The Annals of Statistics
1993, Vol. 21, No. 3, 1272–1317

HYPER MARKOV LAWS IN THE STATISTICAL ANALYSIS OF DECOMPOSABLE GRAPHICAL MODELS

BY A. P. DAWID AND S. L. LAURITZEN

University College London and Aalborg University

This paper introduces and investigates the notion of a *hyper Markov law*, which is a probability distribution over the set of probability measures on a multivariate space that (i) is concentrated on the set of *Markov* probabilities over some decomposable graph, and (ii) satisfies certain conditional independence restrictions related to that graph. A stronger version of this hyper Markov property is also studied.

Our analysis starts by reconsidering the properties of Markov probabilities, using an abstract approach which thereafter proves equally applicable to the hyper Markov case. Next, it is shown constructively that hyper Markov laws exist, that they appear as sampling distributions of maximum likelihood estimators in decomposable graphical models, and also that they form natural conjugate prior distributions for a Bayesian analysis of these models.

As examples we construct a range of specific hyper Markov laws, including the hyper multinomial, hyper Dirichlet and the hyper Wishart and inverse Wishart laws. These laws occur naturally in connection with the analysis of decomposable log-linear and covariance selection models.

1. Introduction. Recent work [Darroch, Lauritzen and Speed (1980), Edwards and Kreiner (1983), Wermuth and Lauritzen (1983), Lauritzen (1989), Whittaker (1990), Edwards (1990) and Wermuth and Lauritzen (1990)] has shown the value and versatility of *graphical models*, namely statistical models embodying a collection of marginal and conditional independences which may be summarized by means of a graph. Such models combine richness in modelling, clarity of interpretation and ease of analysis. Graphical models have also been studied in connection with "probabilistic expert systems" [Lauritzen and Spiegelhalter (1988), Pearl (1988) and Spiegelhalter, Dawid, Lauritzen and Cowell (1993)], where they have been found to provide a powerful tool for specifying complex multivariate distributions, and for simplifying and organizing probability calculations in them.

The underlying graph, which may be directed or undirected, or even a combination of the two, has a node for each variable in the problem. In the directed case, the "parent" nodes, from which edges lead into a given "child"

Received October 1989; revised May 1992.

AMS 1991 *subject classifications*. Primary 62H99; secondary 62E15.

Key words and phrases. Bayesian statistics, covariance selection, collapsibility, contingency tables, cut, decomposable graphs, Dirichlet distribution, expert systems, graphical models, hyper Dirichlet law, hyper inverse Wishart law, hyper matrix F law, hyper matrix t law, hyper Normal law, hyper Multinomial law, hyper Wishart law, inverse Wishart distribution, log-linear models, matrix F distribution, matrix t distribution, multivariate analysis, triangulated graphs, Wishart distribution.

node, are taken to be the only direct influences on that child, which is thus independent of all other possible or indirect influences, conditional on these parents. In the undirected case, a node is taken to be independent of all others, given its immediate neighbours.

To introduce the topic of the present paper, let us consider the particularly simple and well-known example of such a graphical model involving two independent discrete variables A and B with levels i and j. Based upon a multinomial sample of size n represented by a table of counts $\{N_{ij}\}$, the maximum likelihood estimator of the unknown probability distribution p_{ij} is given by

$$\hat{p}_{ij} = N_{i+}N_{+j}/n^2,$$

where $+$ denotes summation over the corresponding index. As the counts are random, the estimator $\hat{p}$ is a random distribution. The distributional law of $\hat{p}$ is an example of what we shall term a *hyper Markov law* in the present paper. This term reflects the fact that, under the assumption of independence of the variables A and B, that is, that $p_{ij} = p_{i+}p_{+j}$, the estimators of the marginal distributions

$$\{\hat{p}_{i+}\} = \{N_{i+}/n\} \quad \text{and} \quad \{\hat{p}_{+j}\} = \{N_{+j}/n\}$$

are stochastically independent according to the law of $\hat{p}$. Thus the Markov property at the model level, embodied in the independence of A and B, is reflected at the higher level represented by the law of the estimator $\hat{p}$.

Another simple example of a hyper Markov law involves three variables X, Y and Z following a trivariate normal distribution with mean zero and a covariance matrix with the xyth-element σ^{xy} of its inverse equal to zero. This is equivalent to X and Y being conditionally independent given Z. If we take a random sample (X_i, Y_i, Z_i), $i = 1, \ldots, n$ from this distribution, it holds—under the given assumption—that [using the conditional independence notation of Dawid (1979a)]

$$\left(\hat{\beta}_{x\cdot z}, \hat{\sigma}^2_{xx\cdot z}\right) \perp\!\!\!\perp \left(\hat{\beta}_{y\cdot z}, \hat{\sigma}^2_{yy\cdot z}\right) \mid \hat{\sigma}^2_{zz},$$

where $\hat{\beta}_{x\cdot z} = \Sigma X_i Z_i / \Sigma Z_i^2$ and $\hat{\sigma}^2_{xx\cdot z} = n^{-1}\{\Sigma X_i^2 - (\Sigma X_i Z_i)^2/\Sigma Z_i^2\}$ are the sample estimators of the parameters of the regression (slope and residual variance) of X on Z and so on. This again reflects the Markov property at a higher level.

As we shall show in the present paper, these are special instances of a general phenomenon, extending to a large class of *decomposable* models, the statistical analysis of which is particularly tractable and useful [Frydenberg and Lauritzen (1989)].

The work on graphical models in expert systems has, for the most part, assumed completely specified probabilities. These might, for example, be subjectively assessed by the same subject-matter expert from whom the appropriate graphical structure is elicited. However, if data can be observed on the variables in question, it should be possible to learn from these data so as to add

to, and perhaps eventually override, the information provided by the expert. This suggests a Bayesian approach, in which the task of the expert is to provide the graphical structure, and a prior distribution expressing his or her uncertainty about its numerical parameters. This approach has been studied by Spiegelhalter and Lauritzen (1990) in the context of a directed graph. A major purpose of the present paper is to explore the details of the Bayesian approach, with special emphasis on the undirected case. It turns out that laws satisfying the hyper Markov property described above prove particularly amenable for use as prior distributions. For example, consider as before two independent discrete variables A and B with levels i and j. One possible prior distribution, reflecting the independence structure, would be to assume

$$p_{ij} = \theta_i \eta_j,$$

where $\theta = (\theta_i)$ and $\eta = (\eta_j)$ are assumed to be independent and Dirichlet distributed. The prior law of p so constructed is a hyper Markov law, indeed a special instance of what we term the *hyper Dirichlet law* on this graph.

The main body of the paper begins by reconsidering a familiar area—the definition and properties of Markov distributions over an undirected graph. Some formal graph-theoretic notions and proofs are deferred to the Appendix. Here and throughout the rest of the paper we restrict ourselves to the case of a decomposable graph. This allows us to develop the theory in a way which is thereafter immediately applicable to our major concern: Defining and investigating the structure of a *hyper Markov law* over a family of Markov distributions. Indeed, throughout this work we emphasize the generality and extremely broad applicability of the particular methods which we use to develop our results. This is accomplished by using a formulation based entirely on simple properties of the relation of conditional independence—properties which, when considered at a suitably abstract axiomatic level, can be seen to apply to several different interpretations.

We then turn our attention to the structure of families of Markov distributions as statistical models, and in particular to the relationship between the parametric (rather than probabilistic) model structure and the underlying graph. Here we introduce the concept of a *meta Markov model*, and show that the theory developed can be applied to yield valuable results about the sampling theory of such models. We next discuss Bayesian analysis of decomposable models, where the concept of a *strong hyper Markov law* satisfying yet stronger independence properties becomes relevant. It turns out to produce an especially simple decomposition of the Bayesian analysis into a collection of subanalyses for smaller problems. It also allows similar localization of the problem of making comparisons between rival candidate models on the basis of empirical data, using Bayesian or semi-Bayesian methods.

Finally we illustrate our theory with detailed investigations of some important special cases, based, respectively, on the normal mean model, the multinomial model and the normal dispersion model.

2. Markov distributions.

2.1. *Conditional independence.* Since we shall make extensive use of properties of *conditional independence*, as discussed in Dawid (1979a, 1980), we begin with a brief description of the most basic issues concerning this notion.

DEFINITION 2.1. If X, Y, Z are random variables on a probability space $(\Omega, \mathscr{A}, P)$, we say that X *is conditionally independent of* Y *given* Z *under* P, and write $X \perp\!\!\!\perp Y \mid Z\ [P]$, if, for any measurable set A in the sample space of X, there exists a version of the conditional probability $P(X \in A \mid Y, Z)$ which is a function of Z alone.

Usually $(\Omega, \mathscr{A}, P)$ will be fixed and P is then omitted from the notation. If Z is trivial we say that X *is independent of* Y, and write $X \perp\!\!\!\perp Y$.

The ternary relation $X \perp\!\!\!\perp Y \mid Z$ has the following properties.

PROPERTY 1. If $X \perp\!\!\!\perp Y \mid Z$, then $Y \perp\!\!\!\perp X \mid Z$.

PROPERTY 2. If $X \perp\!\!\!\perp Y \mid Z$ and U is a function of X, then $U \perp\!\!\!\perp Y \mid Z$.

PROPERTY 3. If $X \perp\!\!\!\perp Y \mid Z$ and U is a function of X, then $X \perp\!\!\!\perp Y \mid (Z, U)$.

PROPERTY 4. If $X \perp\!\!\!\perp Y \mid Z$ and $X \perp\!\!\!\perp W \mid (Y, Z)$, then $X \perp\!\!\!\perp (W, Y) \mid Z$.

(Note that the converse to Property 4 follows from Properties 1–3.)

Another property of the conditional independence relation is often used, namely:

PROPERTY 5. If $X \perp\!\!\!\perp Y \mid Z$ and $X \perp\!\!\!\perp Z \mid Y$, then $X \perp\!\!\!\perp (Y, Z)$.

However Property 5 does not hold universally, but only under additional conditions—essentially that there be no nontrivial logical relationship between Y and Z [Dawid (1979b, 1980)]. These will hold if, for example, the joint density of all variables is everywhere strictly positive. Property 5 is needed for the analysis of the Markov property on arbitrary undirected graphs. We therefore emphasize that in our development, restricting attention to decomposable graphs only, Property 5 is nowhere needed—all our results are consequences of Properties 1–4 alone, which hold universally.

A useful generalization of the idea of conditional independence is to allow one or both of the conditioning variables in the definition to represent a *parameter* of a statistical model. This leads us to express concepts such as sufficiency in terms of conditional independence. The above properties and all their consequences still apply, so long as one is careful never to write a conditional independence property in which the left-hand variable is or contains a nonrandom parameter. (In particular, the symmetry property, Property

1, is thereby greatly restricted.) However, as soon as one is willing to adopt a Bayesian approach, and treat the parameters as random variables, even this restriction may be lifted.

An important point to make about Properties 1–4 is that, treated as purely formal axioms, the apply to other relations than probabilistic conditional independence [Pearl (1988) and Smith (1989)]. Hence any general results which are based on these properties alone will be more widely applicable. We shall take advantage of this in Section 4.

2.2. *Definition and basic properties.* Throughout this paper $\mathscr{G}$ will always denote an undirected graph (V, E) which is assumed to be decomposable. See Appendix A for this and other graph-theoretic notions. Associated with each $v \in V$ we shall have a random variable X_v taking values in a sample space $\mathscr{X}_v$. For $A \subseteq V$ we write X_A for $(X_v)_{v \in A}$, $\mathscr{X}_A$ for its sample space $\times_{v \in A} \mathscr{X}_v$, $X = X_V$ and $\mathscr{X} = \mathscr{X}_V$. By a distribution over A or $\mathscr{G}_A$, we mean a joint distribution for X_A over $\mathscr{X}_A$. If P is a distribution over $U \subseteq V$, and $A, B \subseteq U$, then P_A will denote the marginal distribution of X_A, and $P_{A|B}$ the collection (labelled by x_B) of conditional distributions of X_A given $X_B = x_B$.

In this section we define the Markov property for a decomposable graph, and show that our definition is equivalent to the usual one. We develop a number of consequences of our definition, using only the conditional independence Properties 1–4. Many of these results are already well known. Our purpose in proving them again is twofold.

(i) To demonstrate that all the results are indeed consequences of Properties 1–4 alone, in particular, that there is no need for any of the special conditions such as positivity of densities which are required in the general nondecomposable case.

(ii) To indicate their generality, so that we can to a large extent rely on the same proofs to develop extensions to more complicated situations in later sections.

We shall, in this section, use the notation $A \perp\!\!\!\perp B \mid C$ to denote $X_A \perp\!\!\!\perp X_B \mid X_C$.

DEFINITION 2.2. A distribution P on V is called *Markov* over $\mathscr{G}$ if for any decomposition (A, B) of $\mathscr{G}$

$$A \perp\!\!\!\perp B \,|\, A \cap B \,[P].$$

The directed version of the Markov property pertains to a directed acyclic graph $\mathscr{D}$ with vertex set V [Lauritzen, Dawid, Larsen and Leimer (1990)]:

DEFINITION 2.3. A distribution P on V is *directed Markov* over $\mathscr{D}$ if, for all $v \in V$,

$$(\{v\} \cup \mathrm{pa}(v)) \perp\!\!\!\perp \mathrm{nd}(v) \,|\, \mathrm{pa}(v) \,[P]. \tag{1}$$

NOTE. It would have been equivalent, and simpler, to write (1) as $\{v\} \perp\!\!\!\perp \mathrm{nd}(v) \mid \mathrm{pa}(v)$. However, here and elsewhere, we wish to state definitions and results in a form which will be useful for later generalization. This sometimes results in a way of expressing formulae or arguments which may appear pedantic or unnecessary for the case in hand.

Equivalent to (1) is the corresponding assertion which results on replacing $\mathrm{nd}(v)$ by $\mathrm{pr}(v)$, where $\mathrm{pr}(v)$ denotes the predecessors of v in a *well-numbering* of V, that is, a numbering under which each vertex is assigned a higher number than any of its parents.

If $\mathscr{D}$ is a perfect directed version of the undirected decomposable graph $\mathscr{G}$, then the directed Markov property over $\mathscr{D}$ is the same as the undirected Markov property over $\mathscr{G}$, as shown in various generalities by Wermuth (1980), Wermuth and Lauritzen (1983) and Kiiveri, Speed and Carlin (1984), and restated in Section B of the Appendix, Propositions B.2 and B.6.

Markov distributions on decomposable graphs may be constructed by a simple algorithm, as described subsequently.

Let $A, B \subseteq V$, and suppose we are given distributions Q and R for X_A and X_B, respectively. If there exists a single underlying joint distribution with these distributions as its marginals, then Q and R must be consistent in the following sense:

DEFINITION 2.4. We say that distributions Q over A and R over B are *consistent* if they both yield the same distribution over $A \cap B$.

LEMMA 2.5. *Suppose that the distributions Q over A and R over B are consistent. Then there exists a unique distribution P over $A \cup B$ such that* (i) $P_A = Q$, (ii) $P_B = R$ *and* (iii) $A \perp\!\!\!\perp B \mid A \cap B[P]$.

PROOF. We construct P by specifying its margin over A to be Q, and its conditional distributions over B given A to be the same as those over B given $A \cap B$ calculated from R. [These requirements are necessary if (i)–(iii) are to hold.] The conditional distributions in this construction are only determined *modulo* a subset of $\mathscr{X}_{A \cap B}$ having probability zero under R, hence also under Q; whence all choices will lead to the identical distribution P, which is thus uniquely determined. Then (i) and (ii) hold by construction and (ii) by the consistency condition. □

We shall denote the set of distributions satisfying condition (iii) above by $M(A, B)$. We further call P, satisfying the conditions of Lemma 2.5, the *Markov combination* of Q and R, and write $P = Q \star R$. If P, Q and R have density functions p, q and r, respectively, then we have

$$p(x) = \frac{q(x_A)r(x_B)}{q_{A \cap B}(x_{A \cap B})},$$

where the denominator could equally well have been written as $r_{A\cap B}(x_{A\cap B})$. In particular, $P \in M(A, B)$ if and only if we have

$$p(x) = \frac{p_A(x_A)p_B(x_B)}{p_{A\cap B}(x_{A\cap B})}. \tag{2}$$

We extend the above construction to the case of a general decomposable graph $\mathscr{G}$, as follows.

Let $\mathscr{C}$ be the set of cliques of $\mathscr{G}$, and suppose that we are given a pairwise consistent collection of distributions $\{Q_C: C \in \mathscr{C}\}$, Q_C being a distribution over C. Let the cliques be perfectly numbered as $(C_1, \ldots, C_k)$, see Section A of the Appendix.

The obvious attempt to define a Markov distribution P having the $\{Q_C\}$ as its margins on cliques is recursively to define

$$P_{C_1} = Q_{C_1}, \tag{3}$$

$$P_{H_{i+1}} = P_{H_i} \star Q_{C_{i+1}}. \tag{4}$$

A simple inductive argument, using (2), establishes that the distribution P satisfying (3) and (4) has density

$$p(x) = \frac{\prod_{i=1}^{k} p_{C_i}(x_{C_i})}{\prod_{i=2}^{k} p_{S_i}(x_{S_i})}. \tag{5}$$

We remark that the separators (S_i) are, apart from order, the same for any perfect numbering of the cliques. Each such separator S will be repeated $\nu(S)$ times in any sequence (S_i), where $\nu(S)$ is a combinatorial index, related to the number of disconnected components of $\mathscr{G}_{V\setminus S}$, see Darroch, Lauritzen and Speed (1980) or Lauritzen, Speed and Vijayan (1984) for details. If we denote by $\mathscr{S}$ the collection of separators incorporating $\nu(S)$ repetitions of each S, then (5) may be written as

$$p(x) = \frac{\prod_{C\in\mathscr{C}} p_C(x_C)}{\prod_{S\in\mathscr{S}} p_S(x_S)}. \tag{6}$$

THEOREM 2.6. *The distribution constructed above is the unique Markov distribution over $\mathscr{G}$ having the given consistent distributions as its clique marginals.*

PROOF. Proposition B.1 together with Corollary A.8 implies that, for any Markov distribution, we must have

$$C_{i+1} \perp\!\!\!\perp H_i \mid S_{i+1}, \tag{7}$$

so that (3) and (4) must hold; and clearly these together determine a unique P.

