

HYPERS AND STRUCTURAL MARKOV LAWS FOR GRAPHICAL MODELS

Simon Byrne

Clare College
and
Statistical Laboratory,
Department of Pure Mathematics and Mathematical Statistics

This dissertation is submitted for the degree of
Doctor of Philosophy



University of Cambridge

September 2011

Declaration

This dissertation is the result of my own work and includes nothing which is the outcome of work done in collaboration except where specifically indicated in the text.

Acknowledgements

I would like to thank everyone who helped in the writing of this thesis, especially my supervisor Philip Dawid for the thoughtful discussions and advice over the past three years. I would also like to thank my colleagues in the Statistical Laboratory for their support and friendship. Finally I would like to thank my family, especially Rebecca, without whom I would never have been able to complete this work.

Summary

My thesis focuses on the parameterisation and estimation of graphical models, based on the concept of *hyper and meta Markov properties*. These state that the parameters should exhibit conditional independencies, similar to those on the sample space. When these properties are satisfied, parameter estimation may be performed locally, *i.e.* the estimators for certain subsets of the graph are determined entirely by the data corresponding to the subset.

Firstly, I discuss the applications of these properties to the analysis of case-control studies. It has long been established that the maximum likelihood estimates for the odds-ratio may be found by logistic regression, in other words, the “incorrect” prospective model is equivalent to the correct retrospective model. I use a generalisation of the hyper Markov properties to identify necessary and sufficient conditions for the corresponding result in a Bayesian analysis, that is, the posterior distribution for the odds-ratio is the same under both the prospective and retrospective likelihoods. These conditions can be used to derive a parametric family of prior laws that may be used for such an analysis.

The second part focuses on the problem of inferring the structure of the underlying graph. I propose an extension of the meta and hyper Markov properties, which I term *structural Markov properties*, for both undirected decomposable graphs and directed acyclic graphs. Roughly speaking, it requires that the structure of distinct components of the graph are conditionally independent given the existence of a separating component. This allows the analysis and comparison of multiple graphical structures, while being able to take advantage of the common conditional independence constraints. Moreover, I show that these properties characterise exponential families, which form conjugate priors under sampling from compatible Markov distributions.

Contents

Declaration	iii
Acknowledgements	v
Summary	vii
I Hyper Markov properties	1
1 Graphical models and hyper Markov properties	3
1.1 Conditional independence	3
1.2 Separoids and Graphical models	4
1.3 Hyper Markov properties	7
1.4 Constructing hyper Markov laws	9
1.5 Variation independence and meta Markov properties	10
1.6 Gaussian graphical models and Hyper inverse Wishart laws .	11
1.7 Contingency tables and Hyper Dirichlet laws	15
1.8 Notes and other developments	19
2 Logistic regression and case-control studies	21
2.1 Notation and definitions	23
2.2 Maximum likelihood estimators	23
2.3 Bayesian analysis of case-control studies	25
2.4 Strong hyper Markov laws for logistic regression	29
2.5 Stratified case-control studies	33
2.6 Discussion	36

II Structural Markov properties	37
3 Background	39
4 Undirected decomposable graphical models	43
4.1 Motivation and definition	43
4.2 Projections and products	44
4.3 Structural meta Markov property	47
4.4 Compatible distributions and laws	48
4.5 Clique vector	50
4.6 Clique exponential family	52
4.7 Marginalisation	56
4.8 Computation	58
5 Directed acyclic graphical models	63
5.1 Ordered directed structural Markov property	63
5.2 Markov equivalence and Dagoids	64
5.3 Ancestral sets and remainder dagoids	67
5.4 Structural Markov property	69
5.5 d-Clique vector	71
5.6 Compatibility	74
5.7 Computation	79
6 Discussion	83
A Graph terminology	85
A.1 Undirected graphs	85
A.2 Directed graphs	87
Bibliography	91

List of Figures

4.1	A representation of the structural Markov property for undirected graphs.	44
4.2	Neighbouring graphs on 5 vertices	61
5.1	Four directed acyclic graphs with the same skeleton.	65
5.2	The d-cliques and d-separators of different graphs.	73
5.3	Three Markov equivalent graphs in which the same edge removal will result in a transition to a distinct Markov equivalence class.	79
A.1	The cliques of an undirected decomposable graph.	88

Part I

Hyper Markov properties

Graphical models and hyper Markov properties

In this chapter, we introduce the necessary definitions and theorems for graphical models, in particular the role of conditional independence and the Markov properties of graphs.

The basis of this thesis involves the hyper Markov properties, introduced in a seminal paper by Dawid and Lauritzen (1993). We introduce the necessary terminology and results, but refer the reader to the original paper for more details. Finally, we review some of the notable subsequent developments in hyper Markov theory.

1.1 Conditional independence

One of the most fundamental concepts of graphical models is the notion of conditional independence, which is used to describe the relationship between random variables.

Definition 1.1.1 (Conditional independence). Let X, Y, Z be random variables on a joint probability space (Ω, \mathcal{A}, P) . We say X is *conditionally independent of Y given Z* , written

$$X \perp\!\!\!\perp Y \mid Z \quad [P],$$

if there exists a conditional probability measure for X given Y, Z under P that only depends on Z .

In circumstances where the distribution is implied, we may drop the $[P]$. If Z is trivial, then we have *marginal independence* and may write $X \perp\!\!\!\perp Y$.

Theorem 1.1.1 (Dawid 1979, 1980). *The conditional independence statement $(\cdot \perp\!\!\!\perp \cdot \mid \cdot)$ is a ternary relation on the set of random variables on (Ω, \mathcal{A}, P) , such that for any random variables X, Y, Z, W and measurable function f , we have the following:*

- C0 $X \perp\!\!\!\perp Y \mid X$.
- C1 *If $X \perp\!\!\!\perp Y \mid Z$, then $Y \perp\!\!\!\perp X \mid Z$.*
- C2 *If $X \perp\!\!\!\perp Y \mid Z$, then $f(X) \perp\!\!\!\perp Y \mid Z$.*
- C3 *If $X \perp\!\!\!\perp Y \mid Z$, then $X \perp\!\!\!\perp Y \mid (Z, f(X))$.*
- C4 *If $X \perp\!\!\!\perp Y \mid Z$ and $X \perp\!\!\!\perp W \mid (Y, Z)$, then $X \perp\!\!\!\perp (W, Y) \mid Z$.*

Under certain conditions, such as the lack of any functional relationship between X, Y, Z , we also have the following property:

- C5 *If $X \perp\!\!\!\perp Y \mid Z$ and $X \perp\!\!\!\perp Z \mid Y$, then $X \perp\!\!\!\perp (Y, Z)$,*

though we will not utilise this property further.

1.2 Separoids and Graphical models

Graphical models encode a set of conditional independence properties in the structure of a graph. To facilitate later developments, we describe these conditional independence properties via the abstract *separoid* terminology of Dawid (2001), similar to the *semi-graphoid* of Pearl and Paz (1987); Pearl (1988).

Definition 1.2.1. Let M be a set with elements of the form $\langle A, B \mid C \rangle$ where A, B, C are subsets of a finite set V (that is, M is a ternary relation on V). Then M is a *separoid* if it satisfies the following properties:

- S0 $\langle A, B \mid A \rangle \in M$.
- S1 *If $\langle A, B \mid C \rangle \in M$, then $\langle B, A \mid C \rangle \in M$.*
- S2 *If $\langle A, B \mid C \rangle \in M$ and $D \subseteq A$, then $\langle D, B \mid C \rangle \in M$.*
- S3 *If $\langle A, B \mid C \rangle \in M$ and $D \subseteq A$, then $\langle A, B \mid C \cup D \rangle \in M$.*
- S4 *If $\langle A, B \mid C \rangle \in M$ and $\langle A, D \mid B \cup C \rangle \in M$, then $\langle A, B \cup D \mid C \rangle \in M$,*

Remark. Dawid (2001) actually defines a more general construction on a semilattice, but the above characterisation is sufficient for our purposes.

For each vertex $v \in V$, we define a random variable X_v on a sample space \mathcal{X}_v . Furthermore, for any $A \subseteq V$ we write the vector $X_A = (X_v)_{v \in A}$, the product space $\mathcal{X}_A = \prod_{v \in A} \mathcal{X}_v$, and $X = X_V$ and $\mathcal{X} = \mathcal{X}_V$.

A joint distribution P for X is *Markov* with respect to a separoid M if:

$$\langle A, B | C \rangle \in M \quad \Rightarrow \quad X_A \perp\!\!\!\perp X_B | X_C \quad [P].$$

Specifically, we will focus on the separoids induced by graphs (see Appendix A for the necessary graph terminology). We define the *separoid of an undirected graph* \mathcal{G} :

$$\mathcal{M}(\mathcal{G}) = \{ \langle A, B | C \rangle : A \text{ and } B \text{ are separated by } C \text{ in } \mathcal{G} \}. \quad (1.1)$$

The *separoid of a directed acyclic graph* \mathcal{G} is the set:

$$\mathcal{M}(\mathcal{G}) = \{ \langle A, B | C \rangle : A \text{ and } B \text{ are separated by } C \text{ in } \mathcal{G}_{\text{an}}^{\text{m}}(A \cup B \cup C) \}. \quad (1.2)$$

We say a distribution is Markov with respect to a graph, if it is Markov with respect to its separoid.

We note that properties S0–4 are constructive, that is, each specifies the existence of an element of the separoid. Therefore, by iteratively applying these properties to an arbitrary set N of such triples, we can generate all the elements of the smallest separoid containing N , which we the *separoid closure* of N , and denote by \overline{N} . Furthermore, we say N is a *spanning subset* of \overline{N} .

The link between the conditional independence properties C0–4, and the separoid properties S0–4, implies the following:

Lemma 1.2.1. *Let N be a spanning subset of the separoid M , and P be a distribution for X such that:*

$$\langle A, B | C \rangle \in N \quad \Rightarrow \quad X_A \perp\!\!\!\perp X_B | X_C \quad [P].$$

then P is Markov with respect to M .

The separoid of an undirected decomposable graph can be generated by the decompositions of the graph:

Theorem 1.2.2 (Dawid and Lauritzen 1993, Theorem 2.8). *Let \mathcal{G} be an undirected decomposable graph. Then the set:*

$$\{\langle A, B \mid A \cap B \rangle : (A, B) \text{ is a decomposition of } \mathcal{G}\}$$

is a separoid spanning set for $\mathcal{M}(\mathcal{G})$.

Similarly, the separoid of a directed acyclic graph can be generated by the parent sets of individual vertices:

Theorem 1.2.3 (Lauritzen, Dawid, et al. 1990, Propositions 4 and 5). *Let \mathcal{G} be a directed acyclic graph, with a well-ordering \prec of the vertices. Then the sets:*

$$\{\langle \{v\}, \text{nd}_{\mathcal{G}}(v) \mid \text{pa}_{\mathcal{G}}(v) \rangle : v \in V\} \quad (1.3)$$

$$\{\langle \{v\}, \text{pr}_{\prec}(v) \mid \text{pa}_{\mathcal{G}}(v) \rangle : v \in V\} \quad (1.4)$$

are separoid spanning sets for $\mathcal{M}(\mathcal{G})$.

The sets (1.3) and (1.4) are known as the *local* and *ordered Markov properties*.

We note that for any separoid M , there is a natural projection onto a subset $U \subseteq V$:

$$M_U = \{\langle A, B \mid C \rangle \in M : A, B, C \subseteq U\}$$

One interesting question is under what circumstances does this projection agree with the separoid of the induced subgraph, *i.e.* $\mathcal{M}_U(\mathcal{G}) = \mathcal{M}(\mathcal{G}_U)$?

Theorem 1.2.4 (Asmussen and Edwards 1983, Corollary 2.5). *Let \mathcal{G} be an undirected graph. Then $\mathcal{M}_U(\mathcal{G}) = \mathcal{M}(\mathcal{G}_U)$ if and only if \mathcal{G} is collapsible onto U .*

For the directed case, there is no known characterisation of such sets, however there is the following sufficient condition:

Theorem 1.2.5. *Let \mathcal{G} be a directed acyclic graph. If U is an ancestral set, then $\mathcal{M}_U(\mathcal{G}) = \mathcal{M}(\mathcal{G}_U)$.*

Another useful property is that the separoids of edge subgraphs are in fact larger:

Theorem 1.2.6. *Let $\mathcal{G}, \mathcal{G}'$ be either undirected or directed acyclic graphs on V . Then:*

$$\mathcal{E}(\mathcal{G}) \subseteq \mathcal{E}(\mathcal{G}') \quad \Rightarrow \quad \mathcal{M}(\mathcal{G}) \supseteq \mathcal{M}(\mathcal{G}').$$

In other words, if P is Markov with respect to \mathcal{G} , and \mathcal{G} is an edge subgraph of \mathcal{G}' , then P must also be Markov with respect to \mathcal{G}' .

1.3 Hyper Markov properties

We define a *model* to be a family of probability distributions Θ over a common measurable space. Specifically, we will focus on the case where $\Theta \subseteq \mathfrak{P}(\mathcal{G})$, the set of distributions that are Markov with respect to a graph \mathcal{G} .

For any $\theta \in \Theta$ and $A \subseteq V$, define θ_A to be the marginal distribution of X_A under θ . Moreover, for $A, B \subseteq V$, we define $\theta_{A|B}$ to be the family of conditional distributions of $(X_A | X_B = x_b)_{x_b}$. If we use $\phi \simeq \psi$ to denote the existence of a bijective function between ϕ and ψ , we note that for any $A \subseteq V$, (Dawid and Lauritzen 1993, Lemma 3.1)

$$\theta \simeq (\theta_A, \theta_{V \setminus A | A}).$$

A *law* \mathcal{L} is a probability distribution of a random distribution $\tilde{\theta}$ taking values in Θ . As we primarily focus on Bayesian methodology, we will use laws to describe the prior and posterior distributions for statistical models, though Dawid and Lauritzen (1993) also used laws in the context of sampling distributions of estimators.

A law $\mathcal{L}(\tilde{\theta})$ for an undirected graph \mathcal{G} is (*weak*) *hyper Markov* if for any decomposition (A, B) of \mathcal{G} :

$$\tilde{\theta}_A \perp\!\!\!\perp \tilde{\theta}_B \mid \tilde{\theta}_{A \cap B} \quad [\mathcal{L}] \quad (1.5)$$

We note that both weak hyper Markov properties may be characterised in terms of their separoids:

Theorem 1.3.1. *A $\mathcal{L}(\tilde{\theta})$ is weak hyper Markov with respect to an undirected decomposable or directed acyclic graph \mathcal{G} if and only if:*

$$\tilde{\theta}_{A \cup C} \perp\!\!\!\perp \tilde{\theta}_{B \cup C} \mid \tilde{\theta}_C \quad [\mathcal{L}] \quad (1.6)$$

for all $\langle A, B \mid C \rangle \in \mathcal{M}(\mathcal{G})$.

Proof. By Theorems 1.2.2 and 1.2.3, we simply need to show that there are analogous properties to properties S0–4: that is, for every $A, B, C, D \subseteq V$:

$$\text{H0} \quad \tilde{\theta}_{A \cup A} \perp\!\!\!\perp \tilde{\theta}_{B \cup A} \mid \tilde{\theta}_{A \cup A}.$$

$$\text{H1} \quad \text{If } \tilde{\theta}_{A \cup C} \perp\!\!\!\perp \tilde{\theta}_{B \cup C} \mid \tilde{\theta}_C, \text{ then } \tilde{\theta}_{B \cup C} \perp\!\!\!\perp \tilde{\theta}_{A \cup C} \mid \tilde{\theta}_C.$$

H2 If $\tilde{\theta}_{AUC} \perp\!\!\!\perp \tilde{\theta}_{BUC} \mid \tilde{\theta}_C$ and $D \subseteq A$, then $\tilde{\theta}_{DUC} \perp\!\!\!\perp \tilde{\theta}_{BUC} \mid \tilde{\theta}_C$.

H3 If $\tilde{\theta}_{AUC} \perp\!\!\!\perp \tilde{\theta}_{BUC} \mid \tilde{\theta}_C$ and $D \subseteq A$, then $\tilde{\theta}_{AUCUD} \perp\!\!\!\perp \tilde{\theta}_{BUCUD} \mid \tilde{\theta}_{CUD}$.

H4 If $\tilde{\theta}_{AUC} \perp\!\!\!\perp \tilde{\theta}_{BUC} \mid \tilde{\theta}_C$ and $\tilde{\theta}_{AUBUC} \perp\!\!\!\perp \tilde{\theta}_{DUBUC} \mid \tilde{\theta}_{BUC}$, then $\tilde{\theta}_{AUC} \perp\!\!\!\perp \tilde{\theta}_{DUBUC} \mid \tilde{\theta}_C$.

H0–2 and H4 follow immediately by the properties of conditional independence C0–4. To show H3, we note that $\theta_{CUD} \simeq (\theta_C, \theta_{D|C})$, and hence:

$$\tilde{\theta}_{AUCUD} \perp\!\!\!\perp \tilde{\theta}_{BUC} \perp\!\!\!\perp \tilde{\theta}_{DUC}$$

Furthermore, as $X_D \perp\!\!\!\perp X_B \mid X_C$, then we have $\theta_{BUCUD} \simeq (\theta_{BUC}, \theta_{D|C})$. \square

As a consequence of this and Theorem 1.2.6:

Corollary 1.3.2. *If \mathcal{L} is hyper Markov with respect to \mathcal{G} , and $\mathcal{E}(\mathcal{G}) \subseteq \mathcal{E}(\mathcal{G}')$, then \mathcal{L} is hyper Markov with respect to \mathcal{G}' .*

However, for much of the work in this thesis, we will utilise the stronger property:

Definition 1.3.1 (Strong hyper Markov property). A law $\mathcal{L}(\tilde{\theta})$ is *strong hyper Markov* with respect to an undirected decomposable graph \mathcal{G} if for any decomposition (A, B) of \mathcal{G} :

$$\tilde{\theta}_{B|A} \perp\!\!\!\perp \tilde{\theta}_A \quad [\mathcal{L}]. \quad (1.7)$$

A law $\mathcal{L}(\tilde{\theta})$ is *strong directed hyper Markov* with respect to a directed acyclic graph \mathcal{G} if for every vertex $v \in V$:

$$\tilde{\theta}_{v|\text{pa}(v)} \perp\!\!\!\perp \tilde{\theta}_{\text{nd}(v)} \quad [\mathcal{L}]. \quad (1.8)$$

Interestingly, there is no corresponding property to Corollary 1.3.2: if \mathcal{L} is strong hyper Markov with respect to \mathcal{G} , it need not be strong hyper Markov with respect to $\mathcal{G}' \supseteq \mathcal{G}$ (though it will still be weak hyper Markov), as we will see in Example 1.6.1 below.

One of the key benefits of strong hyper Markov laws is that when used as prior distributions in a Bayesian analysis, the posterior updating may be done locally:

Theorem 1.3.3 (Dawid and Lauritzen 1993, Corollary 5.5). *If the prior law $\mathcal{L}(\tilde{\theta})$ is strong hyper Markov, and x is a completely observed realisation of X , then the posterior $\mathcal{L}(\tilde{\theta} | X = x)$ is strong hyper Markov, and for any clique C*

$$\mathcal{L}(\tilde{\theta}_C | X = x) = \mathcal{L}(\tilde{\theta}_C | X_C = x_C)$$

1.4 Constructing hyper Markov laws

One of the convenient aspects of conditional independence is that it allows us to define distributions in a piecewise manner.

Theorem 1.4.1 (Dawid and Lauritzen 1993, Lemma 2.5). *Let Q be a distribution for X_A and R for X_B , such that $Q_{A \cap B} = R_{A \cap B}$, then there exists a unique distribution P for $X_{A \cup B}$ such that:*

- (i) $P_A = Q$,
- (ii) $P_B = R$, and
- (iii) $X_A \perp\!\!\!\perp X_B | X_{A \cap B} [P]$.

Importantly, we can apply the same procedure to laws:

Theorem 1.4.2 (Dawid and Lauritzen 1993, Lemma 3.3). *Let $\mathcal{M}(\tilde{\theta}_A)$ be a law for X_A , and $\mathcal{N}(\tilde{\theta}_B)$ be a law for X_B such that $\mathcal{M}_{A \cap B} = \mathcal{N}_{A \cap B}$. Then there exists a unique law $\mathcal{L}(\tilde{\theta}_{A \cup B})$ such that:*

- (i) $\mathcal{L}_A = \mathcal{M}$,
- (ii) $\mathcal{L}_B = \mathcal{N}$,
- (iii) $\tilde{\theta}_A \perp\!\!\!\perp \tilde{\theta}_B | \tilde{\theta}_{A \cap B} [\mathcal{L}]$, and
- (iv) $X_A \perp\!\!\!\perp X_B | X_{A \cap B} [\tilde{\theta}]$, almost surely under \mathcal{L} .

Dawid and Lauritzen (1993) term Q and R to be *consistent*, and P to be their *Markov combination*, denoted by the operation $P = Q \star R$. Likewise, \mathcal{M} and \mathcal{N} are *hyper consistent*, and \mathcal{L} is their *hyper Markov combination*, denoted by $\mathcal{L} = \mathcal{M} \odot \mathcal{N}$.

Both of these operations are products in a category-theoretic sense: they are defined uniquely by their marginal projections. This concept of a *conditional product* was explored by Dawid and Studený (1999), who explored its axioms and how they relate to those of conditional independence.

We can use these to construct Markov distributions and hyper Markov laws on undirected decomposable graphs by specifying marginal distributions and laws on cliques, and sequentially applying the above operations. Specifically, if $(P^{(C)})_{C \in \text{cl}(\mathcal{G})}$ are a set of pairwise consistent distributions, then the distribution:

$$\star_{C \in \text{cl}(\mathcal{G})} P^{(C)} \quad (1.9)$$

is Markov with respect to \mathcal{G} . Likewise, if $[\mathcal{L}^{(C)}(\tilde{\theta}_C)]_{C \in \text{cl}(\mathcal{G})}$ are a set of pairwise hyper consistent laws, the law:

$$\odot_{C \in \text{cl}(\mathcal{G})} \mathcal{L}^{(C)} \quad (1.10)$$

will be hyper Markov with respect to \mathcal{G} .

How does one obtain a set of consistent distributions or hyper consistent laws? One method is to simply take an arbitrary joint distribution or law for X_V , and take the marginal distributions or laws on the cliques; these will by necessity be (hyper) consistent

Finally, as much of our focus will be on strong hyper Markov laws, we would like to know under what conditions the resultant law will be strong hyper Markov:

Theorem 1.4.3 (Dawid and Lauritzen 1993, Proposition 3.16). *A weak hyper Markov law $\mathcal{L}(\tilde{\theta})$ is strong hyper Markov if and only if the marginal law $\mathcal{L}(\tilde{\theta}_C)$ for each clique C is strong hyper Markov.*

In other words, (1.10) is strong hyper Markov if and only if each $\mathcal{L}^{(C)}$ is strong hyper Markov with respect to the complete graph on C .

1.5 Variation independence and meta Markov properties

We can obtain similar properties by replacing the probabilistic independence with variation independence:

Definition 1.5.1. Let ϕ, ψ and ω be functions on a common domain D . Then we define the *conditional range* of ϕ given ω to be the image under ϕ of the fibre of ω :

$$\phi[\omega^{-1}(\cdot)].$$

Furthermore, we define ϕ to be conditionally variation independent of ψ given ω , written:

$$\phi \perp\!\!\!\perp \psi \mid \omega \quad [D]$$

if the conditional range of ϕ given (ψ, ω) is constant in ψ . That is, for all $v \in \psi(D), w \in \omega(D)$:

$$\phi[(\psi, \omega)^{-1}(v, w)] = \phi[\omega^{-1}(w)].$$

We note that the relation $(\cdot \perp\!\!\!\perp \cdot \mid \cdot)$ satisfies the same properties C0–4, and hence

If we replace the probabilistic independence of the hyper Markov properties with variation independence, we obtain *meta Markov* properties:

Definition 1.5.2. Let \mathcal{G} be an undirected graph. Then a model $\Theta \subseteq \mathfrak{P}(\Theta)$ is (*weak*) *meta Markov* if:

$$\theta_A \perp\!\!\!\perp \theta_B \mid \theta_{A \cap B} \quad [\Theta]$$

for all decompositions (A, B) of \mathcal{G} . Likewise, Θ is *strong meta Markov* if:

$$\theta_{B|A} \perp\!\!\!\perp \theta_A \quad [\Theta].$$

for all decompositions (A, B) of \mathcal{G} .

As we shall see in the next chapter, the variation independence has an important role in the properties of the maximum likelihood estimators, particularly when the data are obtained from different sampling regimes. Moreover, the support of weak/strong hyper Markov laws will be weak/strong meta Markov. As such, the lack of a meta Markov model will preclude the existence of a corresponding hyper Markov law.

1.6 Gaussian graphical models and Hyper inverse Wishart laws

One of the most common graphical models is the multivariate Gaussian graphical model, also called the covariance selection model (Dempster 1972; Wermuth 1976a). Let $X = \prod_{v \in V} X_v$, then:

$$\theta(X) = \mathcal{N}(0, \Sigma)$$

In particular, θ is Markov with respect to an undirected graph \mathcal{G} if:

$$\{u, v\} \notin \mathcal{E}(\mathcal{G}) \quad \Rightarrow \quad \Lambda_{uv} = 0 \quad (1.11)$$

where $\Lambda = \Sigma^{-1}$ is the *precision matrix*.

For any disjoint $A, B \subseteq V$, the marginal and conditional distributions are:

$$\theta(X_A) = \mathcal{N}(0, \Sigma_{AA}) \quad \text{and} \quad \theta(X_B | X_A = x_A) = \mathcal{N}(\Gamma_{B|A}x_A, \Sigma_{B|A})$$

where $\Gamma_{B|A} = \Sigma_{BA}\Sigma_{AA}^{-1}$, and $\Sigma_{B|A} = \Sigma_{BB} - \Sigma_{BA}\Sigma_{AA}^{-1}\Sigma_{BA}$ is the Schur complement. Therefore we may write:

$$\theta_A \simeq \Sigma_{AA} \quad \text{and} \quad \theta_{B|A} \simeq (\Gamma_{B|A}, \Sigma_{B|A})$$

The conjugate prior law for the complete model is the *inverse Wishart law*, $\mathcal{L}(\tilde{\theta}) = \mathcal{IW}(\delta; \Phi)$, using the notation of Dawid (1981). In particular, this law is strong hyper Markov on the complete graph, since for any disjoint $A, B \subseteq V$:

$$\tilde{\theta}_A \perp\!\!\!\perp \tilde{\theta}_{B|A} \quad [\mathcal{L}]$$

where:

$$\begin{aligned} \mathcal{L}(\tilde{\Sigma}_{AA}) &= \mathcal{IW}(\delta; \Phi_{AA}) \\ \mathcal{L}(\tilde{\Sigma}_{B|A}) &= \mathcal{IW}(\delta + |A|; \Phi_{B|A}) \\ \mathcal{L}(\tilde{\Gamma}_{B|A} | \tilde{\Sigma}_{B|A}) &= \Phi_{BA}\Phi_{AA}^{-1} + \mathcal{N}_{B \times A}(\tilde{\Sigma}_{B|A}, \Phi_{AA}^{-1}) \end{aligned}$$

The hyper Markov law constructed by the hyper Markov combination on the cliques (1.10) is termed the *hyper inverse Wishart*. Moreover, since any clique marginal law is strong hyper Markov with respect to the complete graph, then by Theorem 1.4.3, the hyper inverse Wishart is also strong hyper Markov.

Interestingly, this property is generally unique to the inverse Wishart law:

Theorem 1.6.1 (Geiger and Heckerman 2002, Theorem 7). *Let \mathcal{G} be the complete graph on 3 or more vertices. Then the law for the Gaussian graphical model on \mathcal{G} is strong hyper Markov if and only if it is an inverse Wishart law.*

However, if there are only two vertices in the clique, then the family of strong hyper Markov laws is slightly more general:

Theorem 1.6.2 (Geiger and Heckerman 2002, Theorem 12). *Let \mathcal{G} be the complete graph on 2 vertices. Then the law for the Gaussian graphical model on \mathcal{G} is strong hyper Markov if and only if it has a density on the precision space of the form:*

$$h(\Lambda_{12})|\Lambda|^{-\delta/2-2} \exp\{-\frac{1}{2} \text{tr}(\Phi\Lambda)\},$$

for some arbitrary function h .

These results, when combined with Theorem 1.4.3, severely limit the choice of possible strong hyper Markov laws for the Gaussian model.

For example, Letac and Massam (2007) and Rajaratnam, Massam, and Carvalho (2008) define a more general family of the hyper inverse Wishart law, which they term a ‘‘Type II Wishart’’. This law obeys the directed strong hyper Markov law (for a given perfect ordering of cliques), but due to the above results, will not generally be strong hyper Markov with respect to an undirected graph.

Also of interest is the corresponding law on the precision matrix:

Theorem 1.6.3 (Roverato 2000, Equation 13). *If $\mathcal{L}(\tilde{\Sigma}) = \mathcal{H}\mathcal{I}\mathcal{W}_{\mathcal{G}}(\delta; \Phi)$, then corresponding law for $\tilde{\Lambda} = \tilde{\Sigma}^{-1}$ has a density proportional to:*

$$|\Lambda|^{(\delta-2)/2} \exp\{-\frac{1}{2} \text{tr}(\Phi\Lambda)\}$$

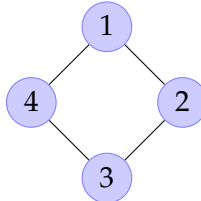
for $\Lambda \in \mathcal{P}_{\mathcal{G}(0)}^+$, the set of positive definite matrices satisfying (1.11).

