equivalent subset of the ingenuous parameters. This is a result of the fact that both the κ-parametrization and the generalized Möbius parameters are based on fixed tail states, whereas each ingenuous parameter involves probabilities for all possible tail states.

7 Parsimonious Modelling with Marginal Log-Linear Parameters

The number of parameters in a model associated with a sparse graph containing bidirected edges can, in some cases, be relatively large. In a purely bidirected graph, the number of

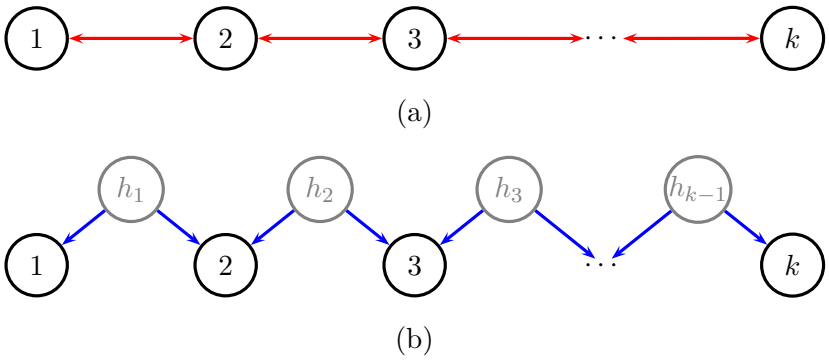


Figure 5: (a) A bidirected k-chain and (b) a DAG with latent variables $(h_1, \ldots, h_{k-1})$ generating the same conditional independence structure.

parameters depends upon the number of connected sets of vertices; in the case of a chain of bidirected edges such as that shown in Figure 5(a), this means that the number of parameters grows quadratically with the length of the chain.

The parametrization of Richardson (2009), and its special case for purely bidirected graphs in Drton and Richardson (2008a), does not present us with any obvious method of reducing the parameter count whilst preserving the conditional independence structure.

In contrast, there are well established methods for sparse modelling with other classes of graphical models. In the case of an undirected graph with binary random variables, restricting to one parameter for each vertex and each edge leads to 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.

7.1 Simulated Data

Consider the DAG with latent variables shown in Figure 5(b); over the observed variables, the only conditional independences which hold are the same as those in the bidirected chain in Figure 5(a).

We randomly generated 1,000 distributions from this DAG model with $k = 6$, where each latent variable was given three states, and each observed variable two. The probability of each observed variable being zero, conditional on each state of its parents, was an independent uniform random draw on $(0, 1)$; latent states were fixed to occur with equal probability. For each distribution, a sample size of 10,000 was drawn, and the bidirected chain model was fitted to it by maximum likelihood estimation. For each of the 1,000 data sets, we then measured the increase in deviance associated with removing the higher level parameters

The histogram in Figure 6(a) demonstrates that the deviance increase from setting the 5- and 6-way interaction parameters to zero (a total of three parameters) was not distinguishable from that which would be observed if these parameters contributed nothing to the model. The deviance increase from setting the 4-, 5- and 6-way interactions to zero appeared to have a slightly heavier tail than the associated χ^2-distribution, as suggested by the outliers in Figure 6(b). Removing the 3-way interactions in addition to this caused a dramatic increase in the