Now let $\mathscr{D}$ be a perfect directed version of $\mathscr{G}$ compatible with the clique numbering. Then the distribution so constructed is directed Markov over $\mathscr{D}$. This follows from the fact that for any $v \in V$, $\{v\} \cup \mathrm{pa}(v)$ is complete, and thus $\{v\} \cup \mathrm{pa}(v) \subseteq C_{i+1}$ for some clique C_{i+1}, where we suppose $i + 1$ to have been chosen as small as possible. Then $S_{i+1} \subseteq \mathrm{pa}(v)$ and $\mathrm{nd}(v) = \mathrm{pa}(v) \cup H_i$. We

can then manipulate (7) to derive

$$(\{v\} \cup \mathrm{pa}(v)) \perp\!\!\!\perp \mathrm{nd}(v) \not\vdash \mathrm{pa}(v).$$

The theorem now follows from Proposition B.2. □

With an alternative formulation we have:

COROLLARY 2.7. *If P is Markov over $\mathscr{G}$, then $P = P_A \star P_B$ for any decomposition (A, B) of $\mathscr{G}$.*

PROOF. This is a direct consequence of Theorem 2.6 and Proposition B.7. □

Finally we have the important result.

THEOREM 2.8. *If P is Markov over $\mathscr{G}$, then*

$$A \perp\!\!\!\perp B \mid S[P]$$

whenever S separates A from B.

PROOF. See Corollary B.4. □

The property of P expressed in Theorem 2.8 is the so-called *global* Markov property, which is more usual as a definition of the Markov property on $\mathscr{G}$ [Speed (1979)]. Since it is clear that the global Markov property implies that of Definition 2.2, we have thus shown that, for decomposable graphs, the two definitions are equivalent and—importantly—this holds without restricting the densities to be positive.

3. Hyper Markov laws.

3.1. *Definition and basic properties.* In the present section we introduce various types of distribution laws for a quantity θ which takes values in the set $M(\mathscr{G})$ of Markov probabilities over a given undirected decomposable graph $\mathscr{G}$. Such laws occur, for example, as sampling distributions of maximum likelihood estimators, but can also be used as prior or posterior distributions when it is known only that the distribution of the data is Markov over $\mathscr{G}$.

To avoid confusion, we shall generally refer to a distribution of θ over $M(\mathscr{G})$ as a *law* for θ, and use the notation $\pounds(\theta)$ to denote the law of θ. We shall also consider corresponding notions for directed graphs, although our emphasis is on the undirected case. By a law over A or $\mathscr{G}_A$ we shall mean a law over $M(\mathscr{G}_A)$. [We shall generally only deal with cases where $\mathscr{G}$ is collapsible onto A, in which case $\phi \in M(\mathscr{G}_A)$ if and only if $\phi = \theta_A$ for some $\theta \in M(\mathscr{G})$.]

Also in this section, and in contrast to Section 2, we shall interpret the notation

$$A \perp\!\!\!\perp B \mid C \ [\pounds]$$

to mean

$$\theta_A \perp\!\!\!\perp \theta_B \mid \theta_C \ [\mathfrak{L}]$$

under the law $\mathfrak{L}$ (omitted from the notation if clear from the context) for θ. We shall make extensive use of the fact that much of the analysis of Section 2 remains meaningful under this reinterpretation.

We first note the following, where an expression of the form $\alpha \simeq \beta$ means that each of α and β is a function of the other.

LEMMA 3.1. *It holds that*:

(i) *If* $A \subseteq V$, *then*

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

(ii) *If* $\mathscr{C}$ *is the set of cliques of* $\mathscr{G}$, *then*

$$\theta \simeq \{\theta_C : C \in \mathscr{C}\}.$$

(iii) *If* $\mathscr{G}$ *is collapsible onto* $U \subseteq V$, *and* (A, B) *is a decomposition of* $\mathscr{G}_U$, *then*

$$\theta_U \simeq (\theta_A, \theta_B).$$

PROOF. The assertion (i) is universally true, (ii) follows from Theorem 2.6 and (iii) from Proposition B.5 and Corollary 2.7. □

We note that, while it is always true that $X_{A \cup B} \simeq (X_A, X_B)$, the corresponding statement for θ, namely,

$$\theta_{A \cup B} \simeq (\theta_A, \theta_B)$$

is false in general: We can only assert, as implied by (i), that (θ_A, θ_B) is determined by $\theta_{A \cup B}$. However, (iii) gives a sufficient condition for the converse to be true.

It follows from (i) that given a law $\mathfrak{L}$ for θ, the law of θ_A is determined for any $A \subseteq V$. We shall denote this law by $\mathfrak{L}(\theta_A)$ or $\mathfrak{L}_A$. Similarly, we use $\mathfrak{L}(\theta_{A \mid B})$ or $\mathfrak{L}_{A \mid B}$ to denote the induced law of $\theta_{A \mid B}$. [Note that this latter is really a "joint law" for the collection of distributions $\{\theta(\cdot \mid X_B = x_B)\}$.]

We shall need to generalize the idea of Markov combination, as follows.

DEFINITION 3.2. We say that laws $\mathscr{M}$ over A and $\mathscr{N}$ over B are *hyperconsistent* if they both induce the same law over $A \cap B$.

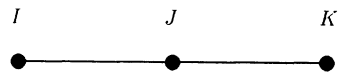
Clearly, if $\mathscr{M}$ and $\mathscr{N}$ are both induced by appropriate marginalization from a common underlying law, they must be hyperconsistent. Note that if $A \cap B = \varnothing$, any pair of laws is hyperconsistent.

LEMMA 3.3. *Suppose that we are given hyperconsistent laws $\mathscr{M}$ over A and $\mathscr{N}$ over B. Then there exists a unique law $\mathfrak{L}$ over $A \cup B$ such that*:

(i) $\mathfrak{L}$ *is concentrated on* $M(A, B)$,
(ii) $\mathfrak{L}_A = \mathscr{M}$,
(iii) $\mathfrak{L}_B = \mathscr{N}$ *and*
(iv) $\theta_A \perp\!\!\!\perp \theta_B \mid \theta_{A \cap B}$.

PROOF. By (i) and Lemma 3.1(iii), it is enough to specify a joint law for (θ_A, θ_B), ensuring that this gives probability one to the event that θ_A and θ_B are consistent. By (ii), the marginal for θ_A has to be $\mathscr{M}$, while, by (iii) and (iv), the conditional law of θ_B given θ_A under $\mathfrak{L}$ must be the same as the conditional law of θ_B given $\theta_{A \cap B}$ under $\mathscr{N}$. This construction specifies a unique joint law $\mathfrak{L}$ for (θ_A, θ_B), these being automatically consistent; and hyperconsistency then ensures that (iii) holds and thus all requirements are satisfied. □

EXAMPLE 3.4. Consider three variables I, J, K related as in the graph $\mathscr{G}$ below



corresponding to $M(\mathscr{G})$ consisting of those θ satisfying

$$\theta_{ijk} = \theta_{ij+}\theta_{+jk}/\theta_{+j+}. \tag{8}$$

If we specify the laws of the marginal distributions $\{\theta_{ij+}\}$ and $\{\theta_{+jk}\}$ as Dirichlet distributions with parameters $\{\alpha_{ij}\}$ and $\{\beta_{jk}\}$, these are hyperconsistent if and only if

$$\alpha_{+j} = \beta_{j+}$$

in which case a law for $\{\theta_{ijk}\}$ exists with these marginals, and with (8) satisfied almost surely. Note that it is not totally immediate that hyperconsistency of the laws is enough to ensure this existence. If we further assume the law to have

$$\{\theta_{ij+}\} \perp\!\!\!\perp \{\theta_{+jk}\} \mid \{\theta_{+j+}\}[\mathfrak{L}],$$

it is uniquely determined from the specified laws of the marginal distributions.

The law $\mathfrak{L}$ satisfying the conditions of Lemma 3.3 will be termed the *hyper Markov combination* of $\mathscr{M}$ and $\mathscr{N}$, and will be denoted by $\mathfrak{L} = \mathscr{M} \odot \mathscr{N}$.

We now wish, in analogy with the Markov case, to extend this construction to build up a law over $\mathscr{G}$ specified only by its clique-marginals $\{\mathfrak{L}_C : C \in \mathscr{C}\}$. In order to do so, we need to impose on $\mathfrak{L}$ a requirement parallel to the Markov requirement on the distribution θ of X.

DEFINITION 3.5. A law $\mathfrak{L}(\theta)$ on $M(\mathscr{G})$ is called (weak) *hyper Markov* over $\mathscr{G}$ if for any decomposition (A, B) of $\mathscr{G}$

$$\theta_A \perp\!\!\!\perp \theta_B \mid \theta_{A \cap B}. \tag{9}$$

NOTE. We remark that the condition (9) is equivalent to either of $\theta_{A \mid B} \perp\!\!\!\perp \theta_B \mid \theta_{A \cap B}$ or $\theta_{A \mid B} \perp\!\!\!\perp \theta_{B \mid A} \mid \theta_{A \cap B}$. However it is truly different from the corresponding pointwise property $\theta_A(x_A) \perp\!\!\!\perp \theta_B(x_B) \mid \theta_{A \cap B}(x_{A \cap B})$ for all x. Neither of these conditions implies the other.

EXAMPLE 3.6. Another hyper Markov law which is associated with the graph in Example 3.4 is the law of

$$\hat{\theta}_{ijk} = \frac{N_{ij+}N_{+jk}}{N_{+j+}n},$$

when $\{N_{ijk}\}$ is assumed to follow a multinomial distribution with expectation $\{n\theta_{ijk}\}$ with $\theta_{ijk} = \theta_{ij+}\theta_{+jk}/\theta_{+j+}$. That is, this law satisfies

$$\left\{\hat{\theta}_{ij+}\right\} \perp\!\!\!\perp \left\{\hat{\theta}_{+jk}\right\} \mid \left\{\hat{\theta}_{+j+}\right\} [\mathfrak{L}]. \tag{10}$$

This will be shown to hold in a more general setting for other maximum likelihood estimators in Section 4.

If there were further restrictions on θ, for example, if I and K had the same state space and it was known that $\theta_{ij+} = \theta_{+ji}$, then we would estimate θ as

$$\tilde{\theta}_{ijk} = \frac{(N_{ij+} + N_{+ji})(N_{kj+} + N_{+jk})}{4N_{+j+}n}$$

which would not satisfy (10) since $\tilde{\theta}_{ij+} = \tilde{\theta}_{+ji}$ for all i, j.

Because the formal properties of laws over subsets of V, with respect to conditional independence and hyper Markov combination, are essentially identical to those of distributions over subsets of V with respect to conditional independence and Markov combination, we are able to reuse much of the development of Section 2 to derive parallel results for the hyper Markov case. The key ideas and results are given below, together with any necessary changes and comments relating to the new setting.

First let $\mathscr{D}$ be a directed acyclic graph and let θ range over the class $M(\mathscr{D})$ of directed Markov distributions over $\mathscr{D}$.

DEFINITION 3.7. A law $\mathfrak{L}(\theta)$ on $M(\mathscr{D})$ is (weak) *directed hyper Markov* over $\mathscr{D}$ if for all $v \in V$

$$\theta_{\{v\} \cup \mathrm{pa}(v)} \perp\!\!\!\perp \theta_{\mathrm{nd}(v)} \mid \theta_{\mathrm{pa}(v)}. \tag{11}$$

Note that the use of $\{v\} \cup \mathrm{pa}(v)$ instead of v alone is essential here; the latter would not be equivalent to the definition as given, since we do not necessarily have $\theta_{\{v\} \cup \mathrm{pa}(v)} \simeq (\theta_v, \theta_{\mathrm{pa}(v)})$. An alternative statement that *is*

equivalent to the definition is

$$\theta_{\{v\}\,|\,\mathrm{pa}(v)} \perp\!\!\!\perp \theta_{\mathrm{nd}(v)} \,|\, \theta_{\mathrm{pa}(v)}.$$

Once again, for any well-numbering of V, (11) is equivalent to the corresponding assertion resulting on replacing $\theta_{\mathrm{nd}(v)}$ by $\theta_{\mathrm{pr}(v)}$.

If $\mathscr{D}$ is a perfect directed version of the undirected decomposable graph $\mathscr{G}$ then, as we have seen, the directed Markov property over $\mathscr{D}$ is the same as the undirected Markov property over $\mathscr{G}$. Correspondingly, we can show equivalence of the weak directed and undirected hyper Markov properties for a law $\mathfrak{L} \in M(\mathscr{D}) = M(\mathscr{G})$.

PROPOSITION 3.8. *Let $\mathscr{D}$ be a perfect directed acyclic graph and $\mathscr{G}$ its undirected version. Then, if* (11) *holds, $\mathfrak{L}(\theta)$ is a (weak) hyper Markov law over $\mathscr{G}$.*

PROOF. The proof is essentially the same as for Proposition B.2. The only points that need care relate to the translation into the new context of the statements $X_A \simeq (X_{\{\lambda\} \cup \mathrm{pa}(\lambda)}, X_{A^*})$ and $X_B \simeq (X_{B^*}, X_S)$.

But these continue to hold when X is replaced by θ, the former by Lemma 3.1(i), and the latter on applying (iii) of the same lemma to the inductive hypothesis, noting that $\mathscr{G}$ is collapsible onto B and (B^*, S) is a decomposition of $\mathscr{G}_B$. □

The converse to Proposition 3.8 follows by direct analogy to Proposition B.5 and Proposition B.6.

Next we show that hyper Markov laws exist and can be constructed by essentially the same algorithm as Markov distributions.

Let $\mathscr{C}$ be the set of cliques of $\mathscr{G}$, and suppose that we are given a pairwise hyperconsistent collection of laws $\{\mathscr{M}_C : C \in \mathscr{C}\}$, $\mathscr{M}_C$ being a law over C. Again let the cliques be perfectly numbered as $(C_1, \ldots, C_k)$ and let $\mathscr{D}$ be a perfect directed version of $\mathscr{G}$ compatible with the clique numbering.

A hyper Markov distribution $\mathfrak{L}$ over $\mathscr{G}$ having the $\{\mathscr{M}_C\}$ as its margins on cliques must satisfy

$$\mathfrak{L}_{C_1} = \mathscr{M}_{C_1}, \tag{12}$$

$$\mathfrak{L}_{H_{i+1}} = \mathfrak{L}_{H_i} \odot \mathscr{M}_{C_{i+1}}. \tag{13}$$

THEOREM 3.9. *The distribution defined by* (12) *and* (13) *is the unique hyper Markov law over $\mathscr{G}$ with the given hyperconsistent laws $\{\mathscr{M}_C\}$ over clique marginals.*

PROOF. Again the proof parallels that of Theorem 2.6, except that additional care is needed to see that $\theta_{\mathrm{pr}(v)} \simeq (\theta_{\mathrm{pa}(v)}, \theta_{H_i})$. □

We can now show the following important consequence of the hyper Markov property.

THEOREM 3.10. *If £ is hyper Markov over $\mathscr{G}$, then*

$$\theta_A \perp\!\!\!\perp \theta_B \mid \theta_S \ [£]$$

whenever S separates A from B in $\mathscr{G}$.

PROOF. Follows as in Proposition B.3 and Corollary B.4. □

The hyper Markov property is preserved under collapsible marginalization:

PROPOSITION 3.11. *If $\mathscr{G}$ is collapsible onto A and £ is hyper Markov over $\mathscr{G}$, then $£_A$ is hyper Markov over $\mathscr{G}_A$.*

PROOF. Same as for Proposition B.5. □

Note that there is no hope of getting such a result for general sets A, since the Markov property of the marginal distribution θ_A itself is preserved if and only if $\mathscr{G}$ is collapsible onto A [Frydenberg (1990)].

3.2. *The strong hyper Markov property.* For the Bayesian analysis of graphical models it is convenient to consider laws of θ that have still stronger independence properties than those expressed in Definition 3.5.

DEFINITION 3.12. A law $£(\theta)$ on $M(\mathscr{G})$ is called *strong hyper Markov* over $\mathscr{G}$ if for any decomposition (A, B) of $\mathscr{G}$

$$\theta_{B \mid A} \perp\!\!\!\perp \theta_A.$$

NOTE. This condition implies that $£(\theta)$ is weak hyper Markov. It is strictly stronger than the corresponding pointwise property

$$\theta_{B \mid A}(x_B \mid x_A) \perp\!\!\!\perp \theta_A(x_A) \quad \text{for all } x.$$

Note also that we could have $B = V$ and A complete in the definition. If we restrict the decomposition to be proper, we get a less strong hyper Markov property. This has also some interest, but we abstain from discussing this further in the present paper.

PROPOSITION 3.13. *A law $£(\theta)$ on $M(\mathscr{G})$ is strong hyper Markov over $\mathscr{G}$ if and only if, under £,*

$$\perp\!\!\!\perp \{\theta_{A \mid B}, \theta_{B \mid A}, \theta_{A \cap B}\} \tag{14}$$

whenever $A \cap B$ is complete and separates A from B.

Here (14) expresses *mutual independence*, itself definable in terms of the conditional independence relation satisfying Properties 1–4 [Dawid (1979a)].

PROOF. We can, if necessary, extend A and B to form a decomposition of $\mathscr{G}$. If $\mathfrak{L}(\theta)$ is strong hyper Markov, $\theta_{A|B} \perp\!\!\!\perp \theta_B$. Since $\theta_B \simeq (\theta_{A\cap B}, \theta_{B|A})$ and $\theta_{A\cap B} \perp\!\!\!\perp \theta_{B|A}$, we have (14). The converse is clear. □

Many results about strong hyper Markov laws are straightforward or can be shown by methods similar to, if not identical with, those for the weak hyper Markov case.

DEFINITION 3.14. A law $\mathfrak{L}$ on $M(\mathscr{D})$, where $\mathscr{D}$ is a directed graph, is *strong directed hyper Markov* over $\mathscr{D}$ if, under $\mathfrak{L}$, for all $v \in V$ we have

$$\theta_{v|\mathrm{pa}(v)} \perp\!\!\!\perp \theta_{\mathrm{nd}(v)}. \tag{15}$$

Again, (15) is equivalent to the corresponding assertion resulting on replacing $\theta_{\mathrm{nd}(v)}$ by $\theta_{\mathrm{pr}(v)}$ for a well-numbering of V. Further, (15) is equivalent to the mutual independence of $\{\theta_{v|\mathrm{pa}(v)}\colon v \in V\}$, termed "global independence" by Spiegelhalter and Lauritzen (1990). (We do not here make any use of the concept of "local independence" discussed in that paper.)