In particular, we note that this is exactly proportional to the density of the Wishart law $\mathcal{W}(\delta + p - 1; \Phi^{-1})$, albeit concentrated on $\mathcal{P}_{\mathcal{G}(0)}^+$. This means that the hyper inverse Wishart law may be obtained from the inverse Wishart by appropriately conditioning on the precision matrix:

Corollary 1.6.4. *If $\mathcal{L}(\tilde{\Sigma}) = \mathcal{I}\mathcal{W}(\delta; \Phi)$, then:*

$$\mathcal{L}(\tilde{\Sigma} | \tilde{\Lambda}_{uv} = 0, \{u, v\} \notin \mathcal{E}(\mathcal{G})) = \mathcal{H}\mathcal{I}\mathcal{W}_{\mathcal{G}}(\delta; \Phi)$$

Corollary 1.6.4 suggests a way that we might extend the definition of the hyper inverse Wishart law to non-decomposable graphs. For example, with the graph:



we can define the hyper inverse Wishart law as the conditional law given $\lambda_{13} = \lambda_{24} = 0$. This approach was investigated by Roverato (2002) and Atay-Kayis and Massam (2005). However, we note that the corresponding weak hyper Markov property $\mathcal{M}(\mathcal{G})$, namely:

$$\tilde{\theta}_{\{1,2,3\}} \perp\!\!\!\perp \tilde{\theta}_{\{1,3,4\}} \mid \tilde{\theta}_{\{1,3\}} \quad \text{and} \quad \tilde{\theta}_{\{1,2,4\}} \perp\!\!\!\perp \tilde{\theta}_{\{2,3,4\}} \mid \tilde{\theta}_{\{2,4\}},$$

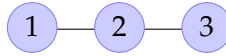
does not hold. Interestingly, there are still some strong hyper Markov-type properties, for example if C is a clique, then:

$$\tilde{\theta}_{V|C} \perp\!\!\!\perp \tilde{\theta}_C.$$

Furthermore, the normalisation constant of such a density generally does not have a closed-form solution.

Finally, we note that unlike the Markov and weak hyper Markov properties, if a graph is strong hyper Markov with respect to \mathcal{G} , it need not be strong hyper Markov with respect to \mathcal{G}' , where $\mathcal{E}(\mathcal{G}') \supseteq \mathcal{E}(\mathcal{G})$:

Example 1.6.1. Suppose we have the Gaussian model on 3 vertices, that is Markov with respect to the graph:



Then we can parameterise this model by the incomplete covariance matrix:

$$\Sigma^* = \begin{bmatrix} \sigma_{11} & \sigma_{12} & * \\ \sigma_{12} & \sigma_{22} & \sigma_{23} \\ * & \sigma_{23} & \sigma_{33} \end{bmatrix}.$$

In the completion of this matrix, the missing element $* = \sigma_{13} = \sigma_{12}\sigma_{23}/\sigma_{22}$.

We now investigate the partition of the parameters into $(\theta_{\{1,3\}}, \theta_{2|\{1,3\}})$. Note that:

$$|\sigma_{12}| \leq \sqrt{\sigma_{11}\sigma_{22}} \quad \text{and} \quad |\sigma_{23}| \leq \sqrt{\sigma_{22}\sigma_{33}} \quad \Rightarrow \quad |\sigma_{13}| \leq \sqrt{\sigma_{11}\sigma_{33}}$$

and so $\theta_{\{1,3\}} \simeq \Sigma_{\{1,3\}}$ may be any positive-semidefinite 2×2 matrix. Furthermore, the regression coefficients of 2 on $\{1, 3\}$ will be:

$$\begin{aligned} \Gamma_{2|\{1,3\}} &= \begin{bmatrix} \sigma_{12} \\ \sigma_{23} \end{bmatrix}^\top \begin{bmatrix} \sigma_{11} & \sigma_{12}\sigma_{23}/\sigma_{22} \\ \sigma_{12}\sigma_{23}/\sigma_{22} & \sigma_{33} \end{bmatrix}^{-1} \\ &= \frac{1}{\sigma_{11}\sigma_{22}\sigma_{33} - \sigma_{12}^2\sigma_{23}^2} [\sigma_{22}\sigma_{12}(\sigma_{22}\sigma_{33} - \sigma_{23}^2), \sigma_{22}\sigma_{23}(\sigma_{11}\sigma_{22} - \sigma_{12}^2)] \end{aligned}$$

Consider the following cases:

- If $\sigma_{13} = 0$, then $\Gamma_{2|\{1,3\}} = (\sigma_{12}/\sigma_{11}, \sigma_{23}/\sigma_{33})$. However this would also imply that either σ_{12} or $\sigma_{23} = 0$, and hence $\Gamma_{2|\{1,3\}}$ can only take values on the axes $\mathbb{R} \times \{0\} \cup \{0\} \times \mathbb{R}$.
- If $\sigma_{13} > 0$, then it follows that $\sigma_{12}\sigma_{23} > 0$, and by examining the terms of the above expression, it follows that $\Gamma_{2|\{1,3\}}$ can only take values in the quadrants $(\mathbb{R}_{>0})^2 \cup (\mathbb{R}_{<0})^2$
- If $\sigma_{13} < 0$, then $\sigma_{12}\sigma_{23} < 0$, and by similar inspection, $\Gamma_{2|\{1,3\}}$ can only take values in the quadrants $(\mathbb{R}_{>0} \times \mathbb{R}_{<0}) \cup (\mathbb{R}_{<0} \times \mathbb{R}_{>0})$.

As the range of $\Gamma_{2|\{1,3\}}$ depends on the value of σ_{13} , then $\theta_{2|\{1,3\}}$ and $\theta_{\{1,3\}}$ are not variation independent. Therefore the model cannot be strong meta Markov on the complete graph, nor can any law with full support, such as the hyper inverse Wishart, be strong hyper Markov on this graph.

1.7 Contingency tables and Hyper Dirichlet laws

The other common graphical model is the contingency table. If the variable X is discrete sample space \mathcal{X}_v is finite, then each cell $x = (x_v)_{v \in V} \in \mathcal{X}$ will have probability $\theta(x)$.

Such models are usually parameterised in log-linear form (see, for example, Darroch, Lauritzen, and Speed 1980), but for our purposes it is easier to work with the clique-marginal distributions. Specifically, for any clique C , we let $\theta(x_C)$ be the marginal probability of any cell $x_C \in \mathcal{X}_C$.

Then the probability of any cell $x \in \mathcal{X}$ is:

$$\theta(x) = \frac{\prod_{C \in \text{cd}(\mathcal{G})} \theta(x_C)}{\prod_{S \in \text{sep}(\mathcal{G})} \theta(x_S)^{\nu_{\mathcal{G}}(S)}}$$

where $\nu_{\mathcal{G}}(S)$ denotes the multiplicity of a separator S .

The standard conjugate prior on the complete graph is the *Dirichlet law* $\mathcal{L}(\tilde{\theta}) = \mathcal{D}(\alpha)$, where the parameter $\alpha : \mathcal{X} \rightarrow \mathbb{R}_{>0}$. This is a strong hyper Markov law, and:

$$\begin{aligned} \mathcal{L}(\tilde{\theta}_A) &= \mathcal{D}(\alpha_A) \\ \mathcal{L}(\tilde{\theta}_{B|A}(\cdot | x_A)) &= \mathcal{D}(\alpha_{A \cup B}(\cdot, x_A)) \end{aligned}$$

where $\alpha_A(x_A) = \sum_{x': x'_A = x_A} \alpha(x')$

The hyper Markov combination of a set of consistent Dirichlet laws is the *hyper Dirichlet law*, which by Theorem 1.4.3, must also be strong hyper Markov.

Interestingly, this is not the only strong hyper Markov law. Consider a 2×2 table, with $\theta_{xy} = \theta(x, y)$. Geiger and Heckerman (1997, equation 10) note that a law is strong hyper Markov if and only if it has a density of the form:

$$h \left(\frac{\theta_{00}\theta_{11}}{\theta_{01}\theta_{10}} \right) \theta_{00}^{\alpha_{00}-1} \theta_{01}^{\alpha_{01}-1} \theta_{10}^{\alpha_{10}-1} \theta_{11}^{\alpha_{11}-1} \quad (1.12)$$

for some function h . In the case where h is constant, this is simply a Dirichlet distribution.

We can extend sufficient condition to larger tables:

Theorem 1.7.1. *If a law $\mathcal{L}(\tilde{\theta})$ for a 2-way contingency table $X \times Y$ on $\mathcal{X} \times \mathcal{Y}$ has density:*

$$h \left(\left[\frac{\theta_{xy}\theta_{x^*y^*}}{\theta_{xy^*}\theta_{x^*y}} \right]_{x \neq x^*, y \neq y^*} \right) \prod_{x,y} \theta_{xy}^{\alpha_{xy}-1}, \quad (1.13)$$

then it is strong hyper Markov.

Proof. The Jacobian determinant of the transformation $\theta_{xy} \mapsto (\theta_{+y}, \theta_{x|y})$ is:

$$\left| \frac{d\theta_{xy}}{d(\theta_{+y}, \theta_{x|y})} \right| = \prod_y \theta_{+y}^{|\mathcal{X}|-1}$$

(see, for example, Heckerman, Geiger, and Chickering 1995, Theorem 10), which gives the joint density for $(\theta_{+y}, \theta_{x|y})$:

$$\prod_y \theta_{+y}^{\alpha_{+y}-1} h \left(\left[\frac{\theta_{x|y}\theta_{x^*|y^*}}{\theta_{x|y^*}\theta_{x^*|y}} \right]_{x \neq x^*, y \neq y^*} \right) \prod_{x,y} \theta_{x|y}^{\alpha_{xy}-1} \quad (1.14)$$

which factorises into a term involving only θ_{+y} terms, and another involving only $\theta_{x|y}$ terms. Therefore $\tilde{\theta}_Y \perp\!\!\!\perp \tilde{\theta}_{X|Y}$. By symmetry, the same argument holds in the other direction. \square

It is unclear if the converse is true: the corresponding result in (1.12) relies on results from functional equations, and it is unclear if these arguments can be extended directly to higher dimensions.

The form of the density in Theorem 1.7.1 assumes the law has full support. However, as we shall demonstrate in the next chapter, we can obtain

strong hyper Markov laws on submanifolds by conditioning on the odds-ratio parameter. Furthermore, the form of the density in (1.14) gives the following:

Corollary 1.7.2. *If a law $\mathcal{L}(\tilde{\theta})$ satisfies the conditions of Theorem 1.7.1, the marginal laws are:*

$$\mathcal{L}(\tilde{\theta}_X) = \mathcal{D}(\alpha_X) \quad \text{and} \quad \mathcal{L}(\tilde{\theta}_Y) = \mathcal{D}(\alpha_Y).$$

Example 1.7.1. One way to construct such a law is through a mixture of Dirichlet laws $\mathcal{L}(\tilde{\theta} | \tilde{\alpha}) = \mathcal{D}(\tilde{\alpha})$, where the law for the mixing parameter $\mathcal{L}(\tilde{\alpha})$ has constant marginals:

$$\mathcal{L}(\tilde{\alpha}_{x+} = a_{x+}) = 1 \quad \text{and} \quad \mathcal{L}(\tilde{\alpha}_{+y} = a_{+y}) = 1.$$

By the properties of the Dirichlet law, we have:

$$\tilde{\theta}_Y \perp\!\!\!\perp \tilde{\theta}_{X|Y} | \tilde{\alpha} \quad \text{and} \quad \tilde{\theta}_X \perp\!\!\!\perp \tilde{\theta}_{Y|X} | \tilde{\alpha} \quad [\mathcal{L}].$$

Furthermore, the constant marginal laws imply that $\tilde{\theta}_Y \perp\!\!\!\perp \tilde{\alpha}$ and $\tilde{\theta}_X \perp\!\!\!\perp \tilde{\alpha}$, and so:

$$\tilde{\theta}_Y \perp\!\!\!\perp (\tilde{\theta}_{X|Y}, \tilde{\alpha}) \quad \text{and} \quad \tilde{\theta}_X \perp\!\!\!\perp (\tilde{\theta}_{Y|X}, \tilde{\alpha}) \quad [\mathcal{L}].$$

Therefore $\mathcal{L}(\tilde{\theta})$ is strong hyper Markov. Now define $a_{xy} = a_{x+}a_{+y}/a_{++}$, and:

$$\tilde{\eta}_{xy} = \tilde{\alpha}_{xy} - a_{xy}, \quad x \neq x^*, y \neq y^*$$

Furthermore, note that:

$$\begin{aligned} \tilde{\alpha}_{x^*y} &= \tilde{\alpha}_{+y} - \sum_{x \neq x^*} \tilde{\alpha}_{xy} = a_{x^*y} - \sum_{x \neq x^*} \tilde{\eta}_{xy}, \quad y \neq y^* \\ \tilde{\alpha}_{xy^*} &= \tilde{\alpha}_{x+} - \sum_{y \neq y^*} \tilde{\alpha}_{xy} = a_{xy^*} - \sum_{y \neq y^*} \tilde{\eta}_{xy}, \quad x \neq x^* \\ \tilde{\alpha}_{x^*y^*} &= \tilde{\alpha}_{+y^*} - \sum_{x \neq x^*} \tilde{\alpha}_{xy^*} = a_{x^*y^*} + \sum_{x \neq x^*, y \neq y^*} \tilde{\eta}_{xy}. \end{aligned}$$

Therefore, $\tilde{\eta}$ completely characterises the mixing vector. $\mathcal{L}(\tilde{\theta})$ has a density of the form:

$$\pi(\theta) = \mathbb{E}_{\mathcal{L}(\tilde{\alpha})}[\pi(\theta|\tilde{\alpha})] = \mathbb{E}_{\tilde{\alpha}} \left[\frac{1}{B(\tilde{\alpha})} \prod_{x,y} \theta_{xy}^{\tilde{\alpha}_{xy}-1} \right].$$

This may be re-expressed as:

$$\pi(\theta) = \prod_{x,y} \theta_{xy}^{a_{xy}-1} \mathbb{E}_{\mathcal{L}(\tilde{\alpha}, \tilde{\eta})} \left[\frac{1}{B(\tilde{\alpha})} \prod_{x \neq x^*, y \neq y^*} \left(\frac{\theta_{xy} \theta_{x^*y^*}}{\theta_{xy^*} \theta_{x^*y}} \right)^{\tilde{\eta}_{xy}} \right] \quad (1.15)$$

which is of the same form as the density in Theorem 1.7.1.

Unfortunately, for most functions h , the normalisation constant of the density (1.13) will usually not have an analytic form. Even for Example 1.7.1, the occurrence of the beta function inside the integral in (1.15) will usually mean that the precise form of the density is intractable, other than for finite mixtures.

Theorem 1.7.1 may be extended to higher-order tables, but the arbitrary function h is limited to the highest order interaction term:

Theorem 1.7.3. *If a law $\mathcal{L}(\tilde{\theta})$ for an n -way contingency table $X = \prod_{v \in V} X_v$ has density of the form:*

$$h \left(\left[\prod_{B \subseteq V} \theta(x_B^*, x_{V \setminus B})^{(-1)^{|V \setminus B|}} \right]_{x: x_v \neq x_v^*} \right) \prod_x \theta(x)^{\alpha(x) - 1} \quad (1.16)$$

then it is strong hyper Markov.

Proof. For any $\emptyset \subset A \subset V$, let $A^c = V \setminus A$, and note that the first product term in (1.16) may be written as:

$$\prod_{C \subseteq A} \prod_{D \subseteq A^c} \theta(x_C^*, x_{A \setminus C}, x_D^*, x_{A^c \setminus D})^{(-1)^{|A \setminus C| + |A^c \setminus D|}}.$$

This may be rewritten as:

$$\prod_{C \subseteq A} \left[\theta(x_C^*, x_{A \setminus C}, x_{A^c}^*) \prod_{D \subseteq A^c} \theta(x_C^*, x_{A \setminus C}, x_D^*, x_{A^c \setminus D})^{(-1)^{|A^c \setminus D|}} \right]^{(-1)^{|A \setminus C|}} \quad (1.17)$$

Recall that any finite, non-empty set has an equal number of even and odd size subsets, therefore:

$$\sum_{D \subseteq A^c} (-1)^{|A^c \setminus D|} = -1,$$

and so (1.17) may be expressed as:

$$\prod_{C \subseteq A} \left[\prod_{D \subseteq A^c} \left(\frac{\theta(x_C^*, x_{A \setminus C}, x_D^*, x_{A^c \setminus D})}{\theta(x_C^*, x_{A \setminus C}, x_{A^c}^*)} \right)^{(-1)^{|A^c \setminus D|}} \right]^{(-1)^{|A \setminus C|}}$$

By the same argument over C , we obtain:

$$\prod_{C \subseteq A} \prod_{D \subseteq A^c} \left(\frac{\theta(x_C^*, x_{A \setminus C}, x_D^*, x_{A^c \setminus D})}{\theta(x_C^*, x_{A \setminus C}, x_{A^c}^*)} \frac{\theta(x_A^*, x_{A^c}^*)}{\theta(x_A^*, x_D^*, x_{A^c \setminus D})} \right)^{(-1)^{|A \setminus C| + |A^c \setminus D|}}$$

Note that the term inside the parenthesis is of the same form as the fraction in (1.13), and hence satisfies the conditions of Theorem 1.7.1 (with $x = x_A$ and $y = x_{A^c}$), and therefore $\tilde{\theta}_A \perp\!\!\!\perp \tilde{\theta}_{V \setminus A}$. \square

Furthermore, this link to Theorem 1.7.1 means that we have a similar result to Corollary 1.7.2:

Corollary 1.7.4. *If a law $\mathbb{L}(\tilde{\theta})$ satisfies Theorem 1.7.3, then for any $A \subset V$, the marginal law $\mathbb{L}(\tilde{\theta}_A) = \mathcal{D}(\alpha_A)$.*

1.8 Notes and other developments

The Dirichlet process (Ferguson 1973) is a law on an arbitrary measurable space, and may be regarded as an infinite-dimensional extension of the Dirichlet law. It has many interesting properties, notably that the resulting probability measure is almost surely discrete.

Asci, Nappo, and Piccioni (2006) and Heinz (2009) independently develop the *hyper Dirichlet process*, defined as a hyper Markov combination of consistent Dirichlet processes on the cliques. Unlike the hyper Dirichlet law however, it is generally not strong hyper Markov. In fact, under quite general conditions—such as the base measures being continuous and the graph \mathcal{G} being connected—it will simply be a Dirichlet process whose base measure is the Markov combination of the clique base measures, in other words:

$$\bigodot_{C \in \text{cl}(\mathcal{G})} \mathcal{D}\mathcal{P}(\mu_C, A) = \mathcal{D}\mathcal{P}\left(\bigstar_{C \in \text{cl}(\mathcal{G})} \mu_C, A\right).$$

This is due to the inherent discrete nature of the Dirichlet process. If θ is drawn from a Dirichlet process $\mathcal{D}\mathcal{P}(\mu, A)$ on some product space $\mathcal{X} \times \mathcal{Y} \times \mathcal{Z}$, then it will (almost surely) have a representation of the form:

$$\theta = \sum_i a_i \delta_{(x_i, y_i, z_i)}$$

where the coordinates (x_i, y_i, z_i) are drawn i.i.d. from μ , and δ is the Dirac measure. If μ is continuous, then the probability of there being two distinct coordinates (x_i, y_i, z_i) and (x_j, y_j, z_j) such that $y_i = y_j$ will be zero. Therefore, if a triple (X, Y, Z) is drawn from θ , then $\theta(X = x_i, Z = z_i \mid Y = y_i) = 1$, and hence, somewhat trivially, we have the Markov property:

$$X \perp\!\!\!\perp Z \mid Y \quad [\theta].$$

Finally, we note that although the strong hyper Markov property is very restrictive, such laws can form useful building blocks in constructing more

general laws. One common approach is to specify the law in an hierarchical manner, by specifying a family of strong hyper Markov laws as well as a mixing distribution over this family (often called a *hyperprior*).

We note that the resultant marginal law will usually *not* be strong, or even weak, hyper Markov (Example 1.7.1 being an exception). Nevertheless, such an approach can still be advantageous, as the conditional independence properties may still be exploited for computational purposes.

For example, suppose that $\mathcal{L}(\tilde{\theta} | \tilde{\alpha})$ is a family of hyper Dirichlet laws, and $\mathcal{L}(\tilde{\alpha})$ describes the mixing law. Then by exploiting the conditional hyper Markovity and the fact that:

$$X \perp\!\!\!\perp \tilde{\alpha} | \tilde{\theta} \quad [\mathcal{L}]$$

we could construct a Markov chain Monte Carlo algorithm to obtain a sample from the posterior by alternating the following steps:

- a) For each clique C_i , independently sample:

$$\theta_{C_i|S_i}^{(n+1)} \sim \mathcal{L}(\tilde{\theta}_{C_i|S_i} | \alpha^{(n)}, X_C = x_c),$$

where S_i is the i th separator in a perfect ordering C_1, \dots, C_k .

- b) Sample:

$$\alpha^{(n+1)} \sim \mathcal{L}(\tilde{\alpha} | \theta^{(n+1)}).$$

In this case, step (a) may be performed in parallel on up to k processors. This is particularly useful in the case where the evaluation of the likelihood function is computationally intensive. Furthermore, each processor would only require the data X_{C_i} of the corresponding clique.

Logistic regression and case-control studies

An interesting application of the meta Markov and hyper Markov properties arises in analysis of case-control studies.

If one wishes to determine particular risk factors for a disease (or any other binary outcome), there are two basic approaches:

Prospective or cohort study Select subjects from the population based on their risk factors, and observe them at the end of a fixed time period to determine if the disease arises.

Case-control or retrospective study Choose a random sample of subjects from the population with the disease (cases), and another sample from the population without (controls). Compare the relative frequencies of the risk factors in the two samples.

Let Y be the response variable taking values in $\{0, 1\}$, corresponding to the absence or presence of disease, respectively (the following results may be extended to the multinomial case, but for sake of simplicity we only pursue the binomial). Let X be the covariates (risk factors) taking values in $\mathcal{X} \subseteq \mathbb{R}^k$. In a prospective study we are observing the conditional distribution of Y given X , and so under a proportional odds assumption, we obtain the model for logistic regression:

$$p(y | x, \alpha, \beta) = \frac{e^{y(\alpha + \beta^\top x)}}{1 + e^{\alpha + \beta^\top x}}, \quad \alpha \in \mathbb{R}, \beta \in \mathbb{R}^k. \quad (2.1)$$

On the other hand, a case-control study will result in observations from the conditional distribution of X given Y . In this case, specifying a probabilistic model becomes much more difficult, particularly if \mathcal{X} is infinite.

Despite these difficulties, case-control studies are often desirable—or in some cases unavoidable—particularly where the disease is relatively rare or

the time until diagnosis may be particularly long, as the costs of obtaining a sufficient sample size for a prospective study are likely to be prohibitive.

The classic result of Prentice and Pyke (1979) showed that the maximum likelihood estimate and asymptotic covariance for the log-odds ratio parameter β could simply be found by logistic regression. In other words, we can use the prospective model to analyse data gathered retrospectively. This particular result has been widely applied in epidemiology and other areas.

In this chapter, we identify the analogous result for the Bayesian case: that is, the conditions under which the posterior distribution for β may be computed using the prospective likelihood instead of the retrospective.

The simplest model of a single binary covariate has been well explored in literature: Zelen and Parker (1986), Nurminen and Mutanen (1987), Marshall (1988) and Ashby, Hutton, and McGee (1993) have all characterised such an analysis, which consists of computing the posterior log odds ratio of a 2×2 contingency table under a Dirichlet prior.

In the case where the covariates are categorical, that is where \mathcal{X} is finite, Seaman and Richardson (2004) identified a class of improper priors that satisfy the desired properties. This class was further expanded by Staicu (2010).

We show that the basis of this prospective–retrospective symmetry is due to “independence” of the parameters: the original result of Prentice and Pyke (1979) can be explained through the variation independence of the parameter space, and that the corresponding Bayesian result will occur when the prior law exhibits analogous probabilistic independence. Furthermore, we arrive at the same class of prior laws as Staicu (2010) via a different route, and demonstrate how they might be extended to stratified designs.

However it should be noted that this is not the only approach for Bayesian analysis of case-control data. With the advent of computational tools such as MCMC, the retrospective likelihood need not present such an obstacle. Indeed this path has been well followed in the literature, as reviewed in Mukherjee, Sinha, and Ghosh (2005). For example, Müller and Roeder (1997), Seaman and Richardson (2001) and Gustafson, Le, and Vallée (2002) have pursued this approach. In particular, Gustafson, Le, and Vallée (2002) note that in general the prospective posterior can serve as a useful approximation to the retrospective posterior, and use this as the basis of an importance sampling scheme.

2.1 Notation and definitions

Throughout the chapter, (X, Y) will be a single joint observation from the specified model, and $(X^{(n)}, Y^{(n)})$ to be a sequence of n such independent observations. p will be the density of the model (with respect to the appropriate measure), with variables indicating the context.

Lemma 2.1.1. *For the above logistic model, we have:*

$$\theta_{Y|X} \simeq (\alpha, \beta) \quad \text{and} \quad \theta_{X|Y} \simeq (\theta_{X|Y=0}, \beta) \quad (2.2)$$

Proof. The first is determined by (2.1), and the second by Bayes theorem:

$$\frac{d\theta_{X|Y=1}}{d\theta_{X|Y=0}}(x) = \frac{\theta_{Y|X=x}(1) \theta_Y(0)}{\theta_{Y|X=x}(0) \theta_Y(1)} \propto e^{\beta^\top x} \quad \square$$

2.2 Maximum likelihood estimators

Prentice and Pyke (1979) showed that the maximum likelihood odds-ratios obtained from a case-control study have the same properties as those arising from a prospective study, and hence may be found via logistic regression. This can be elegantly demonstrated by the strong meta Markov property.

Lemma 2.2.1. *Let Θ_X be the family of all probability distributions over \mathcal{X} , and let $\Theta_{Y|X}$ be the family of conditional distributions with densities of the form in (2.1). Then the corresponding family of joint distributions Θ is strong meta Markov, that is:*

$$\theta_X \ddagger (\alpha, \beta) \quad \text{and} \quad \theta_Y \ddagger (\theta_{X|Y=0}, \beta)$$

Proof. These properties are essentially a reformulation of Müller and Roeder (1997, Lemmas 1 and 2). By definition $\theta_X \ddagger \theta_{Y|X}$. It remains to show variation independence in the opposite direction.

For any θ_X and $\theta_{Y|X}$, the joint distribution θ has a density of the form:

$$p(x, y | \theta) = \frac{e^{y(\alpha + \beta^\top x)}}{1 + e^{\alpha + \beta^\top x}} p(x | \theta_X) \quad (2.3)$$

Therefore the marginal distribution θ_Y is Bernoulli, with parameter γ taking values on the interval $(0, 1)$, where:

$$\gamma = p(y = 1 | \theta_Y) = \int_{\mathcal{X}} \frac{e^{\alpha + \beta^\top x}}{1 + e^{\alpha + \beta^\top x}} p(x | \theta_X) dx \quad (2.4)$$

and the conditional distribution of X given Y has density of the form:

$$p(x|y, \theta_{X|Y}) = \frac{p(x, y | \theta)}{\gamma^y (1 - \gamma)^{1-y}} = \frac{e^{y(\alpha - \log \frac{\gamma}{1-\gamma} + \beta^\top x)}}{(1 - \gamma)(1 + e^{\alpha + \beta^\top x})} p(x | \theta_X). \quad (2.5)$$

Now for any $\gamma' \in (0, 1)$, we may define $\theta' \simeq (\theta'_X, \theta'_{Y|X})$, where:

$$\theta'_{Y|X} \simeq (\alpha', \beta) \in \Theta_{Y|X} \quad \text{such that} \quad \alpha' = \alpha - \log \frac{\gamma}{1 - \gamma} + \log \frac{\gamma'}{1 - \gamma'}, \quad (2.6)$$

and θ'_X has density:

$$p(x | \theta'_X) = \frac{(1 - \gamma')(1 + e^{\alpha' + \beta^\top x})}{(1 - \gamma)(1 + e^{\alpha + \beta^\top x})} p(x | \theta_X)$$

By the definition of γ in (2.4), it can be shown that this integrates to 1, hence $\theta'_X \in \Theta_X$. Furthermore, by matching terms in (2.5), then $\theta_{X|Y} = \theta'_{X|Y}$. Since $\theta'_Y \simeq \gamma'$ may be chosen arbitrarily, it follows that $\theta_Y \not\perp \theta_{Y|X}$. \square

The logistic model has other variation independence properties:

Corollary 2.2.2. *Under the logistic model of Lemma 2.2.1, then:*

$$(\theta_X, \theta_Y) \not\perp \beta$$

Proof. We have $\theta_X \not\perp (\alpha, \beta)$, and for any θ_Y , we can choose α' as in (2.6). \square

Theorem 2.2.3. *Suppose we have a joint model as in Lemma 2.2.1. Then the profile likelihood function for the odds ratio β is the same for both the retrospective model $\Theta_{X|Y}$ and the prospective model $\Theta_{Y|X}$, up to proportionality.*

Proof. This proof follows a similar argument as Dawid and Lauritzen (1993, Lemma 4.10). The joint density for the model θ may be written as:

$$p(x, y | \theta) = p(x | \theta_X) p(y | x, \alpha, \beta) = p(y | \theta_Y) p(x | y, \theta_{X|Y=0}, \beta) \quad (2.7)$$

Therefore the profile likelihood for the joint model may be written in terms of the prospective model:

$$L_p^{\text{joint}}(\beta) = \max_{\alpha, \theta_X} p(x | \theta_X) p(y | x, \theta_{Y|X}) \quad (2.8)$$

By Lemma 2.2.1, the variation independence α and θ_X the factors of (2.8) may be profiled separately, and hence:

$$L_p^{\text{joint}}(\beta) \propto \max_{\alpha} p(y | x, \alpha, \beta) = L_p^{\text{pro}}(\beta)$$

where L_p^{pro} denotes the profile likelihood of the prospective model. The same argument applies to the retrospective profile likelihood $L_p^{\text{ret}}(\beta)$:

$$L_p^{\text{joint}}(\beta) \propto \max_{\theta_{X|Y=0}} p(x | y, \theta_{X|Y=0}, \beta) = L_p^{\text{ret}}(\beta) \quad \square$$

From this we obtain the result of Prentice and Pyke (1979):

Corollary 2.2.4. *For data observed in a case control study, the maximum likelihood estimate of the log odds parameter $\hat{\beta}$ and its asymptotic covariance may be computed as if the data were observed prospectively, that is, using logistic regression.*

Proof. The maximum likelihood estimator is a function of the profile likelihood, as is the asymptotic covariance (see Patefield 1985). \square

The same argument may be extended trivially to any penalised logistic regression estimator of the form:

$$\arg \max_{\alpha, \beta} [\log p(y | x, \alpha, \beta) + \phi(\beta)].$$

Examples of such estimators include ridge regression, where $\phi(\beta) \propto \|\beta\|_2$, and lasso, where $\phi(\beta) \propto \|\beta\|_1$. Such methods have proven successful in genome-wide association studies (GWAS), which involve case-control data with extremely high-dimensional covariates (Park and Hastie 2008; Wu et al. 2009).