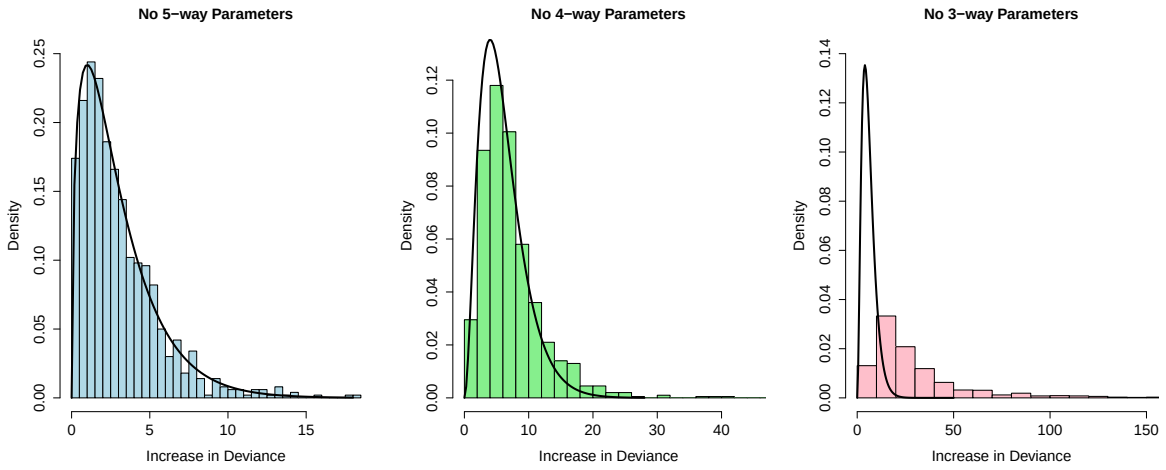


Figure 6: Histograms showing the increase in deviance caused by setting to zero (a) the 5- and 6-way interaction parameters; (b) the 4-, 5- and 6-way interaction parameters; (c) the 3-, 4-, 5- and 6-way interaction parameters. The datasets were generated from the DAG in Figure 5(b). The plotted densities are χ^2 with 3, 6 and 10 degrees of freedom respectively.

deviance, as may be observed from the heavy tail of the histogram in Figure 6(c).

Note that under the process which generated these models, each of these parameters was non-zero almost surely. As the sample size increases the power of a likelihood ratio test for a fixed distribution tends to one, so it must be the case that a simulation such as the above would, for large enough data sets, show significant deviation from the associated χ^2 distributions. However, even at a fairly large sample sizes of 10,000, a limited effect was observed in Figures 6(a) and (b), and the following example with real data suggests that higher order interactions are not particularly useful in practice.

7.2 Example: Trust Data

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. Can most people be trusted?

Helpful. Do you think most people are usually helpful?

MemUn, MemCh. Are you a member of a labour union / church?

ConLegis, ConClerg, ConBus. Do you have confidence in congress / organized religion / business?

In that paper, the model given by the graph in Figure 7 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

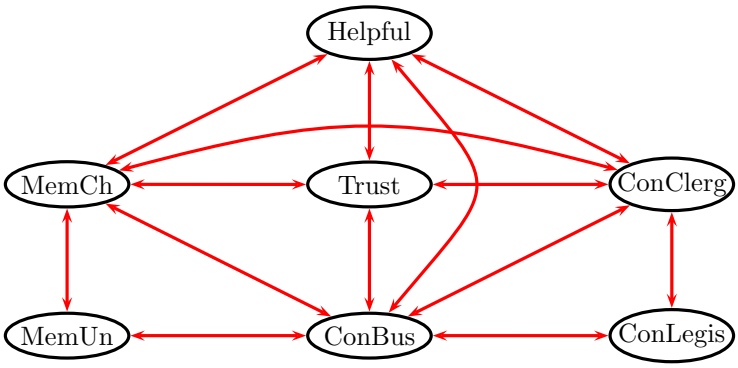


Figure 7: Markov model for trust data given in Drton and Richardson (2008a).

than the graph in Figure 7, and yet has 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 may consider the responses to a questionnaire to be jointly affected by unmeasured characteristics of the respondent, such as their 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: if we remove an edge between two vertices a and b, then this corresponds to requiring that $\lambda_A^V = 0$ for every effect A containing both a and b. Removing that edge in a bidirected model may correspond merely to setting $\lambda_{ab}^{ab} = 0$ and nothing else, depending upon the other edges present. Using the ingenuous parametrization, it is easy to constrain higher order terms to be zero.

Starting with the model in Figure 7 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. Furthermore, removing 21 of the remaining 25 three-way interaction terms increases the deviance to 111.48 on 102 degrees of freedom; using an aymptotic χ^2 approximation gives a p-value of 0.755, so this model is not contradicted by the data. 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.