In the strong case, the identity between the directed and undirected definitions no longer holds. Instead we have the following:

PROPOSITION 3.15. *Let $\mathscr{G}$ be a decomposable graph. Then $\mathfrak{L}(\theta)$ is strong hyper Markov over $\mathscr{G}$ if and only if $\mathfrak{L}(\theta)$ is strong directed hyper Markov over $\mathscr{D}$ for every perfect directed version $\mathscr{D}$ of $\mathscr{G}$.*

PROOF. That the strong hyper Markov property over $\mathscr{G}$ implies the strong directed Markov property over any such $\mathscr{D}$ follows as in the proof of Proposition B.6.

For the converse let (A, B) be a decomposition of $\mathscr{G}$. Choosing $\mathscr{D}$ such that A is ancestral in $\mathscr{D}$, the strong directed hyper Markov property over $\mathscr{D}$ ensures that we have $\theta_{B|A} \perp\!\!\!\perp \theta_A$ and the result follows. □

Since there is only one hyper Markov law over $\mathscr{G}$ with given clique-marginal laws, whether or not this is strong hyper Markov must depend on properties of those marginal laws. In fact we have the following:

PROPOSITION 3.16. *Let $\mathfrak{L}$ be hyper Markov over $\mathscr{G}$. Then $\mathfrak{L}$ is strong hyper Markov if and only if, for all cliques C of the graph $\mathscr{G}$ and all subsets A of C we have*

$$\theta_{C\setminus A|A} \perp\!\!\!\perp \theta_A\ [\mathfrak{L}]. \tag{16}$$

PROOF. Since in this case A is complete, (V, A) forms a decomposition of $\mathscr{G}$, so that, when $\mathfrak{L}$ satisfies Definition 3.12, (16) must hold. The converse follows by close analogy to the proof of Theorem 3.9. □

EXAMPLE 3.17. Consider again the three variables I, J, K related as in Example 3.4 and Example 3.6. Since the marginal Dirichlet laws satisfy

$$\{\theta_{ij+}/\theta_{+j+}\} \perp\!\!\!\perp \{\theta_{+j+}\} \quad \text{and} \quad \{\theta_{+jk}/\theta_{+j+}\} \perp\!\!\!\perp \{\theta_{+j+}\},$$

the hyper Markov law constructed in Example 3.4 is a strong law. However this is not the case for the law constructed in Example 3.6, since, with $\theta_{i|j} = \theta_{ij+}/\theta_{+j+}$, we have

$$\hat{\theta}_{i|j} = N_{ij+}/N_{+j+},$$

which is not independent of $\hat{\theta}_{+j+} = N_{+j+}/n$ as $n\hat{\theta}_{i|j}\hat{\theta}_{+j+}$ must be an integer.

There can be no result for the strong case analogous to Proposition B.3 and therefore also not a global property as in Theorem 2.8. This can be seen from Proposition 3.16 since, for example, if $\mathscr{G}^*$ were the complete graph, the strong hyper Markov property of $\mathscr{G}^*$ could only hold in the very special case that the conditional distributions $\theta_{B|A}$ are independent of the marginal distribution θ_A for all subsets A of V. Similar to Proposition 3.11, we have the following:

PROPOSITION 3.18. *If $\mathscr{G}$ is collapsible onto A and $\mathcal{L}(\theta)$ is strong hyper Markov over $\mathscr{G}$, then $\mathcal{L}(\theta_A)$ is strong hyper Markov over $\mathscr{G}_A$ and $\theta_{V\setminus A|A} \perp\!\!\!\perp \theta_A$.*

PROOF. That $\mathcal{L}(\theta_A)$ is strong hyper Markov is a direct consequence of Propositions 3.11 and 3.16. If $B_1, \ldots, B_k$ are the connected components of $V \setminus A$, then $\{\theta_{B_i|V\setminus B_i} = \theta_{B_i|A}\}$ are mutually independent and independent of θ_A by Proposition 3.13. Further these determine $\theta_{V\setminus A|A}$ whereby the result follows. □

4. Meta Markov models.

4.1. *Definition and basic properties.* In many cases it is of interest to consider statistical models where the model is restricted further than just through the restriction $\theta \in M(\mathscr{G})$. We may wish to assume, for example, that θ is a multivariate normal distribution such as in the covariance selection models of Dempster (1972). In general we would wish to restrict θ to lie in some subfamily $\mathscr{P} \subseteq M(\mathscr{G})$.

If a law $\mathcal{L}(\theta)$ confined to $\mathscr{P} \subseteq M(\mathscr{G})$ is strong hyper Markov, then the property $\theta_A \perp\!\!\!\perp \theta_{B|A}$ for a decomposition (A, B) implies, in particular, that with probability 1 under $\mathcal{L}(\theta)$, θ_A and $\theta_{B|A}$ are variation independent, that is, any value that can be taken by θ_A is logically compatible with any value that can be taken by $\theta_{B|A}$. Similar logical relations hold in the weak hyper Markov case. This leads us to consider the structure of models satisfying these logical relations. It turns out that such models have a number of interesting statistical properties, many of which are close parallels of those already studied in the Markov and hyper Markov cases. The notions developed are extensions of what

Barndorff-Nielsen (1978) terms a cut: The statistic $t(x) = x_A$ is a *cut* if θ_A, describing the marginal distribution of $t(X)$, and $\theta_{B|A}$, describing the conditional distribution of X given $t(X)$, are variation independent parameters.

Formally, for $U \subseteq V$ let $\mathscr{A}_U$ denote the family of all probability distributions on $\mathscr{X}_U$. A *model* $\mathscr{P}$ on U is a subfamily of $\mathscr{A}_U$. For $A, B \subseteq U$ we write $\mathscr{P}_A = \{\theta_A\colon\ \theta \in \mathscr{P}\}$, and $\mathscr{P}_{A|B} = \{\theta_{A|B}\colon\ \theta \in \mathscr{P}\}$. We call $\mathscr{P}_A$ the *marginal* model of $\mathscr{P}$ over A, and similarly $\mathscr{P}_{A|B}$ the *conditional* model of $\mathscr{P}$ over A given B.

For $\mathscr{P}$ a model on $U \subseteq V$, let ϕ, χ, ω be parameter functions with domain $\mathscr{A}_U$. We define the *conditional range of* ϕ *given* $\omega = w$ *under* $\mathscr{P}$ to be $\{\phi \circ P\colon\ P \in \mathscr{P}$ and $\omega \circ P = w\}$.

DEFINITION 4.1. We say that ϕ is *variation independent of* χ *given* ω *under* $\mathscr{P}$ and write $\phi \ddagger \chi \mid \omega\ [\mathscr{P}]$ if, for any $(x, w) \in (\chi, \omega) \circ \mathscr{P}$, the conditional range under $\mathscr{P}$ of ϕ given $(\chi, \omega) = (x, w)$ depends only on w; equivalently, if it is the same as the conditional range of ϕ given $\omega = w$.

Another equivalent requirement, symmetric as between ϕ and χ, is that, whenever $(f, w) \in (\phi, \omega) \circ \mathscr{P}$ and $(x, w) \in (\chi, \omega) \circ \mathscr{P}$, then $(f, x, w) \in (\phi, \chi, \omega) \circ \mathscr{P}$. In the case that ω is trivial, we write $\phi \ddagger \chi$, meaning that the range of values allowed for ϕ is unrestricted by specifying the value of χ; or, equivalently, that the range of (ϕ, χ) is a product space.

The variation independence expressed by the above definition has much in common with ordinary probabilistic independence.

LEMMA 4.2. *Properties* 1–4 *continue to hold when* $\perp\!\!\!\perp$ *is replaced by* $\ddagger$.

PROOF. This is straightforward to check. □

In exact parallel with the definition of a hyper Markov law, we can now introduce the concept of a *meta Markov model*:

DEFINITION 4.3. A model $\mathscr{P}$ over V is (*weak*) *meta Markov* with respect to the decomposable graph $\mathscr{G}$ if $\mathscr{P} \subseteq M(\mathscr{G})$ and, for any decomposition (A, B) of $\mathscr{G}$, $\theta_A \ddagger \theta_B \mid \theta_{A \cap B}\ [\mathscr{P}]$. The model $\mathscr{P}$ is *strong meta Markov* if $\theta_A \ddagger \theta_{B|A}\ [\mathscr{P}]$.

We note that $\mathscr{P}$ is strong meta Markov if and only if X_A is a cut in $\mathscr{P}$ for all decompositions (A, B). It follows from Lemma 2.5 that the full model $M(\mathscr{G})$ is strong meta Markov.

When θ has a weak or strong hyper Markov law, the support of this law is a weak or strong meta Markov model.

We could similarly define the weak and strong directed meta Markov properties, in complete analogy to Definitions 2.3 and 3.14. Again, the support of a directed hyper Markov law is a directed meta Markov model.

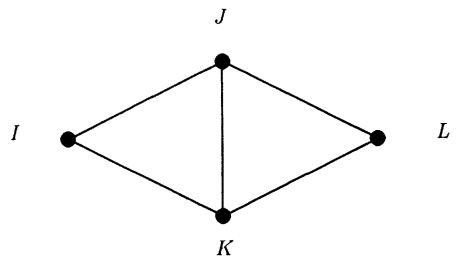
EXAMPLE 4.4. Consider a four-dimensional contingency table with variables I, J, K, L and the hierarchical log-linear model with generating class

$$\{\{I, J\}, \{I, K\}, \{J, K\}, \{J, L\}, \{K, L\}\},$$

meaning that the logarithm of the probabilities has an expansion as

$$\log \theta_{ijkl} = \alpha_{ij} + \beta_{ik} + \gamma_{jk} + \delta_{jl} + \varepsilon_{kl}.$$

The model has independence graph $\mathscr{G}$ as in the following picture



and it is clear that $\mathscr{P} \neq M(\mathscr{G})$, that is, it is a proper submodel of the full graphical model, since no three factor interaction terms are allowed.

But this restriction makes the model strong meta Markov, since the implied model for the marginal distribution of a clique, say θ_{ijk}, is the hierarchical model of no three factor interaction among (I, J, K), letting

$$\log \theta_{ijk} = \mu_{ij} + v_{jk} + w_{ik}$$

and the conditional distribution of L given these varies independently thereof in the set having an expansion as

$$\log \theta_{l|ijk} = \log \theta_{l|jk} = \lambda_{l|j} + \mu_{l|k} - c_{jk},$$

where c_{jk} is the log-normalizing constant

$$c_{jk} = \log \sum_{jl} \exp\{\lambda_{l|j} + \mu_{l|k}\},$$

that is, effects of J and K on the level of L are additive.

To continue the analogy between hyper Markov laws and meta Markov models, we need to define an appropriate concept of *meta Markov combination*. Let $A, B \subseteq V$, and let $\mathscr{P}$ and $\mathscr{Q}$ be models over A and B, respectively. We call $\mathscr{P}$ and $\mathscr{Q}$ *metaconsistent* if $\mathscr{P}_{A\cap B} = \mathscr{Q}_{A\cap B}$. In this case it may be seen that there exists a unique submodel $\mathscr{R}$ of $M(A, B)$ such that $\mathscr{R}_A = \mathscr{P}$, $\mathscr{R}_B = \mathscr{Q}$, and $\theta_A \ddagger \theta_B \mid \theta_{A\cap B}[\mathscr{R}]$. In fact $\mathscr{R} = \{\theta \in M(A, B)\colon \theta_A \in \mathscr{P}, \theta_B \in \mathscr{Q}\}$. We call $\mathscr{R}$ the *meta Markov combination* of $\mathscr{P}$ and $\mathscr{Q}$.

We now remark that all the results of Section 3 may be translated directly into the present context, simply by replacing "law" by "model" and "hyper" by "meta" throughout. The proofs use only Properties 1–4 and the concept of hyper Markov combination, and all translate directly. In particular, we draw attention to the following properties of meta Markov models which may be derived in this way.

PROPOSITION 4.5. *Let $\mathscr{D}$ be a perfect directed acyclic graph and $\mathscr{G}$ its undirected version. Then $\mathscr{P}$ is weak meta Markov over $\mathscr{G}$ if and only if it is weak directed meta Markov over $\mathscr{D}$.*

$\mathscr{P}$ is strong meta Markov with respect to $\mathscr{G}$ if and only if it is strong directed meta Markov with respect to every perfect directed version of $\mathscr{G}$.

THEOREM 4.6. *Let $\mathscr{C}$ be the set of cliques of a decomposable graph $\mathscr{G}$, and let there be given a pairwise metaconsistent collection of models $\{\mathscr{P}_C\colon C \in \mathscr{C}\}$, $\mathscr{P}_C$ being defined over C. Then there exists a unique model $\mathscr{P}$ which is meta Markov with respect to $\mathscr{G}$ and which has the given models as its marginals over the cliques.*

We note that $\mathscr{P}$ may be constructed sequentially as a directed meta Markov model, in which, for each node $v \in V$, the conditional range of $\theta_{v \mid \mathrm{pa}(v)}$ given $\theta_{\mathrm{pa}(v)}$ is that derived from the marginal model over $A = \{v\} \cup \mathrm{pa}(v)$ of the given $\mathscr{P}_C$, for any clique C containing the complete set A.

COROLLARY 4.7. *Let $\mathscr{P}$ be meta Markov with respect to $\mathscr{G}$. If, for all $C \in \mathscr{C}$ and all subsets A of C, X_C is a cut in $\mathscr{P}_C$, then $\mathscr{P}$ is strong meta Markov and conversely.*

THEOREM 4.8. *If $\mathscr{P}$ is meta Markov with respect to $\mathscr{G}$, then*

$$\theta_A \ddagger \theta_B \mid \theta_S$$

whenever S separates A from B.

PROPOSITION 4.9. *If $\mathscr{P}$ is (weak or strong) meta Markov with respect to $\mathscr{G}$, and $\mathscr{G}$ is collapsible onto A, then $\mathscr{P}_A$ is (weak or strong) meta Markov with respect to $\mathscr{G}_A$. Moreover, in the strong case $\theta_A \ddagger \theta_{V \setminus A \mid A}$ (so that X_A is a cut in $\mathscr{P}$).*

It may be noted that Proposition 4.9 extends the result of Asmussen and Edwards (1983) that, when $\mathscr{G}$ is collapsible onto A, X_A is a cut in $M(\mathscr{G})$.

4.2. *Sampling properties for strong meta Markov models.* We now present a number of properties associated with the distribution of the maximum likelihood estimator of a distribution θ in a model $\mathscr{P}$. Throughout the whole of this section we suppose that $\mathscr{P}$ is strong meta Markov with respect to a decomposable graph $\mathscr{G}$. Also we here apply the convention mentioned in

subsection 2.1, that the conditional independence symbol is allowed to contain parameters on its right-hand sides that are not random variables. So, for example,

$$X \perp\!\!\!\perp \theta \,|\, (Y, \psi(\theta))$$

expresses that the conditional distributions of X given Y only depend on θ through $\psi(\theta)$.

LEMMA 4.10. *Suppose that $\mathscr{G}$ is collapsible onto A. Let $X^{(n)} = (X^1, \ldots, X^n)$ be a random sample from some $\theta \in \mathscr{P}$, and let $\hat{\theta}$ be the maximum likelihood estimator of $\theta \in \mathscr{P}$ based on $X^{(n)}$. Then $\hat{\theta}_A$ is a function only of $X_A^{(n)} = (X_A^1, \ldots, X_A^n)$.*

PROOF. The joint density of $X^{(n)}$ factorises in the form

$$p(x^{(n)} \,|\, \theta) = p\big(x_A^{(n)} \,|\, \theta_A\big) p\big(x_{V \setminus A}^{(n)} \,|\, x_A^{(n)}, \theta_{V \setminus A \,|\, A}\big).$$

Since by Proposition 4.9 $\theta_A \ddagger \theta_{V \setminus A \,|\, A}$, that is, θ_A and $\theta_{V \setminus A \,|\, A}$ are variation independent, the two terms may be maximised separately. □

COROLLARY 4.11. *The maximum likelihood estimator of θ_A is the same, whether based on $X^{(n)}$ or on $X_A^{(n)}$.*

COROLLARY 4.12. *The sampling distribution of $\hat{\theta}_A$ is a law over $M(\mathscr{G}_A)$ which depends on $\theta \in \mathscr{P}$ only through θ_A.*

Note that in symbols the result in this corollary can be expressed as $\hat{\theta}_A \perp\!\!\!\perp \theta_{B \,|\, A} \,|\, \theta_A$.

For certain models, the maximum likelihood estimator of θ may possess additional sufficiency properties, in which case further simplifications ensue.

DEFINITION 4.13. We say that a model $\mathscr{P}$, parametrised by θ, is *good* with respect to $\mathscr{G}$ if, for any complete subset C of $\mathscr{G}$, $\hat{\theta}_C$ is sufficient for $\theta_C \in \mathscr{P}_C$ based on $X_C^{(n)}$.

Note that, if $\mathscr{P}$ is good, and C is complete, then $\hat{\theta}_C$ is in fact minimal sufficient for θ based on $X_C^{(n)}$.

LEMMA 4.14. *If $\mathscr{P}$ is good with respect to $\mathscr{G}$, then $\hat{\theta}_A$ is sufficient for $\theta_A \in \mathscr{P}$ based on $X_A^{(n)}$ whenever $\mathscr{G}$ is collapsible onto A.*

PROOF. Since, when $\mathscr{G}$ is collapsible onto A, $\mathscr{P}_A$ is strong meta Markov with respect to $\mathscr{G}_A$, it is enough to prove the result in the case $A = V$. It holds by definition when V is complete. Otherwise, let (A, B) be a proper decomposition of $\mathscr{G}$, and suppose that the result has been established for $\mathscr{G}_A$ and $\mathscr{G}_B$. Then for each of the cases $K = A$, B or $A \cap B$, $\hat{\theta}_K$ is a function of $X_K^{(n)}$, and is a sufficient statistic for $\theta_K \in \mathscr{P}_K$ based on $X_K^{(n)}$.