2.3 Bayesian analysis of case-control studies

We now investigate how these results correspond to a Bayesian analysis. We will use π to denote the density of the prior law, and π^{pro} and π^{ret} to denote the densities of the posterior laws \mathcal{L}^{pro} and \mathcal{L}^{ret} under prospective and retrospective likelihoods, respectively:

$$\begin{aligned} \pi^{\text{pro}}(\alpha, \beta | x^{(n)}, y^{(n)}) &\propto \pi(\alpha, \beta) p(y^{(n)} | x^{(n)}, \alpha, \beta) \\ \pi^{\text{ret}}(\theta_{X|Y=0}, \beta | x^{(n)}, y^{(n)}) &\propto \pi(\theta_{X|Y=0}, \beta) p(x^{(n)} | y^{(n)}, \theta_{X|Y=0}, \beta) \end{aligned}$$

Furthermore, we will use \bar{p} to denote the density of the *marginal model*, where parameters have been integrated out (using the prior law), for example:

$$\bar{p}(y^{(n)} | x^{(n)}, \beta) = \int p(y^{(n)} | x^{(n)}, \alpha, \beta) \pi(\alpha | \beta) d\alpha \quad (2.9)$$

In other words, when interpreted as a function of β , $\bar{p}(y^{(n)}|x^{(n)}, \beta)$ is the *marginal likelihood* for β .

We now present the key result of this section:

Theorem 2.3.1. *Let $\mathcal{L}(\tilde{\theta})$ be a prior law for the joint parameters of the logistic model. Then the posterior marginal law for $\tilde{\beta}$ is the same under both prospective and retrospective likelihood for all possible observations $(x^{(n)}, y^{(n)})$, if and only if:*

$$\tilde{\beta} \perp\!\!\!\perp \tilde{\theta}_X \quad \text{and} \quad \tilde{\beta} \perp\!\!\!\perp \tilde{\theta}_Y \quad [\mathcal{L}] \quad (2.10)$$

Proof. Firstly, note that the marginal posterior densities for $\tilde{\beta}$ may be written as:

$$\begin{aligned} \pi^{\text{pro}}(\beta | x^{(n)}, y^{(n)}) &\propto \pi(\beta) \bar{p}(y^{(n)} | x^{(n)}, \beta) \\ \pi^{\text{ret}}(\beta | x^{(n)}, y^{(n)}) &\propto \pi(\beta) \bar{p}(x^{(n)} | y^{(n)}, \beta) \end{aligned}$$

where \bar{p} denotes the marginal model. Hence the marginal posteriors are equal if and only if the retrospective and prospective marginal likelihoods for β are proportional (for $\pi(\beta) > 0$). In other words, whenever there exists a function k such that:

$$\bar{p}(x^{(n)} | y^{(n)}, \beta) = \bar{p}(y^{(n)} | x^{(n)}, \beta) k(x^{(n)}, y^{(n)}). \quad (2.11)$$

These models are also related through the joint model:

$$\bar{p}(x^{(n)} | y^{(n)}, \beta) \bar{p}(y^{(n)} | \beta) = \bar{p}(y^{(n)} | x^{(n)}, \beta) \bar{p}(x^{(n)} | \beta),$$

therefore (2.11) is equivalent to:

$$\bar{p}(x^{(n)} | \beta) = \bar{p}(y^{(n)} | \beta) k(x^{(n)}, y^{(n)}). \quad (2.12)$$

Since $X^{(n)} \perp\!\!\!\perp \tilde{\beta} | \tilde{\theta}_X$, we may write the marginal model for $X^{(n)} | \tilde{\beta}$ as:

$$\bar{p}(x^{(n)} | \beta) = \int_{\Theta_X} \left[\prod_{i=1}^n p(x_i | \theta_X) \right] \pi(\theta_X | \beta) d\theta_X \quad (2.13)$$

Therefore, if $\tilde{\theta}_X \perp\!\!\!\perp \tilde{\beta}$, then $\bar{p}(x^{(n)} | \beta)$ must be constant in β , and the same for $\bar{p}(x^{(n)} | \beta)$ if $\tilde{\theta}_Y \perp\!\!\!\perp \tilde{\beta}$, hence (2.10) implies (2.12).

To show the converse, suppose that (2.12) holds for all values of $(x^{(n)}, y^{(n)})$. As $\bar{p}(x^{(n)} | \beta)$ is a density, it must be proportional to $k(x^{(n)}, y_0^{(n)})$, for any fixed $y_0^{(n)}$, and so $X^{(n)}$ is independent of $\tilde{\beta}$.

Note that $\bar{p}(x^{(n)} | \beta)$ is the density of a mixture of i.i.d. variables, and recall that the mixing measure of any infinite sequence is almost surely unique (see, for example, Aldous 1985, Lemma 2.15). As (2.13) must hold for all possible values of $x^{(n)}$, and n may be arbitrarily large, it follows that $\pi(\theta_X | \beta)$ must also be invariant of β , and hence $\tilde{\theta}_X \perp\!\!\!\perp \tilde{\beta}$. The same argument holds for $\tilde{\theta}_Y$. \square

Several authors have identified similar results. Notably, Müller and Roeder (1997) appear to have almost identified the conditions in (2.10), but then incorrectly claim that the “argument about the retrospective likelihood only carries over to posterior inference on β if α and β are independent and θ_X is not otherwise constrained”. This misconception appears to be due to the fact that although there is a one-to-one mapping between α and θ_Y , this mapping is itself dependent on β , through (2.4). Unfortunately, this means that the Dirichlet process mixture they propose does not satisfy the required properties.

Example 2.3.1. A simple example of a law $\mathcal{L}(\tilde{\theta})$ satisfying Theorem 2.3.1 would be any with the property:

$$(\tilde{\theta}_X, \tilde{\theta}_Y) \perp\!\!\!\perp \tilde{\beta} \quad [\mathcal{L}].$$

One method of constructing such a law would be to take two arbitrary laws $\mathcal{L}_m(\tilde{\theta})$ and $\mathcal{L}_o(\tilde{\theta})$, and take \mathcal{L} to be the product law of their projections $\mathcal{L}_m(\tilde{\theta}_X, \tilde{\theta}_Y)$ and $\mathcal{L}_o(\tilde{\beta})$. By Corollary 2.2.2, there will exist a $\tilde{\theta}$ with these marginals, and since:

$$\tilde{\theta} \simeq (\tilde{\theta}_X, \alpha(\tilde{\theta}_X, \tilde{\theta}_Y, \tilde{\beta}), \tilde{\beta}) \simeq (\tilde{\theta}_X, \tilde{\theta}_Y, \tilde{\beta}),$$

such a law would be uniquely defined.

Unfortunately, such a law would probably not be all that useful, as it would still require computing the integral:

$$\bar{p}(y | x, \beta) = \int_{\Theta_X \times \Theta_Y} \frac{e^{\alpha(\beta, \theta_X, \theta_Y) + \beta^\top x}}{1 + e^{\alpha(\beta, \theta_X, \theta_Y) + \beta^\top x}} d\mathcal{L}_m(\theta_X, \theta_Y), \quad (2.14)$$

which may not be any easier than the retrospective likelihood.

One method of avoiding the need to compute such an integral is to require $\tilde{\alpha}$ and $\tilde{\theta}_X$ to be independent, as occurs under strong hyper Markov laws:

Corollary 2.3.2. *If $\mathcal{L}(\tilde{\theta})$ is strong hyper Markov, that is if:*

$$(\tilde{\alpha}, \tilde{\beta}) \perp\!\!\!\perp \tilde{\theta}_X \quad \text{and} \quad (\tilde{\theta}_{X|Y=0}, \tilde{\beta}) \perp\!\!\!\perp \tilde{\theta}_Y \quad [\mathcal{L}], \quad (2.15)$$

then the posterior law for $\tilde{\beta}$ is the same under both the prospective and retrospective likelihood.

We note that a directly equivalent result was identified by Staicu (2007, Theorem 1) for the case where \mathcal{X} is finite. Unfortunately, this elegant formulation was modified in the published version of the manuscript to the more complicated Staicu (2010, Theorem 2).

The problem of model comparison for case-control studies has received comparatively little attention in the literature, particularly for Bayesian analyses. However we note that we may derive a similar result to that of Theorem 2.3.1:

Theorem 2.3.3. *If $\mathcal{L}_1(\tilde{\theta})$ and $\mathcal{L}_2(\tilde{\theta})$ have the same marginal laws for $\tilde{\theta}_X$ and $\tilde{\theta}_Y$, then the Bayes factor between the prospective models is equal to the Bayes factor between the retrospective models.*

Proof. One argument is to construct a law $\mathcal{L}^*(\tilde{\theta}, \tilde{M})$ that is a mixture of \mathcal{L}_1 and \mathcal{L}_2 , where \tilde{M} is a variable indicating the mixture component. Then the conditions of the theorem are equivalent to:

$$\tilde{M} \perp\!\!\!\perp \tilde{\theta}_X \quad \text{and} \quad \tilde{M} \perp\!\!\!\perp \tilde{\theta}_Y \quad [\mathcal{L}^*].$$

By the same argument as Theorem 2.3.1, the posterior probabilities, and hence the Bayes factors, must be equal.

Alternatively, let \bar{p}_1 and \bar{p}_2 denote the marginal models under the respective priors. Then:

$$\frac{\bar{p}_1(\mathbf{y}^{(n)} | \mathbf{x}^{(n)})}{\bar{p}_2(\mathbf{y}^{(n)} | \mathbf{x}^{(n)})} = \frac{\bar{p}_1(\mathbf{y}^{(n)} | \mathbf{x}^{(n)}) \bar{p}_1(\mathbf{x}^{(n)})}{\bar{p}_2(\mathbf{y}^{(n)} | \mathbf{x}^{(n)}) \bar{p}_2(\mathbf{x}^{(n)})} = \frac{\bar{p}_1(\mathbf{x}^{(n)} | \mathbf{y}^{(n)}) \bar{p}_1(\mathbf{y}^{(n)})}{\bar{p}_2(\mathbf{x}^{(n)} | \mathbf{y}^{(n)}) \bar{p}_2(\mathbf{y}^{(n)})} = \frac{\bar{p}_1(\mathbf{x}^{(n)} | \mathbf{y}^{(n)})}{\bar{p}_2(\mathbf{x}^{(n)} | \mathbf{y}^{(n)})}$$

since $\bar{p}_1(\mathbf{x}^{(n)}) = \bar{p}_2(\mathbf{x}^{(n)})$ and $\bar{p}_1(\mathbf{y}^{(n)}) = \bar{p}_2(\mathbf{y}^{(n)})$. □

The requirement that the laws have the same marginals may seem restrictive, but there is a simple way we may construct such laws:

Proposition 2.3.4. *Suppose $\mathcal{L}(\tilde{\theta})$ satisfies the conditions of Theorem 2.3.1. Then the law on the submodel defined by $\mathcal{L}_0(\tilde{\theta}) = \mathcal{L}(\tilde{\theta} | \tilde{\beta}_j = 0)$ will also satisfy the conditions of Theorem 2.3.1, and \mathcal{L} and \mathcal{L}_0 will satisfy Theorem 2.3.3.*

Proof. This follows from (2.10) by noting that:

$$\tilde{\beta} \perp\!\!\!\perp \tilde{\theta}_X \mid \tilde{\beta}_j \quad \text{and} \quad \tilde{\beta} \perp\!\!\!\perp \tilde{\theta}_Y \mid \tilde{\beta}_j \quad [\mathcal{L}]. \quad \square$$

2.4 Strong hyper Markov laws for logistic regression

Given the results of Corollary 2.3.2, we now investigate various strong hyper Markov laws for use as prior laws in case-control studies.

A single binary covariate

In the case of a single binary covariate we may take $\mathcal{X} = \{0, 1\}$, then the logistic model is a reparameterisation of a 2×2 contingency table.

Example 2.4.1. The simplest strong hyper Markov law for this model is the Dirichlet law $\mathcal{L}(\tilde{\theta}) = \mathcal{D}(a_{xy})$. This law has been well explored in the literature, in particular Altham (1969), who investigated log odds ratio parameter; and was later used in the context of case-control studies by Zelen and Parker (1986), Nurminen and Mutanen (1987), Marshall (1988) and Ashby, Hutton, and McGee (1993).

The Dirichlet law has density:

$$\pi(\theta) = \frac{1}{B(\theta_{00}, \theta_{01}, \theta_{10}, \theta_{11})} \theta_{00}^{a_{00}-1} \theta_{01}^{a_{01}-1} \theta_{10}^{a_{10}-1} \theta_{11}^{a_{11}-1}.$$

By reparameterising $\theta_{xy} = \frac{e^{y(\alpha+\beta x)}}{1+e^{\alpha+\beta x}} \theta_{0+}^{1-x} \theta_{1+}^x$, we find $\mathcal{L}(\tilde{\theta}_{x+}) = \mathcal{B}(a_{0+}, a_{1+})$, and:

$$\pi(\alpha, \beta) = \frac{e^{\alpha a_{01}} e^{(\alpha+\beta) a_{11}}}{(1+e^{\alpha})^{a_{0+}} (1+e^{\alpha+\beta})^{a_{1+}}} \quad (2.16)$$

Recall from the previous chapter that there is actually a more general family of strong hyper Markov laws on 2×2 tables. Specifically, a law with density of the form (1.12), in which case the density of $\mathcal{L}(\tilde{\alpha}, \tilde{\beta})$ would be:

$$\pi(\alpha, \beta) = g(\beta) \frac{e^{\alpha a_{01}} e^{(\alpha+\beta) a_{11}}}{(1+e^{\alpha})^{a_{0+}} (1+e^{\alpha+\beta})^{a_{1+}}}$$

where $g(\beta) = h(e^{\beta})$.

Finite covariate space

We now investigate the more general case where \mathcal{X} is larger, but still finite. Prior specification is not so simple: the proportional odds constraint implies that the logistic model will be nested within a sub-manifold of the probability simplex of the full $|\mathcal{X}| \times 2$ contingency table.

We solve this problem by adapting the conditioning procedure from Dawid and Lauritzen (2001, section 4) for constructing laws on nested models:

1. Choose an arbitrary strong hyper Markov law $\mathcal{L}'(\tilde{\theta})$ for the saturated model on $\mathcal{X} \times \{0, 1\}$.
2. Construct the law \mathcal{L} from \mathcal{L}' conditional on $\tilde{\theta}$ satisfying the proportional odds requirement.

With regards to the second point above, as Dawid and Lauritzen (2001) emphasised, the Borel–Kolmogorov paradox means that there is no unique way to perform such a conditioning operation. Furthermore, in selecting the method of conditioning, we need to ensure that it preserves the strong hyper Markov property.

Without loss of generality, we can assume that there exists $x_1, \dots, x_{k+1} \in \mathcal{X}$ such that $(1, x_1), (1, x_2), \dots, (1, x_{k+1})$ are linearly independent (otherwise \mathcal{X} exists on some affine subspace of \mathbb{R}^k , and so β is not identifiable). We may reparameterise the saturated model as:

$$p(y | x, \alpha, \beta, \eta) = \frac{e^{y(\alpha + \beta^\top x + \eta_x)}}{1 + e^{\alpha + \beta^\top x + \eta_x}} \quad (2.17)$$

where $\eta_x = 0$ if $x = x_1, \dots, x_{k+1}$. As such we may write $\theta_{Y|X} \simeq (\alpha, \beta, \eta)$ and $\theta_{X|Y} \simeq (\theta_{X|Y=0}, \beta, \eta)$, and hence by the strong hyper Markov property:

$$(\tilde{\alpha}, \tilde{\beta}, \tilde{\eta}) \perp\!\!\!\perp \tilde{\theta}_X \quad \text{and} \quad (\tilde{\theta}_{X|Y=0}, \tilde{\beta}, \tilde{\eta}) \perp\!\!\!\perp \tilde{\theta}_Y \quad [\mathcal{L}'].$$

Note that the logistic model is on the manifold defined by $\eta = 0$, and that:

$$(\tilde{\alpha}, \tilde{\beta}) \perp\!\!\!\perp \tilde{\theta}_X \mid \tilde{\eta} \quad \text{and} \quad (\tilde{\theta}_{X|Y=0}, \tilde{\beta}) \perp\!\!\!\perp \tilde{\theta}_Y \mid \tilde{\eta} \quad [\mathcal{L}'],$$

and hence $\mathcal{L}(\tilde{\theta}) = \mathcal{L}'(\tilde{\theta} \mid \tilde{\eta} = 0)$ is strong hyper Markov.

Example 2.4.2. We know from Theorem 1.7.1 that densities of the form:

$$h \left(\left[\begin{array}{c} \theta_{x_1} \theta_{x^*0} \\ \theta_{x_0} \theta_{x^*1} \end{array} \right]_{x \neq x^*} \right) \prod_{x \in \mathcal{X}} \theta_{x_0}^{a_{x_0}-1} \theta_{x_1}^{a_{x_1}-1}, \quad (2.18)$$

for some arbitrary $x^* \in \mathcal{X}$, are strong hyper Markov for the full $|\mathcal{X}| \times 2$ contingency table model.

The Jacobian determinant of the above transformation is:

$$\left| \frac{d\theta_{Y|X}}{d(\alpha, \beta, \eta)} \right| \propto \prod_{x \in \mathcal{X}} \frac{e^{\alpha + \beta^\top x + \eta_x}}{(1 + e^{\alpha + \beta^\top x + \eta_x})^2} \quad (2.19)$$

and hence the density for $\mathcal{L}'(\tilde{\alpha}, \tilde{\beta}, \tilde{\eta})$ is of the form:

$$h \left(\left[e^{\beta^\top (x - x^*) + \eta_x - \eta_{x^*}} \right]_{x \neq x^*} \right) \prod_{x \in \mathcal{X}} \frac{e^{(\alpha + \beta^\top x + \eta_x) a_{x1}}}{(1 + e^{\alpha + \beta^\top x + \eta_x})^{a_{x+}}}.$$

By conditioning on $\eta_x = 0$ for all $x \in \mathcal{X}$, we obtain the density of $\mathcal{L}(\tilde{\alpha}, \tilde{\beta})$:

$$g(\beta) \prod_{x \in \mathcal{X}} \frac{e^{(\alpha + \beta^\top x) a_{x1}}}{(1 + e^{\alpha + \beta^\top x})^{a_{x+}}} \quad (2.20)$$

where $g(\beta) = h \left(\left[e^{\beta^\top (x - x^*)} \right]_{x \neq x^*} \right)$.

The Jacobian of the transformation in terms of the retrospective parameters is:

$$\left| \frac{d(\alpha, \beta, \theta_X)}{d(\theta_{X|0}, \beta, \gamma)} \right| = \frac{(1 - \gamma)^{|\mathcal{X}|-1}}{\gamma} \prod_{x \in \mathcal{X}} (1 + e^{\alpha + \beta^\top x}) \quad (2.21)$$

and so the density of $\mathcal{L}(\tilde{\theta}_{X|0}, \tilde{\beta})$ is:

$$g(\beta) \frac{\prod_{x \in \mathcal{X}} \theta_{x|0}^{a_{x+}-1} e^{a_{x1} \beta^\top x}}{\left[\sum_{x \in \mathcal{X}} e^{\beta^\top x} \theta_{x|0} \right]^{a_{+1}}}. \quad (2.22)$$

There are other ways to perform such a conditioning operation, such as using the (non-log) odds ratio, but η has the desirable property of being invariant of the choice of x^* and x_1, \dots, x_{k+1} .

We note that the prior from Staicu (2010, Example 2) may be obtained by rewriting (2.20) as:

$$g^*(\beta) e^{\alpha a_{+1}} \prod_{x \in \mathcal{X}} (1 + e^{\alpha + \beta^\top x})^{-a_{x+}} \quad (2.23)$$

where $g^*(\beta) = g(\beta) \exp \left\{ \sum_{x \in \mathcal{X}} \beta^\top x a_{x1} \right\}$. Furthermore, by taking the limit as $a_{+1} \rightarrow 0$, we also obtain the improper prior of Seaman and Richardson (2004) and Staicu (2010, Example 1).

However, we argue that the form of (2.20) is more easily interpreted: it may be thought of as the product of an improper prior with density

$g(\beta) d\beta d\alpha$ and a logistic likelihood function, where the a_{xy} represent “pseudo-counts”. This has the further benefit of being able to easily adapt existing computational methods: for example, a Laplace approximation can be found using regular logistic regression software.

Although x appears in the density of $\mathcal{L}(\tilde{\alpha}, \tilde{\beta})$, we disagree with Staicu (2010) that this constitutes a covariate dependent prior, such as the g -priors of Zellner (1986): it is dependent on the *a priori* expected frequency of the covariates, and not the observed frequency of the covariates in the data.

We also note that this law may itself be constructed as the posterior of a beta prior law:

Proposition 2.4.1. *For each $x \in \mathcal{X}$, let:*

$$\tau_x = \frac{e^{\alpha + \beta^\top x_i}}{1 + e^{\alpha + \beta^\top x_i}}$$

For some $x_1, \dots, x_{k+1} \in \mathcal{X}$ such that $(1, x_1), (1, x_2), \dots, (1, x_{k+1})$ are linearly independent, let $\mathcal{L}'(\tilde{\theta})$ be the product law of the marginal laws:

$$\mathcal{L}'(\tilde{\tau}_{x_i}) = \mathcal{B}(a_{x_i 0}, a_{x_i 1}).$$

For all other $x \neq x_1, \dots, x_{k+1}$, let:

$$\mathcal{L}'(Z_x | \tilde{\theta}) = \text{Binomial}(a_{x+}, \tau_x).$$

Then the posterior law $\mathcal{L}'(\tilde{\theta} | Z_x = a_{x1})$ will have density of the form (2.20), where g constant.

Proof. The prior law $\mathcal{L}'(\tilde{\alpha}, \tilde{\beta})$ will have density proportional to:

$$\prod_{x=x_1, \dots, x_k} \frac{e^{(\alpha + \beta^\top x) a_{x1}}}{(1 + e^{\alpha + \beta^\top x})^{a_{x+}}}.$$

Likewise the likelihood of $(Z_x = a_{x1})_{x \neq x_1, \dots, x_{k+1}}$ will be proportional to:

$$\prod_{x \neq x_1, \dots, x_{k+1}} \frac{e^{(\alpha + \beta^\top x) a_{x1}}}{(1 + e^{\alpha + \beta^\top x})^{a_{x+}}}. \quad \square$$

This is particularly useful for implementing such procedures in generic Bayesian MCMC packages such as WinBUGS, OpenBUGS and JAGS, and note that these packages happily accept non-integer values for binomial counts. Furthermore, arbitrary functions g may be included by use of the

“zero Poisson” trick (see Spiegelhalter et al. 2003, “Specifying a new sampling distribution”).

Unfortunately, this method is somewhat impractical for large numbers of covariates. In particular, we note that the size of \mathcal{X} increases exponentially with its dimensionality k . Furthermore, as \mathcal{X} increases, $\tilde{\beta}$ will tend to concentrate around 0. To compensate for this, the values of (a_{xy}) can be chosen closer to 0, but unfortunately, the above software packages tend not work well, if at all, for very small values.

Extension to Dirichlet processes

A natural question is how to extend the above laws to the case where \mathcal{X} is infinite, for example where a covariate is continuous. One obvious choice would be to replace the Dirichlet law $\mathcal{D}(a_{x+})$ for $\mathcal{L}(\tilde{\theta}_X)$ with a Dirichlet process $\mathcal{D}\mathcal{P}(\mu, A)$. In this case, the form of the densities in equations (2.20) and (2.22) suggests the following:

Conjecture 2.4.2. *Let μ_0, μ_1 be measures on \mathcal{X} , $A_0, A_1 > 0$ and $\bar{\mu} = (A_0\mu_0 + A_1\mu_1)/(A_0 + A_1)$. Define a law $\mathcal{L}(\tilde{\theta})$ such that:*

$$\tilde{\theta}_X \perp\!\!\!\perp \tilde{\theta}_{Y|X} \quad [\mathcal{L}]$$

where $\mathcal{L}(\tilde{\theta}_X) = \mathcal{D}\mathcal{P}(\bar{\mu}, A_+)$, and $\mathcal{L}(\tilde{\alpha}, \tilde{\beta})$ has a density (with respect to the Lebesgue measure on \mathbb{R}^{k+1}):

$$g(\beta) \exp \left\{ A_1(\alpha + \beta^\top \mathbb{E}_{\mu_1(X)}[X]) - (A_0 + A_1) \mathbb{E}_{\bar{\mu}(X)} [\log(1 + e^{\alpha + \beta^\top X})] \right\}, \quad (2.24)$$

then $\mathcal{L}(\tilde{\theta})$ is strong hyper Markov, with $\mathcal{L}(\tilde{\theta}_Y) = \mathcal{B}(A_0, A_1)$, and $\mathcal{L}(\tilde{\theta}_{X|Y=0}, \tilde{\beta})$ has density with respect to a product measure of a Dirichlet process $\mathcal{D}\mathcal{P}(\bar{\mu}, A_+)$ and Lebesgue on \mathbb{R}^k of:

$$g(\beta) \exp \left\{ A_1 \beta^\top \mathbb{E}_{\mu_1(X)}[X] \right\} \left(\mathbb{E}_{\theta_{X|Y=0}(X)}[e^{\beta^\top X}] \right)^{-A_1}.$$

Unfortunately, the expectation terms in (2.24) means that we can’t easily apply the standard Dirichlet process machinery of taking projections onto finite partitions of \mathcal{X} , and appealing to the Kolmogorov extension theorem.

2.5 Stratified case-control studies

A more complicated *stratified* or *matched* case-control studies, in which participants are selected by both the outcome Y and an additional stratum vari-

able S . Such a design can estimate the odds-ratio of interest with much greater efficiency than an unstratified study.

The model is similar to that above, but with an intercept parameter that varies by strata, such that the prospective model is:

$$p(y | x, s, \alpha, \beta) = \frac{e^{\alpha_s + \beta^\top x}}{1 + e^{\alpha_s + \beta^\top x}}$$

Unfortunately, this additional complication makes the estimates more difficult. As the number of strata will increase with the sample size n , the usual maximum likelihood estimator is no longer consistent.

Instead, the standard classical approach seeks to maximise the *conditional likelihood*:

$$\ell_c(\beta) = \prod_{s \in \mathcal{S}} \frac{\prod_{i \in I_s} e^{y_i \beta^\top x_x}}{\sum_{\rho} \prod_{i \in I_s} e^{y_{\rho(i)} \beta^\top x_x}}$$

where $I_s = \{i : s_i = s\}$, and the summation in the denominator is over the possible permutations of $(y_i)_{i \in I_s}$.

Note that the number of terms in the denominator: if there are a cases and b controls in each stratum—called $a : b$ matching—the sum will have $\binom{a+b}{a}$ terms. Most studies use 1 : 1 or 1 : m matching, but if larger strata are used, this sum can quickly become computationally intractable.

In a Bayesian analysis however, the conditional likelihood does not have a direct interpretation. Rice (2004, Theorem 1) showed there will exist a law such that the marginal retrospective likelihood $\bar{p}(x | y, s, \beta)$ will be proportional to the conditional likelihood. However such a law will depend on the matching scheme: *e.g.* a 1 : 1 matched design will require a different law than a 1 : 2 matched design.

Instead, we can extend Theorem 2.3.1 to find conditions under which we may use the prospective likelihood under *any* matching scheme:

Theorem 2.5.1. *Let $\mathcal{L}(\tilde{\theta}_{XY|S})$ be a prior law for the parameters of the stratified logistic model. Then the posterior marginal law for $\tilde{\beta}$ is the same under both prospective and retrospective likelihood for all possible observations $(x^{(n)}, y^{(n)}, s^{(n)})$, if and only if:*

$$\tilde{\beta} \perp\!\!\!\perp \tilde{\theta}_{X|S} \quad \text{and} \quad \tilde{\beta} \perp\!\!\!\perp \tilde{\theta}_{Y|S} \quad [\mathcal{L}]. \quad (2.25)$$

Proof. The argument is essentially the same as that of Theorem 2.3.1, noting that $\theta_{X|S}$ and $\theta_{Y|S}$ are the joint distributions for the random vectors $(X|S = s)_{s \in \mathcal{S}}$ and $(Y|S = s)_{s \in \mathcal{S}}$, respectively. \square

One way of constructing such a law is to use a conditioning procedure similar to that in the previous section:

1. For each stratum s , let $\mathcal{L}_s(\tilde{\theta}_{XY|S=s})$ be a law satisfying Theorem 2.3.1, where $\theta_{Y|X,S=s} \simeq (\alpha_s, \beta_s)$.
2. Let $\mathcal{L}^*(\tilde{\theta}_{XY|S})$ be the product law $\prod_s \mathcal{L}_s$, and therefore:

$$\tilde{\theta}_{X|S} \perp\!\!\!\perp (\tilde{\beta}_s)_{s \in \mathcal{S}} \quad \text{and} \quad \tilde{\theta}_{Y|S} \perp\!\!\!\perp (\tilde{\beta}_s)_{s \in \mathcal{S}} \quad [\mathcal{L}^*].$$

3. Reparameterise $(\beta_s)_{s \in \mathcal{S}} \simeq [\beta, (\tau_s)_{s \neq s^*}]$, where $\beta = \beta_{s^*}$ for some stratum s^* , and $\tau_s = \beta_s - \beta$ for each $s \neq s^*$.
4. Condition on $\tau_s = 0$. Since:

$$\tilde{\theta}_{X|S} \perp\!\!\!\perp \tilde{\beta} \mid (\tilde{\tau}_s)_{s \neq s^*} \quad \text{and} \quad \tilde{\theta}_{Y|S} \perp\!\!\!\perp \tilde{\beta} \mid (\tilde{\tau}_s)_{s \neq s^*} \quad [\mathcal{L}^*],$$

it follows that $\mathcal{L}(\tilde{\theta}_{XY|S}) = \mathcal{L}^*(\tilde{\theta}_{XY|S} \mid \tilde{\tau} = 0)$ will satisfy the conditions of Theorem 2.5.1.

