Learning Essential Graph Markov Models from Data

Robert Castelo[1] and Michael D. Perlman[2]

[1] Grup de Recerca en Informàtica Biomèdica, Departament de Ciències
Experimentals i de la Salut, Universitat Pompeu Fabra,
Psg. Marítim 37–49, 08003 Barcelona, Spain
robert.castelo@cexs.upf.es

[2] Department of Statistics, University of Washington,
Box 354322, Seattle WA 98195–4322, USA
michael@ms.washington.edu

Abstract. In a model selection procedure where many models are to be compared, computational efficiency is critical. For acyclic digraph (ADG) Markov models (aka DAG models or Bayesian networks), each ADG Markov equivalence class can be represented by a unique chain graph, called an *essential graph (EG)*. This parsimonious representation might be used to facilitate selection among ADG models. Because EGs combine features of decomposable graphs and ADGs, a scoring metric can be developed for EGs with categorical (multinomial) data. This metric may permit the characterization of local computations directly for EGs, which in turn would yield a learning procedure that does not require transformation to representative ADGs at each step for scoring purposes, nor is the scoring metric constrained by Markov equivalence.

1 Introduction

Model selection, or statistical learning, among ADG models is constrained due to rapid growth in the number of ADGs as the number of nodes increases, so computational efficiency becomes critical. To this end, Dawid and Lauritzen (1993, eqn. (37)) show that the Bayes factor (the ratio of the two marginal likelihoods, cf. Cowell et al. (1999, pg. 250)) between two decomposable (DEC) graphical Markov models that differ in one single adjacency can be expressed in terms of at most four cliques from both junction trees, thereby facilitating selection among such models. Analogously, for the case of acyclic digraph Markov models, it suffices to compare the local scores for the variable whose parent set changes.

From a non-causal perspective we are interested in learning Markov equivalence classes of ADGs. Andersson et al. (1997) showed that each (possibly large) ADG Markov equivalence class can be represented by a *single* unique chain graph, called an *essential graph (EG)*, which combines features of decomposable graphs and ADGs. In this study we utilize this economical representation to facilitate selection among non-causal ADG models by directly scoring EG models. In particular, a scoring metric can be developed for EGs with categorical (multinomial) data. This metric may permit the characterization of local computations directly for EGs, which in turn would yield a learning procedure that does not require transformation to representative ADGs at each step for scoring purposes.

This characterization may depend, for instance, on a particular class of *local transformations* for EGs proposed by Perlman (2001) that require at most two edge changes in the EG or an accompanying ADG, whose local scoring functions can be obtained. This yields a learning procedure that does not require transformation to representative ADGs at each step for scoring purposes. Furthermore, exploiting the recent works by Chickering (2002b) and Auvray and Wehenkel (2002) may lead to a characterization of local computations for our scoring metric that respects the graphical Markov model inclusion order - cf. (Castelo and Kočka, 2002, Section 3).

After reviewing terminology in Section 2, in Section 3 we present the factorization for pdfs that are Markov with respect to an EG G. The scoring metric for EGs with categorical data is given in Section 4, and related to those for decomposable graphs and ADGs in Section 5. The local transformations and corresponding local computations for EGs are discussed in Section 6, while concluding remarks are in Section 7.

2 Background Concepts

A random variable is denoted by an upper case letter indexed by a number or a lowercase letter, e.g. X_1 or X_i. A set or a vector of random variables is denoted by an upper case letter, e.g. X, which may be indexed by an uppercase letter, e.g. X_A. For notational convenience, we sometimes abbreviate X_i by i, $X_A \equiv \{X_i | i \in A\}$ by A, etc. All random variables in this article will be discrete.

A graph G is a pair (V, E) where V is a set of vertices and E is a set of edges. When every edge in E is undirected, G is an *undirected graph* (UG). When every edge in E is directed, G is a *directed graph.* If a directed graph G has no directed cycles, then G is an *acyclic digraph* (ADG).