We have

$$p(x^{(n)} \mid \theta) = \frac{p\left(x_A^{(n)} \mid \theta_A\right) p\left(x_B^{(n)} \mid \theta_B\right)}{p\left(x_{A \cap B}^{(n)} \mid \theta_{A \cap B}\right)}. \tag{17}$$

By the Fisher–Neyman factorization theorem, each of the three terms on the right-hand side of (17) has the form

$$p\left(x_K^{(n)} \mid \theta_K\right) = a\left(x_K^{(n)}\right) b\left(\hat{\theta}_K, \theta_K\right).$$

Hence, since θ is determined by θ_A and θ_B, and similarly for $\hat{\theta}$, the same form applies when $K = V$. □

LEMMA 4.15. *Suppose that $\mathscr{P}$ is good with respect to $\mathscr{G}$. If $\mathscr{G}$ is collapsible onto A, and $C \subseteq A$ is complete, then*

$$X_A^{(n)} \perp\!\!\!\perp \theta \mid \left(\hat{\theta}_C, \theta_{A \mid C}\right).$$

PROOF. Definition 4.13 implies $X_C^{(n)} \perp\!\!\!\perp \theta \mid \hat{\theta}_C$, whence

$$X_C^{(n)} \perp\!\!\!\perp \theta \mid \left(\hat{\theta}_C, \theta_{A \mid C}\right). \tag{18}$$

Also, from the very meaning of $\theta_{A \mid C}$, $X_A^{(n)} \perp\!\!\!\perp \theta \mid (X_C^{(n)}, \theta_{A \mid C})$, so that

$$X_A^{(n)} \perp\!\!\!\perp \theta \mid \left(X_C^{(n)}, \hat{\theta}_C, \theta_{A \mid C}\right). \tag{19}$$

The result follows from (18) and (19). □

We shall also need to consider a stronger sufficiency concept than that of Definition 4.13. Recall [Lehmann and Scheffé (1950)] that a family $\mathscr{P}$ of distributions for a random variable X is called *boundedly complete* if the conditions (i) U is a bounded real function of X, and (ii) $\mathbf{E}(U \mid \theta) = 0$ for every $\theta \in \mathscr{P}$, imply that $U = 0$ almost surely for every $\theta \in \mathscr{P}$. A function W of X is called boundedly complete if the family $\mathscr{P}_W$ of induced distributions for W is. Here, however, in order to avoid confusion with the concept of a complete subset in a graph, we shall use the term *saturated* instead of "boundedly complete."

DEFINITION 4.16. We say that $\mathscr{P}$ is *very good* with respect to $\mathscr{G}$ if $\mathscr{P}$ is good with respect to $\mathscr{G}$, and $\hat{\theta}_C$ is saturated for any complete C.

Note that, since any function of a saturated statistic is itself saturated, it is enough to require that this condition hold for every clique C of $\mathscr{G}$.

LEMMA 4.17. *A sufficient condition for $\mathscr{P}$ to be very good is that $\mathscr{P}$ be weak meta Markov, and that, for any complete set S in $\mathscr{G}$, the model $\mathscr{P}_S$ form a full exponential family.*

PROOF. In this case $\hat{\theta}_S$ is sufficient and saturated in $\mathscr{P}_S$, and is a cut in $\mathscr{P}_C$, whenever $S \subseteq C$ with C a clique of $\mathscr{G}$ [Barndorff-Nielsen (1978)]. □

We require a general result on saturation, extending Theorem 2 of Basu (1955).

LEMMA 4.18. *Let $\mathscr{P}$, parametrised by θ, be a family of distributions for a random variable X. Let T, Y and S be functions of X such that T is saturated and sufficient for $\mathscr{P}$, S is sufficient based on Y, and S is a function of T. Then $Y \perp\!\!\!\perp (T, \theta) \mid S$.*

PROOF. Let $J \subseteq \mathscr{Y}$, the sample space for Y, and define

$$Z = P(Y \in J \mid T, \theta) - P(Y \in J \mid S, \theta).$$

By sufficiency, neither term on the right-hand side depends on θ. Thus Z is a bounded function of T, and $E(Z \mid \theta) = 0$ for all θ. By saturation, $Z = 0$ almost surely for every θ. Since this holds for all measurable $J \subseteq \mathscr{Y}$, the result follows. □

COROLLARY 4.19. *If $\mathscr{P}$ is very good with respect to $\mathscr{G}$, and $C \subseteq A$ where A is complete in $\mathscr{G}$, then*

$$X_C^{(n)} \perp\!\!\!\perp \left(\hat{\theta}_A, \theta\right) \mid \hat{\theta}_C.$$

A number of useful conditional independence properties enjoyed by very good models can now be developed.

PROPOSITION 4.20. *If $\mathscr{P}$ is very good with respect to $\mathscr{G}$, and C is complete in $\mathscr{G}$, then*

$$X_C^{(n)} \perp\!\!\!\perp (\hat{\theta}, \theta) \mid \hat{\theta}_C.$$

PROOF. The result holds if V is complete by Corollary 4.19. Suppose it holds for all decomposable proper subgraphs of $\mathscr{G}$, and let (A, B) be a proper decomposition of $\mathscr{G}$. Without loss of generality, we suppose $C \subseteq A$. By the Markov property of the distribution θ of X we have $X_A^{(n)} \perp\!\!\!\perp X_B^{(n)} \mid (X_{A \cap B}^{(n)}, \theta)$, whence

$$X_A^{(n)} \perp\!\!\!\perp \hat{\theta}_B \mid \left(X_{A \cap B}^{(n)}, \hat{\theta}_{A \cap B}, \theta\right). \tag{20}$$

Also, by assumption,

$$X_{A \cap B}^{(n)} \perp\!\!\!\perp \hat{\theta}_B \mid \left(\hat{\theta}_{A \cap B}, \theta\right). \tag{21}$$

From (20) and (21) we obtain $X_A^{(n)} \perp\!\!\!\perp \hat{\theta}_B \mid (\hat{\theta}_{A \cap B}, \theta)$, whence

$$X_C^{(n)} \perp\!\!\!\perp \hat{\theta}_B \mid \left(\hat{\theta}_A, \theta\right). \tag{22}$$

Again by assumption,

$$X_C^{(n)} \perp\!\!\!\perp \hat{\theta}_A \mid \left(\hat{\theta}_C, \theta\right). \tag{23}$$

Together, (22) and (23) imply $X_C^{(n)} \perp\!\!\!\perp (\hat{\theta}_A, \hat{\theta}_B) \mid (\hat{\theta}_C, \theta)$, and the result follows since, by sufficiency, $X_C^{(n)} \perp\!\!\!\perp \theta \mid \hat{\theta}_C$. □

PROPOSITION 4.21. *Suppose that $\mathscr{P}$ is very good with respect to $\mathscr{G}$. Let (A, B) be a decomposition of $\mathscr{G}$. Then*

$$X_B^{(n)} \perp\!\!\!\perp \left(\hat{\theta}_A, \theta_A\right) \mid \left(\hat{\theta}_{A\cap B}, \theta_{B\mid A}\right).$$

PROOF. From Proposition 4.20 we obtain

$$X_{A\cap B}^{(n)} \perp\!\!\!\perp \left(\hat{\theta}_A, \theta_A\right) \mid \left(\hat{\theta}_{A\cap B}, \theta_{B\mid A}\right). \tag{24}$$

Also, from the Markov property and the definition of $\theta_{B\mid A}$,

$$X_B^{(n)} \perp\!\!\!\perp \left(X_A^{(n)}, \theta_A\right) \mid \left(X_{A\cap B}^{(n)}, \theta_{B\mid A}\right),$$

whence

$$X_B^{(n)} \perp\!\!\!\perp \left(\hat{\theta}_A, \theta_A\right) \mid \left(X_{A\cap B}^{(n)}, \theta_{B\mid A}\right). \tag{25}$$

Since $\hat{\theta}_{A\cap B}$ is a function of $X_{A\cap B}^{(n)}$, which in turn is a function of $X_B^{(n)}$, (24) and (25) together yield the result. □

As an immediate consequence we get the main result of this section.

THEOREM 4.22. *If $\mathscr{P}$ is very good with respect to $\mathscr{G}$, then the sampling distribution of $\hat{\theta}$, given any $\theta \in \mathscr{P}$, is a (weak) hyper Markov law with respect to $\mathscr{G}$.*

PROOF. From Proposition 4.21 it follows immediately that

$$\hat{\theta}_B \perp\!\!\!\perp \hat{\theta}_A \mid \left(\hat{\theta}_{A\cap B}, \theta\right),$$

which is the hyper Markov property. □

The model in Example 4.4 is very good by Lemma 4.17, since both clique-marginal models are hierarchical log-linear models and therefore full exponential families. Hence it follows that

$$\left\{\hat{\theta}_{l\mid jk}\right\} \perp\!\!\!\perp \left\{\hat{\theta}_{ijk}\right\} \mid \left\{\hat{\theta}_{jk}\right\}.$$

As a consequence, if we set constraints to make the λ and u parameters uniquely defined, we have, for example, that

$$\hat{\lambda}_{l\mid j} \perp\!\!\!\perp \hat{u}_{ij} \mid N_{+jk+},$$

where $\{N_{ijkl}\}$ is the table of multinomial counts.

5. Hyper Markov laws for Bayesian inference. In this section we investigate the properties of hyper Markov laws when used as prior distributions for the unknown distribution θ of data X. In particular, we shall see that the strong hyper Markov property permits considerable simplification of the prior-to-posterior analysis.

5.1. *Complete observation.* Let X denote an observation from a distribution θ supposed to be in a meta Markov model over $\mathscr{G}$. If θ is assigned the prior law £, a joint distribution is thereby created for the pair (X, θ). We shall exploit the following properties of this joint distribution.

PROPOSITION 5.1. *If the prior law £(θ) is hyper Markov over $\mathscr{G}$ then the joint distribution of (X, θ) satisfies, for any decomposition (A, B) of $\mathscr{G}$,*

$$(X_A, \theta_A) \perp\!\!\!\perp (X_B, \theta_B) \,|\, (X_{A\cap B}, \theta_{A\cap B}). \tag{26}$$

If £(θ) is strong hyper Markov, it also satisfies

$$(X_A, \theta_A) \perp\!\!\!\perp (X_B, \theta_{B|A}) \,|\, X_{A\cap B}. \tag{27}$$

PROOF. Let (A, B) be a decomposition of $\mathscr{G}$. It follows from the very meaning of θ_A and $\theta_{B|A} = \theta_{B|A\cap B}$ that

$$X_A \perp\!\!\!\perp \theta_{B|A} \,|\, \theta_A \tag{28}$$

and

$$X_B \perp\!\!\!\perp (X_A, \theta_A) \,|\, (X_{A\cap B}, \theta_{B|A}). \tag{29}$$

Combining (28) with the hyper Markov property $\theta_A \perp\!\!\!\perp \theta_{B|A} \,|\, \theta_{A\cap B}$, we obtain the relation $(X_A, \theta_A) \perp\!\!\!\perp \theta_{B|A} \,|\, \theta_{A\cap B}$, whence

$$(X_A, \theta_A) \perp\!\!\!\perp \theta_{B|A} \,|\, (X_{A\cap B}, \theta_{A\cap B}). \tag{30}$$

From (29) we deduce

$$(X_A, \theta_A) \perp\!\!\!\perp X_B \,|\, (X_{A\cap B}, \theta_{B|A}, \theta_{A\cap B}),$$

which combines with (30) to give (26). The corresponding result in the strong case is similar. □

By further conditioning with X_A and X_B in (26) and (27) we obtain the following:

COROLLARY 5.2. *If the prior law of θ is hyper Markov, so is the posterior law obtained by conditioning on complete data $X = x$. If the prior law is strong hyper Markov, so is the posterior.*

We note that the result extends to the case in which the data consist of a random sample of size n from the distribution θ, since we may introduce the observations one at a time.

Corollary 5.2 shows that the family of hyper Markov laws and the family of strong hyper Markov laws each forms a *conjugate family* for the sampling family $M(\mathscr{G})$ of Markov models over $\mathscr{G}$. But strong hyper Markov laws have the further advantage that the updating can be performed locally, as we shall now show.

PROPOSITION 5.3. *Suppose the prior law* $\mathfrak{L}(\theta)$ *is strong hyper Markov over* $\mathscr{G}$, *and let* $\mathscr{G}$ *be collapsible onto* A. *Then*

$$(X_A, \theta_A) \perp\!\!\!\perp (X_{V\setminus A}, \theta_{V\setminus A \mid A}) \mid X_{\mathrm{bd}(V\setminus A)}.$$

PROOF. Essentially the same argument as needed for Proposition 5.1. □

COROLLARY 5.4. *In this case the posterior distribution of* θ_A *based on* X *is the same as that based on* X_A.

PROOF. From Proposition 5.3 we deduce $\theta_A \perp\!\!\!\perp X_{V\setminus A} \mid X_A$. □

As a consequence of Corollary 5.4 we have the following:

COROLLARY 5.5. *If the prior law* $\mathfrak{L}(\theta)$ *is strong hyper Markov, the posterior law of* θ *is the unique (strong) hyper Markov law* $\mathfrak{L}^*$ *specified by the clique-marginal laws* $\{\mathfrak{L}_C^*: C \in \mathscr{C}\}$, *where* $\mathscr{L}_C^*$ *is the posterior distribution of* θ_C *based on its prior law* $\mathfrak{L}_C$ *and the clique-specific data* $X_C = x_C$. *When densities exist,* $\pi(\theta_C \mid x) \propto \pi(\theta_C) p(x_C \mid \theta_C)$.

Thus, when using a strong hyper Markov law as prior distribution, one can localize the calculation of the posterior and restrict attention to one clique at a time, updating the law of its marginal distribution using as data the values observed for the variables in that clique only. Again, this result extends to the case that the data form a random sample from the distribution θ.

We emphasize that the corresponding result in general is false if the prior law is only weakly hyper Markov. In that case, information about θ_C can feed in from $x_{V\setminus C}$.

There are analogous results in the strong directed hyper Markov cases [Spiegelhalter and Lauritzen (1990)]; we omit the details. However, the weak directed hyper Markov property is not generally preserved under sampling from a directed Markov distribution unless the underlying directed graph is perfect, in which case, as we have seen, the problem is identical with an undirected one.

5.2. *Marginal data distributions.* Another simplification special to the strong hyper Markov case relates to the marginal data-distribution of the observables X (i.e., their distribution not conditioned on θ). Thus we have the following:

PROPOSITION 5.6. *If the prior law of* θ *is strong hyper Markov, then the marginal distribution of* X *is Markov.*

PROOF. Follows from Proposition 5.1. □

We note that this result will continue to apply in the case that X is a random sample $X = X^{(n)} = (X^1, X^2, \ldots, X^n)$ of n observations from θ, since

the Markov property conditional on θ holds for $X^{(n)}$. However, after marginalization over θ the (X^i) will no longer be independent, but only exchangeable. In this case the predictive distribution of X^{n+1} given $X^{(n)}$ is the expectation of its sampling distribution θ under the posterior law given $X^{(n)}$. This will always be Markov, although changing from one observation to another.

The following proposition is a strengthening of Theorem 4.22.

PROPOSITION 5.7. *Let $\mathscr{P}$ be very good with respect to $\mathscr{G}$, and suppose that θ is assigned a strong hyper Markov prior law over $\mathscr{P}$. Then if (A, B) is a decomposition of $\mathscr{G}$, in the joint distribution of $\hat{\theta}$ and θ we have*

$$\left(\hat{\theta}_{B\,|\,A}, \theta_{B\,|\,A}\right) \perp\!\!\!\perp \left(\hat{\theta}_A, \theta_A\right) \,|\, \hat{\theta}_{A\cap B}.$$

PROOF. By Corollary 4.12 $\hat{\theta}_A \perp\!\!\!\perp \theta_{B\,|\,A} \,|\, \theta_A$. Here and in the following we exploit that when we have established conditional independences involving parameters θ that are not random, these remain true for random θ, whatever prior distributions are assigned.

Also, with a strong hyper Markov prior, $\theta_{B\,|\,A} \perp\!\!\!\perp \theta_A$. We deduce $\theta_{B\,|\,A} \perp\!\!\!\perp (\hat{\theta}_A, \theta_A)$, whence

$$\theta_{B\,|\,A} \perp\!\!\!\perp \left(\hat{\theta}_A, \theta_A\right) \,|\, \hat{\theta}_{A\cap B}. \tag{31}$$

From Proposition 4.21,

$$\hat{\theta}_{B\,|\,A} \perp\!\!\!\perp \left(\hat{\theta}_A, \theta_A\right) \,|\, \left(\hat{\theta}_{A\cap B}, \theta_{B\,|\,A}\right). \tag{32}$$

The result now follows from (31) and (32). □

COROLLARY 5.8. *Under the conditions of Proposition* 5.7, *the marginal law of $\hat{\theta}$ (not conditioned on θ) is (weak) hyper Markov over $\mathscr{G}$.*

An interesting consequence is the following statement establishing a kind of converse to Corollary 5.5: Conjugate priors for strong meta Markov models must be strong hyper Markov. Define the notion of a conjugate prior as in Barndorff-Nielsen (1978). Then

PROPOSITION 5.9. *Suppose that the conditions in Lemma* 4.17 *hold. Let* £ *be a hyper Markov law such that, for any clique C, the law of θ_C is a conjugate prior distribution for the model $\mathscr{P}_C$. Then* £ *is strong hyper Markov. In particular, the result of Corollary* 5.8 *applies.*

PROOF. This follows from Barndorff-Nielsen [(1978), page 149] which guarantees that, in this case, $\theta_S \perp\!\!\!\perp \theta_{C\setminus S\,|\,S}$ for any $S \subseteq C$. □

We further have when considering a subset A of the variables:

PROPOSITION 5.10. *Suppose £(θ) is strong hyper Markov. If $\mathscr{G}$ is collapsible onto $A \subseteq V$, then* (i) $P_A = \mathbf{E}(\theta_A)$ *and* (ii) $P_{V\setminus A\,|\,A} = \mathbf{E}(\theta_{V\setminus A\,|\,A})$, *where P*

denotes the marginal distribution of X and $\mathbf{E}$ *denotes expectation under the law* £.

PROOF. (i) is immediate. For (ii), note that $X_A \perp\!\!\!\perp \theta_{V\setminus A\,|\,A} \,|\, \theta_A$, by definition, while $\theta_{V\setminus A\,|\,A} \perp\!\!\!\perp \theta_A$ by Proposition 3.18. Thus $\theta_{V\setminus A\,|\,A} \perp\!\!\!\perp (X_A, \theta_A)$ and so $\theta_{V\setminus A\,|\,A} \perp\!\!\!\perp X_A$. Then for any measurable set K,

$$P\big(X_{V\setminus A} \in K \mid X_A\big) = \mathbf{E}\big(\theta_{V\setminus A\,|\,A}\big(X_{V\setminus A} \in K \mid X_A\big)\big|\, X_A\big)$$
$$= \mathbf{E}\big(\theta_{V\setminus A\,|\,A}\big(X_{V\setminus A} \in K \mid X_A\big)\big),$$

which gives (ii). □

The above result also yields the same relations between the densities of P and θ when these exist.

Again closely analogous results hold in the strong directed case, and we omit the details.