Example 2.5.1. If we let each $\mathcal{L}_s(\tilde{\alpha}_s, \tilde{\beta}_s)$ be of the form in Example 2.4.2, the density for the law $\mathcal{L}^*(\tilde{\alpha}, \tilde{\beta}, \tilde{\tau})$ will be of the form:

$$\prod_{s \in \mathcal{S}} g_s(\beta + \tau_s) \prod_{x \in \mathcal{X}} \frac{e^{(\alpha_s + (\beta + \tau_s)^\top x) a_{x1s}}}{(1 + e^{\alpha_s + (\beta - \tau_s)^\top x}) a_{x+s}}.$$

Conditioning on $\tilde{\tau} = 0$ gives a density for $\mathcal{L}(\tilde{\alpha}, \tilde{\beta})$ as:

$$g(\beta) \prod_{(x,s) \in \mathcal{X} \times \mathcal{S}} \frac{e^{(\alpha_s + \beta^\top x) a_{x1s}}}{(1 + e^{\alpha_s + \beta^\top x}) a_{x+s}}$$

Interestingly, this is of the same form as the density (2.20), where the strata are simply treated as an additional categorical covariate in the model.

Note that we haven't specified of any type of model for the stratum variable S , as we have assumed all data are observed conditional on S . However, we note that under the additional assumption:

$$\tilde{\theta}_{XY|S} \perp\!\!\!\perp \tilde{\theta}_S \quad [\mathcal{L}],$$

we can treat the data as if they were randomly sampled from the population, as it would be in a cross-sectional study.

2.6 Discussion

We have illustrated the role of parameter independence, both variational and probabilistic, for making inference about parameters under different sampling regimes. In particular, we have shown the importance of these considerations when selecting prior laws for such models, in order to avoid introducing incorrect information into the posterior law.

There is potential for these techniques to be successfully applied to other models. In particular, the stratified case-control model is closely related to the Rasch model, commonly used in psychometrics for measuring ability or attitudes of individuals based on tests and questionnaires.

Part II

Structural Markov properties

Background

In the remainder of this thesis, we consider the problem of inferring the structure of a graphical model from data.

Initial approaches to the problem utilised sequential hypothesis tests. The first such approach appears in the context of estimating covariance matrices, where Dempster (1972) noted the usefulness of imposing sparsity on the precision matrix, and proposed a simple forward selection procedure based on a likelihood ratio test. Later work by Wermuth (1976a,b) identified the similarities in the multiplicative structure of contingency tables and covariance matrix selection, and proposed a backwards selection procedure for such models. Sundberg (1975); Frydenberg and Lauritzen (1989) further showed that for decomposable graphs, the likelihood ratio between such nested models differing by an edge may be computed locally.

However any such approach will suffer from the fact that the test statistics are not independent, therefore making it difficult to correct for multiple comparisons problems. Recently, Drton and Perlman (2008) propose a method to control the overall error rate via simultaneous testing.

Somewhat similar approaches are based on *scoring*: each graph is assigned some numerical measure of fit, and the graph with the largest score is selected. In the case of a small number of vertices, it is possible to evaluate the score on all possible graphs, but for larger models a heuristic search procedure is required. Such approaches include Buntine (1991), Cooper and Herskovits (1992), Heckerman, Geiger, and Chickering (1995), Spirtes, Glymour, and Scheines (2000), Chickering (2002, 2003). In certain cases, these may be the mode of a posterior graph law, commonly called the maximum *a posteriori* (MAP) estimate. However such methods fail to quantify any other aspects of the posterior and so, we would argue, cannot really be considered Bayesian.

More recently, *graphical lasso* approaches have become popular. These are based on the “lasso” shrinkage estimators, used for model selection in regression problems (see Tibshirani 1996), which seek to maximise the log-likelihood with a ℓ_1 penalty on the coefficients. As a consequence of the peaked form of the penalty, variables with little predictive power will have their coefficients shrunk to zero. Yuan and Lin (2007); Banerjee, El Ghaoui, and d’Aspremont (2008); Friedman, Hastie, and Tibshirani (2008); Rothman et al. (2008) apply this approach to Gaussian graphical models, with an estimator for the precision matrix of the form:

$$\arg \max_{\Lambda} [\log \det \Lambda - \text{tr}(S\Lambda) - \rho \|\Lambda\|_1],$$

where $\|\Lambda\|_1$ denotes the sum of absolute values of elements of Λ (some authors exclude diagonal elements from this sum), and ρ is a the adjustable penalty term. As in the regression case, this will shrink some of the off-diagonal elements of Λ to zero, corresponding to missing edges of the estimated graph.

We instead approach the problem in a fully Bayesian manner: utilising a prior law for the structure of the graph itself, and the parameters each graphical model. With the advent of Bayesian computational techniques such as Markov chain Monte Carlo, such approaches have become computationally feasible. For example, Madigan and York (1995); Madigan, Andersson, et al. (1996); Giudici and Castelo (2003); Friedman and Koller (2003); Ellis and Wong (2008); Mukherjee and Speed (2008) investigated such approaches for directed acyclic graphs, and Giudici and Green (1999); Dellaportas and Forster (1999); Brooks, Giudici, and Roberts (2003) for undirected decomposable graphs.

However, very little work has focused on the choice of prior for the graph itself. Most authors utilise either a simple uniform prior, or some modification of a Erdős–Rényi random graph, where the existence of each edge is independent of the others.

Part of the problem is the difficulty of specifying a stochastic process over the set of graphs under consideration. In the case of undirected decomposable graphs, the only other examples appear to be McMorris and Scheinerman (1991), in which the vertices correspond to random subtrees of a fixed tree, and Lunagomez (2009), in which the vertices are generated by some point process in Euclidean space, and connectivity is determined by

simplicial complexes. Unfortunately, both these methods rely on stochastic processes over auxiliary structures, which makes their exact graphical properties difficult to determine. For directed graphs, the problem is even more difficult, due to the problems of Markov equivalence.

In the subsequent chapters, we investigate how the meta and hyper Markov properties may be extended to the case where the structure of the graph is itself unknown. We develop the theory for undirected and directed graphs separately, as the characterisations of these properties differ in their construction.

Undirected decomposable graphical models

We propose a method of extending the undirected meta and hyper Markov properties to allow for the case where the graph is itself a random quantity.

4.1 Motivation and definition

Recall that a law $\mathcal{L}(\tilde{\theta})$ over $\mathfrak{P}(\mathcal{G})$, the set of Markov distributions with respect to \mathcal{G} , is (weak) hyper Markov if for any decomposition (A, B) :

$$\tilde{\theta}_A \perp\!\!\!\perp \tilde{\theta}_B \mid \tilde{\theta}_{A \cap B} \quad [\mathcal{L}]. \quad (4.1)$$

We now consider the case where the graph itself is not fixed, but is instead a random variable $\tilde{\mathcal{G}}$. As the graph is a parameter in the model, we will term its distribution a *graph law*, and will usually denote this by $\mathfrak{G}(\tilde{\mathcal{G}})$. In particular, we would like to develop hyper Markov-type properties for $\tilde{\mathcal{G}}$.

Consider the case where the $\mathfrak{G}(\tilde{\mathcal{G}})$ is defined over some family \mathfrak{F} of undirected decomposable graphs, in which (A, B) is a common decomposition of all $\mathcal{G} \in \mathfrak{F}$. Recall that in (4.1), $\tilde{\theta}_A$ will take values in $\mathfrak{P}(\mathcal{G}_A)$, and similarly for $\tilde{\theta}_B$ and $\tilde{\theta}_{A \cap B}$: that is, $\tilde{\mathcal{G}}_A$ influences the support of $\tilde{\theta}_A$. One way to extend the hyper Markov property in this case would be to require that:

$$\tilde{\mathcal{G}}_A \perp\!\!\!\perp \tilde{\mathcal{G}}_B \mid \tilde{\mathcal{G}}_{A \cap B} \quad [\mathfrak{G}]$$

Note that the term $\tilde{\mathcal{G}}_{A \cap B}$ is redundant: if (A, B) is a decomposition of \mathcal{G} , then $\mathcal{G}_{A \cap B}$ must be complete, and so we are left with a statement of marginal independence $\tilde{\mathcal{G}}_A \perp\!\!\!\perp \tilde{\mathcal{G}}_B$.

A more general question remains: how might this property be extended to an arbitrary family of graphs, such as those without a common decomposition? This motivates the following property:

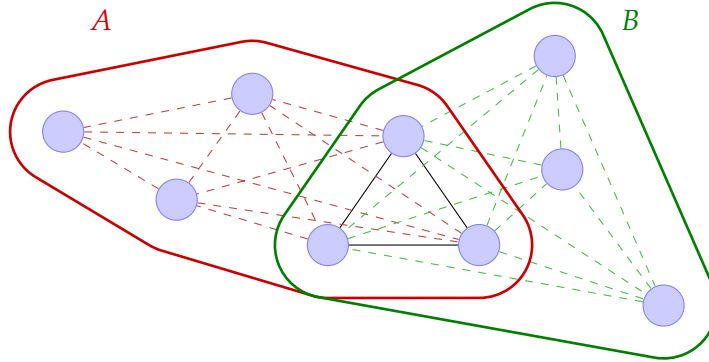


Figure 4.1: A representation of the structural Markov property for undirected graphs. Conditional on (A, B) being a decomposition, the existence of the remaining edges in $\tilde{\mathcal{G}}_A$ (---) are independent of those in $\tilde{\mathcal{G}}_B$ (---).

Definition 4.1.1 (Structural Markov property). A *covering pair* (of V) is any pair of sets (A, B) such that $A \cup B = V$. A graph law $\mathfrak{G}(\tilde{\mathcal{G}})$ over the set \mathfrak{U} of undirected decomposable graphs on V is *structurally Markov* if for any covering pair (A, B) , we have:

$$\tilde{\mathcal{G}}_A \perp\!\!\!\perp \tilde{\mathcal{G}}_B \mid \{\tilde{\mathcal{G}} \in \mathfrak{U}(A, B)\} \quad [\mathfrak{G}], \quad (4.2)$$

where $\mathfrak{U}(A, B)$ is the set of decomposable graphs for which (A, B) is a decomposition.

In essence, the structural Markov property states that the structure of different parts of the graph are conditionally independent given that they are in separate parts of a decomposition. See Figure 4.1 for depiction.

Unlike the Markov and hyper Markov properties however, the conditional independence is defined with respect to the *event* $\tilde{\mathcal{G}} \in \mathfrak{U}(A, B)$, and not a random variable. In other words, we do not assume $\tilde{\mathcal{G}}_A \perp\!\!\!\perp \tilde{\mathcal{G}}_B \mid \tilde{\mathcal{G}} \notin \mathfrak{U}(A, B)$.

4.2 Projections and products

Recall that Theorems 1.4.1 and 1.4.2, allow us to specify distributions and laws in a piecewise manner over the cliques, via a conditional product operation. We can apply the same arguments to graphs:

Proposition 4.2.1. *Let \mathcal{H} and \mathcal{J} be two graphs on A and B respectively, such that both $\mathcal{H}_{A \cap B}$ and $\mathcal{J}_{A \cap B}$ are complete. Then there exists a unique graph \mathcal{G} on $A \cup B$ such that:*

- (i) $\mathcal{G}_A = \mathcal{H}$,
- (ii) $\mathcal{G}_B = \mathcal{J}$, and
- (iii) $A \cap B$ separates A and B in \mathcal{G} .

Proof. To satisfy (i) and (ii), the edge set must contain $\mathcal{E}(\mathcal{H}) \cup \mathcal{E}(\mathcal{J})$. It cannot contain any additional edges $\{u, v\}$, as this would violate either (i), if $\{u, v\} \subseteq A$; (ii), if $\{u, v\} \subseteq B$, or (iii), if $u \in A \setminus B$ and $v \in B \setminus A$. \square

We define the resulting graph to be the *graph product*, denoted by:

$$\mathcal{G} = \mathcal{H} \otimes \mathcal{J}.$$

Furthermore, the completeness requirement on the intersection $A \cap B$ implies that (A, B) will be a decomposition of \mathcal{G} . We note that a more general operator could be defined by only requiring that \mathcal{H} and \mathcal{J} be collapsible onto $A \cap B$, but this is not needed in the following theory.

The graph product provides a very useful characterisation of the structural Markov property:

Proposition 4.2.2. *A graph law \mathfrak{G} is structurally Markov if and only if for every covering pair (A, B) , and every $\mathcal{G}, \mathcal{G}' \in \mathfrak{U}(A, B)$,*

$$\pi(\mathcal{G}) \pi(\mathcal{G}') = \pi(\mathcal{G}_A \otimes \mathcal{G}'_B) \pi(\mathcal{G}'_A \otimes \mathcal{G}_B) \quad (4.3)$$

where π is the density of \mathfrak{G} with respect to the counting measure on \mathfrak{U} .

Proof. Note that both $\mathcal{G}_A \otimes \mathcal{G}'_B, \mathcal{G}'_A \otimes \mathcal{G}_B \in \mathfrak{U}(A, B)$. Furthermore, the density of a structural Markov law is of the form:

$$\pi(\mathcal{G} \mid \mathfrak{U}(A, B)) = \pi(\mathcal{G}_A \mid \mathfrak{U}(A, B)) \pi(\mathcal{G}_B \mid \mathfrak{U}(A, B)).$$

The result follows by substitution into (4.3). \square

Creating a natural projection operation for a graph law is considerably more difficult: as we will shall see in Section 4.7, the structural Markov property is generally not preserved under various forms of marginalisation. But it is preserved conditional on a decomposition.

Lemma 4.2.3. *Let (A, B) be a decomposition of a graph \mathcal{G} , and (S, T) a covering pair of A with $A \cap B \subseteq T$. Then (S, T) is a decomposition of \mathcal{G}_A if and only if $(S, T \cup B)$ is a decomposition of \mathcal{G} .*

Proof. Recall that W separates U and V in \mathcal{G} if and only if $\langle U, V \mid W \rangle \in \mathcal{M}(\mathcal{G})$.

Since (S, T) is a covering pair of A , $\langle S \cup T, B \mid S \cap B \rangle \in \mathcal{M}(\mathcal{G})$, and hence $\langle S, B \mid T \rangle \in \mathcal{M}(\mathcal{G})$. If (S, T) is a decomposition of \mathcal{G}_A , then $\langle S, T \mid S \cap T \rangle \in \mathcal{M}(\mathcal{G}_A)$, which implies that $\langle S, B \cup T \mid T \cap S \rangle \in \mathcal{M}(\mathcal{G})$. Since $\mathcal{G}_{(S \cup B) \cap T} = \mathcal{G}_{T \cap S}$ is complete, $(S \cup B, T)$ is a decomposition of \mathcal{G} .

The converse result follows by the reverse argument. \square

Theorem 4.2.4. *Let $\mathfrak{G}(\tilde{\mathcal{G}})$ be a structurally Markov graph law: then the conditional law for $\tilde{\mathcal{G}}_A \mid \tilde{\mathcal{G}} \in \mathfrak{U}(A, B)$ is also structurally Markov.*

Proof. Let (S, T) be a covering pair of A : If we restrict $\tilde{\mathcal{G}} \in \mathfrak{U}(A, B)$, then $\tilde{\mathcal{G}}_{A \cap B}$ must be complete. As we are only interested in the case where (S, T) is a decomposition of $\tilde{\mathcal{G}}_A$, then $A \cap B$ must be a subset of either S or T : without loss of generality, we may assume $A \cap B \subseteq T$.

$(S, T \cup B)$ is a covering pair of V , so by the structural Markov property:

$$\tilde{\mathcal{G}}_S \perp\!\!\!\perp \tilde{\mathcal{G}}_{T \cup B} \mid \{\tilde{\mathcal{G}} \in \mathfrak{U}(S, T \cup B)\}.$$

If $\mathbb{1}_E$ is the indicator variable of an event E , we may can write:

$$\tilde{\mathcal{G}}_S \perp\!\!\!\perp (\tilde{\mathcal{G}}_T, \mathbb{1}_{\tilde{\mathcal{G}}_{T \cup B} \in \mathfrak{U}(T, B)}) \mid \{\tilde{\mathcal{G}} \in \mathfrak{U}(S, T \cup B)\}.$$

By the axioms of conditional independence, the term $\mathbb{1}_{\tilde{\mathcal{G}}_{T \cup B} \in \mathfrak{U}(T, B)}$ may be moved to the right-hand side. Furthermore, we are only interested in the case where it equals 1, hence we can write:

$$\tilde{\mathcal{G}}_S \perp\!\!\!\perp \tilde{\mathcal{G}}_T \mid \{\mathcal{G}_{T \cup B} \in \mathfrak{U}(T, B)\}, \{\tilde{\mathcal{G}} \in \mathfrak{U}(S, T \cup B)\}.$$

By Lemma 4.2.3, $\tilde{\mathcal{G}}_{T \cup B} \in \mathfrak{U}(T, B)$ if and only if $\tilde{\mathcal{G}} \in \mathfrak{U}(S \cup T, B) = \mathfrak{U}(A, B)$.

$$\tilde{\mathcal{G}}_S \perp\!\!\!\perp \tilde{\mathcal{G}}_T \mid \{\tilde{\mathcal{G}} \in \mathfrak{U}(A, B)\}, \{\tilde{\mathcal{G}} \in \mathfrak{U}(S, T \cup B)\}.$$

Again, by Lemma 4.2.3, $\tilde{\mathcal{G}} \in \mathfrak{U}(S, T \cup B)$ if and only if $\tilde{\mathcal{G}}_A \in \mathfrak{U}(S, T)$, hence:

$$\tilde{\mathcal{G}}_S \perp\!\!\!\perp \tilde{\mathcal{G}}_T \mid \{\tilde{\mathcal{G}} \in \mathfrak{U}(A, B)\}, \{\tilde{\mathcal{G}}_A \in \mathfrak{U}(S, T)\}. \quad \square$$

4.3 Structural meta Markov property

We can also define a similar property by replacing probabilistic conditional independence $(\cdot \perp\!\!\!\perp \cdot \mid \cdot)$ with variation independence $(\cdot \dagger \cdot \mid \cdot)$, analogous to the relationship between the hyper Markov and meta Markov properties.

Definition 4.3.1 (Structural meta Markov property). For a family of undirected decomposable graphs \mathfrak{F} and a covering pair (A, B) , let $\mathfrak{F}(A, B) = \mathfrak{F} \cap \mathfrak{U}(A, B)$. Then \mathfrak{F} is *structurally meta Markov* if for every covering pair (A, B) :

$$\mathcal{G}_A \dagger \mathcal{G}_B \mid \{\mathcal{G} \in \mathfrak{F}(A, B)\}.$$

Equivalently, we can characterise it in terms of the graph product operation:

Theorem 4.3.1. *A family of undirected decomposable graphs \mathfrak{F} is structurally meta Markov if and only if $\mathcal{G}_A \otimes \mathcal{G}'_B \in \mathfrak{F}$ for all $\mathcal{G}, \mathcal{G}' \in \mathfrak{F}(A, B)$.*

Therefore, if a family of graphs is characterised by a specific property, we can show that it is structurally meta Markov if this property is preserved under the graph product operation.

Example 4.3.1. The set of undirected decomposable graphs whose cliques size is bounded above by some integer n is structurally meta Markov. To see this, note that a clique of $\mathcal{G}_A \otimes \mathcal{G}'_B$ must be a clique of either \mathcal{G}_A or \mathcal{G}'_B (and hence of either \mathcal{G} or \mathcal{G}'). Therefore, the graph product operation cannot increase the size of the largest clique.

An interesting special case is $n = 2$, which is the set of trees on V .

Example 4.3.2. For two graphs $\mathcal{G}^L, \mathcal{G}^U \in \mathfrak{U}$ such that $\mathcal{E}(\mathcal{G}^L) \subseteq \mathcal{E}(\mathcal{G}^U)$, the “sandwich” set between the two graphs:

$$\{\mathcal{G} \in \mathfrak{U} : \mathcal{E}(\mathcal{G}^L) \subseteq \mathcal{E}(\mathcal{G}) \subseteq \mathcal{E}(\mathcal{G}^U)\}.$$

is structurally meta Markov.

Theorem 4.3.2. *The support of a structurally Markov graph law is a structurally meta Markov family.*

Proof. Let \mathfrak{F} be the support of the structurally Markov graph law \mathfrak{G} with density π . By Proposition 4.2.2, if $\mathcal{G}, \mathcal{G}' \in \mathfrak{F}(A, B)$ and both $\pi(\mathcal{G})$ and $\pi(\mathcal{G}')$

are non-zero, then $\pi(\mathcal{G}_A \otimes \mathcal{G}'_B)$ must also be non-zero, and hence in $\mathfrak{F}(A, B)$. Therefore, by Theorem 4.3.1, \mathfrak{F} is structurally meta Markov. \square

4.4 Compatible distributions and laws

We now investigate how the structural Markov property interacts with the Markov and hyper Markov properties.

Definition 4.4.1. Let $X = (X_v)_{v \in V}$ be a random variable, and $\theta = \{\theta^{(\mathcal{G})} : \mathcal{G} \in \mathfrak{U}\}$ be a family of probability distributions for X . We write $X \sim \theta$ if, given $\tilde{\mathcal{G}} = \mathcal{G}$, $X \sim \theta^{(\mathcal{G})}$. Then θ is *compatible* if:

- (i) For each $\mathcal{G} \in \mathfrak{U}$, X is Markov with respect to \mathcal{G} under $\theta^{(\mathcal{G})}$, and
- (ii) $\theta_C^{(\mathcal{G})} = \theta_C^{(\mathcal{G}')}$ whenever $C \subseteq V$ induces a complete subgraph in both $\mathcal{G}, \mathcal{G}' \in \mathfrak{U}$.

Likewise, let $\mathfrak{L} = \{\mathfrak{L}^{(\mathcal{G})} : \mathcal{G} \in \mathfrak{U}\}$ be a family of laws for the parameters $\tilde{\theta}$ of a family of distributions on X . Again, we can write $\tilde{\theta} \sim \mathfrak{L}$ if, given $\tilde{\mathcal{G}} = \mathcal{G}$, $\tilde{\theta} \sim \mathfrak{L}^{(\mathcal{G})}$. Then \mathfrak{L} is *hyper compatible* if:

- (i) For all $\mathcal{G} \in \mathfrak{U}$, $\mathfrak{L}^{(\mathcal{G})}$ is a weak hyper Markov law on \mathcal{G} , and
- (ii) $\mathfrak{L}_C^{(\mathcal{G})} = \mathfrak{L}_C^{(\mathcal{G}')}$ if C induces a complete subgraph in both $\mathcal{G}, \mathcal{G}' \in \mathfrak{U}$.

Remark. Dawid and Lauritzen (1993, section 6.2) originally used the term compatible to refer to what we term the hyper compatible case: we introduce the distinction so as to extend the terminology to the non-hyper case.

In the above definition, both θ and \mathfrak{L} will be characterised entirely by $\theta^{(\mathcal{G}^{(V)})}$ and $\mathfrak{L}^{(\mathcal{G}^{(V)})}$ respectively, where $\mathcal{G}^{(V)}$ is the complete graph on V .

A graph law $\mathfrak{G}(\tilde{\mathcal{G}})$ combined with a compatible set of distributions θ defines a joint distribution (\mathfrak{G}, θ) on $(\tilde{\mathcal{G}}, X)$ under which $X \mid \tilde{\mathcal{G}} = \mathcal{G} \sim \theta^{(\mathcal{G})}$. Likewise, \mathfrak{G} combined with a set of hyper compatible laws \mathfrak{L} defines a joint law $(\mathfrak{G}, \mathfrak{L})$ on $(\tilde{\mathcal{G}}, \tilde{\theta})$, and so a joint distribution on $(\tilde{\mathcal{G}}, \tilde{\theta}, X)$.

The key conditional independence property of any such joint distribution or law can be characterised as follows:

Proposition 4.4.1. *If $\tilde{\mathcal{G}}$ has a graph law \mathfrak{G} , and $X \sim \theta$ for a compatible family θ , then:*

$$X_A \perp\!\!\!\perp \tilde{\mathcal{G}}_B \mid \tilde{\mathcal{G}}_A, \{\tilde{\mathcal{G}} \in \mathfrak{U}(A, B)\} \quad [\mathfrak{G}, \theta]$$

Similarly, if $\tilde{\mathcal{G}}$ has a graph law \mathfrak{G} , and $\tilde{\theta} \sim \mathfrak{L}$ for a hyper compatible family \mathfrak{L} , then:

$$\tilde{\theta}_A \perp\!\!\!\perp \tilde{\mathcal{G}}_B \mid \tilde{\mathcal{G}}_A, \{\tilde{\mathcal{G}} \in \mathfrak{U}(A, B)\} \quad [\mathfrak{G}, \mathfrak{L}]$$

Proof. If $\mathcal{G} \in \mathfrak{U}(A, B)$, then \mathcal{G}_A is uniquely determined by its cliques. Therefore the distribution of X_A and law of $\tilde{\theta}_A$ are each fixed. \square

When combined with the structural Markov property, we obtain some useful results:

Theorem 4.4.2. *If $\tilde{\mathcal{G}}$ has a structurally Markov graph law \mathfrak{G} , and X has a distribution from a compatible set θ , then:*

$$(X_A, \tilde{\mathcal{G}}_A) \perp\!\!\!\perp (X_B, \tilde{\mathcal{G}}_B) \mid X_{A \cap B}, \{\tilde{\mathcal{G}} \in \mathfrak{U}(A, B)\} \quad [\mathfrak{G}, \theta]$$

Proof. The Markov property states that under $[\mathfrak{G}, \theta]$:

$$X_A \perp\!\!\!\perp X_B \mid X_{A \cap B}, \tilde{\mathcal{G}}, \{\tilde{\mathcal{G}} \in \mathfrak{U}(A, B)\} \quad (4.4)$$

Since $(\tilde{\mathcal{G}}, \{\tilde{\mathcal{G}} \in \mathfrak{U}(A, B)\}) \simeq (\tilde{\mathcal{G}}_A, \tilde{\mathcal{G}}_B, \{\tilde{\mathcal{G}} \in \mathfrak{U}(A, B)\})$, we can rewrite (4.4) as:

$$X_A \perp\!\!\!\perp X_B \mid X_{A \cap B}, \tilde{\mathcal{G}}_A, \tilde{\mathcal{G}}_B, \{\tilde{\mathcal{G}} \in \mathfrak{U}(A, B)\} \quad (4.5)$$

A trivial consequence of Proposition 4.4.1:

$$X_A \perp\!\!\!\perp \tilde{\mathcal{G}}_B \mid X_{A \cap B}, \tilde{\mathcal{G}}_A, \{\tilde{\mathcal{G}} \in \mathfrak{U}(A, B)\} \quad (4.6)$$

By (4.5) and (4.6):

$$X_A \perp\!\!\!\perp (X_B, \tilde{\mathcal{G}}_B) \mid X_{A \cap B}, \tilde{\mathcal{G}}_A, \{\tilde{\mathcal{G}} \in \mathfrak{U}(A, B)\} \quad (4.7)$$

Furthermore, by the structural Markov property and Proposition 4.4.1:

$$\tilde{\mathcal{G}}_A \perp\!\!\!\perp (X_B, \tilde{\mathcal{G}}_B) \mid \{\tilde{\mathcal{G}} \in \mathfrak{U}(A, B)\}, \quad (4.8)$$

and we can further condition on $X_{A \cap B}$. The result follows from this and (4.7). \square

Corollary 4.4.3. *If $\tilde{\mathcal{G}}$ has a structurally Markov graph law, and X has a distribution from a compatible set θ , then the posterior graph law for $\tilde{\mathcal{G}}$ is structurally Markov.*

Proof. By Theorem 4.4.2 and the axioms of conditional independence, we easily obtain:

$$\tilde{\mathcal{G}}_A \perp\!\!\!\perp \tilde{\mathcal{G}}_B \mid X, \{\tilde{\mathcal{G}} \in \mathfrak{U}(A, B)\}. \quad \square$$

We may also apply similar arguments apply at the hyper level:

Theorem 4.4.4. *If $\tilde{\mathcal{G}}$ has a structurally Markov graph law \mathfrak{G} , and θ has a law from a hyper compatible set \mathfrak{L} , then:*

$$(\theta_A, \tilde{\mathcal{G}}_A) \perp\!\!\!\perp (\theta_B, \tilde{\mathcal{G}}_B) \mid \theta_{A \cap B}, \{\tilde{\mathcal{G}} \in \mathfrak{U}(A, B)\} \quad [\mathfrak{G}, \mathfrak{L}]$$

Furthermore, if each law $\mathcal{L}^{(\mathcal{G})} \in \mathfrak{L}$ is strong hyper Markov with respect to \mathcal{G} , then:

$$(\theta_A, \tilde{\mathcal{G}}_A) \perp\!\!\!\perp (\theta_{B|A}, \tilde{\mathcal{G}}_B) \mid \{\tilde{\mathcal{G}} \in \mathfrak{U}(A, B)\} \quad [\mathfrak{G}, \mathfrak{L}]$$

Proof. The proof for the first case is the same as in Theorem 4.4.2. The proof for the strong case follows similar steps, except starts with the strong hyper Markov property:

$$\theta_A \perp\!\!\!\perp \theta_{B|A} \mid \tilde{\mathcal{G}}, \{\tilde{\mathcal{G}} \in \mathfrak{U}(A, B)\} \quad \square$$

Hyper compatible sets of strong hyper Markov laws have the additional advantage that the posterior graph law will also be structurally Markov: this follows from Theorem 4.4.2 and Dawid and Lauritzen (1993, Proposition 5.6), which states that the marginal distribution of the data under a strong hyper Markov law is Markov. Furthermore, the posterior family of graph laws $\{\mathcal{L}^{(\mathcal{G})}(\cdot \mid X) : \mathcal{G} \in \mathfrak{U}\}$ will maintain hyper compatibility.