8 Proofs

Proof of Proposition 6.1. Recalling that $M = L \cup N$,

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

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

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

Alternatively, if $x_N = y_N$, then

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

which is independent of y_M. Thus

$$\kappa_{L|N}(x_L \,|\, x_N) = \frac{1}{|\mathfrak{X}_L|} \sum_{y_L \in \mathfrak{X}_L} \log p_M(y_L, x_N) \prod_{v \in L} \left(|\mathfrak{X}_v| \mathbb{I}_{\{x_v = y_v\}} - 1 \right),$$

since $\mathfrak{X}_M = \mathfrak{X}_L \times \mathfrak{X}_N$. □

Proof of Lemma 2.9. Using the independence, we have

$$p_{ABC}(x_{ABC}) = p_{AC}(x_{AC}) \cdot p_{B|C}(x_B \,|\, x_C).$$

Thus

$$\lambda_{AD}^{ABC}(x_{AD}) = \frac{1}{|\mathfrak{X}_{ABC}|} \sum_{y_{ABC} \in \mathfrak{X}_{ABC}} \left(\log p_{AC}(y_{AC}) + \log p_{B|C}(y_B \,|\, y_C) \right) \prod_{v \in A \cup D} \left(|\mathfrak{X}_v| \mathbb{I}_{\{x_v = y_v\}} - 1 \right).$$

We can split this sum into terms involving $p_{AC}(y_{AC})$ and those involving $p_{B|C}(y_B \,|\, y_C)$. For the first of these,

$$\begin{aligned}
&\frac{1}{|\mathfrak{X}_{ABC}|} \sum_{y_{ABC} \in \mathfrak{X}_{ABC}} \log p_{AC}(y_{AC}) \prod_{v \in A \cup D} \left(|\mathfrak{X}_v| \mathbb{I}_{\{x_v = y_v\}} - 1 \right) \\
&= \frac{1}{|\mathfrak{X}_{AC}|} \sum_{y_{AC} \in \mathfrak{X}_{AC}} \log p_{AC}(y_{AC}) \prod_{v \in A \cup D} \left(|\mathfrak{X}_v| \mathbb{I}_{\{x_v = y_v\}} - 1 \right) \\
&= \lambda_{AD}^{AC}(x_{AC}),
\end{aligned}$$

because the summand has no dependence on y_B. For the latter,

$$\begin{aligned}
&\frac{1}{|\mathfrak{X}_{ABC}|} \sum_{y_{ABC} \in \mathfrak{X}_{ABC}} \log p_{B|C}(y_B \,|\, y_C) \prod_{v \in A \cup D} \left(|\mathfrak{X}_v| \mathbb{I}_{\{x_v = y_v\}} - 1 \right) \\
&= \frac{1}{|\mathfrak{X}_{ABC}|} \sum_{y_{BC} \in \mathfrak{X}_{BC}} \log p_{B|C}(y_B \,|\, y_C) \sum_{y_A \in \mathfrak{X}_A} \prod_{v \in A \cup D} \left(|\mathfrak{X}_v| \mathbb{I}_{\{x_v = y_v\}} - 1 \right) \\
&= 0,
\end{aligned}$$

because the inner sum is zero. This gives the result. □

Proof of Lemma 4.4. First we show that we can construct all the local $\log |L|$-way interaction parameters.

Let $N \equiv M \setminus L$, and pick some $x_L \in \tilde{\mathfrak{X}}_L$ and $x_N \in \mathfrak{X}_N$; for $A \subseteq \{1, \ldots, |M|\}$, let $\mathbf{1}_A$ be a vector of length $|L|$ with a 1 in position j if $j \in A$, and 0 otherwise. Then consider

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

We collect terms containing $\log p_M(y_M)$ for some y_M. If for some $w \in L$, $y_w \notin \{x_w, x_w + 1\}$, then the inner sum

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

because the value of the outer indicator function is 0 in both terms when $v = w$. Alternatively, if $y_w \in \{x_w, x_w + 1\}$ for all $w \in L$, then the identity

$$B(A) = \{v \in L \,|\, x_v + \mathbb{I}_{\{v \in A\}} = y_v\}$$

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

$$\begin{aligned}
&\sum_{A\subseteq L}(-1)^{|L\setminus A|}\prod_{v\in L}\Big(|\mathfrak{X}_v|^{\mathbb{I}_{\{x_v+\mathbb{I}_{\{v\in A\}}=y_v\}}}-1\Big)\\
&\qquad=(-1)^{\|x_L-y_L\|}\sum_{B\subseteq L}(-1)^{|L\setminus B|}\prod_{v\in L}\big(|\mathfrak{X}_v|^{\mathbb{I}_{\{v\in B\}}}-1\big)\\
&\qquad=(-1)^{\|x_L-y_L\|}\prod_{v\in L}|\mathfrak{X}_v|\\
&\qquad=(-1)^{\|x_L-y_L\|}|\mathfrak{X}_L|,
\end{aligned}$$