5.3. *Partial observation*. When data have only been observed on a subset of the variables, neither of the hyper Markov properties is preserved in general under sampling. An exception is when the observed variables are of the form X_D, where $\mathscr{G}$ is collapsible onto D. Then we have the following:

PROPOSITION 5.11. *If the prior law* £(θ) *is hyper Markov over* $\mathscr{G}$ *and* $\mathscr{G}$ *is collapsible onto* D, *then the posterior law* £$(\theta \mid x_D)$ *is hyper Markov over* $\mathscr{G}$.

PROOF. Let (A, B) be a decomposition of $\mathscr{G}$, and $S = A \cap B$. Define $A^* = A \cap D$, $B^* = B \cap D$, $S^* = S \cap D$ $(= A^* \cap B^*)$. We have, by definition, $X_{A^*} \perp\!\!\!\perp \theta \,|\, \theta_{A^*}$, whence

$$X_{A^*} \perp\!\!\!\perp \theta_{B\,|\,A} \,|\, \theta_A. \tag{33}$$

Now (A^*, B^*) is a decomposition of $\mathscr{G}_D$, while by collapsibility, θ_D is a Markov distribution over $\mathscr{G}_D$. Hence $X_{B^*} \perp\!\!\!\perp (X_{A^*}, \theta) \,|\, (X_{S^*}, \theta_{B^*\,|\,S^*})$, whence, since $\theta_{B^*\,|\,S^*}$ is a function of $\theta_B \simeq (\theta_{B\,|\,A}, \theta_S)$,

$$X_{B^*} \perp\!\!\!\perp (X_{A^*}, \theta_A) \,|\, (X_{S^*}, \theta_{B\,|\,A}, \theta_S). \tag{34}$$

From (33) and (34) we can argue as in the proof of Proposition 5.1 to deduce

$$(X_{A^*}, \theta_A) \perp\!\!\!\perp (X_{B^*}, \theta_B) \,|\, (X_{S^*}, \theta_S),$$

and the result follows. □

COROLLARY 5.12. *If* $\mathscr{D}$ *is a perfect directed acyclic graph, with* D *being ancestral in* $\mathscr{D}$, *and the prior law* £(θ) *is weak directed hyper Markov, then so is the posterior law* £$(\theta \mid x_D)$.

Note that the strong hyper Markov property is not preserved under sampling with partial observation. A counter example is provided by the case

$V = \{a, b\}$ with $\mathscr{G}$ the complete graph on V. If X_a and X_b are binary, then we simply have a 2×2 contingency table. A Dirichlet prior law for the unknown probabilities will be strong hyper Markov (see subsection 7.2); however it may be checked that the posterior law based on the partial observation $X_a = x_a$ will not exhibit the strong hyper Markov requirement $\theta_b \perp\!\!\!\perp \theta_{a|b}$.

We mention that the strong directed hyper Markov property is preserved under sampling from ancestral sets [Spiegelhalter and Lauritzen (1990)].

6. Comparison of models.

6.1. *Generalities.* Let $x^{(n)}$ be observations on $X^{(n)} = (X^1, X^2, \ldots, X^n)$, a random sample from a distribution θ. We may entertain several competing hypotheses about θ, and wish to choose between them in the light of the data. In our context, a typical hypothesis $\mathscr{H}_{\mathscr{G}}$ might imply that θ is Markov with respect to some decomposable graph $\mathscr{G}$ with vertex set V.

Comparing alternative models for a given set of data can, within the scope of the present paper, be identified with comparing different graphs. In the sampling theory framework this would involve considering likelihood ratio statistics for a hypothesis $\mathscr{H}_{\mathscr{G}^*}$, assuming $\mathscr{H}_{\mathscr{G}}$, where $\mathscr{G}^*$ is a decomposable graph, obtained from $\mathscr{G}$ by deleting one or more edges. We shall not discuss the details of this here, but mention that Frydenberg and Lauritzen (1989) give results about the decomposition of these statistics into components, each of which depends only on the data through certain clique marginals and therefore can be locally computed.

If, for each such graph $\mathscr{G}$, we also specify a prior law $\mathfrak{L}_{\mathscr{G}}$ for θ over $M(\mathscr{G})$, we obtain a specialization $\overline{\mathscr{H}}_{\mathscr{G}}$ of $\mathscr{H}_{\mathscr{G}}$ asserting that θ is distributed over $M(\mathscr{G})$ according to the law $\mathfrak{L}_{\mathscr{G}}$. One partially Bayesian approach to choosing between the hypotheses $\{\mathscr{H}_{\mathscr{G}}\}$ is to choose instead between the $\{\overline{\mathscr{H}}_{\mathscr{G}}\}$. Since each $\overline{\mathscr{H}}_{\mathscr{G}}$ induces a marginal data distribution for $X^{(n)}$, the observations $x^{(n)}$ yield *marginal likelihoods* for the competing $\{\overline{\mathscr{H}}_{\mathscr{G}}\}$. These may then be used in a direct likelihood comparison of the hypotheses, or, in a fully Bayesian analysis, further combined with a prior probability for each hypothesis to yield posterior probabilities.

Apart from the intuitively reasonable nature of such a marginal likelihood procedure, general considerations [Dawid (1992)] indicate that it will often have desirable sampling properties. For example, under suitable smoothness conditions, we can expect that, as $n \to \infty$, the marginal likelihood ratio in favour of $\overline{\mathscr{H}}_{\mathscr{G}_1}$ as against $\overline{\mathscr{H}}_{\mathscr{G}_2}$ will tend to infinity almost surely whenever *either* θ satisfies $\mathscr{H}_{\mathscr{G}_1}$ but not $\mathscr{H}_{\mathscr{G}_2}$, *or* θ satisfies $\mathscr{H}_{\mathscr{G}_1}$ which is a lower-dimensional submodel of $\mathscr{H}_{\mathscr{G}_2}$. Under suitable further conditions on the collection $\{\mathscr{H}_{\mathscr{G}}\}$ (e.g., that it is finite and closed under intersection) this property will imply that the marginal likelihood method will yield consistent choice of the true model as $n \to \infty$.

6.2. *Compatibility.* For the above analysis we need not suppose any relationship between the various laws $\{\mathfrak{L}_{\mathscr{G}}\}$ associated with the different hypothe-

ses $\{\mathscr{H}_{\mathscr{G}}\}$. Since these are over different spaces they are necessarily different. However, in the absence of genuine competing expert views, it would seem appropriate that these laws should be made, in some sense, as similar as possible, so as to ensure that the marginal likelihood comparison of the $\{\overline{\mathscr{H}}_{\mathscr{G}}\}$ will reflect real differences in the ability of the different $\{\mathscr{H}_{\mathscr{G}}\}$ to describe the data, rather than differences relating to the incorporation of different prior assumptions.

One way of approaching this is as follows. Let £ be a fixed law for θ over the space of all distributions of X. For any decomposable graph $\mathscr{G}$ with vertex set V, let the law $£_{\mathscr{G}}$ over $M(\mathscr{G})$ be that unique law which is hyper Markov over $\mathscr{G}$ and which induces the same marginal law for each θ_C as does £, where C ranges over the set $\mathscr{C}$ of cliques of $\mathscr{G}$.

A family of laws over different decomposable graphical models which can be constructed in this way starting from a common overall law £ will be called *compatible*. Note that $£_{\mathscr{G}_1}$ determines a unique compatible $£_{\mathscr{G}_2}$ if and only if the edges of $\mathscr{G}_2$ form a subset of those in $\mathscr{G}_1$. In this case, Proposition 3.16 shows that $\mathscr{L}_{\mathscr{G}_2}$ will be strong hyper Markov if $£_{\mathscr{G}_1}$ is. In particular if under £, $\theta_A \perp\!\!\!\perp \theta_{V \setminus A \mid A}$ for every $A \subseteq V$, then every $£_{\mathscr{G}}$ in the compatible family associated with £ will be strong hyper Markov. In this case, the marginal likelihood comparison of the various models may be simplified.

Without the above inclusion condition, it is not obvious how to choose a $£_{\mathscr{G}_2}$ compatible with a given $£_{\mathscr{G}_1}$. One suggestion is discussed in Spiegelhalter, Dawid, Lauritzen and Cowell (1993).

6.3. *Strong hyper Markov comparisons.* Let $x^{(n)}$ be data on $X^{(n)}$, a random sample from the distribution θ, and suppose that we wish to perform the marginal likelihood comparison of various competing hypotheses $\{\overline{\mathscr{H}}_{\mathscr{G}}\}$, where all the graphs $\{\mathscr{G}\}$ are decomposable and the associated prior laws are compatible and strong hyper Markov. By Lemma A.10, it is enough to find the marginal likelihood ratio as between two neighbouring hypotheses $\overline{\mathscr{H}}_{\mathscr{G}}$ and $\overline{\mathscr{H}}_{\mathscr{G}^*}$, where $\mathscr{G}^*$ is obtained from $\mathscr{G}$ by deleting a single edge, and $£_{\mathscr{G}^*}$ is the unique law over $M(\mathscr{G}^*)$ compatible with the law $£_{\mathscr{G}}$ over $M(\mathscr{G})$.

By Lemma A.9, the deleted edge, (u, v) say, must belong to a single clique C of $\mathscr{G}$. We can then form a perfect sequence $(C_1, C_2, \ldots, C_k)$ of the cliques of $\mathscr{G}$, starting from $C_1 = C$. Let $H_j = \bigcup_{i=1}^{j} C_i$, $S_{j+1} = C_{j+1} \cap H_j$, $R_{j+1} = C_{j+1} \setminus H_j$. Then each S_j is complete and $\mathscr{G}$ is collapsible onto each H_j.

Now the marginal distribution P of the data under $\overline{\mathscr{H}}_{\mathscr{G}}$ is Markov, by Proposition 5.6, and thus has joint density p of the form

$$p(x^{(n)}) = p_C\left(x_C^{(n)}\right) \prod_{j=2}^{k} p_{R_j \mid S_j}\left(x_{R_j}^{(n)} \mid x_{S_j}^{(n)}\right). \tag{35}$$

After the edge deletion, C will not be a clique of $\mathscr{G}^*$. However, since $\{u, v\}$ is not a subset of any S_j, each S_j remains complete in $\mathscr{G}^*$. It is then not hard to see that $\mathscr{G}^*$ is collapsible onto each H_j, and to deduce that (35) continues to hold when p is replaced by p^*, the density of the marginal data distribution

P^* under $\overline{\mathscr{H}}_{\mathscr{G}^*}$. Furthermore, for $j \geq 2$, $p_{R_j \mid S_j}(x_{R_j}^{(n)} \mid x_{S_j}^{(n)})$ is a function of $P_{C_j} = \mathbf{E}_{\mathfrak{L}_{\mathscr{G}}}(\theta_{C_j})$. But by compatibility $\mathbf{E}_{\mathfrak{L}_{\mathscr{G}}}(\theta_{C_j}) = \mathbf{E}_{\mathfrak{L}_{\mathscr{G}^*}}(\theta_{C_j})$, whence $P_{C_j} = P_{C_j}^*$, and so all terms after the first in (35) are unchanged when p is replaced by p^*. We deduce the following:

PROPOSITION 6.1. *The marginal likelihood ratio for* $\overline{\mathscr{H}}_{\mathscr{G}^*}$ *as against* $\overline{\mathscr{H}}_{\mathscr{G}}$ *is*

$$\Lambda(\mathscr{G}^* : \mathscr{G}) = p_C^*\big(x_C^{(n)}\big)/p_C\big(x_C^{(n)}\big), \tag{36}$$

where C *is the clique of* $\mathscr{G}$ *containing the edge* (u, v).

This localization of the marginal likelihood model comparison into a single clique of $\mathscr{G}$ is very similar to the corresponding result for maximised likelihood ratio comparisons [Frydenberg and Lauritzen (1989)]. In effect the problem reduces to deciding, on the basis of an otherwise unstructured model for the distribution θ_C of X_C, whether or not, under θ_C, $X_u \perp\!\!\!\perp X_v \mid X_{C \setminus \{u, v\}}$.

We can analyse (36) further. Let $C_u = C \setminus \{v\}$, $C_v = C \setminus \{u\}$, $C_0 = C \setminus \{u, v\}$. Then, in the marginal data distribution under $\overline{\mathscr{H}}_{\mathscr{G}^*}$, $X_u \perp\!\!\!\perp X_v \mid X_{C_0}$. Thus

$$p_C^*\big(x_C^{(n)}\big) = p_{C_u}^*\big(x_{C_u}^{(n)}\big) p_{C_v}^*\big(x_{C_v}^{(n)}\big)/p_{C_0}^*\big(x_{C_0}^{(n)}\big).$$

Moreover, each of C_u, C_v and C_0 is complete in both $\mathscr{G}$ and $\mathscr{G}^*$, and so, again by compatibility, the associated data distributions over these sets are the same under both P and P^*. We thus obtain

$$\Lambda(\mathscr{G}^* : \mathscr{G}) = \frac{p_{C_u}\big(x_{C_u}^{(n)}\big) p_{C_v}\big(x_{C_v}^{(n)}\big)}{p_{C_0}\big(x_{C_0}^{(n)}\big) p_C\big(x_C^{(n)}\big)}. \tag{37}$$

Equation (37) can also be expressed as

$$\Lambda(\mathscr{G}^* : \mathscr{G}) = \frac{p_{u \mid C_0}\big(x_u^{(n)} \mid x_{C_0}^{(n)}\big)}{p_{u \mid C_v}\big(x_u^{(n)} \mid x_{C_v}^{(n)}\big)}, \tag{38}$$

or the corresponding equation with u and v interchanged.

Both (37) and (38) are intuitively reasonable formulae for investigating the conditional independence of X_u and X_v given X_{C_0} in the marginal data distribution. We note from Proposition 5.10 that, if ϕ is the density function corresponding to the distribution θ, then we can use $p_{C_0}(x_{C_0}^{(n)}) = \mathbf{E}_{\mathfrak{L}}(\phi_{C_0}(x_{C_0}^{(n)}))$ in (37), $p_{u \mid C_0}(x_u^{(n)} \mid x_{C_0}^{(n)}) = \mathbf{E}_{\mathfrak{L}}(\phi_{u \mid C_0}(x_u^{(n)} \mid x_{C_0}^{(n)}))$ in (38) and similarly for the other terms.

7. Some special cases.

7.1. *Location models.* It is instructive to investigate the particular simple case of models where the unknown parameter is a location parameter. Let $\mathscr{G} = (V, E)$ be decomposable, and let $Y = (Y_v : v \in V)$ have a fixed distribution $\theta_0 \in M(\mathscr{G})$ over $\mathscr{X} = \times_{v \in V} \mathscr{X}_v$, where each $\mathscr{X}_v$ is the real line—or, more generally, $\mathscr{X}_v$ could be a vector space. With any $\mu \in \mathscr{X}$ we can associate the

distribution $\theta_\mu \in M(\mathscr{G})$ defined to be that of the transformed variables $X = \mu + Y$. This yields a model $\mathscr{P}$ which is easily seen to be meta Markov.

When will $\mathscr{P}$ be strong meta Markov? If this holds we must have

$$\theta_{V\setminus\{v\}\,|\,v} \ddagger \theta_v \,[\mathscr{P}]$$

for all $v \in V$. Now for $\theta = \theta_\mu \in \mathscr{P}$, $\theta_{V\setminus\{v\}\,|\,v}$ comprises the labelled family $(D_x^\mu\colon x \in \mathscr{X}_v)$, D_x^μ being the conditional distribution for $X_{V\setminus\{v\}}$ given $X_v = x$, when X has distribution θ. Equivalently, D_x^μ is the conditional distribution of $\mu_{V\setminus\{v\}} + Y_{V\setminus\{v\}}$ given $Y_v = x - \mu_v$, when Y has distribution θ_0.

Now it will typically be the case that the distribution D_y^0 of $Y_{V\setminus\{v\}}$ given $Y_v = y$ will depend on y in such a way that, for some y^* and all $y \neq y^*$, there do *not* exist any constants $\lambda_{V\setminus\{v\}}$ for which we can obtain D_y^0 as the distribution of $\lambda_{V\setminus\{v\}} + Y_{V\setminus\{v\}}$ when $Y_{V\setminus\{v\}}$ has distribution $D_{y^*}^0$. This will be the case, for example, if for some $u \neq v$ the conditional variance σ_y^2 of Y_u given $Y_v = y$ takes the value $\sigma_{y^*}^2$ only when $y = y^*$. If this property holds, then knowledge of the labelled family $(D_x^\mu\colon x \in \mathscr{X}_v)$ will enable us to determine that unique value x^* for which $x^* - \mu_v = y^*$, and hence to determine μ_v. Thus μ_v, and hence θ_v, will be determined by $\theta_{V\setminus\{v\}\,|\,v}$, and hence $\mathscr{P}$ cannot be strong meta Markov.

7.1.1. *The meta normal model.* An important exception to the above analysis occurs when θ_0 is a multivariate normal distribution in $M(\mathscr{G})$. For simplicity, suppose that its dispersion matrix H is nonsingular, and specifies all and only those conditional independences implied by the requirement $\theta_0 \in M(\mathscr{G})$. We shall also suppose that $\mathscr{G}$ is connected. Without loss of generality we take $\mathbf{E}(Y) = 0$ under θ_0.

In this case we find that D_x^μ is a normal distribution with mean

$$\left(\mu_{V\setminus\{v\}} - H_{V\setminus\{v\},v} h_{vv}^{-1} \mu_v\right) + H_{V\setminus\{v\},v} h_{vv}^{-1} x_v$$

and dispersion independent of x. Hence in this case $(D_x^\mu\colon x \in \mathscr{X})$ does not determine μ_v, but only $(\mu_{V\setminus\{v\}} - H_{V\setminus\{v\},v} h_{vv}^{-1} \mu_v)$. In fact it is easy to see that this normal family is strong meta Markov. We call it the *meta normal* model $\mathscr{MN}(H)$.

It is straightforward to see (e.g., from Lemma 4.17) that the model $\mathscr{MN}(H)$ is very good with respect to $\mathscr{G}$. Consequently, by Theorem 4.22 the sampling distribution of the maximum likelihood estimator $\hat{\theta}$ of θ is a weak hyper Markov law.

But in fact more is true. If $X^{(n)}$ is a random sample from $\theta = \theta_\mu \in \mathscr{MN}(H)$, the maximum likelihood estimator of μ is $\hat{\mu} = n^{-1}\sum_{i=1}^n X^i$. The sampling distribution of $\hat{\mu}$ for given μ is normal with mean μ and dispersion matrix $K = n^{-1}H$. Since $K \propto H$, the results below imply that this induces a strong hyper Markov law for the sampling distribution of $\hat{\theta}$.