Let G be a graph with vertex set V. The collection of random variables $X_V \equiv \{X_v | v \in V\}$ arises as a categorical dataset and follows a multinomial distribution P defined on a product space $\mathcal{X} = \times(\mathcal{X}_i | i \in V)$. We will use the term *level* to refer to a particular member $x \in \mathcal{X}$. In the present context, a graphical Markov model (GMM), denoted by $\mathbf{M}(G)$, is a family $\{P_\theta\}$ of multinomial distributions on $\mathcal{X}$ that satisfy the Markov properties determined by the graph G. In the Bayesian formulation, the parameter θ is itself random and follows a prior distribution, or *law*, usually denoted by π. It is common to assume that π is Dirichlet, the natural conjugate prior for the multinomial family, under which the posterior distribution of θ remains Dirichlet.

We shall use standard graph-theoretic terminology, such as *pa* for *parents*, *nd* for *nondescendants*, *bd* for *boundary*, and *nb* for *neighbors*. Refer to Lauritzen (1996) for standard GMM results and terminology.

A *chain graph* $G = (E, V)$ is a graph that may have both directed and undirected edges but is *adicyclic*, that is, has no fully or partially directed cycles. Let $\mathcal{T} \equiv \mathcal{T}(G)$ denote the set of *chain components* of G, i.e., the connected components of the graph obtained by removing all arrows from G. Let $\mathcal{D} \equiv \mathcal{D}(G)$ be the acyclic digraph with vertex set $\mathcal{T}$ and where, for $\tau_1, \tau_2 \in \mathcal{T}$, $\tau_1 \to \tau_2 \in \mathcal{D}$ iff $t_1 \to t_2$ for at least one pair $t_1 \in \tau_1$, $t_2 \in \tau_2$. (By adicyclicity, this is well defined.)

Let G be an undirected chordal (decomposable) graph with a set of cliques $\mathcal{C}$ and a set of separators $\mathcal{S}$. Note that some $S \in \mathcal{S}$ may occur in $\mathcal{S}$ more than once because it may repeat in the perfect numbering sequence of the cliques - cf. (Dawid and Lauritzen, 1993, pg. 1278, 1311). Dawid and Lauritzen (1993, Theorem 2.6) showed that the unique Markov distribution over G having a specified consistent familiy $\{f_C \mid C \in \mathcal{C}\}$ of pdfs for its clique marginals, is given by the pdf

$$f(x) = \frac{\prod_{C\in\mathcal{C}} f_C(x_C)}{\prod_{S\in\mathcal{S}} f_S(x_S)} \;, \tag{1}$$

where the separators set $\mathcal{S}$ incorporates $\nu(S)$ repetitions of $S \in \mathcal{S}$ that may occur in any given perfect numbering of the cliques $C \in \mathcal{C}$.

3 Factorization of a Multivariate Distribution According to an Essential Graph

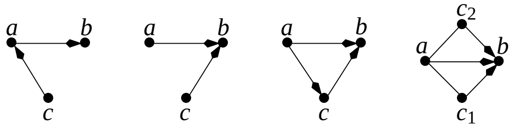
Let D be an ADG with vertex set V and let $[D]$ denote its Markov equivalence class, that is, the set of all ADGs D' over V that determine the same Markov properties as D. Verma and Pearl (1990) showed that $D' \in [D]$ if and only if D and D' have the same *skeleton* ($\equiv$ underlying undirect graph) and *immoralities* ($\equiv$ induced subgraphs of the form $a \to b \leftarrow c$). Andersson et al. (1997) showed that the entire Markov equivalence class $[D]$ is uniquely determined by the *essential graph* (EG) D^* defined as follows:

$$D^* = \cup\{D' \mid D' \in [D]\} \;.$$

Here the union is obtained by the rule that for any pair of vertices (a, b), the union of directed edges between a and b of like orientation is the common directed edge, while the union of directed edges between a and b with at least one opposite orientation is an undirected edge. Furthermore, they characterized those graphs that may occur as essential graphs by the following theorem:

Theorem 01 *A graph $G = (V, E)$ is the essential graph D^* for some ADG D with vertex set V iff G satisfies the following four conditions:*

(i) *G is a chain graph.*
(ii) *For each chain component $\tau \in \mathcal{T}(G)$, the UG G_τ is chordal.*
(iii) *G has no flags, i.e., no induced subgraphs of the form $a \to b - c$.*
(iv) *Each arrow $a \to b$ in G is "strongly protected", that is, occurs in at least one of the following four configurations as an induced subgraph of G:*



(a) *(b)* *(c)* *(d)*

Andersson et al. (2001) provide the following definitions for the LWF and AMP block-recursive Markov properties for general chain graphs, hence in particular for essential graphs.

Definition 01 *A probability measure P on a product probability space $\mathcal{X} \equiv \times(\mathcal{X}_v | v \in V)$, where V indexes a vector of random variables $X = (X_1, \ldots, X_n)$, is said to be* LWF (resp., AMP) block-recursive G-Markovian *if P satisfies the following conditions C1, C2, and C3 (resp., C1, C2, and C3*):*

(C1) $\forall \tau \in \mathcal{T} : X_\tau \perp\!\!\!\perp X_{(nd_{\mathcal{D}}(\tau) \setminus pa_{\mathcal{D}}(\tau))} \,|\, X_{pa_{\mathcal{D}}(\tau)}[P]$, *i.e., P is local $\equiv$ global $\mathcal{D}$-Markovian on $\mathcal{X}$.*

(C2) $\forall \tau \in \mathcal{T}$ *: the conditional distribution $P_{\tau | pa_{\mathcal{D}}(\tau)}$ is global G_τ-Markovian on $\mathcal{X}_\tau$.*

(C3) $\forall \tau \in \mathcal{T}$, $\forall \sigma \subseteq \tau : X_\sigma \perp\!\!\!\perp X_{(pa_{\mathcal{D}}(\tau) \setminus pa_G(\sigma))} \,|\, X_{pa_G(\sigma) \cup nb_G(\sigma)}[P]$.

(C3*) $\forall \tau \in \mathcal{T}$, $\forall \sigma \subseteq \tau : X_\sigma \perp\!\!\!\perp X_{(pa_{\mathcal{D}}(\tau) \setminus pa_G(\sigma))} \,|\, X_{pa_G(\sigma)}[P]$.

Note that if the chain graph G is in fact an essential graph, then by condition (iii) and Theorem 4 of Andersson et al. (2001), its AMP and LWF block-recursive Markov properties coincide. It follows from C1 and Lemma 4.1 of Andersson et al. (2001) that if a probability distribution P is LWF or AMP block-recursive G-Markovian for a chain graph G, then its pdf f admits the following recursive factorization:

$$f(x|G, \theta_G) = \prod_{\tau \in \mathcal{T}} f(x_\tau | x_{pa_G(\tau)}, \theta_\tau) \ . \tag{2}$$

(Here, since we shall consider P to be a member of a parametric statistical model, we include θ_τ to denote the parameters occurring in the conditional pdf and set $\theta_G = (\theta_\tau | \tau \in \mathcal{T})$.)

Note that $\theta_\tau = (\theta_{\tau,\rho} | \rho \in \mathcal{X}_{pa_G(\tau)})$, and therefore the parametrization in (2) involves only those parameters $\theta_{\tau,\rho}$ for which $\rho = x_{pa_G(\tau)}$:

$$f(x|G, \theta_G) = \prod_{\tau \in \mathcal{T}} f(x_\tau | x_{pa_G(\tau)}, \theta_{\tau,\rho}) \ . \tag{3}$$

(See the discussion in (Cowell et al., 1999, Section 9.2).)