4.5 Clique vector

Definition 4.5.1. Define the *completeness vector* of a graph to be the function $c : \mathfrak{U} \rightarrow \{0, 1\}^{2^V}$, such that for each $A \subseteq V$:

$$c_A(\mathcal{G}) = \begin{cases} 1 & \text{if } \mathcal{G}_A \text{ is complete,} \\ 0 & \text{otherwise.} \end{cases}$$

Furthermore, define the *clique vector* of a graph $t : \mathfrak{U} \rightarrow \mathbb{Z}^{2^V}$ to be the Möbius inverse of c by *superset* inclusion:

$$t_B(\mathcal{G}) = \sum_{A \supseteq B} (-1)^{|A \setminus B|} c_A(\mathcal{G}). \quad (4.9)$$

In the language of Studený 2005b, c and t are both *imsets* (integer-valued multisets).

c can likewise be expressed in terms of t :

Proposition 4.5.1. *For any $\mathcal{G} \in \mathfrak{U}$, we have:*

$$c_A(\mathcal{G}) = \sum_{B \supseteq A} t_B(\mathcal{G}), \quad A \subseteq V. \quad (4.10)$$

Proof. This follows from the Möbius inversion theorem (see, for example, Lauritzen 1996, Lemma A.2). \square

Both c and t decompose elegantly:

Lemma 4.5.2. *If $\mathcal{G} \in \mathfrak{U}(A, B)$, then:*

$$c(\mathcal{G}) = [c(\mathcal{G}_A)]^0 + [c(\mathcal{G}_B)]^0 - [c(\mathcal{G}_{A \cap B})]^0, \quad \text{and} \quad (4.11)$$

$$t(\mathcal{G}) = [t(\mathcal{G}_A)]^0 + [t(\mathcal{G}_B)]^0 - [t(\mathcal{G}_{A \cap B})]^0. \quad (4.12)$$

where $[\cdot]^0$ denotes the expansion of a vector with zeroes to the required coordinates.

Proof. $U \subseteq V$ induces a complete subgraph of $\mathcal{G} \in \{\mathfrak{U}(A, B)\}$ if and only if it induces a complete subgraph from \mathcal{G}_A , \mathcal{G}_B or both. (4.11) follows by the inclusion–exclusion principle. (4.12) may then be obtained by substitution into (4.9). \square

Theorem 4.5.3. *For any decomposable graph $\mathcal{G} \in \mathfrak{U}$ and $A \subseteq V$:*

$$t_A(\mathcal{G}) = \begin{cases} 1 & \text{if } A \in \text{cl}(\mathcal{G}), \\ -\nu_{\mathcal{G}}(A) & \text{if } A \in \text{sep}(\mathcal{G}), \text{ and} \\ 0 & \text{otherwise;} \end{cases}$$

where $\text{cl}(\mathcal{G})$ are the cliques of \mathcal{G} , and $\text{sep}(\mathcal{G})$ are the clique separators, and each separator S has multiplicity $\nu_{\mathcal{G}}(S)$.

Proof. For any $C \subseteq V$, let $\mathcal{G}^{(C)}$ be the graph on V whose edge set is $\binom{C}{2}$ (that is, complete on C and empty elsewhere). Then it is straightforward to see that:

$$t_A(\mathcal{G}_C^{(C)}) = \begin{cases} 1 & \text{if } A = C, \\ 0 & \text{otherwise.} \end{cases}$$

Now let C_1, \dots, C_k be a perfect ordering of the cliques of G , and S_2, \dots, S_k be the corresponding separators. By Lemma 4.5.2, it follows that:

$$t(\mathcal{G}) = \sum_{i=1}^k t(\mathcal{G}_{C_i}^{(C_i)}) - \sum_{i=2}^k t(\mathcal{G}_{S_i}^{(S_i)}). \quad \square$$

Objects similar to the clique vector have arisen in several contexts. Notably, it appears to be equivalent to the index v of Lauritzen, Speed, and Vijayan (1984, Definition 5), which is characterised in a combinatorial manner.

Another similar construction is the *standard imset* of Studený (2005b), which is equal to:

$$t(\mathcal{G}^{(V)}) - t(\mathcal{G})$$

where $\mathcal{G}^{(V)}$ is the complete graph.

The algorithm of Wormald (1985) for the enumeration of decomposable graphs is based on a generating function for the vector $\mathbb{R}^{|V|}$ that he termed the “maximal clique vector”, which is defined as:

$$\text{mcv}_k(\mathcal{G}) = \sum_{A \in \binom{V}{k}} t_A(\mathcal{G}), \quad k = 1, \dots, |V|$$

Proposition 4.5.4. *For any $\mathcal{G} \in \mathfrak{U}$, the vector $t(\mathcal{G})$ has the following properties:*

(i)

$$\sum_{A \subseteq V} t_A(\mathcal{G}) = 1$$

(ii) For each $v \in V$:

$$\sum_{A \ni v} t_A(\mathcal{G}) = 1$$

(iii)

$$\sum_{A \subseteq V} |A| t_A(\mathcal{G}) = |V|$$

(iv)

$$\sum_{A \subseteq V} \binom{|A|}{2} t_A(\mathcal{G}) = |\mathcal{E}(\mathcal{G})|$$

Proof. These all follow from Theorem 4.5.3 and the inclusion–exclusion principle. \square

4.6 Clique exponential family

Definition 4.6.1. The *clique exponential family* is the exponential family of graph laws over $\mathfrak{F} \subseteq \mathfrak{U}$, with t as a natural statistic (with respect to the uniform measure on \mathfrak{U}). That is, laws in the family have densities of the form:

$$\pi_\omega(\mathcal{G}) = \frac{1}{Z(\omega)} \exp\{\omega \cdot t(\mathcal{G})\}, \quad \mathcal{G} \in \mathfrak{F}, \quad \omega \in \mathbb{R}^{2^V},$$

where $Z(\omega)$ is the normalisation constant, which will generally be unknown.

Equivalently, the distribution can be parameterised in terms of c :

$$\pi_\omega(\mathcal{G}) = \frac{1}{Z(\omega)} \exp \left\{ \left(\sum_{B \subseteq A} (-1)^{|A \setminus B|} \omega_A \right)_{A \subseteq V} \cdot c(\mathcal{G}) \right\},$$

but t is more useful due to the fact that it is sparse (by Theorem 4.5.3) and, as we shall see, is the natural statistic for posterior updating.

Note that this distribution is over-parameterised: by Proposition 4.5.4 (i) and (ii), there are $|V| + 1$ linear relationships in $t(G)$. For the purpose of identifiability, we could define a normalised vector ω^* as:

$$\omega_A^* = \omega_A + (|A| - 1)\omega_\emptyset - \sum_{v \in A} \omega_{\{v\}}$$

such that $\pi_\omega = \pi_{\omega^*}$, and $\omega_{\{v\}}^* = \omega_\emptyset^* = 0$ for all $v \in V$.

Theorem 4.6.1. *Let \mathfrak{G} be a graph law whose support is \mathfrak{L} . Then \mathfrak{G} is structurally Markov if and only if it is a member of the clique exponential family.*

Proof. For any $C \subseteq V$, define $\mathcal{G}^{(C)}$ as in Theorem 4.5.3, and let \mathfrak{G} have density π .

Suppose that \mathfrak{G} is structurally Markov. For any $\mathcal{G} \in \mathfrak{L}$, let C_1, \dots, C_k be a perfect ordering of the cliques, and let S_2, \dots, S_k be the corresponding separators, and H_1, \dots, H_k be the histories. Furthermore, recursively define the graphs:

$$\mathcal{G}^{*(j)} = \begin{cases} \mathcal{G}^{(C_1)} & \text{if } j = 1, \\ \mathcal{G}_{H_{j-1}}^{*(j-1)} \otimes \mathcal{G}_{(V \setminus H_{j-1}) \cup S_j}^{(C_j)} & \text{if } j = 2, \dots, k. \end{cases}$$

By Proposition 4.2.2, for each $j = 2, \dots, k$:

$$\pi(\mathcal{G}^{*(j)}) \pi(\mathcal{G}^{(S_j)}) = \pi(\mathcal{G}^{*(j-1)}) \pi(\mathcal{G}^{(C_j)})$$

Note that $\mathcal{G}^{*(k)} = \mathcal{G}$, then by induction it follows that:

$$\pi(\mathcal{G}) = \frac{\prod_{j=1}^k \pi(\mathcal{G}^{(C_j)})}{\prod_{j=2}^k \pi(\mathcal{G}^{(S_j)})} \propto \exp\{\omega \cdot t(\mathcal{G})\},$$

by Theorem 4.5.3, where $\omega_C = \log \pi(\mathcal{G}^{(C)})$.

To show the converse let $(\omega)_A = (\omega_S)_{S \subseteq A}$. By Lemma 4.5.2, we have:

$$\begin{aligned} \pi(\mathcal{G}_A | \mathcal{G}_B, \{\mathcal{G} \in \mathfrak{U}(A, B)\}) &\propto \exp \{(\omega)_A \cdot t(\mathcal{G}_A) + (\omega)_B \cdot t(\mathcal{G}_B) - (\omega)_{A \cap B} \cdot t(\mathcal{G}_{A \cap B})\} \\ &\propto \exp \{(\omega)_A \cdot t(\mathcal{G}_A) - (\omega)_{A \cap B} \cdot t(\mathcal{G}_{A \cap B})\} \\ &\propto \pi(\mathcal{G}_A | \{\mathcal{G} \in \mathfrak{U}(A, B)\}). \quad \square \end{aligned}$$

Remark. The requirement that the family have full support is not strictly necessary: we may use exactly the same argument applies to any family \mathfrak{F} with the property that if $\mathcal{G} \in \mathfrak{F}$ and C is a clique of \mathcal{G} , then $\mathcal{G}^{(C)} \in \mathfrak{F}$. This includes Example 4.3.1, and Example 4.3.2 if \mathcal{G}^L is the sparse graph.

We conjecture that this could hold for any structurally hyper Markov law, but have yet to identify a proof.

A very similar family was proposed by Bornn and Caron (2011), however their family allows the use of different parameters for cliques and separators, which will generally not be structurally Markov.

Example 4.6.1 (Giudici and Green 1999; Brooks, Giudici, and Roberts 2003, section 8). The simplest example of such a distribution is the uniform distribution over \mathfrak{U} , corresponding to $\omega = 0$.

Example 4.6.2 (Madigan and Raftery 1994; Jones et al. 2005). Another common approach is to use a set of $\binom{|V|}{2}$ independent Bernoulli variables with probability ψ to indicate edge inclusion (*i.e.* an Erdős–Rényi random graph), conditional on $\tilde{\mathcal{G}}$ being decomposable. The density of such a law is of the form:

$$\pi(\mathcal{G}) \propto \psi^{|\mathcal{E}(\mathcal{G})|} (1 - \psi)^{\binom{p}{2} - |\mathcal{E}(\mathcal{G})|} \propto \left(\frac{\psi}{1 - \psi} \right)^{|\mathcal{E}(\mathcal{G})|}$$

By Proposition 4.5.4 (iv), it follows that this distribution has:

$$\omega_A = \binom{|A|}{2} \log \left(\frac{\psi}{1 - \psi} \right)$$

Furthermore, if each edge e has its own probability ψ_e , then:

$$\omega_A = \sum_{e \in \binom{A}{2}} \log \left(\frac{\psi_e}{1 - \psi_e} \right)$$

Example 4.6.3 (Armstrong et al. 2009). For comparison, it is useful to consider a non-structurally Markov graph law. Define the distribution over the

number of edges to be uniform, and the conditional distribution over the set of graphs with a fixed number of edges to be uniform. This has density of the form:

$$\pi(\mathcal{G}) = \frac{1}{\binom{p}{2} + 1} \frac{1}{|\{\mathcal{G}' \in \mathfrak{U} : |\mathcal{E}(\mathcal{G}')| = |\mathcal{E}(\mathcal{G})|\}|}$$

Now consider the case $V = \{1, 2, 3\}$, then:

$$\begin{aligned} \pi\left(\begin{array}{ccc} \textcircled{1} & \textcircled{2} & \textcircled{3} \end{array}\right) &= \frac{1}{4} \\ \pi\left(\begin{array}{ccc} \textcircled{1} & \text{---} & \textcircled{2} & \textcircled{3} \end{array}\right) &= \frac{1}{12} \\ \pi\left(\begin{array}{ccc} \textcircled{1} & \textcircled{2} & \text{---} & \textcircled{3} \end{array}\right) &= \frac{1}{12} \\ \pi\left(\begin{array}{ccc} \textcircled{1} & \text{---} & \textcircled{2} & \text{---} & \textcircled{3} \end{array}\right) &= \frac{1}{12} \end{aligned}$$

From this it follows that $\tilde{\mathcal{G}}_{\{1,2\}} \not\perp \tilde{\mathcal{G}}_{\{2,3\}} \mid \tilde{\mathcal{G}} \in \mathfrak{U}(\{\{1,2\}, \{2,3\}\})$, and hence the law cannot be structurally Markov.

Posterior updating

We saw in Corollary 4.4.3 that if the sampling distributions are compatible, then posterior updating will preserve the structural Markov property. We now show that this updating may be performed locally, with the exponential clique family forming a conjugate prior for a family of compatible models.

Let θ be a family of compatible distributions for X (such as the marginal model of a strong hyper Markov law), with density p with respect to some product measure. Then:

$$\pi(X|\mathcal{G}) = \prod_{A \subseteq V} p_A(X_A)^{t(\mathcal{G})_A},$$

and thus the posterior law is:

$$\pi(\mathcal{G}|X) \propto \exp \left\{ \left[\omega + (\log p_A(X_A))_{A \subseteq V} \right] \cdot t(\mathcal{G}) \right\}.$$

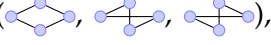
A key benefit of this conjugate formation is that we can describe the posterior law with a parameter of dimension $2^{|V|}$ (strictly speaking, we only need $2^{|V|} - |V| - 1$, due to the over-parameterisation). This is much smaller than an arbitrary law over the set of undirected decomposable graphs, which would require a parameter of length approximately $2^{\binom{|V|}{2}}$.

4.7 Marginalisation

We now consider how to marginalise a graph law. Specifically, for a given a graph law \mathfrak{G} over \mathcal{U} , how might we define the graph law over the set of undirected decomposable graphs on $A \subseteq V$?

We have already proposed one such method in Theorem 4.2.4: however this also required the specification of a decomposition, and multiple such decompositions exist for any given proper subset A .

Below we propose three alternative methods for constructing such a marginal law. However we also demonstrate that none of these preserve the structural Markov property.

Example 4.7.1. The graph law we consider is one of the simplest non-trivial structurally Markov graph law: the uniform law on the set of undirected decomposable graphs on 4 vertices $V = \bullet \bullet \bullet \bullet$. There are $\binom{4}{2} = 6$ possible edges, and hence $2^6 = 64$ possible undirected graphs; of these, 3 are non-decomposable 4-cycles () , so the probability of any one graph is $\frac{1}{61}$.

From this law, we aim seek to construct a graph law over $A = \bullet \bullet$. We show that none of the marginal laws proposed below satisfy the necessary cross-product property of Proposition 4.2.2 for the structural Markov properties. Specifically, if we take $\mathcal{G} = \bullet \bullet$ and $\mathcal{G}' = \bullet \bullet$, then (4.3) states that we require:

$$\pi_A(\bullet \bullet) \pi_A(\bullet \bullet) = \pi_A(\bullet \bullet) \pi_A(\bullet \bullet). \quad (4.13)$$

Unfortunately, this lack of a structural Markov-preserving marginalisation procedure rules out any type of straightforward “self-similarity” type property, such as that exhibited by the Wishart, inverse-Wishart and Dirichlet laws. This also means there is no obvious way to construct a structurally Markov graph law from a convenient infinite-dimensional stochastic process, such as the Chinese restaurant process for generating samples from a Dirichlet process (Aldous 1985, Section 11.19).

Induced subgraph

The simplest method of marginalisation is the induced subgraph $\tilde{\mathcal{G}}_A$. Note that $\tilde{\mathcal{G}}$ may not always be collapsible onto A , in which case $\mathcal{M}(\tilde{\mathcal{G}}_A)$ may not

be contained in $\mathcal{M}(\mathcal{G})$. The probabilities of the graphs in (4.13) are:

$$\begin{aligned} \pi_A(\text{graph 1}) &= \pi(\text{graph 1}, \text{graph 2}, \text{graph 3}, \text{graph 4}, \text{graph 5}, \text{graph 6}, \text{graph 7}, \text{graph 8}, \text{graph 9}) = \frac{8}{61} \\ \pi_A(\text{graph 10}) &= \pi(\text{graph 10}, \text{graph 11}, \text{graph 12}, \text{graph 13}, \text{graph 14}, \text{graph 15}, \text{graph 16}, \text{graph 17}, \text{graph 18}) = \frac{8}{61} \\ \pi_A(\text{graph 19}) &= \pi(\text{graph 19}, \text{graph 20}, \text{graph 21}, \text{graph 22}, \text{graph 23}, \text{graph 24}, \text{graph 25}, \text{graph 26}) = \frac{8}{61} \\ \pi_A(\text{graph 27}) &= \pi(\text{graph 27}, \text{graph 28}, \text{graph 29}, \text{graph 30}, \text{graph 31}, \text{graph 32}, \text{graph 33}, \text{graph 34}) = \frac{7}{61} \end{aligned}$$

Marginal graph

An alternative method of marginalising a graph law is to map to each graph to one that preserves its conditional independencies on the set of vertices of interest.

For an undirected graph \mathcal{G} on V and a subset $A \subseteq V$, we define the *marginal graph* $\mathcal{G}_A^{\mathcal{M}}$ to be the graph on A such that $\{u, v\} \in \mathcal{E}(\mathcal{G}_A^{\mathcal{M}})$ if there exists a path from u to v in $V \setminus (A \cup \{u, v\})$, in other words, if $A \setminus \{u, v\}$ does not separate u and v in \mathcal{G} .

This construction preserves the conditional independence properties of \mathcal{G} on A :

Theorem 4.7.1 (Studený 1997, Lemma 3.1). *If \mathcal{G} is an undirected graph on V , and $A \subseteq V$, then:*

$$\mathcal{M}(\mathcal{G}_A^{\mathcal{M}}) = \mathcal{M}_A(\mathcal{G}).$$

Note that this projection may destroy other properties that can't be expressed in terms of conditional independence, e.g. a full exponential family may map to a curved exponential family.

The graph law induced by this projection does not preserve the structural Markov property:

$$\begin{aligned} \pi_A(\text{graph 1}) &= \pi(\text{graph 1}, \text{graph 2}, \text{graph 3}, \text{graph 4}) = \frac{4}{61} \\ \pi_A(\text{graph 5}) &= \pi(\text{graph 5}, \text{graph 6}, \text{graph 7}, \text{graph 8}, \text{graph 9}, \text{graph 10}) = \frac{6}{61} \\ \pi_A(\text{graph 11}) &= \pi(\text{graph 11}, \text{graph 12}, \text{graph 13}, \text{graph 14}, \text{graph 15}, \text{graph 16}) = \frac{6}{61} \\ \pi_A(\text{graph 17}) &= \pi(\text{graph 17}, \text{graph 18}, \text{graph 19}, \text{graph 20}, \text{graph 21}, \text{graph 22}, \text{graph 23}, \text{graph 24}) = \frac{8}{61} \end{aligned}$$

Conditional on being collapsible

Finally, we consider the graph law for the induced subgraph $\tilde{\mathcal{G}}_A$, conditional on $\tilde{\mathcal{G}}$ being collapsible onto A . The relevant probabilities are proportional to:

$$\begin{aligned} \pi_A(\text{two nodes}) &\propto \pi(\text{two nodes}, \text{two nodes}, \text{two nodes}, \text{two nodes}, \text{two nodes}) = \frac{4}{61} \\ \pi_A(\text{edge}) &\propto \pi(\text{edge}, \text{edge}, \text{edge}, \text{edge}, \text{edge}) = \frac{5}{61} \\ \pi_A(\text{two nodes, edge}) &\propto \pi(\text{two nodes, edge}, \text{two nodes, edge}, \text{two nodes, edge}, \text{two nodes, edge}, \text{two nodes, edge}) = \frac{5}{61} \\ \pi_A(\text{triangle}) &\propto \pi(\text{triangle}, \text{triangle}, \text{triangle}, \text{triangle}, \text{triangle}, \text{triangle}) = \frac{6}{61} \end{aligned}$$

However, we note that by Theorem 4.2.4, if we further condition on the separators, then the structural Markov property will be preserved.

4.8 Computation

The difficulties involved in enumerating \mathfrak{U} as well as computing the normalisation constant of the exponential family mean that some sort of numerical approximation will usually be required.

The most common approach is to construct a Markov chain Monte Carlo (MCMC) algorithm that moves between graphs by perturbing their edge set. The simplest approach, proposed by Giudici and Green (1999), relies on making single edge additions and removals. A key difficulty with such an approach is to characterise which edge modifications will result in the graph remaining decomposable.

For any graph $\mathcal{G} \in \mathfrak{U}$, we define $\mathcal{N}^-(\mathcal{G})$ and $\mathcal{N}^+(\mathcal{G})$ to be the set of undirected decomposable graphs that may be obtained by removing or adding, respectively, a single edge from \mathcal{G} . We call these the *lower* and *upper neighbours* of \mathcal{G} .

Fortunately, there are previous results that characterise the neighbours:

Theorem 4.8.1 (Frydenberg and Lauritzen 1989, Lemma 3). *Let \mathcal{G} be a decomposable graph, where $\{u, v\} \in \mathcal{E}(\mathcal{G})$. Then the graph \mathcal{G}^- obtained by removing $\{u, v\}$ is decomposable if and only if $\{u, v\}$ is a subset of exactly one clique C of \mathcal{G} .*

As a consequence, the set of lower neighbours $\mathcal{N}^-(\mathcal{G})$ can be partitioned according to the clique of \mathcal{G} which contained the removed edge.

For any such edge removal, the change in the clique vector t is characterised in a simple manner by this clique and the two vertices of the edge:

Theorem 4.8.2. *For the edge removal operation in Theorem 4.8.1, the change in the clique vector is:*

$$t_A(\mathcal{G}^-) - t_A(\mathcal{G}) = \begin{cases} -1 & \text{if } A = C \text{ or } C \setminus \{u, v\}, \\ +1 & \text{if } A = C \setminus \{u\} \text{ or } C \setminus \{v\}, \\ 0 & \text{otherwise.} \end{cases}$$

Proof. Firstly, consider the effect on the completeness vector c . The removal of the edge $\{u, v\}$ will only change the B th component ($B \subseteq V$), if:

- (i) the edge appears in \mathcal{G}_B : that is, if $\{u, v\} \subseteq B$, and
- (ii) \mathcal{G}_B is complete, which can only occur with (i) if B is a subset of the clique C .

Therefore:

$$c_B(\mathcal{G}^-) - c_B(\mathcal{G}) = \begin{cases} -1 & \text{if } \{u, v\} \subseteq B \subseteq C \\ 0 & \text{otherwise.} \end{cases}$$

As for the clique vector, it follows that from the definition of t in (4.9) that if $A \not\subseteq C$ then $t_A(\mathcal{G}^-) - t_A(\mathcal{G})$ will be zero, and if $A \subseteq C$ then:

$$\begin{aligned} t_A(\mathcal{G}^-) - t_A(\mathcal{G}) &= \sum_{B \supseteq A} (-1)^{|B \setminus A|} [c_B(\mathcal{G}^-) - c_B(\mathcal{G})] \\ &= \sum_{B: A \cup \{u, v\} \subseteq B \subseteq C} (-1)^{|B \setminus A| + 1} \\ &= \sum_{H: C \setminus (A \cup \{u, v\})} (-1)^{|H| + 3 - |A \cap \{u, v\}|} \end{aligned}$$

Recall that any finite, non-empty set has an equal number of even- and odd-cardinality subsets. Therefore $t_A(\mathcal{G}^-) - t_A(\mathcal{G})$ will be zero unless $C = A \cup \{u, v\}$. Moreover, it will be -1 if $A \cap \{u, v\}$ is even, and $+1$ if it is odd. \square

Likewise, there exists a similar characterisation of edge addition:

Theorem 4.8.3 (Giudici and Green 1999, Theorem 2). *Let \mathcal{G} be a decomposable graph, where the pair of vertices $\{u, v\} \notin \mathcal{E}(\mathcal{G})$. Then the graph \mathcal{G}^+ obtained by the addition of the edge $\{u, v\}$ is decomposable if and only if there exist cliques (of \mathcal{G}) $C_u \ni u$ and $C_v \ni v$ such that $S = C_u \cap C_v$ is a separator of C_u and C_v in \mathcal{G} .*

A consequence of this theorem is that the set of upper neighbours $\mathcal{N}^+(\mathcal{G})$ can be partitioned according to the separators of \mathcal{G} by which they are separated.

Similarly, we can also characterise the change in t by the two vertices and this separator:

Theorem 4.8.4. *For the edge addition operation in Theorem 4.8.3, the change in the clique vector is:*

$$t_A(\mathcal{G}^+) - t_A(\mathcal{G}) = \begin{cases} +1 & \text{if } A = S \text{ or } S \cup \{u, v\}, \\ -1 & \text{if } A = S \cup \{u\} \text{ or } S \cup \{v\}, \\ 0 & \text{otherwise.} \end{cases}$$

Proof. The proof is similar to that of Theorem 4.8.2. For $B \subseteq V$, c_B will only change if

- (i) $\{u, v\} \subseteq B$, and
- (ii) \mathcal{G}_B^+ is complete: this can only with (i) if both $\mathcal{G}_{B \setminus \{u\}}$ and $\mathcal{G}_{B \setminus \{v\}}$ are complete. Since S separates u and v in \mathcal{G} , it follows that $B \setminus \{u, v\} \subseteq S$.

Therefore:

$$c_B(\mathcal{G}^+) - c_B(\mathcal{G}) = \begin{cases} 1 & \text{if } \{u, v\} \subseteq B \subseteq S \cup \{u, v\} \\ 0 & \text{otherwise.} \end{cases}$$

The result follows using the same argument as the proof of Theorem 4.8.2. \square

Finally, it is necessary to show that it is possible to move between any two graphs by such individual edge additions and removals:

Theorem 4.8.5 (Frydenberg and Lauritzen 1989, Lemma 5). *For any two graphs $\mathcal{G}^{(0)}, \mathcal{G}^{(k)} \in \mathfrak{U}$, such that $\mathcal{E}(\mathcal{G}^{(0)}) \subset \mathcal{E}(\mathcal{G}^{(k)})$ and $|\mathcal{E}(\mathcal{G}^{(k)})| - |\mathcal{E}(\mathcal{G}^{(0)})| = k$, there exists a sequence of graphs $\mathcal{G}^{(1)}, \dots, \mathcal{G}^{(k-1)} \in \mathfrak{U}$ such that:*

$$\mathcal{E}(\mathcal{G}^{(0)}) \subset \mathcal{E}(\mathcal{G}^{(1)}) \subset \dots \subset \mathcal{E}(\mathcal{G}^{(k-1)}) \subset \mathcal{E}(\mathcal{G}^{(k)}).$$

Note that we may move between any two arbitrary graphs in at most $\binom{|V|}{2}$ moves by choosing $\mathcal{G}^{(0)}$ to be the sparse graph, or $\mathcal{G}^{(k)}$ to be the complete graph.

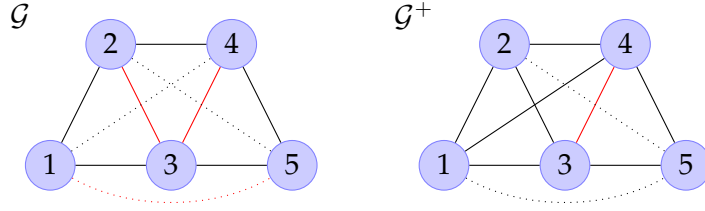


Figure 4.2: Neighbouring graphs on 5 vertices: solid lines (—) indicate edges, dotted lines (⋯) for missing edges. Red lines (— ⋯) are those whose removal/addition will result in a non-decomposable graph. In \mathcal{G} , only 7 of the 10 edges may be modified, whereas in \mathcal{G}^+ (obtained by adding the edge $\{1,4\}$), 9 of the 10 edges may be modified.

We can therefore use these results to construct an MCMC algorithm for sampling from a structurally Markov graph law. Specifically, we can easily construct a Metropolis–Hastings algorithm, with the following transition kernel: given our current graph $\mathcal{G}^{(t)}$, we select a pair of distinct vertices $u, v \in V$. If $\{u, v\} \in \mathcal{E}(\mathcal{G}^{(t)})$, and satisfies Theorem 4.8.1, the vertex is deleted. If $\{u, v\} \notin \mathcal{E}(\mathcal{G}^{(t)})$, and satisfies Theorem 4.8.3, then the edge is added. Otherwise, we stay at the current state.