7.1.2. *Prior laws.* A prior law £ for θ confined to $\mathscr{P}$ may be specified by assigning a prior distribution to μ over $\mathscr{X}$. When will £ be hyper Markov?

Since θ_A is determined by μ_A for any $A \subseteq V$, Theorem 3.10 shows that it is necessary and sufficient that the prior distribution of μ be ordinary Markov.

When will £ be strong hyper Markov? Since the support of £ would then have to be strong meta Markov, the above analysis shows that this will not usually be possible in a nontrivial way. However, it *is* possible for the meta normal model $\mathscr{MN}(H)$. Indeed, we can construct a multivariate normal prior distribution for μ which will induce a strong hyper Markov law £. Suppose that the dispersion matrix K of μ is positive definite. In order for £ to be even weak hyper Markov, this must be such as to make the distribution of μ Markov over $\mathscr{G}$. Let now C be a clique of $\mathscr{G}$, and take $v \in C$, $A = C \setminus \{v\}$. The strong hyper Markov property requires that $\theta_{v|A} \perp\!\!\!\perp \theta_A$, which is equivalent to having zero correlation between $(\mu_v - H_{vA}H_{AA}^{-1}\mu_A)$ and μ_A under the dispersion structure specified by K. This gives

$$K_{vA} - H_{vA}H_{AA}^{-1}K_{AA} = 0,$$

whence

$$K_{vA}K_{AA}^{-1} = H_{vA}H_{AA}^{-1}. \tag{39}$$

Letting now $\Delta = K^{-1}$, $\Gamma = H^{-1}$, (39) is equivalent to

$$\delta_{vv}^{-1}\Delta_{vA} = \gamma_{vv}^{-1}\Gamma_{vA}.$$

This will hold for every $v \in V$ if and only if, for all $u, v \in C$, $\delta_{vv}^{-1}\delta_{vu} = \gamma_{vv}^{-1}\gamma_{vu}$. In particular, since $\gamma_{vu} \neq 0$ for all $v, u \in C$, the same holds for δ_{vu}. Then δ_{vu}/γ_{vu} depends only on v, and similarly depends only on u. It easily follows that for some $\alpha_C > 0$, $\delta_{vu} = \alpha_C\gamma_{vu}$ for all $u, v \in C$. Since $\mathscr{G}$ is connected, the overlap between cliques ensures that all the (α_C) must be equal. Hence we have shown that in order for £ to be strong hyper Markov, K must be a scalar multiple of H. It is easy to see that this condition is also sufficient. Such a normal distribution for μ may be called a *hyper normal* prior law with respect to the model $\mathscr{MN}(H)$.

Note that, since the marginal distribution of $\hat{\mu}$ then has dispersion matrix $K + n^{-1}H$, the law of $\hat{\theta}$ is itself strong hyper Markov when $K \propto H$.

The case that $\mathscr{G}$ is disconnected is similar, except that now H and K are block-diagonal, and it is necessary and sufficient that proportionality hold within each block, the constants for different blocks being possibly different.

7.2. *Multinomial models and the hyper Dirichlet law.* Suppose that all the variables $(X_v)_{v \in V}$ are discrete-valued, that is, they take values in finite sets $(\mathscr{I}_v)_{v \in V}$. The model $M(\mathscr{G})$ is then a decomposable graphical log-linear model such as described in, for example, Darroch, Lauritzen and Speed (1980). Let $\mathscr{I}$ denote the set of possible configurations of X:

$$\mathscr{I} = \bigtimes_{v \in V} \mathscr{I}_v.$$

Then, by (6), an arbitrary distribution θ in $M(\mathscr{G})$ is determined by the

clique marginal probability tables $\theta_{\mathscr{C}} = (\theta_C)_{C \in \mathscr{C}}$ as

$$\theta(i) = \frac{\prod_{C \in \mathscr{C}} \theta_C(i_C)}{\prod_{S \in \mathscr{S}} \theta_S(i_S)} \quad \text{for } i \in \mathscr{I},$$

where $\mathscr{C}$ is the set of cliques of $\mathscr{G}$ and $\mathscr{S}$ is the system of separators in a perfect ordering of these. Note that the same set S may appear several times in the expression. For $S = C \cap D$ where C and D are cliques, θ_S can be calculated by marginalization either from θ_C or from θ_D.

7.2.1. *Sampling theory.* Suppose now that observations

$$x^{(n_0)} = (x^1, x^2, \ldots, x^{n_0})$$

of a random sample $X^{(n_0)} = (X^1, X^2, \ldots, X^{n_0})$ from the Markov distribution θ are obtained. A sufficient statistic is $n = (n(i))_{i \in \mathscr{I}}$, the contingency table of counts, with $n(i)$ the number of observations (x^r) having the particular configuration i. Similarly let n_A denote the counts $n(i_A)$ in the marginal table $\mathscr{I}_A$, obtained by taking into account only the variables in A.

The maximum likelihood estimator $\hat{\theta}$ of $\theta \in M(\mathscr{G})$ is a Markov distribution, with $\hat{\theta}_C = n_C/n_0$ for $C \in \mathscr{C}$ by Corollary 4.11. Thus when all terms are positive we have (Darroch, Lauritzen and Speed, 1980)

$$n_0 \hat{\theta}(i) = \frac{\prod_{C \in \mathscr{C}} n_C(i_C)}{\prod_{S \in \mathscr{S}} n_S(i_S)}.$$

We shall call the sampling distribution of $\hat{m} = n_0 \hat{\theta}$ the *hyper multinomial law*, $\mathscr{HM}_{\mathscr{G}}(n_0, \theta)$. We observe that for $C \in \mathscr{C}$, the induced distribution of $\hat{m}_C = n_C$ is the multinomial distribution with index n_0 and probabilities θ_C. Since the model $M(\mathscr{G})$ is very good with respect to $\mathscr{G}$, we obtain from Theorem 4.22:

PROPOSITION 7.1. *The law $\mathscr{HM}_{\mathscr{G}}(n_0, \theta)$ for $\hat{m}$ determines a hyper Markov law for $\hat{\theta} = \hat{m}/n_0$ over $\mathscr{G}$.*

We emphasize that this law is not strong hyper Markov, since we do not have the independence $\hat{\theta}_{C \setminus A \mid A} \perp\!\!\!\perp \hat{\theta}_A$ for $A \subseteq C \in \mathscr{C}$. Without stretching language greatly, we shall refer to $\mathscr{HM}_{\mathscr{G}}(n_0, \theta)$ as a hyper Markov law for $\hat{m}$. The explicit expression for the density of this law was derived by Sundberg (1975).

7.2.2. *Prior laws.* For each clique $C \in \mathscr{C}$, let

$$\lambda_C = (\lambda_C(i_C))_{i_C \in \mathscr{I}_C}$$

be a given table of arbitrary positive numbers and let $\mathscr{D}(\lambda_C)$ denote the *Dirichlet distribution* for θ_C with density

$$\pi(\theta_C \mid \lambda_C) \propto \prod_{i_C \in \mathscr{I}_C} \theta_C(i_C)^{\lambda_C(i_C) - 1}$$

on the set where $\Sigma_{i_C}\theta_C(i_C) = 1$ and $\theta_C(i_C) > 0$. We recall the following well-known properties of the Dirichlet distribution which all follow easily from the representation of a Dirichlet random variable θ as $\theta(i) = Y(i)/\Sigma_i Y(i)$, where $Y(i)$ are independent and gamma distributed with common scale and shape parameters $\lambda(i)$ [although (ii) and (iii) do not seem explicit in the literature]. See, for example, Johnson and Kotz (1972) or Wilks (1962) for standard results.

LEMMA 7.2. *If* $\mathcal{L}(\theta) = \mathcal{D}(\lambda)$, $A \subseteq V$ *and* $B = V \setminus A$, *then*:

(i) $\mathcal{L}(\theta_A) = \mathcal{D}(\lambda_A)$,
(ii) $\theta_{B|A}(\cdot \mid i_A)$ *are all independent and distributed as* $\mathcal{D}(\lambda_{B|A}(\cdot \mid i_A))$,
(iii) $\theta_A \perp\!\!\!\perp \theta_{B|A}$.

Here we are defining $\lambda_A(i_A) = \Sigma_{j:\, j_A = i_A}\lambda(j)$, $\lambda_{B|A}(i_B \mid i_A) = \lambda(i)$ and so on.

It follows from (i) that the collection of specifications

$$\mathcal{L}(\theta_C) = \mathcal{D}(\lambda_C), \qquad C \in \mathscr{C}$$

will be pairwise hyperconsistent so long as for any two cliques C and D with $C \cap D \neq \emptyset$ we have

$$(40)\quad \lambda_C(i_{C\cap D}) = \sum_{j_C:\, j_{C\cap D} = i_{C\cap D}} \lambda_C(j_C) = \sum_{j_D:\, j_{C\cap D} = i_{C\cap D}} \lambda_D(j_D) = \lambda_D(i_{C\cap D}).$$

In particular, if $\mathscr{G}$ is connected, $\Sigma\lambda_C(i_C)$ will not depend on C. Henceforth we restrict attention to the case of a connected graph $\mathscr{G}$; general graphs are easily handled by considering their connected components separately. When (40) holds, it is possible to find a (nonunique) $\lambda = (\lambda(i))_{i \in \mathscr{I}}$ with C marginal equal to λ_C for all $C \in \mathscr{C}$. To see that this is true one can, for example, take

$$\lambda(i) = \frac{\Pi_{C \in \mathscr{C}}\lambda_C(i_C)}{\Pi_{S \in \mathscr{S}}\lambda_S(i_S)}.$$

We obtain from Theorem 3.9 that given any such hyperconsistent collection $\lambda_{\mathscr{C}} = (\lambda_C)_{C \in \mathscr{C}}$ there exists a unique *hyper Dirichlet* law for θ, denoted by $\mathcal{HD}(\lambda_{\mathscr{C}})$ or $\mathcal{HD}_{\mathscr{G}}(\lambda)$, which is hyper Markov over $\mathscr{G}$ and has $\mathcal{L}(\theta_C) = \mathcal{D}(\lambda_C)$ for all $C \in \mathscr{C}$. Moreover, by (iii) and Proposition 3.16, this law is in fact strong hyper Markov. This is also a consequence of Proposition 5.9, since by Lemma 4.17, $M(\mathscr{G})$ is very good, and the Dirichlet distributions are conjugate to the multinomial model.

If we confine attention to θ_C with prior law $\mathcal{D}(\lambda_C)$, and the data n_C from the marginal table corresponding to clique C, the posterior law for θ_C given n_C will be $\mathcal{D}(\lambda_C + n_C)$.

The marginal counts $(n_C)_{C \in \mathscr{C}}$ automatically satisfy the consistency conditions in (40). It then follows from Corollary 5.5 that, if the prior law of θ was $\mathcal{HD}_{\mathscr{G}}(\lambda)$, then the posterior law will be $\mathcal{HD}_{\mathscr{G}}(\lambda + n)$. The family of hyper Dirichlet laws for θ is thus closed under sampling from the graphical model $M(\mathscr{G})$. We can regard λ as an "equivalent prior sample" characterizing the prior law. The corresponding equivalent posterior sample λ^* is then simply $\lambda^* = \lambda + n$.

7.2.3. *Data distributions.* We also obtain a law $\mathscr{HMD}_{\mathscr{G}}(n_0, \lambda)$, the *hyper multinomial-Dirichlet* law, defined as the marginal distribution of $\hat{m}$ when θ is assigned the prior law $\mathscr{HD}_{\mathscr{G}}(\lambda)$. This will be hyper Markov by Corollary 5.8, the induced distribution for $\hat{m}_C = n_C$ ($C \in \mathscr{C}$) being multinomial-Dirichlet.

Consider now the marginal distribution of the full data-set $X^{(n_0)}$. This is Markov by Proposition 5.6. Within any clique or complete set C we have

$$p_C\big(x_C^{(n_0)}\big) = \mathbf{E}\Big(\prod_{i_C \in \mathscr{I}_C} \theta_{i_C}^{n_{i_C}}\Big),$$

where the expectation is with respect to the law $\mathscr{D}(\lambda_C)$ of θ_C, and we are writing θ_{i_C} for $(\theta_C)_{i_C}$ and so on. This gives

$$p_C\big(x_C^{(n_0)}\big) = \left(\frac{\Gamma(\lambda_0)}{\Gamma(\lambda_0^*)}\right) \prod_{i_C \in \mathscr{I}_C} \left(\frac{\Gamma(\lambda_{i_C}^*)}{\Gamma(\lambda_{i_C})}\right), \tag{41}$$

where $\lambda^* = \lambda + n$ and $\lambda_0 = \sum_{i \in \mathscr{I}} \lambda_i$.

This expression can now be used with (6) to give the full density $p(x^{(n_0)})$. It may be noted that, since

$$p(x^{(n_0)}) = \mathbf{E}\Big(\prod_{i \in \mathscr{I}} \theta_i^{n_i}\Big),$$

we thereby obtain the mixed moments of θ under its prior law $\mathscr{HD}_{\mathscr{G}}(\lambda)$.

From (41) or directly, we find that for $n_0 = 1$,

$$p_C = \lambda_C / \lambda_0.$$

Hence the marginal probability function for a single observation is

$$p = \frac{\prod_{C \in \mathscr{C}} \lambda_C}{\lambda_0 (\prod_{S \in \mathscr{S}} \lambda_S)}. \tag{42}$$

(In the disconnected case we must multiply terms such as the above across the different connected components.) The predictive distribution for X^{n_0+1} given $X^{(n_0)}$ may then be found by substituting for λ in (42) its posterior version $\lambda^* = \lambda + n$.

Expression (41) can also be inserted into (37) to yield the marginal likelihood ratio $\Lambda(\mathscr{G}^* : \mathscr{G})$. For the case of a single observation in cell j, the formula obtained is

$$\Lambda(\mathscr{G}^* : \mathscr{G}) = \frac{\lambda_{j_{C_u}} \lambda_{j_{C_v}}}{\lambda_{j_{C_0}} \lambda_{j_C}}. \tag{43}$$

Likewise the incremental factor inserted into $\Lambda(\mathscr{G}^* : \mathscr{G})$ by the observation of x^{n_0+1} in cell j after already obtaining data $x^{(n_0)}$ summarized by the counts n is given by expression (43), where λ is replaced by $\lambda^* = \lambda + n$. These successive individual factors can be used for continuous monitoring of the relative evidence in favour of $\mathscr{G}^*$ as against $\mathscr{G}$.

7.3. *Covariance selection and the hyper inverse Wishart law.*

7.3.1. *Sampling theory.* Suppose that all variables are continuous and assumed to be jointly multivariate normal with means all equal to zero and

unknown covariance matrix Σ (here assumed positive definite). The intersection $\mathscr{C}\mathscr{S}_{\mathscr{G}}$ of this model with $M(\mathscr{G})$ is the family of *covariance selection models* with respect to $\mathscr{G}$ [Dempster (1972) and Wermuth (1976)]. Such a distribution is determined by the clique-marginal covariance matrices $\{\Sigma^C: C \in \mathscr{C}\}$, which are arbitrary subject only to the consistency requirement that if $S = C \cap C^*$, then the submatrices of Σ^C and of Σ^{C^*} relating to the variables in S must be identical (with common value Σ^S say). When this holds, we can show, for example, from (6) that

$$K = \sum_{C \in \mathscr{C}} [K^C]^0 - \sum_{S \in \mathscr{S}} [K^S]^0, \tag{44}$$

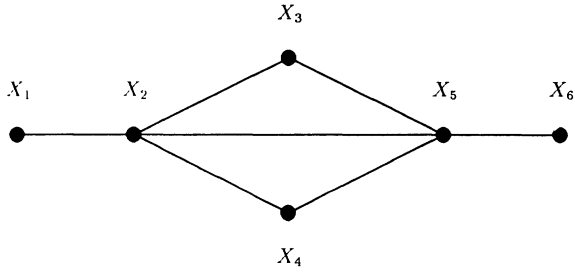
where $K = \Sigma^{-1}$ is the *concentration matrix* corresponding to Σ, and similarly for K^C and K^S; and $[A]^0$ is obtained from the matrix A by extending it with zeros so as to give it the correct dimensions.

Now for $C \in \mathscr{C}$, the model $(\mathscr{C}\mathscr{S}_{\mathscr{G}})_C$ is just the family of all zero-mean multivariate normal distributions for X_C, which has the property that for $A \subseteq C$, any specification of the parameters of the multivariate regression of $X_{C \setminus A}$ on X_A is compatible with any specification of the marginal distribution for X_A. Using this it is easy to see that the model $\mathscr{C}\mathscr{S}_{\mathscr{G}}$ is strong meta Markov. It is also very good, as again follows from Lemma 4.17.

Suppose now that we observe as data a random sample $X^{(n)}$ from a distribution in $\mathscr{C}\mathscr{S}_{\mathscr{G}}$ specified by Σ. If S is the sum-of-products matrix formed from $X^{(n)}$, then in the sampling distribution, $\mathfrak{L}(S \mid \Sigma) = \mathscr{W}(n; \Sigma)$. A sufficient statistic for Σ is $S^{\mathscr{C}} = \{S^C: C \in \mathscr{C}\}$, where $S^C = S_{CC}$. In clique C we have S^C sufficient and $\mathfrak{L}(S^C \mid \Sigma) = \mathscr{W}(n; \Sigma^C)$.

Let $\hat{\Sigma}$ be the maximum likelihood estimator of Σ in the above covariance selection model. The sampling distribution of $\hat{\Sigma}$ given Σ defines a law over $\mathscr{C}\mathscr{S}_{\mathscr{G}}$. By Theorem 4.22, this law will be weak hyper Markov. The marginal distribution of $S^C = n\hat{\Sigma}^C$ for $C \in \mathscr{C}$ is Wishart $\mathscr{W}(n; \Sigma^C)$. We call the distribution of $\hat{S} = n\hat{\Sigma}$ the *hyper Wishart* law $\mathscr{H}\mathscr{W}_{\mathscr{G}}(n; \Sigma)$ and note that it is not generally strong hyper Markov unless Σ is diagonal.

EXAMPLE 7.3. To give an indication some of the more powerful aspects of the theory in the present paper, we study a covariance selection model with mean zero and graph



corresponding to the inverse covariance matrix having zero entries at positions (1, 3), (1, 4), (1, 5), (1, 6), (2, 6), (3, 6) and (4, 6).