where $\|x_L-y_L\|$ is just the number of entries in which x_L and y_L differ. Then

$$\begin{aligned}
\sum_{A\subseteq L}(-1)^{|L\setminus A|}\kappa_{L|N}(x_L+\mathbf{1}_A\,|\,x_N) &= \sum_{\substack{y_M\in\mathfrak{X}_M\\ y_N=x_N\\ y_w\in\{x_w,x_w+1\}}}(-1)^{\|x_L-y_L\|}\log p_M(y_M)\\
&= \sum_{\substack{y_M\in\mathfrak{X}_M\\ y_N=x_N\\ y_w\in\{x_w,x_w+1\}}}(-1)^{\|x_L-y_L\|}\left[\log p_{L|N}(y_L\,|\,y_N)+\log p_N(y_N)\right]\\
&= \sum_{\substack{y_L\in\mathfrak{X}_L\\ y_w\in\{x_w,x_w+1\}}}(-1)^{\|x_L-y_L\|}\log p_{L|N}(y_L\,|\,x_N)
\end{aligned}$$

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). □

Lemma 8.1. *Let $\mathcal{G}$ be an ADMG containing at least one head of size 3 or more. Then $\mathcal{G}$ also contains two heads of the form $\{v_1,v_2\}$ and $\{v_2,v_3\}$, where $\{v_1,v_2,v_3\}$ is barren.*

Proof. Suppose not; let $\mathcal{G}$ be an ADMG which violates this condition, and let H be a head in $\mathcal{G}$ of size $k\geq 3$. Pick 3 vertices $\{w_1,w_2,w_3\}$ in H. By the definition of a head, we can pick a bidirected path ρ through $\mathrm{an}_{\mathcal{G}}(H)$, connecting w_1, w_2 and w_3. Assume that ρ contains no other verices of H, otherwise we can shorten the path until it contains precisely 3 elements, and redefine w_1,w_2,w_3 appropriately. We also assume that ρ is of the form $w_1\leftrightarrow\cdots\leftrightarrow w_2\leftrightarrow\cdots\leftrightarrow w_3$, else we can relabel the vertices so that w_2 is 'in the middle'.

According to our assumption that the theorem 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, and let π be the restriction of ρ from w_1 to w_2. This implies that the bidirected path π from w_1 to w_2 must pass through at least one vertex v which is not an ancestor of $\{w_1,w_2\}$. If there is more than one such vertex, then choose one which has no distinct descendants on the path π. By the construction of π we have $v\in\mathrm{an}_{\mathcal{G}}(H)\setminus H$.

Then letting W be the set of vertices in π, let $H^*=\mathrm{barren}_{\mathcal{G}}(W)$. Since W is $\leftrightarrow$-connected, H^* must be a head, and $\{w_1,w_2,v\}\subseteq H^*$. Thus we have created a head distinct from H, of size at least 3, which is contained in the set of ancestors of H.

The assumption we have made implies that we must be able to repeat this process indefinitely, with each head being contained in the ancestors of the previous head. To see that we

never obtain the same head twice, note that there is a non-empty directed path from $v \in H^*$ to H; but H is contained within the ancestors of any previous heads in the sequence, so if H^* had appeared before, this would imply that H^* was not barren.

Then since H has a finite set of ancestors, the apparently infinite recursion of distinct heads is a contradiction. □

Proof of Theorem 5.5. ($\Leftarrow$). 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 then show it to be equivalent to the ingenuous parametrization.

Let $\mathbb{M}_i \subseteq \mathbb{M}$ be the 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. This also implies that $H_k \cup T_k \setminus \{i\} = \text{mb}(i, [i])$, where $\text{mb}(v, A)$ is the Markov blanket of v in the ancestral set A (see Richardson, 2003, for the definition).

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

$$\begin{aligned} \lambda_C^{A_k} &= \lambda_C^{T_k} && \text{for any } H_k \subseteq C \subseteq H_k \cup T_k \\ \lambda_C^{A_k} &= 0 && \text{for any } \{i\} \subset C \nsubseteq H_k \cup T_k, \end{aligned}$$

where first equality follows from the independence and Lemma 2.9, and the second from the above independence and Lemma 2.7.