Let $\mathcal{N}(\mathcal{G}) = \mathcal{N}^-(\mathcal{G}) \cup \mathcal{N}^+(\mathcal{G})$, then the proposal kernel would be:

$$Q(\mathcal{G}^{(t)}, \mathcal{G}') = \begin{cases} \binom{|V|}{2}^{-1} & \text{for } \mathcal{G}' \in \mathcal{N}(\mathcal{G}^{(t)}), \text{ and} \\ 1 - |\mathcal{N}(\mathcal{G}^{(t)})| \binom{|V|}{2}^{-1} & \text{for } \mathcal{G}' = \mathcal{G}^{(t)}. \end{cases}$$

Since the proposal kernel is symmetric, *i.e.* $Q(\mathcal{G}^{(t)}, \mathcal{G}') = Q(\mathcal{G}', \mathcal{G}^{(t)})$, the acceptance probabilities simply depends on the relative densities. By Theorems 4.8.2 and 4.8.4, this is simply $\min(\alpha, 1)$, where:

$$\alpha = \begin{cases} \exp\{\omega_{C \setminus \{u\}} + \omega_{C \setminus \{v\}} - \omega_{C \setminus \{u,v\}} - \omega_C\} & \text{if } \{u, v\} \in \mathcal{E}(\mathcal{G}^{(t)}), \\ \exp\{\omega_S + \omega_{S \cup \{u,v\}} - \omega_{S \cup \{u\}} - \omega_{S \cup \{v\}}\} & \text{if } \{u, v\} \notin \mathcal{E}(\mathcal{G}^{(t)}). \end{cases}$$

This means that at each step, the acceptance probability can be evaluated locally, utilising only four elements of the parameter vector: this is particularly useful when sampling from a posterior distribution, as we only then need to evaluate the marginal likelihood of four subsets of V .

Remark. It is not explicitly stated in Giudici and Green (1999), but should the proposed edge modification not result in a decomposable graph, it is necessary to record an observation from the current state—and not just sample

another edge—as the calculation of the acceptance ratio would then require finding the cardinality of $\mathcal{N}(\mathcal{G}^{(t)})$, which is quite difficult to calculate (see Thomas and Green 2009a for discussion of this problem).

One practical issue is the construction of an appropriate data structure to represent the graph in computer memory. Although Theorems 4.8.1 and 4.8.3 characterise the possible edge removals and additions, it is far from obvious how to efficiently determine if a proposed edge satisfies these criteria. It is worth noting that simply storing a graph as a set of vertices and edges is clearly inefficient, as this would require recomputing the cliques at each step. The results of Thomas and Green (2009a,b) indicate that a list of cliques stored in a perfect ordering or some representation of a clique tree could be useful for this purpose.

Another problem is the rate of mixing of the Markov chain. Due to the extremely large size of the space \mathcal{U} and the restriction on staying within the space of decomposable graphs, it can take an extremely long time to transition between two graphs. Kijima et al. (2007, 2008) show that for a uniform graph law, certain starting graphs will result in a mixing time exponential in $|V|$.

One possible solution is to propose larger jumps. Green and Thomas (2011) suggest a slight modification of the above scheme in which multiple edges may be removed or added. Another alternative would be to completely separate a vertex from the graph and reconnect it in some other way. However as the sample space for such a proposal scheme would be considerably larger— $|V| \times 2^{|V|-1}$ instead of $\binom{|V|}{2}$ —a uniform proposal distribution could result in frequently proposing moves to non-decomposable or low probability graphs, giving a poor acceptance ratio. This could possibly be improved by an adaptive sampling scheme, however it is far from clear how this could be efficiently constructed. Furthermore, we could lose the benefits of the local computation of the acceptance ratio.

Due to these difficulties, Jones et al. (2005) and Scott and Carvalho (2008) propose non-MCMC “stochastic search” algorithms for obtaining a representative sample of graphs from the posterior distribution. Although the empirical results of these methods seem promising, their accuracy and theoretical properties remain unknown.

Directed acyclic graphical models

We now investigate how the structural Markov property might be extended to directed acyclic graphical models (DAGs). Let \mathfrak{D} be the set of directed acyclic graphs on V vertices.

5.1 Ordered directed structural Markov property

Firstly, we consider a law for a random graph $\tilde{\mathcal{G}}$ over the set \mathfrak{D}^{\prec} : the set of directed acyclic graphs that respect a fixed well ordering \prec on V .

The set \mathfrak{D}^{\prec} is fairly easy to characterise: if an edge exists, its direction is determined by \prec . Furthermore, any subset of the set of pairs of vertices $\binom{V}{2}$ will uniquely characterise a graph in \mathfrak{D}^{\prec} , therefore:

$$|\mathfrak{D}^{\prec}| = 2^{\binom{|V|}{2}}.$$

So how might we develop a structural Markov property for such a graph? Recall that by the strong directed hyper Markov property:

$$\tilde{\theta}_{v|\text{pa}(v)} \perp\!\!\!\perp \tilde{\theta}_{\text{pr}(v)}. \quad (5.1)$$

Both $\text{pr}(v)$ and $\text{pr}(v) \cup \{v\}$ are ancestral sets in any such graph, then by Theorem 1.2.5, the projections of the separoid are equal to those of the induced subgraphs. Furthermore, by Theorem 1.2.3, $\mathcal{M}_{\text{pr}(v) \cup \{v\}}(\mathcal{G})$ is spanned by the set:

$$\{\langle \{u\}, \text{pr}_{\prec}(u) \mid \text{pa}_{\mathcal{G}}(u) \rangle : u \preceq v\}$$

and hence, also by the set:

$$\{\langle \{v\}, \text{pr}_{\prec}(v) \mid \text{pa}_{\mathcal{G}}(v) \rangle\} \cup \mathcal{M}(\mathcal{G}_{\text{pr}(v)}). \quad (5.2)$$

Note that $\langle \{v\}, \text{pr}_{\prec}(v) \mid \text{pa}_{\mathcal{G}}(v) \rangle$ only depends on \mathcal{G} through the parent set of v . The correspondence of (5.2) to (5.1) leads to the following definition of an *ordered directed structural Markov property*:

$$\text{pa}_{\tilde{\mathcal{G}}}(v) \perp\!\!\!\perp \tilde{\mathcal{G}}_{\text{pr}(v)}$$

Since this applies for all $v \in V$, we have:

$$\prod_{v \in V} \text{pa}_{\tilde{\mathcal{G}}}(v).$$

As the parent sets of each vertex will uniquely determine the graph, we may easily write the density of such a law as an exponential family whose natural statistic is parent set of each vertex:

$$\pi(\mathcal{G}) \propto \exp \left\{ \sum_{v \in V} \sum_{A \subseteq \text{pr}(v)} \omega_{v|A} \mathbb{1}_{\text{pa}_{\mathcal{G}}(v)=A} \right\}.$$

For the remainder of this chapter, we develop the structural Markov property to the set of all directed acyclic graphs on V . Unfortunately the above approach cannot be applied directly. For one thing, parent sets cannot be independent: if u is a parent of v , then to maintain acyclicity v cannot be a parent of u .

Firstly, we need to explore two key concepts: Markov equivalence and ancestral sets.

5.2 Markov equivalence and Dagoids

Unlike undirected graphs, there is not a bijective mapping between the graph and its separoid. That is, two or more distinct DAGs may have identical conditional independence properties, as in Figure 5.1.

Definition 5.2.1 (Markov equivalence). Let \mathcal{G} and \mathcal{G}' be directed acyclic graphs such that $\mathcal{M}(\mathcal{G}) = \mathcal{M}(\mathcal{G}')$. Then \mathcal{G} and \mathcal{G}' are *Markov equivalent*, which we write as:

$$\mathcal{G} \stackrel{\mathcal{M}}{\sim} \mathcal{G}'.$$

So when specifying a law for directed acyclic graphs, we are left with the question of whether or not we should treat Markov equivalent graphs as the same model. In other words, whether the model is defined by the graph or

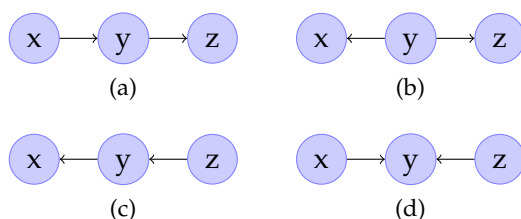


Figure 5.1: Four directed acyclic graphs with the same skeleton. Graphs (a), (b) and (c) are Markov equivalent, and encode the property $x \perp\!\!\!\perp z \mid y$. Graph (d) has the property $x \perp\!\!\!\perp z$.

the set of conditional independence statements which it encodes. We take the latter view.

To simplify notation, we define a *dagoid* to be a Markov equivalence class of directed acyclic graphs. Furthermore, we can define the *complete* and *sparse* dagoids to be the Markov equivalence classes of a complete and sparse DAGs, respectively. We will define \mathfrak{D}^M to be the set of dagoids on V .

A further advantage of working with equivalence classes is that a smaller number of models need be considered. However this may not be as beneficial as one may initially hope: Castelo and Kočka (2004) observed empirically that the ratio of the number DAGs to the number of equivalence classes appears to converge to approximately 3.7, as the number of vertices increases.

Numerous methods of characterising Markov equivalence have arisen:

Skeleton and immoralities

The *skeleton* of a DAG is the undirected graph obtained by substituting the directed edges for undirected ones. A triplet (a, b, c) of vertices is an *immorality* of a DAG \mathcal{G} if the induced graph $\mathcal{G}_{\{a,b,c\}}$ is of the form $a \rightarrow b \leftarrow c$.

Theorem 5.2.1 (Verma and Pearl 1990, Theorem 1, 1992, Corollary 3.2; Frydenberg 1990, Theorem 5.6; Andersson, Madigan, and Perlman 1997a, Theorem 2.1). *Directed acyclic graphs \mathcal{G} and \mathcal{G}' are Markov equivalent if and only if they have the same skeleton and the same immoralities.*

Essential graphs

An essential graph is a unique graphical representation of an equivalence class. An edge of a DAG \mathcal{G} is *essential* if it has the same direction in all Markov equivalent DAGs. The *essential graph* of \mathcal{G} is the graph in which all non-essential edges are replaced by undirected edges.

Although not explored further in this work, the essential graph is a type of *chain graph*, a class of graphs which may have both directed and undirected edges. For further details on chain graphs, in particular their Markov properties and how they relate to undirected and directed acyclic graphs, see Frydenberg (1990) and Andersson, Madigan, and Perlman (1997b).

Theorem 5.2.2 (Andersson, Madigan, and Perlman 1997a, Proposition 4.3). *Directed acyclic graphs \mathcal{G} and \mathcal{G}' are Markov equivalent if and only if they have the same essential graph.*

Unfortunately, there is no simple criteria for determining whether or not an edge of a given DAG is essential—Andersson, Madigan, and Perlman (1997a) proposed an iterative algorithm—which limits their usefulness.

Covered edge reversals

A convenient characterisation of Markov equivalence can be given in terms of edge reversals. An edge $a \rightarrow b$ of a DAG \mathcal{G} is *covered* if $\text{pa}(b) = \text{pa}(a) \cup \{a\}$.

Theorem 5.2.3 (Chickering 1995, Theorem 2). *Directed acyclic graphs \mathcal{G} and \mathcal{G}' are Markov equivalent if and only if there exists a sequence of DAGs:*

$$\mathcal{G} = \mathcal{G}_0, \mathcal{G}_1, \dots, \mathcal{G}_{k-1}, \mathcal{G}_k = \mathcal{G}'$$

such that each $(\mathcal{G}_{i-1}, \mathcal{G}_i)$ differ only by the reversal of one covered edge.

As we shall see, this result is particularly useful for identifying properties that are preserved under Markov equivalence, as we only need to show it is preserved under a covered edge reversal.

Imsets

The *standard imset* of a directed acyclic graph \mathcal{G} is:

$$u_{\mathcal{G}} = \delta_V - \delta_{\emptyset} + \sum_{v \in V} \left[\delta_{\text{pa}_{\mathcal{G}}(v)} - \delta_{\text{pa}_{\mathcal{G}}(v) \cup \{v\}} \right]$$

See Studený (2005b, Page 135).

Theorem 5.2.4 (Studený 2005b, Corollary 7.1). *Directed acyclic graphs \mathcal{G} and \mathcal{G}' are Markov equivalent if and only if $u_{\mathcal{G}} = u_{\mathcal{G}'}$.*

Studený and Vomlel (2009) give details of the relationship between the imset and the essential graph of a DAG, and how one may be obtained from the other.

5.3 Ancestral sets and remainder dagoids

Ancestral sets play a key role in the theory of directed acyclic graphical models. In particular, we note the separoid $\mathcal{M}(\mathcal{G})$ is defined in (1.2) in terms of ancestral sets. However ancestral sets are not preserved under Markov equivalence, that is, an ancestral set in one graph \mathcal{G} need not be ancestral in another Markov equivalent graph \mathcal{G}' . For example, in Figure 5.1, $\{x, y\}$ is ancestral in (a) and (b), but not in (c).

Recall from Theorem 1.2.5 that subgraphs induced by ancestral sets preserve the projection of the separoid. A somewhat trivial consequence is the following:

Proposition 5.3.1. *Let $\mathcal{G} \stackrel{\mathcal{M}}{\sim} \mathcal{G}'$, and $A \subseteq V$ be ancestral in both \mathcal{G} and \mathcal{G}' . Then $\mathcal{G}_A \stackrel{\mathcal{M}}{\sim} \mathcal{G}'_A$.*

This leads to our definition of an ancestral set for a dagoid:

Definition 5.3.1. A set $A \subseteq V$ is *ancestral* in a dagoid \mathcal{D} if it is ancestral for some graph $\mathcal{G} \in \mathcal{D}$. For any such A , define \mathcal{D}_A , the *subdagoid induced by A* , to be the Markov equivalence class of \mathcal{G}_A .

We further define $\mathfrak{D}(A) \subseteq \mathfrak{D}^{\mathcal{M}}$ to be the set of dagoids in which A is an ancestral set.

We note that this property is not quite as strong as the collapsibility property in undirected graphs, in that non-ancestral sets may also preserve the separoid of the induced subgraph. For example, in Figure 5.1 (d), the set $\{x, y\}$ is not ancestral, but induced subgraph preserves the (trivial) separoid.

Definition 5.3.2 (Ancestral insertion). Let \mathcal{G} be a directed acyclic graph on V , of which A is an ancestral set, and let \mathcal{H} be a directed acyclic graph on

A. Then the *insertion of \mathcal{H} into \mathcal{G}* , written:

$$\mathcal{H} \times \mathcal{G}$$

is the directed acyclic graph on V with edge set:

$$\mathcal{E}(\mathcal{H}) \cup [\mathcal{E}(\mathcal{G}) \setminus A^2].$$

In other words, the edges between elements of A are determined by \mathcal{H} , and all other edges are determined by \mathcal{G} .

The graph insertion operation preserves Markov equivalence:

Lemma 5.3.2. *Let \mathcal{G} and \mathcal{G}' be Markov equivalent graphs in which A is an ancestral set, and \mathcal{H} and \mathcal{H}' be Markov equivalent graphs on A . Then:*

$$\mathcal{H} \times \mathcal{G} \stackrel{\mathcal{M}}{\sim} \mathcal{H}' \times \mathcal{G}'$$

Proof. Both graphs must have the same skeleton. Let (a, b, c) be an immorality in $\mathcal{H} \times \mathcal{G}$. Then if $b \in A$, then (a, b, c) must be an immorality of \mathcal{H} , and hence also an immorality of \mathcal{H}' , and so also of $\mathcal{H}' \times \mathcal{G}'$.

Otherwise if $b \notin A$, and at least one of a or c is not in A , then (a, b, c) must be an immorality of \mathcal{G} , and hence an immorality of \mathcal{G}' and $\mathcal{H}' \times \mathcal{G}'$.

Finally, if $b \notin A$ and $a, c \in A$, then $\{a, c\}$ must not be an edge in the skeleton \mathcal{H} , nor an edge in the skeleton of \mathcal{H}' . Hence it must also be an immorality of $\mathcal{H}' \times \mathcal{G}'$. \square

As a consequence of Lemma 5.3.2, for a dagoid \mathcal{D} with ancestral set A , we can define the *ancestral insertion* of a dagoid \mathcal{K} on A into \mathcal{D} as:

$$\mathcal{K} \times \mathcal{D} = [\mathcal{H} \times \mathcal{G}]_{\mathcal{M}}$$

where $\mathcal{G} \in \mathcal{D}$ is a directed acyclic graph with an ancestral set A , and $\mathcal{H} \in \mathcal{K}$.

We can use the idea of an insertion to partition the separoid of a directed acyclic graph.

Definition 5.3.3. Let A be an ancestral set of a directed acyclic graph \mathcal{G} . A directed acyclic graph $\mathcal{G}_{V|A}$ is a *remainder graph of \mathcal{G} given A* if:

$$\mathcal{G}_{V|A} = \mathcal{C}^{(A)} \times \mathcal{G}$$

where $\mathcal{C}^{(A)}$ is a complete dagoid on A .

By Lemma 5.3.2, the remainder graph must be unique up to Markov equivalence. Hence for a dagoid $\mathcal{D} \in \mathfrak{D}(A)$, we can uniquely define the *remainder dagoid of \mathcal{D} given A* , denoted by $\mathcal{D}_{V|A}$.

The name comes from the fact that $\mathcal{M}(\mathcal{D}_A)$ and $\mathcal{M}(\mathcal{D}_{V|A})$ form a spanning subset of $\mathcal{M}(D)$.

Theorem 5.3.3. *Let A be an ancestral set of a directed acyclic graph \mathcal{G} . Then:*

$$\mathcal{M}(\mathcal{G}) = \overline{\mathcal{M}(\mathcal{G}_A) \cup \mathcal{M}(\mathcal{G}_{V|A})}$$

where \overline{S} denotes the Markov closure of a set of conditional independence statements S .

Proof. Recall that $\mathcal{M}(\mathcal{G})$ is spanned by the set of elements of the form:

$$\langle \{v\}, \text{pr}_{\prec}(v) \mid \text{pa}_{\mathcal{G}}(v) \rangle \quad (5.3)$$

where \prec is a well-ordering in which the elements of A precede those of $V \setminus A$. If $v \in A$, then (5.3) will be an element of $\mathcal{M}(\mathcal{G}_A)$, otherwise if $v \notin A$, it will be an element of $\mathcal{M}(\mathcal{G}_{V|A})$. \square

Furthermore, the induced and remainder dagoids are variation independent:

Theorem 5.3.4. *For any $A \subseteq V$, we have:*

$$\mathcal{D}_A \ddagger \mathcal{D}_{V|A} \mid \{\mathcal{D} \in \mathfrak{D}(A)\}$$

Proof. For any $\mathcal{D}, \mathcal{D}' \in \mathfrak{D}A$, we can construct $\mathcal{D}^* = \mathcal{D}_A \times \mathcal{D}'_{V|A}$. This will have the required properties that $\mathcal{D}_A^* = \mathcal{D}_A$ and $\mathcal{D}_{V|A}^* = \mathcal{D}'_{V|A}$. \square

5.4 Structural Markov property

Recall the strong hyper Markov property for the law $\mathcal{L}(\tilde{\theta})$ may be expressed as:

$$\prod_{v \in V} \tilde{\theta}_{v \mid \text{pa}(v)} \quad [\mathcal{L}]$$

For any ancestral set A of \mathcal{G} , we can write:

$$\theta_A \simeq (\theta_{v \mid \text{pa}(v)})_{v \in A} \quad \text{and} \quad \theta_{V|A} \simeq (\theta_{v \mid \text{pa}(v)})_{v \notin A}$$

Therefore, an alternative characterisation of the strong hyper Markov property is:

$$\tilde{\theta}_A \perp\!\!\!\perp \tilde{\theta}_{V|A} \quad [\mathcal{L}]$$

for any ancestral set A of \mathcal{G} .

This motivates the following definition:

Definition 5.4.1 (Dagoid structural Markov property). We say a graph law $\mathfrak{G}(\tilde{\mathcal{D}})$ is *structurally Markov* if for any $A \subseteq V$, we have:

$$\tilde{\mathcal{D}}_{V|A} \perp\!\!\!\perp \tilde{\mathcal{D}}_A \mid \{\tilde{\mathcal{D}} \in \mathfrak{D}(A)\} \quad [\mathfrak{G}].$$

As in the undirected case, we can characterise this property via the odds ratio of the density:

Proposition 5.4.1. *A graph law is structurally Markov if and only if for any $\mathcal{D}, \mathcal{D}' \in \mathfrak{D}(A)$, we have:*

$$\pi(\mathcal{D})\pi(\mathcal{D}') = \pi(\mathcal{D}_A \times \mathcal{D}'_{V|A})\pi(\mathcal{D}'_A \times \mathcal{D}_{V|A}). \quad (5.4)$$

Proof. As in Proposition 4.2.2, we may write the density $\pi(\mathcal{D} \mid \mathfrak{D}(A)) = \pi(\mathcal{D}_A \mid \mathfrak{D}(A))\pi(\mathcal{D}_{V|A} \mid \mathfrak{D}(A))$. \square

Example 5.4.1. As in the undirected case, the simplest example of a structurally Markov graph law is the uniform law over $\mathfrak{D}^{\mathcal{M}}$.

However, we note that some simple laws are *not* structurally Markov

Example 5.4.2. Consider the law in which $\pi(\mathcal{D})$ is proportional to $|\mathcal{D}|$, in other words, the uniform law on \mathfrak{D} projected onto $\mathfrak{D}^{\mathcal{M}}$. Then we note the size of the following dagoids:

$$\begin{aligned} [\begin{array}{c} \bullet \bullet \\ \bullet \bullet \end{array}]_{\mathcal{M}} &= \{ \begin{array}{c} \bullet \bullet \\ \bullet \bullet \end{array} \} \\ [\begin{array}{c} \bullet \rightarrow \bullet \\ \bullet \bullet \end{array}]_{\mathcal{M}} &= \{ \begin{array}{c} \bullet \rightarrow \bullet \\ \bullet \bullet \end{array}, \begin{array}{c} \bullet \leftarrow \bullet \\ \bullet \bullet \end{array} \} \\ [\begin{array}{c} \bullet \bullet \\ \swarrow \searrow \\ \bullet \bullet \end{array}]_{\mathcal{M}} &= \{ \begin{array}{c} \bullet \bullet \\ \swarrow \searrow \\ \bullet \bullet \end{array} \} \\ [\begin{array}{c} \bullet \bullet \\ \swarrow \rightarrow \searrow \\ \bullet \bullet \end{array}]_{\mathcal{M}} &= \{ \begin{array}{c} \bullet \bullet \\ \swarrow \rightarrow \searrow \\ \bullet \bullet \end{array}, \begin{array}{c} \bullet \bullet \\ \swarrow \leftarrow \searrow \\ \bullet \bullet \end{array}, \begin{array}{c} \bullet \bullet \\ \rightarrow \swarrow \searrow \\ \bullet \bullet \end{array}, \begin{array}{c} \bullet \bullet \\ \rightarrow \leftarrow \searrow \\ \bullet \bullet \end{array}, \begin{array}{c} \bullet \bullet \\ \leftarrow \swarrow \searrow \\ \bullet \bullet \end{array}, \begin{array}{c} \bullet \bullet \\ \leftarrow \leftarrow \searrow \\ \bullet \bullet \end{array} \} \end{aligned}$$

As a consequence, this law doesn't satisfy Proposition 5.4.1, when $\mathcal{D} = \begin{array}{c} \bullet \bullet \\ \swarrow \searrow \\ \bullet \bullet \end{array}$ and $\mathcal{D}' = \begin{array}{c} \bullet \bullet \\ \bullet \bullet \end{array}$, and A is chosen as the two top vertices.

5.5 d-Clique vector

The equivalence class formulation of a dagoid is difficult to work with both algebraically and computationally. Instead we propose a vector similar to the clique vector of the previous chapter.

Definition 5.5.1. The *d-clique vector* of a directed acyclic graph \mathcal{G} is:

$$t(\mathcal{G}) = \sum_{v \in V} [\delta_{\{v\} \cup \text{pa}_{\mathcal{G}}(v)} - \delta_{\text{pa}_{\mathcal{G}}(v)}] + \delta_{\emptyset} \in \mathbb{Z}^{2^V} \quad (5.5)$$

where:

$$(\delta_A)_I = \begin{cases} 1 & \text{if } I = A \\ 0 & \text{if } I \neq A \end{cases}.$$

Again, we note the similarity of this vector to the imsets of Studený (2005b), specifically the structural imset $t(\mathcal{G}) = \delta_V - u_{\mathcal{G}}$ of in section 5.2

In a similar manner to the undirected case, we can define the *d-completeness vector* to be the Möbius transform of the d-clique vector:

$$c_A(\mathcal{G}) = \sum_{B \supseteq A} t_B(\mathcal{G}), \quad (5.6)$$

and say that a set $A \subseteq B$ is *d-complete* if $c_A(\mathcal{G}) = 1$.

Lemma 5.5.1. Let \prec be a well-ordering of a directed acyclic graph \mathcal{G} . Then for any non-empty set $A \subseteq V$:

$$c_A(\mathcal{G}) = \begin{cases} 1 & \text{if } A \setminus \{a\} \subseteq \text{pa}_{\mathcal{G}}(a), \\ 0 & \text{otherwise,} \end{cases}$$

where a is the maximal element of A under \prec .

Proof. By substituting (5.5) into (5.6):

$$c_A(\mathcal{G}) = \sum_{v \in V} \mathbb{1}_{A \subseteq \text{pa}_{\mathcal{G}}(v) \cup \{v\}} - \mathbb{1}_{A \subseteq \text{pa}_{\mathcal{G}}(v)}.$$

These terms will cancel out unless $v \in A$. Furthermore, $A \subseteq \text{pa}_{\mathcal{G}}(v) \cup \{v\}$ only if all $u \prec v$ for all $u \in A$. Hence:

$$c_A(\mathcal{G}) = \mathbb{1}_{A \subseteq \text{pa}_{\mathcal{G}}(a) \cup \{a\}}. \quad \square$$

This gives the link to the previous chapter, in particular the cliques and complete sets:

Corollary 5.5.2. *If \mathcal{G} is a perfect directed acyclic graph and \mathcal{G}^s is its skeleton, then $c_{\mathcal{G}} = c_{\mathcal{G}^s}$, and hence $t(\mathcal{G}) = t(\mathcal{G}^s)$.*

Most importantly, the d-clique vector is a unique representation of the dagoid:

Theorem 5.5.3. *Let $\mathcal{G}, \mathcal{G}'$ be directed acyclic graphs on V . Then $\mathcal{G} \stackrel{\mathcal{M}}{\sim} \mathcal{G}'$ if and only if $t(\mathcal{G}) = t(\mathcal{G}')$.*

Proof. To show the d-clique vector is preserved under Markov equivalence, by Theorem 5.2.3 it is sufficient to show that it is preserved under a covered edge reversal. If (a, b) is a covered edge of \mathcal{G} , then the contribution of these vertices to the sum (5.5) is:

$$\begin{aligned} t(\mathcal{G}) &= [\delta_{\{a\} \cup \text{pa}_{\mathcal{G}}(a)} - \delta_{\text{pa}_{\mathcal{G}}(a)}] + [\delta_{\{b\} \cup \text{pa}_{\mathcal{G}}(b)} - \delta_{\text{pa}_{\mathcal{G}}(b)}] \\ &\quad + \sum_{v \neq a, b} [\delta_{\{b\} \cup \text{pa}_{\mathcal{G}}(b)} - \delta_{\text{pa}_{\mathcal{G}}(b)}] + \delta_{\emptyset} \end{aligned}$$

By definition $\text{pa}_{\mathcal{G}}(a) \cup \{a\} = \text{pa}_{\mathcal{G}}(b)$, and so the corresponding terms will cancel. If \mathcal{G}^* is obtained from \mathcal{G} by reversing (a, b) , note that:

$$\text{pa}_{\mathcal{G}}(a) = \text{pa}_{\mathcal{G}^*}(b) \quad \text{and} \quad \text{pa}_{\mathcal{G}}(b) \cup \{b\} = \text{pa}_{\mathcal{G}^*}(a) \cup \{a\},$$

and the remaining terms will be unchanged. Hence $t(\mathcal{G}) = t(\mathcal{G}^*)$.

To show that the d-completeness vector (and hence, also the d-clique vector) is unique to the equivalence class, by Theorem 5.2.1 we can show that it determines the skeleton and immoralities. By Lemma 5.5.1, there is an edge between u and v in \mathcal{G} if and only if $c_{\{u,v\}}(\mathcal{G}) = 1$. Likewise, (u, v, w) is an immorality if and only if $c_{\{u,v,w\}}(\mathcal{G}) = 1$ and $c_{\{u,w\}}(\mathcal{G}) = 0$. \square

This cancellation of terms involving covered edges is very useful: as a consequence, the d-clique vector will generally be quite sparse. In line with the clique vector, we term sets $A \subseteq V$ such that $t_A(\mathcal{D}) = 1$ to be a *d-clique*, and the sets where $t_A(\mathcal{D}) < 0$ to be *d-separators*. See Figure 5.2 for examples.

Theorem 5.5.4. *Let A be an ancestral set of a dagoid \mathcal{D} . Then:*

$$t(\mathcal{D}) = [t(\mathcal{D}_A)]^0 + t(\mathcal{D}_{V \setminus A}) - \delta_A,$$

where $[\cdot]^0$ denotes the expansion of the vector with zeroes to the required coordinates.