Based upon a sample of size n, the maximum likelihood estimate of the inverse covariance is given as

$$n^{-1}\hat{\Sigma}^{-1} = \hat{S}^{-1} = \left[\left(S^{\{1,2\}}\right)^{-1}\right]^0 + \left[\left(S^{\{2,3,5\}}\right)^{-1}\right]^0 + \left[\left(S^{\{2,4,5\}}\right)^{-1}\right]^0 + \left[\left(S^{\{5,6\}}\right)^{-1}\right]^0 - \left[\left(S^{\{2\}}\right)^{-1}\right]^0 - \left[\left(S^{\{5\}}\right)^{-1}\right]^0 - \left[\left(S^{(2,5\}}\right)^{-1}\right]^0,$$

where as usual $[A]^0$ is obtained from A by filling up with zero entries to obtain the correct dimension (here 6×6).

The maximum likelihood estimator has a hyper Wishart distribution from which we deduce, for example, that

$$\hat{\Sigma}^{\{1,2\}} \perp\!\!\!\perp \hat{\Sigma}^{\{3,4,5,6\}} \mid S^{\{2\}}$$

since $\hat{\Sigma}^{\{2\}} = S^{\{2\}}/n$.

7.3.2. *Prior laws.* For each clique $C \in \mathscr{C}$, let Φ^C be a fixed positive definite dispersion matrix. We denote transpose by $'$ and the number of elements in C by $|C|$. The inverse Wishart distribution is the distribution of W^{-1}, where $W = \Sigma_{i=1}^n X_i X_i'$, the X_i being independent and multivariate normal with covariance matrix $(\Phi^C)^{-1}$. Using the parametrization of Dawid (1981) we denote this by $\mathscr{IW}(\delta; \Phi^C)$, where $\delta = n - |C| + 1$. Clearly Σ^C has distribution $\mathscr{IW}(\delta; \mathbf{\Phi}^C)$ if and only if $K^C = (\Sigma^C)^{-1}$ has the Wishart distribution $\mathscr{W}\{\delta + |C| - 1; (\mathbf{\Phi}^C)^{-1}\}$.

We recall the following well-known properties of the multivariate normal and inverse Wishart distributions.

LEMMA 7.4. *Let A be a subset of C and let $B = C \setminus A$. If*

$$\mathfrak{L}(X \mid \Sigma) = \mathscr{N}(0, \Sigma),$$

then

$$\mathfrak{L}(X_A \mid \Sigma) = \mathscr{N}(0, \Sigma_{AA}) \quad \text{and} \quad \mathfrak{L}(X_B \mid X_A, \Sigma) = \mathscr{N}(\Gamma_{B|A} X_A, \Sigma_{B|A}),$$

where

$$\Sigma = \begin{pmatrix} \Sigma_{AA} & \Sigma_{AB} \\ \Sigma_{BA} & \Sigma_{BB} \end{pmatrix}, \quad \Gamma_{B|A} = \Sigma_{BA}\Sigma_{AA}^{-1}, \quad \Sigma_{B|A} = \Sigma_{BB} - \Sigma_{BA}\Sigma_{AA}^{-1}\Sigma_{AB}.$$

Further, if $\mathfrak{L}(\Sigma) = \mathscr{IW}(\delta; \mathbf{\Phi})$ and $\mathbf{\Phi}$ is partitioned accordingly, then:

(i) $\mathfrak{L}(\Sigma_{AA}) = \mathscr{IW}(\delta; \mathbf{\Phi}_{AA})$.
(ii) $\mathfrak{L}(\Sigma_{B|A}) = \mathscr{IW}(\delta + |A|; \mathbf{\Phi}_{B|A})$.
(iii) $\mathfrak{L}(\Gamma_{B|A} \mid \Sigma_{B|A}) = \mathscr{N}(\mathbf{\Phi}_{BA}\mathbf{\Phi}_{AA}^{-1}, \Sigma_{B|A} \otimes \mathbf{\Phi}_{AA}^{-1})$.
(iv) $\Sigma_{AA} \perp\!\!\!\perp (\Gamma_{B|A}, \Sigma_{B|A})$.

For any $C \in \mathscr{C}$ a law for θ_C confined to the zero-mean multivariate normal distributions on C may be specified by assigning a distribution (which we may

again term a law, and denote by £) to Σ^C. It follows from (i) of Lemma 7.4 that the collection of laws

$$\pounds(\Sigma^C) = \mathcal{IW}(\delta; \mathbf{\Phi}^C), \quad C \in \mathcal{C}$$

will be pairwise hyperconsistent so long as, if $B \subseteq C_1 \cap C_2$, the submatrices $\Phi_{BB}^{C_1}$ and $\mathbf{\Phi}_{BB}^{C_2}$ of $\mathbf{\Phi}^{C_1}$ and $\mathbf{\Phi}^{C_2}$ respectively are identical. There will then exist a unique hyper Markov law for Σ corresponding to the marginal specifications defined by $\mathbf{\Phi}^{\mathcal{C}} = \{\mathbf{\Phi}^C, C \in \mathcal{C}\}$, which we call the *hyper inverse Wishart law* and denote by $\mathcal{HIW}(\delta; \mathbf{\Phi}^{\mathcal{C}})$ or $\mathcal{HIW}_{\mathcal{G}}(\delta; \mathbf{\Phi})$, where $\mathbf{\Phi}$ is any dispersion matrix having $\mathbf{\Phi}_{CC} = \mathbf{\Phi}^C$, $C \in \mathcal{C}$—for example, that constructed by a formula such as (44).

Now the inverse Wishart prior distributions for Σ are conjugate to the model $\mathcal{N}(0, \Sigma)$ when Σ is unrestricted. Consequently we deduce, by Proposition 5.9 (or direct from (iv) of Lemma 7.4 and Proposition 3.16):

PROPOSITION 7.5. *The hyper inverse Wishart law is strong hyper Markov.*

If a random sample has been observed as previously described, and the prior law of Σ is $\mathcal{HIW}(\delta; \mathbf{\Phi})$, the prior law for Σ^C is $\mathcal{IW}(\delta; \mathbf{\Phi}^C)$ and the posterior is thus $\mathcal{IW}(\delta + n; S^C + \mathbf{\Phi}^C)$. So the full posterior is $\mathcal{HIW}(\delta + n; S^{\mathcal{C}} + \mathbf{\Phi}^{\mathcal{C}})$ or equivalently $\mathcal{HIW}_{\mathcal{G}}(\delta + n; S + \mathbf{\Phi})$, exhibiting the hyper inverse Wishart laws as a conjugate family for the model $\mathcal{CS}_{\mathcal{G}}$. We can now interpret $\mathbf{\Phi}$ as the sum-of-products matrix from an equivalent prior sample of size δ, and then the corresponding posterior quantities are then $\mathbf{\Phi} + S$ and $\delta + n$.

7.3.3. *Data distributions.* As in the previous special cases we can introduce the weak hyper Markov law corresponding to the marginal data distribution of $\hat{S}$. This will induce a distribution for $S^C = \hat{S}_{CC}$ which is matrix F [Dawid (1981)]. This distribution for $\hat{S}$ may be termed the *hyper matrix F* law $\mathcal{HF}_{\mathcal{G}}(n; \delta; \mathbf{\Phi})$.

We turn now to the marginal distribution of the full data $X^{(n)}$. For any $A \subseteq C$ we shall regard the values for $X_A^{(n)}$ as arranged in a $(n \times |A|)$ matrix, with a row for each observation and a column for each variable. It follows from Dawid (1981) that, using the notation there introduced, for C complete, the marginal data distribution of $X_C^{(n)}$ is the *matrix t* distribution $T(\delta; I_n, \mathbf{\Phi}^C)$, with density [Dickey (1967)]

$$(45) \quad p(x_C^{(n)}) = \pi^{-n|C|/2} \frac{\Gamma_{|C|}((\delta + n + |C| - 1)/2)}{\Gamma_{|C|}((\delta + |C| - 1)/2)} (\det \mathbf{\Phi}^C)^{-n/2} \times \left[\det\left\{I_n + x_C^{(n)} (\mathbf{\Phi}^C)^{-1} x_C^{(n)\prime}\right\}\right]^{-(\delta+n+|C|-1)/2}$$

$$(46) \quad = \pi^{-n|C|/2} \frac{\Gamma_{|C|}((\delta + n + |C| - 1)/2)}{\Gamma_{|C|}((\delta + |C| - 1)/2)} (\det \mathbf{\Phi}^C)^{(\delta+|C|-1)/2} \times \left[\det\{\mathbf{\Phi}^C + S^C\}\right]^{-(\delta+n+|C|-1)/2},$$

where

$$\Gamma_p(\lambda) \propto \Gamma(\lambda)\Gamma\left(\lambda - \tfrac{1}{2}\right) \cdots \Gamma\left(\lambda - \tfrac{1}{2}p + \tfrac{1}{2}\right).$$

The density of the overall marginal Markov distribution for $X^{(n)}$ is thus given by using the above expressions in conjunction with (6). We call this the *hyper matrix t* distribution $\mathscr{HT}(\delta; I_n, \mathbf{\Phi}^{\mathscr{C}})$, or $\mathscr{HT}_{\mathscr{G}}(\delta; I_n, \mathbf{\Phi})$.

In the case of a single observation, expressed as a column-vector x, expression (45) becomes

$$(47) \qquad \frac{(\det \mathbf{\Phi}^C)^{-1/2}\Gamma\left(\frac{1}{2}(\delta + |C|)\right)}{\pi^{1/2|C|}\Gamma\left(\frac{1}{2}(\delta)\right)}\left[1 + x'(\mathbf{\Phi}^C)^{-1}x\right]^{-\frac{1}{2}(\delta+|C|)}.$$

Correspondingly the incremental factor appended to expression (45) on observing $X^{n+1} = x$, after already having observed $X^{(n)}$, is obtained from (47) by substituting $\delta + n$ for δ and $\mathbf{\Phi} + S$ for $\mathbf{\Phi}$.

We can again use expression (45) or (46) in formula (37) for the marginal likelihood ratio comparison $\Lambda(\mathscr{G}^* : \mathscr{G})$, obtaining incremental components, if desired, using (47) as above. Alternatively, we can use formula (38) calculating, for example, $p_{u \mid C_v}(x_u^{(n)} \mid x_{C_v}^{(n)})$ as the expectation under the prior law of the sampling density of $x_u^{(n)}$ given $x_{C_v}^{(n)}$ using the formulae of Lemma 7.4. We find that in the marginal data distribution, the conditional distribution of $X_u^{(n)}$ given $X_D^{(n)}$, where $D = C_v$ or C_0, is

$$X_D^{(n)}\mathbf{\Phi}_{DD}^{-1}\mathbf{\Phi}_{Du} + T\left(\delta + p; I_n + X_D^{(n)}\mathbf{\Phi}_{DD}^{-1}X_D^{(n)\prime}, \mathbf{\Phi}_{u \mid D}\right),$$

where $p = |D|$. Using the appropriate variant of (45) now yields the alternative expression for $\Lambda(\mathscr{G}^* : \mathscr{G})$. We omit the details.

APPENDIX

A. Graph theory.

A.1. Notation and terminology. A *graph* is a pair $\mathscr{G} = (V, E)$, where V is a finite set of *vertices* and the set of *edges* E is a subset of the set $V \times V$ of ordered pairs of distinct vertices. Thus our graphs have no multiple edges and no loops.

Edges $(\alpha, \beta) \in E$ with both (α, β) and (β, α) in E are called *undirected*, whereas an edge (α, β) with its *opposite* (β, α) not contained in E is called *directed*. If the graph has only undirected edges, it is an *undirected* graph and if all edges are directed, the graph is said to be *directed*.

For both the directed and the undirected cases, we define a *path* of length $n \geq 0$ from α to β to be a sequence $\alpha = \alpha_0, \ldots, \alpha_n = \beta$ of distinct vertices such that $(\alpha_{i-1}, \alpha_i) \in E$ for all $i = 1, \ldots, n$. If there is a path from α to β we say that α *leads to* β and write $\alpha \mapsto \beta$.

A subset $C \subseteq V$ is said to be an (α, β) *separator* if all paths from α to β intersect C. The subset C is said to *separate* A *from* B if it is an (α, β) separator for every $\alpha \in A$, $\beta \in B$.

An *n cycle* is a path of length n with the modification that $\alpha = \beta$, that is, it begins and ends in the same point. A *directed acyclic graph* (DAG) is a directed graph without cycles.

For an undirected graph, we use the notation $\alpha \sim \beta$ to denote that there is an edge between α and β. In this case α and β are said to be *adjacent* or *neighbours*. Similarly $\alpha \nsim \beta$ denotes that this is not the case. The *boundary* $\mathrm{bd}(A)$ of a subset A of V is the set of vertices in $V \setminus A$ that are neighbours of vertices in A. The closure of A is $\mathrm{cl}(A) = A \cup \mathrm{bd}(A)$.

In a directed graph $\alpha \to \beta$ denotes the presence of an edge (α, β), that is, from α to β and $\alpha \not\to \beta$ the absence of such an edge. If $\alpha \to \beta$, α is said to be a *parent* of β and β a *child* of α. The set of parents of β is denoted by $\mathrm{pa}(\beta)$. The expression $\mathrm{pa}(A)$ denotes the set of parents of vertices in A that are not themselves elements of A: $\mathrm{pa}(A) = \bigcup_{\alpha \in A} \mathrm{pa}(\alpha) \setminus A$. The set of vertices α such that $\alpha \mapsto \beta$ are the *ancestors* $\mathrm{an}(\beta)$ of β and the *descendants* $\mathrm{de}(\alpha)$ of α are the vertices β such that $\alpha \mapsto \beta$. The *nondescendants* of α are $\mathrm{nd}(\alpha) = V \setminus (\mathrm{de}(\alpha) \cup \{\alpha\})$. A subset A is *ancestral* if $\mathrm{an}(\beta) \subseteq A$ for all $\beta \in A$. A vertex is *terminal* if it has no children.

The *undirected version* $\mathscr{G} = \mathscr{D}^{\sim}$ of a directed graph $\mathscr{D}$ is the graph obtained from $\mathscr{D}$ by substituting undirected edges for directed ones. Similarly we say that $\mathscr{D}$ is a *directed version* of $\mathscr{G}$.

If a numbering of the vertex set V of an undirected graph $\mathscr{G}$ is given, the *corresponding directed version* $\mathscr{D} = \mathscr{G}^{<}$ has edges between the same vertices α and β as the original graph $\mathscr{G}$, but the edge is directed from vertices with low numbers to those with high numbers. Clearly $\mathscr{G}^{<}$ is then a DAG.

If $A \subseteq V$ is a subset of the vertex set of a graph $\mathscr{G}$, it induces a subgraph $\mathscr{G}_A = (A, E_A)$, where the edge set $E_A = E \cap A \times A$ is obtained from $\mathscr{G}$ by keeping edges with both endpoints in A.

A graph is *complete* if all vertices are joined. A subset is *complete* if it induces a complete subgraph. A complete subset that is maximal w.r.t. $\subset$ is called a *clique*.

A.2. *Decompositions of graphs.* Here we describe some basic features of decompositions and decomposable graphs. The notions pertain to *undirected* graphs. We refer to Lauritzen, Speed and Vijayan (1984), Golumbic (1980) and Lauritzen (1989) for further references. The central notion is the following:

DEFINITION A.1. A pair (A, B) of subsets of the vertex set V of an undirected graph $\mathscr{G}$ is said to form a *decomposition* of $\mathscr{G}$ if $V = A \cup B$, $A \cap B$ is complete and $A \cap B$ separates A from B.

When this is the case we say that (A, B) *decomposes* $\mathscr{G}$ into the *components* $\mathscr{G}_A$ and $\mathscr{G}_B$. If the sets A and B in (A, B) are both genuine subsets of V, the decomposition is *proper*. A decomposable graph is one that can be successively decomposed into its cliques. This is stated formally as:

DEFINITION A.2. An undirected graph is said to be *decomposable* if it is complete, or if there exists a proper decomposition (A, B) into decomposable subgraphs $\mathscr{G}_A$ and $\mathscr{G}_B$.

The definition makes sense because the decomposition is assumed to be proper, so that both subgraphs $\mathscr{G}_A$ and $\mathscr{G}_B$ have fewer vertices than the original graph $\mathscr{G}$.

A *triangulated* graph is an undirected graph with the property that every cycle of length $n \geq 4$ possesses a *chord*, that is, two nonconsecutive vertices that are neighbours. A classical result, see, for example, Golumbic (1980), states the following:

PROPOSITION A.3. *An undirected graph is decomposable if and only if it is triangulated.*

Closely related to the notion of a decomposition is the notion of a *simplicial subset*, which is a subset A that has complete boundary. When a subset is simplicial the pair $(\mathrm{cl}(A), V \setminus A)$ is a decomposition of $\mathscr{G}$. A vertex α is said to be simplicial if the subset $\{\alpha\}$ is.

A sequence $(C_1, \ldots, C_k)$ of complete sets in $\mathscr{G}$ such that for all $j > 1$, R_j is simplicial in $\mathscr{G}_{H_j}$, where

$$H_j = (C_1 \cup \cdots \cup C_j), \qquad R_j = C_j \setminus H_{j-1},$$

is said to be *perfect*. H_j are the *histories* and R_j the *residuals* of the sequence. A *perfect numbering* of the vertices V of $\mathscr{G}$ is a numbering $(\alpha_1, \ldots, \alpha_k)$ such that

$$\mathrm{bd}(\alpha_j) \cap \{\alpha_1, \ldots, \alpha_{j-1}\}, \qquad j > 1$$

are complete sets, that is, such that $(\{\alpha_1\}, \ldots, \{\alpha_k\})$ is a perfect sequence of sets. A directed graph $\mathscr{D}$ is *perfect* if $\mathrm{pa}(\alpha)$ is a complete set for all $\alpha \in V$.

If the sets $(C_1, \ldots, C_k)$ form a perfect sequence and the vertices $(\alpha_1, \ldots, \alpha_{|V|})$ are numbered with those in R_1 first, then those in R_2 and so on, then the vertex numbering so obtained will be perfect [Leimer (1989)]. The numbering then induces a perfect directed version $\mathscr{D}$ of $\mathscr{G}$.

The connection between perfect sequences, numberings and decomposable graphs is contained in the following.

PROPOSITION A.4. *The following conditions are equivalent for an undirected graph* $\mathscr{G}$:

(i) *The graph* $\mathscr{G}$ *admits a perfect directed version* $\mathscr{D}$.
(ii) *The cliques of* $\mathscr{G}$ *admit a perfect numbering.*
(iii) *The graph* $\mathscr{G}$ *is decomposable.*

PROOF. See Golumbic (1980).

Note that a perfect numbering of the cliques can be chosen to have any clique as C_1.