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 set $A_{k-1} = A_k \setminus \text{de}_{\mathcal{G}}(j_k)$. Then 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 for every subset of $[i]$ containing some descendant of any of $j_1, \ldots, j_k$. Lastly let $A_0 = A_1 \setminus \mathrm{de}_{\mathcal{G}}(j_1)$.

$$\lambda_C^{A_0} = \lambda_C^{T_0} \qquad \text{for any } \{i\} \subseteq C \subseteq \{i\} \cup T_0$$
$$\lambda_C^{A_0} = 0 \qquad \text{for any } \{i\} \subset C \not\subseteq \{i\} \cup T_0.$$

The margins we have added are $A_0 \subset \ldots \subset A_k$, 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 $\mathbb{M}_i = \mathbb{M}_{i-1} \cup \{A_0, \ldots, A_k\}$, then 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. Then by Lemma 8.1, 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, this set is taken to be the tail of H_3 *if there were* a bidirected arrow between v_1 and v_3. Further let $A = \mathrm{an}_{\mathcal{G}}(H)$.

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:

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

v_3 \ v_2	0	1
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 x_A of A.

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. Thus

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

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

Acknowledgements

This research was supported by the U.S. National Science Foundation grant CNS-0855230 and U.S. National Institutes of Health grant R01 AI032475. Our thanks go to Antonio Forcina and Tamás Rudas for helpful discussions, and to the former for the use of his computer programmes.

References

D. H. Ackley, G. E. Hinton, and T. J. Sejnowski. A learning algorithm for Boltzmann machines. *Cognitive Science*, 9:147–169, 1985.

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. *Statist. Sinica*, 17: 691–711, 2007.

W. P. Bergsma and T. Rudas. Marginal models for categorical data. *Ann. Stat.*, 30(1): 140–159, 2002.

I. Csiszár. i-divergence geometry of probability distributions and minimization problems. *Annals of Probability*, 3(1):146–158, 1975.

J. N. Darroch, S. L. Lauritzen, and T. P. Speed. Markov fields and log-linear models for contingency tables. *Ann. Statist.*, 8:522–539, 1980.

A. P. Dawid. Conditional independence in statistical theory (with discussion). *J. Roy. Statist. Soc. Ser. B*, 41:1–31, 1979.

M. Drton. Discrete chain graph models. *Bernoulli*, 15(3):736–753, 2009.

M. Drton and T. S. Richardson. Binary models for marginal independence. *J. Roy. Statist. Soc. Ser. B*, 70(2):287–309, 2008a.

M. Drton and T. S. Richardson. Graphical methods for efficient likelihood inference in Gaussian covariance models. *J. Mach. Learn. Res.*, 9:893–914, 2008b.

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)*, 2010.

A. Forcina, M. Lupparelli, and G. M. Marchetti. Marginal parameterizations of discrete models defined by a set of conditional independencies. *Journal of Multivariate Analysis*, 101:2519–2527, 2010.

G. F. V. Glonek and P. McCullagh. Multivariate logistic models. *J. Roy. Statist. Soc. Ser. B*, 57(3):533–546, 1995.

G. Kauermann. A note on multivariate logistic models for contingency tables. *Austral. J. Statist.*, 39(3):261–276, 1997.

S. L. Lauritzen. *Graphical Models*. Clarendon Press, Oxford, UK, 1996.

M. Lupparelli, G. M. Marchetti, and W. P. Bergsma. Parameterizations and fitting of bidirected graph models to categorical data. *Scand. J. Statist.*, 36:559576, 2009.

G. M. Marchetti and M. Lupparelli. Chain graph models of multivariate regression type for categorical data. arXiv:0906.2098, 2010.

J. Pearl. *Probabilistic Reasoning in Intelligent Systems*. Morgan Kaufmann, 1988.

T. S. Richardson. Markov properties for acyclic directed mixed graphs. *Scand. J. Statist.*, 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*, 2009.

T. S. Richardson and P. Spirtes. Ancestral graph Markov models. *Ann. Statist.*, 30:962–1030, 2002.

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*, 94:1006–1012, 2010.

J. Whittaker. *Graphical models in applied multivariate statistics*. Wiley, 1990.