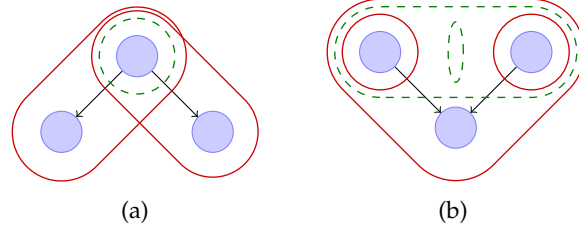


Figure 5.2: The d-cliques (—) and d-separators (---) of different directed acyclic graphs. Note that in the perfect DAG (a), the d-cliques and d-separators are the cliques and separators of the skeleton. However, as in (b), d-separators may contain d-cliques.

Proof. Let $\mathcal{G} \in \mathcal{D}$ in which A is ancestral, and \prec be a well-ordering of \mathcal{G} in which elements of A precede those of $V \setminus A$.

Note that:

$$\text{pa}_{\mathcal{G}}(v) = \begin{cases} \text{pa}_{\mathcal{G}_A}(v) & v \in A, \\ \text{pa}_{\mathcal{G}_{V|A}}(v) & v \notin A. \end{cases}$$

The result then follows after noting that:

$$\sum_{v \in A} [\delta_{(\text{pa}_{\mathcal{G}_{V|A}}(v) \cup \{v\})} - \delta_{\text{pa}_{\mathcal{G}_{V|A}}(v)}] = \delta_A \quad \square$$

We now arrive at the key result of this section: the dagoid structural Markov property characterises an exponential family of graph laws.

Theorem 5.5.5. *Let \mathfrak{G} whose support is $\mathfrak{D}^{\mathcal{M}}$. Then \mathfrak{G} is structurally Markov if and only if it is a member of an exponential family with the d-clique vector as a sufficient statistic. That is, \mathfrak{G} has density:*

$$\pi_{\omega}(\mathcal{D}) \propto \exp\{\omega \cdot t(\mathcal{D})\} \quad (5.7)$$

Proof. If the law is in the exponential family in (5.7), by Theorem 5.5.4, we have:

$$\pi(\mathcal{D}|\mathfrak{D}(A)) \propto \exp\{\omega \cdot [t(\mathcal{D}_A) + t(\mathcal{D}_{V|A})] - \omega_A\} \propto p(\mathcal{D}_A|\mathfrak{D}(A))p(\mathcal{D}_{V|A}|\mathfrak{D}(A))$$

hence the law must be structurally Markov.

For the converse, define $\mathcal{D}^{(A)}$ to be the dagoid in which the induced dagoid on $A \subseteq V$ is complete, but otherwise sparse (in other words, the remainder dagoid $\mathcal{D}_{V|A}^{(\emptyset)}$ of the sparse dagoid $\mathcal{D}^{(\emptyset)}$).

Select some $\mathcal{G} \in \mathcal{D}$, and let v_1, \dots, v_d be a well ordering of V . Recursively define the dagoids:

$$\mathcal{D}^{*(i)} = \begin{cases} \mathcal{D}(\{v_1\}) & \text{if } i = 1, \\ \mathcal{D}_{\text{pr}(v_i)}^{*(i-1)} \times \mathcal{D}_{v_i | \text{pr}(v_i)}^{\{\{v_i\} \cup \text{pa}(v_i)\}} & \text{otherwise.} \end{cases}$$

By Proposition 5.4.1, for $i = 2, \dots, d$:

$$\pi(\mathcal{D}^{*(i-1)})\pi(\mathcal{D}^{\{\{v_i\} \cup \text{pa}(v_i)\}}) = \pi(\mathcal{D}^{*(i)})\pi(\mathcal{D}_{\text{pr}(v_i)}^{\{\{v_i\} \cup \text{pa}(v_i)\}} \times \mathcal{D}_{v_i | \text{pr}(v_i)}^{*(i-1)})$$

However,

$$\mathcal{D}_{\text{pr}(v_i)}^{\{\{v_i\} \cup \text{pa}(v_i)\}} \times \mathcal{D}_{v_i | \text{pr}(v_i)}^{*(i-1)} = \mathcal{D}(\text{pa}(v_i))$$

Therefore, since $\mathcal{D}^{*(d)} = \mathcal{D}$, then:

$$\pi(\mathcal{D}) = \frac{\prod_{i=1}^d \pi(\mathcal{D}^{\{\{v_i\} \cup \text{pa}(v_i)\}})}{\prod_{i=2}^d \pi(\mathcal{D}(\text{pa}(v_i)))}.$$

which is of the form in (5.7), where:

$$\omega_A = \log \pi(\mathcal{D}^{(A)}). \quad \square$$

We note that a similar exponential families were proposed by Mukherjee and Speed (2008), however they treat Markov equivalent graphs as distinct, and allow them to have different probabilities.

5.6 Compatibility

As with the undirected case, a graph law is only part of the story. For each dagoid \mathcal{D} , we also require a method to specify either a Markov sampling distribution, or a law over such sampling distributions.

Definition 5.6.1. Distributions θ and θ' which are Markov with respect to directed acyclic graphs \mathcal{G} and \mathcal{G}' , respectively, are *graph compatible* if for every vertex v where $\text{pa}_{\mathcal{G}}(v) = \text{pa}_{\mathcal{G}'}(v)$, there exists versions of the conditional probability distributions for $X_v | X_{\text{pa}(v)}$ such that:

$$\theta(X_v | X_{\text{pa}(v)}) = \theta'(X_v | X_{\text{pa}(v)}).$$

Furthermore, distributions θ and θ' which are Markov with respect to dagoids \mathcal{D} and \mathcal{D}' , respectively, are (*dagoid compatible*) if they are graph compatible for every pair of graphs $\mathcal{G} \in \mathcal{D}, \mathcal{G}' \in \mathcal{D}'$.

Likewise, laws $\mathcal{L}(\tilde{\theta})$ and $\mathcal{L}'(\tilde{\theta})$, hyper Markov with respect to \mathcal{G} and \mathcal{G}' respectively, are *graph hyper compatible* if for every vertex v where $\text{pa}_{\mathcal{G}}(v) = \text{pa}_{\mathcal{G}'}(v)$, there exists versions of the conditional laws for $\tilde{\theta}_{v|\text{pa}(v)} | \tilde{\theta}_{\text{pa}(v)}$ such that:

$$\mathcal{L}(\tilde{\theta}_{v|\text{pa}(v)} | \tilde{\theta}_{\text{pa}(v)}) = \mathcal{L}'(\tilde{\theta}_{v|\text{pa}(v)} | \tilde{\theta}_{\text{pa}(v)}).$$

Recall that by Theorem 1.3.1 the weak hyper Markov property may be characterised in terms of $\mathcal{M}(\mathcal{G})$, and so the weak hyper Markov property can be defined with respect to a dagoid. Laws $\mathcal{L}(\tilde{\theta})$ and $\mathcal{L}'(\tilde{\theta})$, that are hyper Markov with respect to \mathcal{D} and \mathcal{D}' , respectively, are (*dagoid hyper compatible*) if they are graph compatible for every pair of graphs $\mathcal{G} \in \mathcal{D}, \mathcal{G}' \in \mathcal{D}'$.

As in the undirected case, we can define a family of compatible distributions $\theta = \{\theta^{(\mathcal{G})} : \mathcal{G} \in \mathcal{U}\}$ and a family of hyper compatible laws $\mathcal{L} = \{\mathcal{L}^{(\mathcal{G})} : \mathcal{G} \in \mathcal{U}\}$ if they are pairwise compatible or hyper compatible with respect to the relevant graphs.

Proposition 5.6.1. *Suppose $\mathfrak{G}(\tilde{\mathcal{D}})$ is a graph law over $\mathfrak{D}^{\mathcal{M}}$ and θ is a family of compatible distributions. Then:*

$$X_A \perp\!\!\!\perp \tilde{\mathcal{D}}_{V|A} | \tilde{\mathcal{D}}_A, \{\tilde{\mathcal{D}} \in \mathfrak{D}(A)\} \quad [\theta, \mathfrak{G}] \quad (5.8)$$

and

$$X_{V \setminus A} \perp\!\!\!\perp \tilde{\mathcal{D}}_A | X_A, \tilde{\mathcal{D}}_{V|A}, \{\tilde{\mathcal{D}} \in \mathfrak{D}(A)\} \quad [\theta, \mathfrak{G}]. \quad (5.9)$$

Likewise, if $\mathfrak{G}(\tilde{\mathcal{D}})$ is a graph law over $\mathfrak{D}^{\mathcal{M}}$ and \mathcal{L} is a hyper compatible family of laws, then:

$$\tilde{\theta}_A \perp\!\!\!\perp \tilde{\mathcal{D}}_{V|A} | \tilde{\mathcal{D}}_A, \{\tilde{\mathcal{D}} \in \mathfrak{D}(A)\} \quad [\mathcal{L}, \mathfrak{G}]$$

and

$$\tilde{\theta}_{V \setminus A|A} \perp\!\!\!\perp \tilde{\mathcal{D}}_A | \tilde{\theta}_A, \tilde{\mathcal{D}}_{V|A}, \{\tilde{\mathcal{D}} \in \mathfrak{D}(A)\} \quad [\mathcal{L}, \mathfrak{G}].$$

Proof. This is much the same as Proposition 4.4.1: for (5.8), the distribution for X_A are determined by the parent sets of the vertices in A in some $\mathcal{G} \in \mathfrak{D}$ in which A is ancestral. Likewise, in (5.9), the conditional distribution for $X_{V \setminus A} | X_A$ is determined by the parents sets of vertices in $V \setminus A$. The same argument applies at the hyper level. \square

Note that in the definition of compatibility and hyper compatibility we specifically refer to *versions* of conditional probabilities and laws, as in some cases the conditional distributions/laws will not be uniquely defined, due to the Borel–Kolmogorov paradox.

Example 5.6.1. Suppose the joint distribution P on a triplet of binary variables (X, Y, Z) has $P(X = 1, Y = 1) = 0$, but with $P(X = 1) > 0$ and $P(Y = 1) > 0$. Then the conditional distribution $P(Z = 1 | X = 1, Y = 1)$ is not uniquely defined.

Now consider a compatible distribution P' on the graph:



Then we have $P'(X = 1, Y = 1) = P(X = 1)P(Y = 1) > 0$. Therefore $P'(X = 1, Y = 1, Z = 1)$ may be defined arbitrarily, as for any conditional probability $P'(Z = 1 | X = 1, Y = 1)$, there will exist a corresponding version of $P(Z = 1 | X = 1, Y = 1)$

We could avoid this type of ambiguity in the case of compatible distributions by requiring that the density be positive with respect to some product measure. However the situation isn't so simple at the hyper level:

Example 5.6.2. Consider a law $\mathcal{L}(\tilde{\theta})$ for a triplet of binary variables (X, Y, Z) , and suppose that it is continuous on the full probability simplex.

A hyper compatible law \mathcal{L}' on the graph in (5.10), will have marginal laws $\mathcal{L}'(\tilde{\theta}_X) = \mathcal{L}(\tilde{\theta}_X)$ and $\mathcal{L}'(\tilde{\theta}_Y) = \mathcal{L}(\tilde{\theta}_Y)$. This means the joint law $\mathcal{L}'(\tilde{\theta}_{XY})$ will be their product law, which is concentrated on the manifold $X \perp\!\!\!\perp Y$.

As this manifold will have probability 0 under \mathcal{L} , we may define the conditional law $\mathcal{L}'(\tilde{\theta}_{Z|XY} | \tilde{\theta}_{XY})$ arbitrarily.

It is possible to uniquely define such conditional laws if we impose further conditions, such as the existence of a continuous density for $\mathcal{L}(\tilde{\theta})$. However we can also resolve the problem by insisting on a dagoid form of the strong hyper Markov property:

Definition 5.6.2. A law $\mathcal{L}(\tilde{\theta})$ is over $\mathfrak{B}(\mathcal{D})$ is *strong hyper Markov* with respect to \mathcal{D} if it is strong directed hyper Markov with respect to every $\mathcal{G} \in \mathcal{D}$.

Note that if $\mathcal{G} \in \mathcal{D}$ is perfect, then the dagoid strong hyper Markov property is equivalent to the undirected strong hyper Markov property on the skeleton of \mathcal{G} (see Dawid and Lauritzen 1993, Proposition 3.15).

The notion of hyper compatibility is equivalent to the “parameter modularity” property of Heckerman, Geiger, and Chickering (1995). Likewise, the strong hyper Markov property is equivalent to their “parameter independence”

Example 5.6.3 (Dagoid hyper inverse Wishart law). We may extend the hyper inverse Wishart law in section 1.6 to dagoids. For each vertex v of a directed acyclic graph \mathcal{G} , we define the law for the conditional parameter $\mathcal{L}(\tilde{\theta}_{v|\text{pa}_{\mathcal{G}}(v)})$ to be the same as that of the inverse Wishart $\mathcal{IW}(\delta; \Phi)$. That is:

$$\begin{aligned}\mathcal{L}(\tilde{\Sigma}_{v|\text{pa}_{\mathcal{G}}(v)}) &= \mathcal{IW}(\delta + |\text{pa}_{\mathcal{G}}(v)|; \Phi_{v|\text{pa}_{\mathcal{G}}(v)}) \\ \mathcal{L}(\tilde{\Gamma}_{v|\text{pa}_{\mathcal{G}}(v)} | \tilde{\Sigma}_{v|\text{pa}_{\mathcal{G}}(v)}) &= \Phi_{\{v\}, \text{pa}_{\mathcal{G}}(v)} \Phi_{\text{pa}_{\mathcal{G}}(v)}^{-1} + \mathcal{N}_{\{v\} \times \text{pa}_{\mathcal{G}}(v)}(\tilde{\Sigma}_{v|\text{pa}_{\mathcal{G}}(v)}, \Phi_{\text{pa}_{\mathcal{G}}(v)}^{-1})\end{aligned}$$

By the properties of the inverse Wishart law, it follows that the law derived under a covered edge reversal will be identical, hence may be defined by the dagoid. Furthermore, by the above definition, it is hyper compatible.

Theorem 5.6.2. *If \mathcal{L} is a family of strong hyper Markov hyper compatible laws, then the family of marginal data distributions is compatible*

Proof. The hyper compatibility and the strong hyper Markov property imply that for any two dagoids $\mathcal{D}, \mathcal{D}'$, and any $\mathcal{G} \in \mathcal{D}, \mathcal{G}' \in \mathcal{D}'$, that if $\text{pa}_{\mathcal{G}}(v) = \text{pa}_{\mathcal{G}'}(v)$ for some $v \in V$, then:

$$\mathcal{L}^{(\mathcal{D})}(\tilde{\theta}_{v|\text{pa}}) = \mathcal{L}^{(\mathcal{D}')}(\tilde{\theta}_{v|\text{pa}})$$

Therefore, the family of marginal data distributions $\bar{\theta} = \{\tilde{\theta}^{(\mathcal{D})} : \mathcal{D} \in \mathfrak{D}^{\mathcal{M}}\}$ will have:

$$\bar{\theta}^{(\mathcal{D})}(X_v | X_{\text{pa}_{\mathcal{G}}}) = \mathbb{E}_{\mathcal{L}^{(\mathcal{D})}}[\tilde{\theta}_{v|\text{pa}_{\mathcal{G}}}] = \bar{\theta}^{(\mathcal{D}')} (X_v | X_{\text{pa}_{\mathcal{G}}}) = \mathbb{E}_{\mathcal{L}^{(\mathcal{D}')}}[\tilde{\theta}_{v|\text{pa}_{\mathcal{G}}}]. \quad \square$$

This is particularly useful because, as in the undirected case, the structural Markov property will be preserved in the posterior under compatible sampling:

Theorem 5.6.3. *Suppose $\mathfrak{G}(\tilde{\mathcal{D}})$ is a structurally Markov graph law over $\mathfrak{D}^{\mathcal{M}}$ and θ is a family of compatible distributions. Then the posterior graph law for $\tilde{\mathcal{D}}$ is structurally Markov.*

Proof. By the structural Markov property and (5.8), we have:

$$(X_A, \tilde{\mathcal{D}}_A) \perp\!\!\!\perp \tilde{\mathcal{D}}_{V|A} \mid \{\tilde{\mathcal{D}} \in \mathfrak{D}(A)\}$$

and hence:

$$\tilde{\mathcal{D}}_A \perp\!\!\!\perp \tilde{\mathcal{D}}_{V|A} \mid X_A, \{\tilde{\mathcal{D}} \in \mathfrak{D}(A)\}$$

Combining this with (5.9), we get:

$$\tilde{\mathcal{D}}_A \perp\!\!\!\perp (\tilde{\mathcal{D}}_{V|A}, X_{V \setminus A}) \mid X_A, \{\tilde{\mathcal{D}} \in \mathfrak{D}(A)\}$$

and hence:

$$\tilde{\mathcal{D}}_A \perp\!\!\!\perp \tilde{\mathcal{D}}_{V|A} \mid X, \{\tilde{\mathcal{D}} \in \mathfrak{D}(A)\} \quad \square$$

We can specify a compatible family by a positive density on the complete dagoid:

Theorem 5.6.4. *If the distribution on the complete dagoid has positive density p (with respect to some product measure), then the compatible distribution for any dagoid \mathcal{D} , has density:*

$$p^{(\mathcal{D})}(x) = \prod_{A \subseteq V} [p(x_A)]^{t(\mathcal{D})_A} \quad (5.11)$$

Proof. Let \mathcal{G} be an arbitrary graph in \mathcal{D} . Then by compatibility:

$$p^{(\mathcal{D})}(x) = \prod_{v \in V} p(x_v | x_{\text{pa}(v)}) = \frac{\prod_{i=1}^p p(x_{\{v_i\} \cup \text{pa}(v_i)})}{\prod_{i=2}^p p(x_{\text{pa}(v_i)})} = \prod_{A \subseteq V} [p(x_A)]^{t(\mathcal{D})_A} \quad \square$$

As a consequence, if the graph law has a d -clique exponential family of the form (5.7), and the sampling distributions are compatible with density of the form (5.11), then the posterior graph law will have density:

$$\pi(\mathcal{D} \mid X) \propto \exp\{[\omega + (\log p_A(X_A))_{A \subseteq V}] \cdot t(\mathcal{D})\}.$$

That is, the d -clique exponential family is a conjugate prior under sampling from a compatible family.

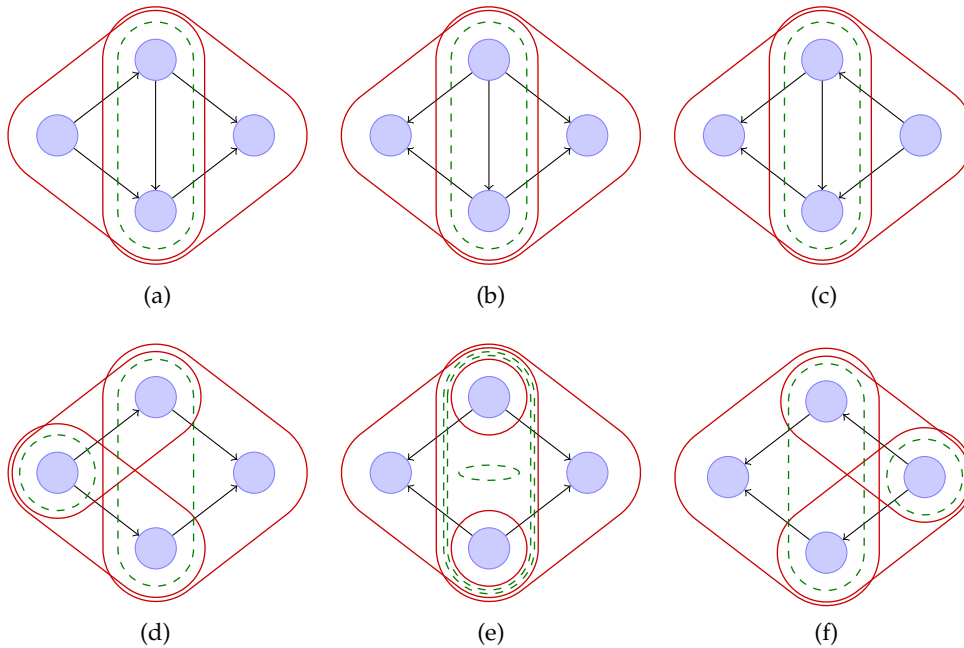


Figure 5.3: Three Markov equivalent graphs, (a), (b) and (c), in which the same edge removal will result in a transition to a distinct Markov equivalence class, (d), (e) and (f), respectively. The d-cliques (—) and d-separators (---) of each graph are also drawn.

5.7 Computation

As in the undirected case, the normalisation constant of the exponential family has no analytic form. Furthermore, as the number of dagoids will increase superexponentially with the number of vertices, simple enumeration of all possible dagoids is generally not be feasible. Thus we will usually need to resort to some approximation method.

Unfortunately, specification of such a MCMC algorithm for dagoids is much more difficult than for the undirected case. Specifically, an individual edge no longer uniquely characterises a neighbouring dagoid, as in Figure 5.3.

If the directed graph \mathcal{G}^+ is obtained from the directed graph \mathcal{G} by the addition of the edge (u, v) , the only term in the summation (5.5) that will

change is those pertaining to the vertex v , in which case:

$$\begin{aligned} t(\mathcal{G}^+) - t(\mathcal{G}) &= (\delta_{\{v\} \cup \text{pa}_{\mathcal{G}^+}(v)} - \delta_{\text{pa}_{\mathcal{G}^+}(v)}) - (\delta_{\{v\} \cup \text{pa}_{\mathcal{G}}(v)} - \delta_{\text{pa}_{\mathcal{G}}(v)}) \\ &= \delta_{\text{pa}_{\mathcal{G}}(v) \cup \{u, v\}} - \delta_{\text{pa}_{\mathcal{G}}(v) \cup \{u\}} - \delta_{\text{pa}_{\mathcal{G}}(v) \cup \{v\}} + \delta_{\text{pa}_{\mathcal{G}}(v)} \end{aligned}$$

In other words, the change in the d-clique vector is determined by the parent set of v in \mathcal{G} . Therefore, to characterise the neighbouring dagoids (defined as the equivalence classes of the neighbouring graphs), we need to know the parent set of v for each $\mathcal{G} \in \mathcal{D}$. Furthermore, as in the undirected case, computing the ratio of probabilities only requires evaluating the parameter, and hence the marginal likelihood, on 4 subsets.

Notably, Chickering (2003), Auvray and Wehenkel (2002) and Studený (2005a) develop methods for characterising the neighbouring dagoids. Unfortunately, all these methods are rather complicated, and it is not readily apparent how they might be utilised in an efficient MCMC approach.

Another approach is to incorporate an auxiliary variable. One such method is the ‘‘Augmented Markov chain Monte Carlo model composition (AMC³)’’ algorithm proposed by Madigan, Andersson, et al. (1996). At each step, the algorithm proposes a random well-ordering \prec of the vertices, with respect to (some graph contained in) the current dagoid \mathcal{D} , and then proposes a random edge addition or removal consistent with this well-ordering to obtain a new dagoid \mathcal{D}' .

However, as with any auxiliary variable method, there are two related difficulties:

- (a) Specification of the proposal distribution for the auxiliary variable: it is generally quite difficult to characterise the well-orderings of a dagoid. The authors propose a mechanism based on maximum cardinality search, however this will fail to reach all possible orderings.
- (b) Evaluation of the acceptance ratio:

$$\frac{\pi(\mathcal{D}' | \prec)}{\pi(\mathcal{D} | \prec)} = \frac{\pi(\mathcal{D}') \pi(\prec | \mathcal{D}')}{\pi(\mathcal{D}) \pi(\prec | \mathcal{D})}$$

The authors propose two methods of approximating this ratio, but it is unclear if there will have an adverse influence on the accuracy of the computation.

Castelo and Kočka (2004) propose a similar method they term ‘‘enhanced MC³’’, which instead uses the graph itself as the auxiliary variable. At each

step, the algorithm performs a sequence of random covered edge reversals, then proposes a move. Again, they require an approximation of the acceptance ratio.

Discussion

We have demonstrated the usefulness of hyper Markov properties, and shown how they might be extended to the case where the structure of the graph is unknown. In particular, we have shown that these structural Markov properties characterise exponential families on the set of undirected decomposable graphs, and the set of Markov equivalence classes of directed acyclic graphs. Furthermore, when used as priors, they are conjugate with a family of compatible sampling distributions.

One of the significant remaining challenges is computation. Although we have suggested some possible MCMC approaches for evaluating the posterior, these methods will generally become impractical as the number of vertices in the graph increases. This problem could be alleviated somewhat by considering a smaller family of graphs, such as limiting the maximum clique size in undirected graphs.

One disadvantage of sample-based approximations such as MCMC is that the posterior is approximated by a large set of graphs. By averaging over a large number of graphs, we lose some of the convenient aspects of the graphical formulation, such as efficient propagation algorithms for inferring marginal distributions. One possible solution would be to find a graph, or small set of graphs, that best represent the posterior in some way.

A common approach is to simply take the graph with the highest posterior probability, the so-called maximum *a posteriori* or MAP estimate. However, this may not necessarily be optimal: Barbieri and Berger (2004) showed that for regression models, the median model (that which incorporates the variables whose marginal probability greater than $\frac{1}{2}$) can be optimal in terms of future predictive ability.

It would be of interest to know if such a result could extend to graphical model determination, perhaps based on edge inclusion probability, though

6. DISCUSSION

this would depend on the choice loss function. Furthermore, in the undirected case, there is no guarantee that the resultant graph would be decomposable, and it is not clear how such an approach could be applied in the case of dagoids.

Graph terminology

We provide a quick summary of graph terminology used throughout the paper. For further details, see Lauritzen (1996) or Cowell et al. (2007).

A *graph* \mathcal{G} is a pair of finite sets:

- $\mathcal{V}(\mathcal{G})$ of *vertices* or *nodes*, and
- $\mathcal{E}(\mathcal{G})$ of *edges*, which are pairs of distinct vertices.

We say a graph \mathcal{G} is *on* V if $\mathcal{V}(\mathcal{G}) = V$.

A.1 Undirected graphs

In an *undirected graph*, the edges are unordered pairs $\{u, v\}$, that is:

$$\mathcal{E}(\mathcal{G}) \subseteq \binom{\mathcal{V}(\mathcal{G})}{2} = \{A \subseteq \mathcal{V}(\mathcal{G}) : |A| = 2\}.$$

We say \mathcal{G} is *sparse* if $\mathcal{E}(\mathcal{G}) = \emptyset$. Conversely, \mathcal{G} is *complete* if it contains an edge between every pair of vertices, that is $\mathcal{E}(\mathcal{G}) = \binom{\mathcal{V}(\mathcal{G})}{2}$.

A graph \mathcal{G}' is a *subgraph* of \mathcal{G} if:

$$\mathcal{V}(\mathcal{G}') \subseteq \mathcal{V}(\mathcal{G}) \quad \text{and} \quad \mathcal{E}(\mathcal{G}') \subseteq \mathcal{E}(\mathcal{G}).$$

Specifically, it is an *edge subgraph* if $\mathcal{V}(\mathcal{G}') = \mathcal{V}(\mathcal{G})$. The *subgraph (of \mathcal{G}) induced by* $A \subseteq \mathcal{V}(\mathcal{G})$ is the graph \mathcal{G}_A on A , with the edges from \mathcal{G} that are between elements of A , that is:

$$\mathcal{E}(\mathcal{G}_A) = \{\{u, v\} \in \mathcal{E}(\mathcal{G}) : u, v \in A\}.$$

A set $B \subseteq \mathcal{V}(\mathcal{G})$ is *complete* in \mathcal{G} if the induced subgraph \mathcal{G}_B is complete.

If $\{u, v\} \in \mathcal{E}(\mathcal{G})$, then u, v are *adjacent*, and u is a *neighbour* of v . We write $\text{ne}_{\mathcal{G}}(v)$ to be the set of neighbours of v . The *boundary* $\text{bd}_{\mathcal{G}}(A)$ of a subset A

of $\mathcal{V}(\mathcal{G})$ is the set of vertices in $\mathcal{V}(\mathcal{G}) \setminus A$ that are neighbours of elements in A .

A *path* (of length k) is a sequence of vertices v_0, v_1, \dots, v_k such that each $\{v_{i-1}, v_i\} \in \mathcal{E}(\mathcal{G})$. A *cycle* is a path that starts and ends at the same vertex.

Vertices $u, v \in \mathcal{V}(\mathcal{G})$ are *connected* if there exists a path from u to v . A *connected component* is a maximal subset C of $\mathcal{V}(\mathcal{G})$ such that every pair $u, v \in C$ is connected. A graph is *connected* if it has exactly one connected component.

A *chord* of a cycle is an edge joining two nonconsecutive vertices. A graph is *triangulated* or *chordal* if any cycle of length ≥ 4 , has a chord.

Vertices a and b are *separated* by a subset $S \subseteq \mathcal{V}(\mathcal{G})$ if every path from a to b passes through S . In this case, we can say that S is an *a-b separator*. Subsets A and B are separated by S if every $a \in A$ and $b \in B$ are separated by S .

\mathcal{G} is *collapsible* onto a subset A of $\mathcal{V}(\mathcal{G})$ if each connected component B_i of $\mathcal{G}_{V \setminus A}$ has a boundary $\text{bd}_{\mathcal{G}}(B_i)$ in \mathcal{G} that is complete.

A *decomposition* of an undirected graph is a pair (A, B) such that:

- (i) $A \cup B = \mathcal{V}(\mathcal{G})$,
- (ii) $\mathcal{G}_{A \cup B}$ is complete, and
- (iii) A and B are separated by $A \cap B$ in \mathcal{G} .

A decomposition is *proper* if both A and B are proper subsets of V .

A graph is *decomposable* if it is complete, or there exists a proper decomposition (A, B) such that \mathcal{G}_A and \mathcal{G}_B are decomposable.

Note. This recursive characterisation is well-defined, as both \mathcal{G}_A and \mathcal{G}_B must have fewer vertices than \mathcal{G} .

We will write \mathfrak{U} as the set of undirected decomposable graphs on V .