A property that turns out to be extremely important in the statistical context is that of collapsibility [Asmussen and Edwards (1983)]:

DEFINITION A.5. An undirected graph $\mathcal{G}$ is *collapsible* onto A if every connected component B_i of $V \setminus A$ has complete boundary in $\mathcal{G}$.

Theorem 2.3 of Asmussen and Edwards (1983) shows that this definition is equivalent to the one they originally gave. When $\mathcal{G}$ is collapsible onto A, B_i are all simplicial in $\mathcal{G}$, so that the pairs $(\mathrm{cl}(B_i), V \setminus B_i)$ form a decomposition of $\mathcal{G}$ for all i.

LEMMA A.6. *The graph $\mathcal{G}$ is collapsible onto A if and only if, for any triple (A_1, A_2, S) of subsets of A,*

$$S \text{ separates } A_1 \text{ and } A_2 \text{ in } \mathcal{G}_A \quad \Rightarrow \quad S \text{ separates } A_1 \text{ and } A_2 \text{ in } \mathcal{G}.$$

PROOF. See the proof of Corollary 2.5 in Asmussen and Edwards (1983).

LEMMA A.7. *If $(C_1, \ldots, C_k)$ is a perfect sequence of sets with $C_1 \cup \cdots \cup C_k = V$, then $\mathcal{G}$ is collapsible onto the history H_i for each i.*

PROOF. This is Theorem 3.3 of Asmussen and Edwards (1983).

If we let S_j denote the *separators* $S_j = H_{j-1} \cap C_j$, we further have the following:

COROLLARY A.8. *For every i, S_i separates R_i from H_{i-1} in $\mathcal{G}$.*

PROOF. The result follows from Lemmas A.6 and A.7. □

The next two lemmas are concerned with edge removals in decomposable graphs.

LEMMA A.9. *Let $\mathcal{G}^*$ and $\mathcal{G}$ both be decomposable with the same vertex set and with $E \subset E^*$, and with $\mathcal{G}^*$ having exactly one more edge than $\mathcal{G}$. Then this edge is contained in exactly one clique c^* of $\mathcal{G}^*$*

PROOF. This is Lemma 3 of Frydenberg and Lauritzen (1989).

LEMMA A.10. *Let $\mathcal{G}^*$ and $\mathcal{G}$ both be decomposable with the same vertex set and with $E \subset E^*$. Then there is an increasing sequence $\mathcal{G} = \mathcal{G}_0 \subset \cdots \subset \mathcal{G}_k = \mathcal{G}^*$ of decomposable graphs that differ by exactly one edge.*

PROOF. This is Lemma 5 of Frydenberg and Lauritzen (1989).
And finally we will need:

LEMMA A.11. *For $S \subseteq V$, let $\mathcal{G}[S]$ be the graph with vertex set V and with an edge joining vertices α and β if and only if either $\alpha \in S$ or $\beta \in S$, or S does not separate α from β in $\mathcal{G}$. Then:*

(i) *if $\mathcal{G}$ is connected then S is complete in $\mathcal{G}[S]$;*
(ii) *if S separates A from B in $\mathcal{G}$, the same holds in $\mathcal{G}[S]$;*
(iii) *$\mathcal{G}[S]$ is decomposable.*

PROOF. Suppose first that $\mathcal{G}$ is connected. Then clearly S is complete in $\mathcal{G}[S]$. If there exists a path π in $\mathcal{G}[S]$ from A to B avoiding S, then any two consecutive vertices of π can be joined by a path in $\mathcal{G}$ avoiding S and so S does not separate A from B in $\mathcal{G}$. To see that (iii) holds, consider a cycle κ in $\mathcal{G}[S]$ of length greater than or equal to 4, and α and β nonconsecutive vertices of κ. If α and β are nonadjacent in $\mathcal{G}[S]$, then neither is in S and S separates α from β in $\mathcal{G}$ and, thus, by (i) in $\mathcal{G}[S]$. Then κ provides two distinct paths in $\mathcal{G}[S]$ from α to β, each of which intersects S at distinct, nonconsecutive vertices. Since S is complete in $\mathcal{G}[S]$, these vertices are adjacent in $\mathcal{G}[S]$ providing a chord for κ.

In the disconnected case, (ii) and (iii) follow easily from the same properties already shown to hold within each connected component. □

B. The Markov property. Formal proofs. This section contains some of the formal resutls and proofs behind the developments in our paper. We have chosen to formulate the results as pertaining to the notion of Markov distributions, whereas they really are of a quite abstract nature, so that we can use them in other sections of the main paper. We also here use the notation $A \perp\!\!\!\perp B \mid C$ to denote $X_A \perp\!\!\!\perp X_B \mid X_C$.

From the definition and Properties 1–4 of conditional independence, we obtain the following:

PROPOSITION B.1. *If P is Markov over $\mathcal{G}$, then*

$$A \perp\!\!\!\perp B \mid A \cap B$$

whenever $A \cap B$ is complete and separates A from B.

PROOF. Each of A and B can be extended to $\tilde{A}$ and $\tilde{B}$ such that these form a full decomposition with $\tilde{A} \cap \tilde{B} = A \cap B$. Then Property 2 applies. □

PROPOSITION B.2. *Let $\mathcal{D}$ be a perfect directed acyclic graph and $\mathcal{G}$ its undirected version. Then, if*

$$(\{v\} \cup \mathrm{pa}(v)) \perp\!\!\!\perp \mathrm{nd}(v) \mid \mathrm{pa}(v) \ [P] \tag{48}$$

holds, P is Markov over $\mathcal{G}$.

PROOF. For $A \cup B = V$, write $x \simeq (x_A, x_B)$ to denote the (obvious) fact that any point $x \in \mathscr{X}$ is determined by its projections x_A and x_B. We show the result by induction on $|V|$, the number of vertices of $\mathscr{G}$. For $|V| = 1$ the statement is trivial. Suppose this implication has been established for $|V| \leq n$ and that we are dealing with the case $|V| = n + 1$. Let λ be a terminal vertex in $\mathscr{D}$. Then, since $\mathscr{D}$ is perfect, $\{\lambda\} \cup \mathrm{pa}(\lambda)$ is complete. Hence, if (A, B) is a decomposition of $\mathscr{G}$ we may without loss of generality suppose $\{\lambda\} \cup \mathrm{pa}(\lambda) \subseteq A$, whence in particular $\lambda \in A$. Let $V^* = V \setminus \{\lambda\}$, $\mathscr{G}^* = \mathscr{G}_{V^*}$, $A^* = A \cap V^*$, $B^* = B \cap V^*$, $S = A \cap B$ and $S^* = S \cap V^*$. Then (A^*, B^*) is a decomposition of $\mathscr{G}^*$ with $A^* \cap B^* = S^*$. Hence, by the induction hypothesis,

$$A^* \perp\!\!\!\perp B^* \mid S^*. \tag{49}$$

Also, from (48), since $\mathrm{pa}(\lambda) \subseteq A^* \subseteq V^* = \mathrm{nd}(\lambda)$, we obtain

$$(\{\lambda\} \cup \mathrm{pa}(\lambda)) \perp\!\!\!\perp V^* \mid A^*,$$

whence

$$(\{\lambda\} \cup \mathrm{pa}(\lambda)) \perp\!\!\!\perp B^* \mid A^*. \tag{50}$$

Since $X_A \simeq (X_{\{\lambda\} \cup \mathrm{pa}(\lambda)}, X_{A^*})$, (49) and (50) are together equivalent to

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

whence, since $S^* \subseteq S \subseteq A$,

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

If $\lambda \notin B$ we have $B^* = B$. Otherwise $X_B \sim (X_{B^*}, X_S)$. In either case we deduce $A \perp\!\!\!\perp B \mid S$ and the induction is established. □

PROPOSITION B.3. *If $\mathscr{G}^*$ is a decomposable graph with the same vertex set as $\mathscr{G}$, but with larger edge set, that is, $E \subset E^*$, then any distribution P which is Markov over $\mathscr{G}$ is also Markov over $\mathscr{G}^*$.*

PROOF. By Lemma A.10 it is enough to consider the case where $\mathscr{G}$ and $\mathscr{G}^*$ differ by exactly one edge, $\{\alpha, \beta\}$, say, in which case Lemma A.9 gives that this edge is a member of exactly one clique C_1 of $\mathscr{G}^*$. Now make a perfect numbering of the cliques of $\mathscr{G}^*$ beginning with this particular clique and let $\mathscr{D}^*$ be a corresponding perfect directed version of $\mathscr{G}^*$. Because $\{\alpha, \beta\} \subseteq C_1$ only, it is not a subset of any of the separator sets S_i. Since S_i is complete and separates R_i from H_i in $\mathscr{G}^*$ for all i, the same holds in $\mathscr{G}$. The Markov property over $\mathscr{G}$ together with Proposition B.1 now gives that P must be directed Markov over $\mathscr{D}^*$, as in the proof of Theorem 2.6. From Proposition B.2 we obtain that P is Markov over $\mathscr{G}^*$. □

COROLLARY B.4. *Theorem 2.8 holds.*

PROOF. Suppose $S = A \cap B$ separates A from B and let $\mathscr{G}^* = \mathscr{G}[S]$ as defined in Lemma A.11. Then, by this lemma, $\mathscr{G}^*$ is decomposable and S separates A from B in $\mathscr{G}^*$. Thus the theorem follows from Proposition B.3. □

The process of forming marginal distributions will, under certain circumstances, preserve the Markov property. More precisely, we have the following:

PROPOSITION B.5. *If $\mathscr{G}$ is collapsible onto A and P is Markov over $\mathscr{G}$, then P_A is Markov over $\mathscr{G}_A$.*

PROOF. This follows immediately from Lemma A.6 and Proposition B.1. See also the definition of collapsibility given by Asmussen and Edwards (1983). □

The result in Proposition B.5 is strongest possible—if $\mathscr{G}$ is not collapsible onto A then there exists a Markov distribution P over $\mathscr{G}$ for which P_A is not Markov over $\mathscr{G}_A$ [Frydenberg (1990)].

A consequence of Proposition B.5 is the converse to Proposition B.2, relating the directed and undirected Markov properties.

PROPOSITION B.6. *Suppose P is Markov over $\mathscr{G}$ and let $\mathscr{D}$ be a perfect directed version of $\mathscr{G}$. Then P is directed Markov over $\mathscr{D}$.*

PROOF. First realize that a decomposable undirected graph $\mathscr{G}$ is collapsible onto a subset A if and only if A is ancestral in some perfect directed version $\mathscr{D}$ of $\mathscr{G}$. In particular, $\mathscr{G}$ is collapsible onto $G_v^- = \{v\} \cup \mathrm{nd}(v)$ for any $v \in V$. Hence, if P is Markov over $\mathscr{G}$, then by Proposition B.5, $P_{G_v^-}$ is Markov over $\mathscr{G}_{G_v^-}$. Since $(\{v\} \cup \mathrm{pa}(v), \mathrm{nd}(v))$ is a decomposition of $\mathscr{G}_{G_v^-}$, (1) follows. □

Another consequence is the following alternative recursive characterization of the Markov property.

PROPOSITION B.7. *Let (A, B) be a decomposition of $\mathscr{G}$. Then P is Markov over $\mathscr{G}$ if and only if*:

(i) *P_A is Markov over $\mathscr{G}_A$,*
(ii) *P_B is Markov over $\mathscr{G}_B$ and*
(iii) *$A \perp\!\!\!\perp B \mid A \cap B\ [P]$.*

PROOF. If P is Markov over $\mathscr{G}$, then (i) and (ii) follow from Proposition B.5, while (iii) holds by definition. Conversely, suppose (i)–(iii) hold. We can construct a perfect directed version $\mathscr{D}_A$ of $\mathscr{G}_A$, and then extend the restriction of $\mathscr{D}_A$ to $A \cap B$ to a perfect directed version $\mathscr{D}_B$ of $\mathscr{G}_B$. Then $\mathscr{D} = \mathscr{D}_A \cup \mathscr{D}_B$ will be a perfect directed version of $\mathscr{G}$. Since, by Proposition B.6, P_A and P_B are directed Markov with respect to $\mathscr{D}_A$ and $\mathscr{D}_B$, it easily follows from (iii) that P is directed Markov over $\mathscr{D}$ and hence by Proposition B.2, Markov over $\mathscr{G}$. That $P = P_A \star P_B$ follows trivially. □

Acknowledgments. This work was supported by the Complex Stochastic Systems Initiative of the Science and Engineering Research Council. We are grateful to David Spiegelhalter for many helpful comments. We should also

like to acknowledge the stimulus to this work given by participants at the joint Royal Statistical Society/SERC 1989 Edinburgh Workshop on Expert Systems and Statistics. In particular we benefited greatly from Glenn Shafer's insights into the level of generality attainable in this kind of analysis.

REFERENCES

ASMUSSEN, S. and EDWARDS, D. (1983). Collapsibility and response variables in contingency tables. *Biometrika* **70** 567–78.

BARNDORFF-NIELSEN, O. E. (1978). *Information and Exponential Families in Statistical Theory*. Wiley, New York.

BASU, D. (1955). On statistics independent of a complete sufficient statistic. *Sankhyā* **15** 377–380.

DARROCH, J. N., LAURITZEN, S. L. and SPEED, T. P. (1980). Markov fields and log linear models for contingency tables. *Ann. Statist.* **8** 522–539.

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

DAWID, A. P. (1979b). Some misleading arguments involving conditional independence. *J. Roy. Statist. Soc. Ser. B* **41** 249–252.

DAWID, A. P. (1980). Conditional independence for statistical operations. *Ann. Statist.* **8** 598–617.

DAWID, A. P. (1981). Some matrix-variate distribution theory: Notational considerations and a Bayesian application. *Biometrika* **68** 265–274.

DAWID, A. P. (1992). Prequential analysis, stochastic complexity and Bayesian inference (with discussion). In *Bayesian Statistics* **4** (J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith, eds.) 109–125. Clarendon Press, Oxford.

DEMPSTER, A. P. (1972). Covariance selection. *Biometrics* **28** 157–175.

DICKEY, J. M. (1967). Matrixvariate generalizations of the multivariate t distribution and the inverted multivariate t distribution. *Ann. Math. Statist.* **38** 511–518.

EDWARDS, D. (1990). Hierarchical interaction models (with discussion). *J. Roy. Statist. Soc. Ser. B* **52** 3–20 and 51–72.

EDWARDS, D. and KREINER, S. (1983). The analysis of contingency tables by graphical models. *Biometrika* **70** 553–562.

FRYDENBERG, M. (1990). Marginalization and collapsibility in graphical interaction models. *Ann. Statist.* **18** 790–805.

FRYDENBERG, M. and LAURITZEN, S. L. (1989). Decomposition of maximum likelihood in mixed interaction models. *Biometrika* **76** 539–555.

GOLUMBIC, M. C. (1980). *Algorithmic Graph Theory and Perfect Graphs*. Academic, London.

JOHNSON, N. L. and KOTZ, S. (1972). *Distributions in Statistics. Continuous Multivariate Distributions*. Wiley, New York.

KIIVERI, H., SPEED, T. P. and CARLIN, J. B. (1984). Recursive causal models. *J. Austral. Math. Soc. Ser. A* **36** 30–52.

LAURITZEN, S. L. (1989). Mixed graphical association models (with discussion). *Scand. J. Statist.* **16** 273–306.

LAURITZEN, S. L., DAWID, A. P., LARSEN, B. N. and LEIMER, H.-G. (1990). Independence properties of directed Markov fields. *Networks* **20** 491–505.

LAURITZEN, S. L., SPEED, T. P. and VIJAYAN, K. (1984). Decomposable graphs and hypergraphs. *J. Austral. Math. Soc. Ser. A* **36** 12–29.

LAURITZEN, S. L. and SPIEGELHALTER, D. J. (1988). Local computations with probabilities on graphical structures and their application to expert systems (with discussion). *J. Roy. Statist. Soc. Ser. B* **50** 157–224.

LEHMANN, E. L. and SCHEFFÉ, H. (1950). Completeness, similar regions and unbiased estimation. Part I. *Sankhyā* **10** 305–340.

LEIMER, H.-G. (1989). Triangulated graphs with marked vertices. In *Graph Theory in Memory of G. A. Dirac* (L. D. Andersen, C. Thomassen, B. Toft and P. D. Vestergaard, eds.) 311—324. North-Holland, Amsterdam. *Ann. Discrete Math.* **41**.

PEARL, J. (1988). *Probabilistic Inference in Intelligent Systems*. Morgan Kaufmann, San Mateo, CA.

SMITH, J. Q. (1989). Influence diagrams for statistical modelling. *Ann. Statist.* **17** 654–672.

SPEED, T. P. (1979). A note on nearest-neighbour Gibbs and Markov probabilities. *Sankhyā Ser. A* **41** 184–197.

SPIEGELHALTER, D. J., DAWID, A. P., LAURITZEN, S. L. and COWELL, R. G. (1993). Bayesian analysis in expert systems with discussion. *Statist. Sci.* **8** 219–283.

SPIEGELHALTER, D. J. and LAURITZEN, S. L. (1990). Sequential updating of conditional probabilities on directed graphical structures. *Networks* **20** 579–605.

SUNDBERG, R. (1975). Some results about decomposable (or Markov-type) models for multidimensional contingency tables: Distribution of marginals and partitioning of tests. *Scand. J. Statist.* **2** 71–79.

WERMUTH, N. (1976). Analogies between multiplicative models in contingency tables and covariance selection. *Biometrics* **32** 95–108.

WERMUTH, N. (1980). Linear recursive equations, covariance selection and path analysis. *J. Amer. Statist. Assoc.* **75** 963–972.

WERMUTH, N. and LAURITZEN, S. L. (1983). Graphical and recursive models for contingency tables. *Biometrika* **70** 537–552.

WERMUTH, N. and LAURITZEN, S. L. (1990). On substantive research hypotheses, conditional independence graphs and graphical chain models (with discussion). *J. Roy. Statist. Soc. Ser. B* **52** 21–72.

WHITTAKER, J. (1990). *Graphical Models in Applied Multivariate Statistics*. Wiley, Chichester.

WILKS, S. S. (1962). *Mathematical Statistics*. Wiley, New York.

DEPARTMENT OF STATISTICAL SCIENCE
UNIVERSITY COLLEGE LONDON
GOWER STREET
LONDON WC1E 6BT
ENGLAND

DEPARTMENT OF MATHEMATICS
AND COMPUTER SCIENCE
AALBORG UNIVERSITY
FREDRIK BAJERS VEJ 7
DK-9220 AALBORG
DENMARK