A *clique* is a subset of vertices $C \subseteq \mathcal{V}(\mathcal{G})$ such that \mathcal{G}_C is complete, and it is maximal with this property. We use $\text{cl}(\mathcal{G})$ to denote the set of cliques of \mathcal{G} .

For a sequence of subsets B_1, \dots, B_k of $\mathcal{V}(\mathcal{G})$ we can define:

$$H_i = B_1 \cup \dots \cup B_i, \quad i = 1, \dots, k$$

termed the *histories*, and:

$$R_i = B_i \setminus H_{i-1}, \quad \text{and} \quad S_i = B_i \cap H_{i-1}, \quad i = 2, \dots, k$$

residuals and *separators*, respectively. This sequence is *perfect* if each \mathcal{G}_{B_i} is complete, and each $S_i \subseteq B_j$ for some $j < i$ (called the “running intersection property”).

A numbering of the vertices v_1, \dots, v_n is *perfect* if the corresponding sequence of sets:

$$B_j = \{v_1, \dots, v_j\} \cap (\text{neg}(v_j) \cup \{v_j\}), \quad j = 1, \dots, n$$

is perfect.

Theorem A.1.1 (Lauritzen 1996, Proposition 2.5 and Proposition 2.17). *For an undirected graph \mathcal{G} , the following conditions are equivalent:*

- (i) \mathcal{G} is decomposable.
- (ii) \mathcal{G} is chordal.
- (iii) Every minimal separator between any two nonadjacent nodes is complete.
- (iv) The vertices of G admit a perfect numbering.
- (v) The cliques of G can be arranged in a perfect sequence.

The *separators* of a decomposable graph are the separators in a perfect sequence of cliques, the set of which we denote by $\text{sep}(\mathcal{G})$, and we note that these will be the minimal separators between non-adjacent vertices. Furthermore, the same separator S may appear multiple times in this sequence: the number of times in which it appears is said to be its *multiplicity*, which we will usually denote by $\nu_{\mathcal{G}}(S)$.

An algorithm known as *maximum cardinality search* (Tarjan and Yannakakis 1984) can determine whether or not a graph is decomposable, and if it is, will provide a perfect numbering. A small modification (Cowell et al. 2007, Algorithm 4.11) will also provide a perfect ordering of cliques.

A.2 Directed graphs

In a *directed graph*, the edges are ordered pairs of distinct vertices (u, v) , and so:

$$\mathcal{E}(\mathcal{G}) \subseteq \{(u, v) : u, v \in \mathcal{V}(\mathcal{G}), u \neq v\}.$$

Some notions transfer directly from undirected graphs, such as a sparse graph, subgraph and induced subgraph.

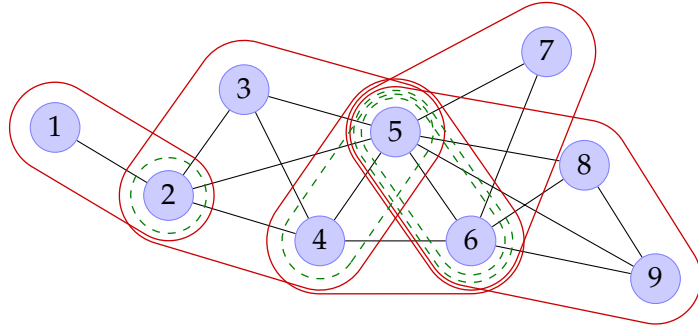


Figure A.1: The cliques (—) and separators (---) of an undirected decomposable graph. Note that the separator $\{5, 6\}$ has mutliplicity 2, as it separates the three cliques $\{4, 5, 6\}$, $\{5, 6, 7\}$ and $\{5, 6, 8, 9\}$.

The *skeleton* of a directed graph \mathcal{G} is the undirected graph \mathcal{G}^s obtained by substituting the directed edges for undirected ones, *i.e.* $\mathcal{V}(\mathcal{G}^s) = \mathcal{V}(\mathcal{G})$ and:

$$\mathcal{E}(\mathcal{G}^s) = \{\{u, v\} : (u, v) \in \mathcal{V}(\mathcal{G})\}.$$

In this case, we say \mathcal{G} is a *directed version* of \mathcal{G}^s . A directed graph is *complete* if its skeleton is complete.

A *directed path* is a sequence of vertices v_0, \dots, v_k such that each pair $(v_{i-1}, v_i) \in \mathcal{E}(\mathcal{G})$. A *trail* is a sequence such that at least one of either (v_{i-1}, v_i) or (v_i, v_{i-1}) is in $\mathcal{E}(\mathcal{G})$, *i.e.* a trail is a path in the skeleton.

An *directed acyclic graph* or DAG is a directed graph without any directed cycles. We note that this precludes the existence of a pair of opposingly directed edges (u, v) and (v, u) in the same graph. We will write \mathfrak{D} as the set of directed acyclic graphs on V .

If $(u, v) \in \mathcal{E}(\mathcal{G})$ we say u is a *parent* of v , and v is a *child* of u . We write the set of parents of v as $\text{pa}_{\mathcal{G}}(v)$. If there exists a directed path from u to v , then u is an *ancestor* of v , and v is a *descendant* of u . We write the set of ancestors and descendants of v as $\text{an}_{\mathcal{G}}(v)$ and $\text{de}_{\mathcal{G}}(v)$, respectively, and note that both of these sets include v itself. Furthermore, for a set $A \subseteq \mathcal{V}(\mathcal{G})$, we write $\text{an}_{\mathcal{G}}(A) = \cup_{v \in A} \text{an}_{\mathcal{G}}(v)$ and $\text{de}_{\mathcal{G}}(A) = \cup_{v \in A} \text{de}_{\mathcal{G}}(v)$. The *non-descendants* of v is the complement set $V \setminus \text{de}_{\mathcal{G}}(v)$, and written $\text{nd}_{\mathcal{G}}(v)$. A subset $A \subseteq \mathcal{V}(\mathcal{G})$ is *ancestral* if $\text{an}_{\mathcal{G}}(A)$.

The *moral graph* \mathcal{G}^m is the undirected graph constructed from the skeleton of \mathcal{G} by adding edges between any two vertices that have a common child.

A *well-ordering* of a directed acyclic graph \mathcal{G} is an proper ordering \prec of the vertices $v_1 \prec \dots \prec v_p$ such that for any edge $(u, v) \in \mathcal{E}(\mathcal{G})$, we have $u \prec v$.

A directed acyclic graph is *perfect* if the subgraph induced by the parent set of each node is complete. In this case, the moral graph is simply the skeleton. Furthermore, a well-ordering of the vertices of a perfect graph will induce a perfect numbering of the skeleton.

Bibliography

- Aldous, David J. (1985). “Exchangeability and related topics”. In: *École d’été de probabilités de Saint-Flour, XIII—1983*. Vol. 1117. Lecture Notes in Mathematics, 1–198. Springer, Berlin. DOI: 10.1007/BFb0099421. MR883646.
- Altham, Patricia M. E. (1969). “Exact Bayesian analysis of a 2×2 contingency table, and Fisher’s “exact” significance test”. *Journal of the Royal Statistical Society. Series B (Methodological)* **31**(2), 261–269. MR0269016.
- Andersson, Steen A., David Madigan, and Michael D. Perlman (1997a). “A characterization of Markov equivalence classes for acyclic digraphs”. *Annals of Statistics* **25**(2), 505–541. DOI: 10.1214/aos/1031833662. MR1439312.
- Andersson, Steen A., David Madigan, and Michael D. Perlman (1997b). “On the Markov equivalence of chain graphs, undirected graphs, and acyclic digraphs”. *Scandinavian Journal of Statistics* **24**(1), 81–102. DOI: 10.1111/1467-9469.00050. MR1436624.
- Armstrong, Helen, Christopher K. Carter, Kin Foon Kevin Wong, and Robert Kohn (2009). “Bayesian covariance matrix estimation using a mixture of decomposable graphical models”. *Statistics and Computing* **19**(3), 303–316. DOI: 10.1007/s11222-008-9093-8.
- Asci, Claudio, Giovanna Nappo, and Mauro Piccioni (2006). “The hyper-Dirichlet process and its discrete approximations: the butterfly model”. *Journal of Multivariate Analysis* **97**(4), 895–924. DOI: 10.1016/j.jmva.2005.08.009. MR2256566.
- Ashby, Deborah, Jane L. Hutton, and Magnus A. McGee (1993). “Simple Bayesian Analyses for Case-Control Studies in Cancer Epidemiology”.

- Journal of the Royal Statistical Society. Series D (The Statistician)* **42**(4), 385–397.
- Asmussen, Søren and David Edwards (1983). “Collapsibility and response variables in contingency tables”. *Biometrika* **70**(3), 567–578. DOI: 10.1093/biomet/70.3.567. MR725370.
- Atay-Kayis, Aliye and Hélène Massam (2005). “A Monte Carlo method for computing the marginal likelihood in nondecomposable Gaussian graphical models”. *Biometrika* **92**(2), 317–335. DOI: 10.1093/biomet/92.2.317. MR2201362.
- Auvray, Vincent and Louis Wehenkel (2002). “On the construction of the inclusion boundary neighbourhood for Markov equivalence classes of Bayesian network structures”. In: *Proceedings of the Eighteenth Annual Conference on Uncertainty in Artificial Intelligence*. (A. Darwiche and N. Friedman, eds.), 26–35. Morgan Kaufmann, San Francisco, CA. ISBN: 1-55860-897-4.
- Banerjee, Onureena, Laurent El Ghaoui, and Alexandre d’Aspremont (2008). “Model selection through sparse maximum likelihood estimation for multivariate Gaussian or binary data”. *Journal of Machine Learning Research* **9**, 485–516. MR2417243.
- Barbieri, Maria Maddalena and James O. Berger (2004). “Optimal predictive model selection”. *Annals of Statistics* **32**(31), 870–897. DOI: 10.1214/009053604000000238. MR2065192.
- Bornn, Luke and François Caron (2011). “Bayesian clustering in decomposable graphs”. *Bayesian Analysis*, To appear.
- Brooks, Stephen P., Paolo Giudici, and Gareth O. Roberts (2003). “Efficient construction of reversible jump Markov chain Monte Carlo proposal distributions”. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)* **65**(1), 3–55. DOI: 10.1111/1467-9868.03711. MR1959092.
- Buntine, Wray (1991). “Theory Refinement on Bayesian Networks”. In: *Proceedings of the Seventh Annual Conference on Uncertainty in Artificial Intelligence*. (B. D’Ambrosio, P. Smets, and P. Bonissone, eds.), 52–60. Morgan Kaufmann, San Mateo, CA. ISBN: 1-55860-203-8.

-
- Castelo, Robert and Tomáš Kočka (2004). “On inclusion-driven learning of Bayesian networks”. *Journal of Machine Learning Research* 4(4), 527–574. MR2072261.
- Chickering, David Maxwell (1995). “A transformational characterization of equivalent Bayesian network structures”. In: *Proceedings of the Eleventh Annual Conference on Uncertainty in Artificial Intelligence*. (P. Besnard and S. Hanks, eds.), 87–98. Morgan Kaufmann, San Francisco, CA. ISBN: 1-55860-385-9. MR1615012.
- Chickering, David Maxwell (2002). “Learning equivalence classes of Bayesian-network structures”. *Journal of Machine Learning Research* 2(3), 445–498. MR1929415.
- Chickering, David Maxwell (2003). “Optimal structure identification with greedy search”. *Journal of Machine Learning Research* 3(3), 507–554. MR1991085.
- Cooper, Gregory F. and Edward Herskovits (1992). “A Bayesian method for the induction of probabilistic networks from data”. *Machine Learning* 9(4), 309–347. DOI: 10.1007/BF00994110.
- Cowell, Robert G., A. Philip Dawid, Steffen L. Lauritzen, and David J. Spiegelhalter (2007). *Probabilistic networks and expert systems*. Springer-Verlag, New York. ISBN: 978-0-387-71823-1. MR1697175.
- Darroch, John N., Steffen L. Lauritzen, and Terence P. Speed (1980). “Markov fields and log-linear interaction models for contingency tables”. *Annals of Statistics* 8(3), 522–539. DOI: 10.1214/aos/1176345006. MR568718.
- Dawid, A. Philip (1979). “Conditional independence in statistical theory”. *Journal of the Royal Statistical Society. Series B (Methodological)* 41(1), 1–31. MR535541.
- Dawid, A. Philip (1980). “Conditional independence for statistical operations”. *Annals of Statistics* 8(3), 598–617. DOI: 10.1214/aos/1176345011. MR568723.
- Dawid, A. Philip (1981). “Some matrix-variate distribution theory: notational considerations and a Bayesian application”. *Biometrika* 68(1), 265–274. DOI: 10.1093/biomet/68.1.265. MR614963.
- Dawid, A. Philip (2001). “Separoids: a mathematical framework for conditional independence and irrelevance”. *Annals of Mathematics and Ar-*

- tificial Intelligence* **32**(1-4). Representations of uncertainty, 335–372. DOI: 10.1023/A:1016734104787. MR1859870.
- Dawid, A. Philip and Steffen L. Lauritzen (1993). “Hyper-Markov laws in the statistical analysis of decomposable graphical models”. *Annals of Statistics* **21**(3), 1272–1317. DOI: 10.1214/aos/1176349260. MR1241267.
- Dawid, A. Philip and Steffen L. Lauritzen (2001). “Compatible prior distributions”. In: *Bayesian Methods with Applications to Science, Policy and Official Statistics*. Proceedings of the 6th World Meeting of the International Society for Bayesian Analysis. (E. I. George, ed.), 109–118. Office for Official Publications of the European Communities, Luxembourg.
- Dawid, A. Philip and Milan Studený (1999). “Conditional products: an alternative approach to conditional independence”. In: *Proceedings of the Seventh International Workshop on Artificial Intelligence and Statistics*. (D. Heckerman and J. Whittaker, eds.), 32–40. Morgan Kaufmann, San Francisco.
- Dellaportas, Petros and Jonathan J. Forster (1999). “Markov chain Monte Carlo model determination for hierarchical and graphical log-linear models”. *Biometrika* **86**(3), 615–633. DOI: 10.1093/biomet/86.3.615. MR1723782.
- Dempster, Arthur P. (1972). “Covariance selection”. *Biometrics* **28**(1), 157–175.
- Drton, Mathias and Michael D. Perlman (2008). “A SINful approach to Gaussian graphical model selection”. *Journal of Statistical Planning and Inference* **138**(4), 1179–1200. DOI: 10.1016/j.jspi.2007.05.035. MR2416875.
- Ellis, Byron and Wing Hung Wong (2008). “Learning causal Bayesian network structures from experimental data”. *Journal of the American Statistical Association* **103**(482), 778–789. DOI: 10.1198/016214508000000193. MR2524009.
- Ferguson, Thomas S. (1973). “A Bayesian analysis of some nonparametric problems”. *Annals of Statistics* **1**(2), 209–230. DOI: 10.1214/aos/1176342360. MR0350949.
- Friedman, Jerome, Trevor Hastie, and Robert Tibshirani (2008). “Sparse inverse covariance estimation with the graphical lasso”. *Biostatistics* **9**(3), 432–441. DOI: 10.1093/biostatistics/kxm045.

- Friedman, Nir and Daphne Koller (2003). "Being Bayesian About Network Structure. A Bayesian Approach to Structure Discovery in Bayesian Networks". *Machine Learning* **50**(1–2), 95–125. DOI: 10.1023/A:1020249912095.
- Frydenberg, Morten (1990). "The chain graph Markov property". *Scandinavian Journal of Statistics* **17**(4), 333–353. MR1096723.
- Frydenberg, Morten and Steffen L. Lauritzen (1989). "Decomposition of maximum likelihood in mixed graphical interaction models". *Biometrika* **76**(3), 539–555. DOI: 10.1093/biomet/76.3.539. MR1040647.
- Geiger, Dan and David Heckerman (1997). "A characterization of the Dirichlet distribution through global and local parameter independence". *Annals of Statistics* **25**, 1344–1369. DOI: 10.1214/aos/1069362752. MR1447755.
- Geiger, Dan and David Heckerman (2002). "Parameter priors for directed acyclic graphical models and the characterization of several probability distributions". *Annals of Statistics* **30**(5), 1412–1440. DOI: 10.1214/aos/1035844981. MR1936324.
- Giudici, Paolo and Robert Castelo (2003). "Improving Markov Chain Monte Carlo Model Search for Data Mining". *Machine Learning* **50**(1-2), 127–158. DOI: 10.1023/A:1020202028934.
- Giudici, Paolo and Peter J. Green (1999). "Decomposable graphical Gaussian model determination". *Biometrika* **86**(4), 785–801. DOI: 10.1093/biomet/86.4.785. MR1741977.
- Green, Peter J. and Alun Thomas (2011). *Sampling decomposable graphs using a Markov chain on junction trees*. arXiv:1104.4079.
- Gustafson, Paul, Nhu D. Le, and Marc Vallée (2002). "A Bayesian approach to case-control studies with errors in covariables". *Biostatistics* **3**(2), 229–243. DOI: 10.1093/biostatistics/3.2.229.
- Heckerman, David, Dan Geiger, and David Maxwell Chickering (1995). "Learning Bayesian networks: The combination of knowledge and statistical data". *Machine Learning* **20**(3), 197–243. DOI: 10.1023/A:1022623210503.
- Heinz, Daniel (2009). "Building hyper Dirichlet processes for graphical models". *Electronic Journal of Statistics* **3**, 290–315. DOI: 10.1214/08-EJS269. MR2495840.

- Jones, Beatrix, Carlos Carvalho, Adrian Dobra, Chris Hans, Chris Carter, and Mike West (2005). "Experiments in stochastic computation for high-dimensional graphical models". *Statistical Science* **20**(4), 388–400. DOI: 10.1214/088342305000000304. MR2210226.
- Kijima, Shuji, Masashi Kiyomi, Yoshio Okamoto, and Takeaki Uno (2007). *On listing, sampling, and counting the chordal graphs with edge constraints*. Tech. rep. Research Institute for Mathematical Sciences, Kyoto University. URL: <http://www.kurims.kyoto-u.ac.jp/preprint/file/RIMS1610.pdf>.
- Kijima, Shuji, Masashi Kiyomi, Yoshio Okamoto, and Takeaki Uno (2008). "On listing, sampling, and counting the chordal graphs with edge constraints". In: *Computing and combinatorics*. Vol. 5092. Lecture Notes in Computer Science, 458–467. Springer, Berlin. DOI: 10.1007/978-3-540-69733-6_45. MR2473446.
- Lauritzen, Steffen L. (1996). *Graphical models*. Vol. 17. Oxford Statistical Science Series. Oxford University Press, New York. ISBN: 0-19-852219-3. MR1419991.
- Lauritzen, Steffen L., A. Philip Dawid, B. N. Larsen, and Hanns-Georg Leimer (1990). "Independence properties of directed Markov fields". *Networks* **20**(5), 491–505. DOI: 10.1002/net.3230200503. MR1064735.
- Lauritzen, Steffen L., Terence P. Speed, and Kaipillil Vijayan (1984). "Decomposable graphs and hypergraphs". *Journal of the Australian Mathematical Society (Series A)* **36**(1), 12–29. DOI: 10.1017/S1446788700027300. MR719998.
- Letac, Gérard and Hélène Massam (2007). "Wishart distributions for decomposable graphs". *Annals of Statistics* **35**(3), 1278–1323. DOI: 10.1214/009053606000001235. MR2341706.
- Lunagomez, Simon (2009). "A Geometric Approach for Inference on Graphical Models". PhD thesis. Duke University.
- Madigan, David, Steen A. Andersson, Michael D. Perlman, and Chris T. Volinsky (1996). "Bayesian model averaging and model selection for markov equivalence classes of acyclic digraphs". *Communications in Statistics. Theory and Methods* **25**(11), 2493–2519. DOI: 10.1080/03610929608831853.

- Madigan, David and Adrian E. Raftery (1994). "Model Selection and Accounting for Model Uncertainty in Graphical Models Using Occam's Window". *Journal of the American Statistical Association* **89**(428), 1535–1546.
- Madigan, David and Jeremy York (1995). "Bayesian Graphical Models for Discrete Data". *International Statistical Review / Revue Internationale de Statistique* **63**(2), 215–232.
- Marshall, Roger J. (1988). "Bayesian analysis of case-control studies". *Statistics in Medicine* **7**(12), 1223–1230. DOI: 10.1002/sim.4780071203.
- McMorris, Frederick R. and Edward R. Scheinerman (1991). "Connectivity threshold for random chordal graphs". *Graphs and Combinatorics* **7**(2), 177–181. DOI: 10.1007/BF01788142. MR1115136.
- Mukherjee, Bhramar, Samiran Sinha, and Malay Ghosh (2005). "Bayesian analysis of case-control studies". In: *Bayesian thinking: modeling and computation*. (D. K. Dey and C. R. Rao, eds.). Vol. 25. Handbook of Statistics, 793–819. Elsevier/North-Holland, Amsterdam. ISBN: 0-444-51539-9. DOI: 10.1016/S0169-7161(05)25027-7. MR2490547.
- Mukherjee, Sach and Terence P. Speed (2008). "Network inference using informative priors". *Proceedings of the National Academy of Sciences* **105**(38), 14313–14318. DOI: 10.1073/pnas.0802272105.
- Müller, Peter and Kathryn Roeder (1997). "A Bayesian semiparametric model for case-control studies with errors in variables". *Biometrika* **84**(3), 523–537. DOI: 10.1093/biomet/84.3.523. MR1603977.
- Nurminen, Markku and Pertti Mutanen (1987). "Exact Bayesian analysis of two proportions". *Scandinavian Journal of Statistics* **14**(1), 67–77. MR895790.
- Park, Mee Young and Trevor Hastie (2008). "Penalized logistic regression for detecting gene interactions". *Biostatistics* **9**(1), 30–50. DOI: 10.1093/biostatistics/kxm010.
- Patefield, W. M. (1985). "Information from the maximized likelihood function". *Biometrika* **72**(3), 664–668. DOI: 10.1093/biomet/72.3.664. MR817581.
- Pearl, Judea (1988). *Probabilistic reasoning in intelligent systems: networks of plausible inference*. Morgan Kaufmann, San Mateo, CA. ISBN: 0-934613-73-7. MR965765.

- Pearl, Judea and Azaria Paz (1987). "Graphoids: a Graph-Based Logic for Reasoning about Relevance Relations". In: *Advances in Artificial Intelligence - II*. (B. D. Boulay, D. Hogg, and L. Steel, eds.), 357–363. North-Holland, Amsterdam.
- Prentice, Ross L. and Ronald Pyke (1979). "Logistic disease incidence models and case-control studies". *Biometrika* **66**(3), 403–411. DOI: 10.1093/biomet/66.3.403. MR556730.
- Rajaratnam, Bala, H el ene Massam, and Carlos M. Carvalho (2008). "Flexible covariance estimation in graphical Gaussian models". *Annals of Statistics* **36**(6), 2818–2849. DOI: 10.1214/08-AOS619. MR2485014.
- Rice, Kenneth M. (2004). "Equivalence between conditional and mixture approaches to the Rasch model and matched case-control studies, with applications". *Journal of the American Statistical Association* **99**(466), 510–522. DOI: 10.1198/016214504000000511. MR2062836.
- Rothman, Adam J., Peter J. Bickel, Elizaveta Levina, and Ji Zhu (2008). "Sparse permutation invariant covariance estimation". *Electronic Journal of Statistics* **2**, 494–515. DOI: 10.1214/08-EJS176. MR2417391.
- Roverato, Alberto (2000). "Cholesky decomposition of a hyper inverse Wishart matrix". *Biometrika* **87**(1), 99–112. DOI: 10.1093/biomet/87.1.99. MR1766831.
- Roverato, Alberto (2002). "Hyper inverse Wishart distribution for non-decomposable graphs and its application to Bayesian inference for Gaussian graphical models". *Scandinavian Journal of Statistics* **29**(3), 391–411. DOI: 10.1111/1467-9469.00297. MR1925566.
- Scott, James G. and Carlos M. Carvalho (2008). "Feature-Inclusion Stochastic Search for Gaussian Graphical Models". *Journal of Computational and Graphical Statistics* **17**(4), 790–808. DOI: 10.1198/106186008X382683.
- Seaman, Shaun R. and Sylvia Richardson (2001). "Bayesian analysis of case-control studies with categorical covariates". *Biometrika* **88**(4), 1073–1088. DOI: 10.1093/biomet/88.4.1073. MR1872220.
- Seaman, Shaun R. and Sylvia Richardson (2004). "Equivalence of prospective and retrospective models in the Bayesian analysis of case-control

- studies". *Biometrika* **91**(1), 15–25. DOI: 10.1093/biomet/91.1.15. MR2050457.
- Spiegelhalter, David, Andrew Thomas, Nicky Best, and Dave Lunn (Jan. 2003). *WinBUGS User Manual*.
- Spirtes, Peter, Clark Glymour, and Richard Scheines (2000). *Causation, prediction, and search*. Second. MIT Press, Cambridge, MA. ISBN: 0-262-19440-6. MR1815675.
- Staicu, Ana-Maria (2007). *On the equivalence of prospective and retrospective likelihood methods in case-control studies*. Tech. rep. 07:13. Statistics group, Department of Mathematics, University of Bristol. URL: <http://www.stats.bris.ac.uk/research/stats/reports/2007/0713.pdf>.
- Staicu, Ana-Maria (2010). "On the equivalence of prospective and retrospective likelihood methods in case-control studies". *Biometrika* **97**(4), 990–996. DOI: 10.1093/biomet/asq054. MR2746168.
- Studený, Milan (1997). "On marginalization, collapsibility and precollapsibility". In: *Distributions with given marginals and moment problems*. (Prague, 1996). (V. Beneš and J. Štěpán, eds.), 191–198. Kluwer Academic Publishers, Dordrecht. ISBN: 0-7923-4573-8. MR1614672.
- Studený, Milan (2005a). "Characterization of inclusion neighbourhood in terms of the essential graph". *International Journal of Approximate Reasoning* **38**(3), 283–309. DOI: 10.1016/j.ijar.2004.05.007. MR2116940.
- Studený, Milan (2005b). *Probabilistic Conditional Independence Structures*. Springer-Verlag, London. ISBN: 978-1-85233-891-6. DOI: 10.1007/b138557.
- Studený, Milan and Jiří Vomlel (2009). "A reconstruction algorithm for the essential graph". *International Journal of Approximate Reasoning* **50**(2), 385–413. DOI: 10.1016/j.ijar.2008.09.001. MR2514506.
- Sundberg, Rolf (1975). "Some results about decomposable (or Markov-type) models for multidimensional contingency tables: distribution of marginals and partitioning of tests". *Scandinavian Journal of Statistics* **2**(2), 71–79. MR0375634.
- Tarjan, Robert E. and Mihalis Yannakakis (1984). "Simple linear-time algorithms to test chordality of graphs, test acyclicity of hypergraphs, and se-

- lectively reduce acyclic hypergraphs". *SIAM Journal on Computing* **13**(3), 566–579. DOI: 10.1137/0213035. MR749707.
- Thomas, Alun and Peter J. Green (2009a). "Enumerating the decomposable neighbors of a decomposable graph under a simple perturbation scheme". *Computational Statistics & Data Analysis* **53**(4), 1232–1238. DOI: 10.1016/j.csda.2008.10.029. MR2657086.
- Thomas, Alun and Peter J. Green (2009b). "Enumerating the junction trees of a decomposable graph". *Journal of Computational and Graphical Statistics* **18**(4), 930–940. DOI: 10.1198/jcgs.2009.07129. MR2598034.
- Tibshirani, Robert (1996). "Regression shrinkage and selection via the lasso". *Journal of the Royal Statistical Society. Series B (Methodological)* **58**(1), 267–288. MR1379242.
- Verma, Tom and Judea Pearl (1990). "Equivalence and Synthesis of Causal Models". In: *Proceedings of the Sixth Annual Conference on Uncertainty in Artificial Intelligence*. (P. Bonissone, M. Henrion, L. Kanal, and J. Lemmer, eds.), 220–227. Elsevier Science, New York, NY.
- Verma, Tom and Judea Pearl (1992). "An Algorithm for Deciding if a Set of Observed Independencies Has a Causal Explanation". In: *Proceedings of the Eighth Annual Conference on Uncertainty in Artificial Intelligence*. (D. Dubois, M. Wellman, B. D'Ambrosio, and P. Smets, eds.), 323–330. Morgan Kaufmann, San Mateo, CA. ISBN: 1-55860-258-5.
- Wermuth, Nanny (1976a). "Analogies between multiplicative models in contingency tables and covariance selection". *Biometrics* **32**(1), 95–108. MR0403088.
- Wermuth, Nanny (1976b). "Model Search among Multiplicative Models". *Biometrics* **32**(2), 253–263.
- Wormald, Nicholas C. (1985). "Counting labelled chordal graphs". *Graphs and Combinatorics* **1**(2), 193–200. DOI: 10.1007/BF02582944. MR951781.
- Wu, Tong Tong, Yi Fang Chen, Trevor Hastie, Eric Sobel, and Kenneth Lange (2009). "Genome-wide association analysis by lasso penalized logistic regression". *Bioinformatics* **25**(6), 714–721. DOI: 10.1093/bioinformatics/btp041.

- Yuan, Ming and Yi Lin (2007). "Model selection and estimation in the Gaussian graphical model". *Biometrika* **94**(1), 19–35. DOI: 10.1093/biomet/asm018. MR2367824.
- Zelen, M. and R. A. Parker (1986). "Case-control studies and bayesian inference". *Statistics in Medicine* **5**(3), 261–269. DOI: 10.1002/sim.4780050307.
- Zellner, Arnold (1986). "On assessing prior distributions and Bayesian regression analysis with g -prior distributions". In: *Bayesian inference and decision techniques*. Essays in honor of Bruno de Finetti. (P. K. Goel and A. Zellner, eds.). Vol. 6. Studies in Bayesian Econometrics and Statistics, 233–243. North-Holland, Amsterdam. ISBN: 0-444-87712-6. MR881437.