For the remainder of this paper, G shall denote an essential graph. By (ii), every chain component $\tau \in \mathcal{T}$ induces a chordal $\equiv$ decomposable graph G_τ, so each term $f(x_\tau | x_{pa_G(\tau)}, \theta_{\tau,\rho})$ admits a further factorization according to the cliques and clique separators of G_τ (cf. (4) in Theorem 02 below) as shown by Dawid and Lauritzen (1993), Theorem 2.6. In fact, we now show that this further factorization is sufficient as well as necessary for a pdf to determine a G-Markovian distribution. (No positivity assumptions are needed - cf. (Dawid and Lauritzen, 1993, pg. 1279).)

Let $\mathcal{C}_\tau$ (resp., $\mathcal{S}_\tau$) denote the set of cliques (resp., clique separators) for the decomposable undirected graph G_τ, $\tau \in \mathcal{T} \equiv \mathcal{T}(G)$. Let P be a distribution on $\mathcal{X}$ that admits a pdf f.

Theorem 02 *Let G be an essential graph. The distribution P is AMP $\equiv$ LWF G-Markovian iff f factorizes as*

$$f(x|G, \theta_G) = \prod_{\tau \in \mathcal{T}} \left[\frac{\prod_{C \in \mathcal{C}_\tau} f(x_C | x_{pa_G(\tau)}, \theta_{C,\rho})}{\prod_{S \in \mathcal{S}_\tau} f(x_S | x_{pa_G(\tau)}, \theta_{S,\rho})} \right] . \tag{4}$$

Proof. As already noted above, if P is AMP $\equiv$ LWF block-recursive G-Markovian then f satisfies (4). Conversely, if (4) holds it is straightforward to show that P satisfies C1 and C2 and that

$$f(x_\tau | x_{pa_G(\tau)}, \theta_{\tau,\rho}) = f(x_\tau | x_{pa_{\mathcal{D}}(\tau)} \, , \theta_{\tau,\rho}) \tag{5}$$

for every $\tau \in \mathcal{T}$, so

$$X_\tau \perp\!\!\!\perp X_{(pa_{\mathcal{D}}(\tau) \setminus pa_G(\tau))} \,|\, X_{pa_G(\tau)} \, [P] \, . \tag{6}$$

But if $\emptyset \neq \sigma \subseteq \tau \in \mathcal{T}$ then $pa_G(\sigma) = pa_G(\tau)$ because G is an essential graph and satisfies (iii), hence C3* holds.

Note that, in the previous theorem, $\theta_{C,\rho}$ and $\theta_{S,\rho}$ are subsets of $\theta_{\tau,\rho}$.

4 Bayesian Scoring Metric for Multinomial Data

Let $D = \{x^{(1)}, x^{(2)}, \ldots, x^{(N)}\}$ be a set of N exchangeable observations sampled from a multinomial distribution comprising the set of random variables $X = \{X_1, \ldots, X_n\}$. From a Bayesian perspective, we wish to choose a model, represented by an essential graph G, according to its posterior probability $p(G|D)$. By Bayes' theorem,

$$p(G|D) \propto f(D|G)p(G) \; , \tag{7}$$

where $f(D|G)$ is the integrated likelihood given by

$$f(D|G) = \int_{\theta_G} f(D|G, \theta_G)\pi(\theta_G) d\theta_G \; , \tag{8}$$

with respect to a prior distribution π. The reason for integrating out the parameters θ_G in (8) is that for the purpose of model comparison, we require the likelihood unconditional on any fixed parameter values. The Bayesian scoring metric for a model G given data D is the logarithm of the right-hand side of (7):

$$\mathrm{sc}(G|D) = \log[f(D|G)p(G)] \ . \tag{9}$$

To calculate the integrated likelihood $f(D|G)$ from (8), we use Theorem 02 and make two independence assumptions about the parameters. First, the parameters $(\theta_\tau|\tau \in \mathcal{T})$ are *a priori* independent. This assumption is analogous to the one of global independence formulated by Spiegelhalter and Lauritzen (1990) for ADG Markov models. For that reason, we will refer to this assumption as the *global independence* assumption hereafter. Under this assumption we may write $\pi(\theta_G)$ as a product of the densities, where each of them involves only the parameters regarding one chain component $\tau \in \mathcal{T}$:

$$\pi(\theta_G) = \prod_{\tau \in \mathcal{T}} \pi(\theta_\tau) \ . \tag{10}$$

Second, for each $\tau \in \mathcal{T}$, the parameters $(\theta_{\tau,\rho}|\rho \in \mathcal{X}_{pa_G(\tau)})$ are also *a priori* independent. This assumption is known as *local independence* (Spiegelhalter and Lauritzen, 1990), and permits a factorization of $\pi(\theta_\tau)$ as

$$\pi(\theta_\tau) = \prod_{\rho \in \mathcal{X}_{pa_G(\tau)}} \pi(\theta_{\tau,\rho}) \ . \tag{11}$$

Because G_τ is decomposable for each τ, we can apply the results of Dawid and Lauritzen (1993). We begin by assuming that $\pi(\theta_{\tau,\rho})$, the prior law for $\theta_{\tau,\rho}$, is *strong hyper Markov*, which allows us to factorize it as

$$\pi(\theta_{\tau,\rho}) = \frac{\prod_{C \in \mathcal{C}_\tau} \pi(\theta_{C,\rho})}{\prod_{S \in \mathcal{S}_\tau} \pi(\theta_{S,\rho})} \ . \tag{12}$$

Therefore, we just need to specify prior laws for each clique $C \in \mathcal{C}_\tau$. For $\pi(\theta_{\tau,\rho})$ being strong hyper Markov, we will consider a *hyper Dirichlet* law for each $\theta_{C,\rho}$, denoted by $\mathcal{D}(\theta_{C,\rho};\vartheta_{C,\rho})$, which is specified by a hyperparameter set of positive numbers $\vartheta_{C,\rho} = (N'_{C1},\ldots,N'_{Cq(C)})$ and a Dirichlet distribution with density

$$\pi(\theta_{C,\rho}) = \pi(\theta_{C,\rho}|\vartheta_{C,\rho}) = \frac{\Gamma(\sum_{k=1}^{q(C)} N'_{Ck})}{\prod_{k=1}^{q(C)} \Gamma(N'_{Ck})} \prod_{k=1}^{q(C)} \theta_C^{N'_{Ck}-1} \ , \tag{13}$$

where $q(C)$ is the cardinality of the product space $\mathcal{X}_C$.

The above collection of hyper Dirichlet laws will determine a unique hyper Dirichlet law for $\theta_{\tau,\rho}$, which is strong hyper Markov, provided that the collection $\mathcal{D}(\theta_{C,\rho};\vartheta_{C,\rho}), C \in \mathcal{C}$ is specified such that they are (pairwise) *hyperconsistent*. Two laws, $\mathcal{L}_\mathcal{A}$ for θ_A and $\mathcal{L}_\mathcal{B}$ for θ_B, are hyperconsistent if they both induce the same law for $A \cap B$ (Dawid and Lauritzen, 1993, pg. 1280). In the case of hyper Dirichlet laws, they are hyperconsistent as long as for any two cliques C and D such that $C \cap D \neq \emptyset$ and any given lth level $x_{\{C \cap D\}l} \in \mathcal{X}_{\{C \cap D\}}$, their corresponding hyperparameters ϑ_C and ϑ_D satisfy (Dawid and Lauritzen, 1993, pg. 1304)

$$\sum_{x_{Ck} \supset x_{\{C \cap D\}l}} N'_{Ck} = \sum_{x_{Dk} \supset x_{\{C \cap D\}l}} N'_{Dk} \ . \tag{14}$$

For a saturated multinomial model under a hyper Dirichlet prior law, the marginal probability distribution of a dataset $D_C = \{x_C^{(1)}, x_C^{(2)}, \ldots, x_C^{(N)}\}$, where $x_C^{(m)} \in \mathcal{X}_C$, $m = 1, \ldots, N$, and $\mathcal{X}_C = \times(\mathcal{X}_i | i \in C)$, is given by

$$f_C(D_C | \vartheta_C) = \frac{\Gamma(N'_C)}{\Gamma(N'_C + N)} \prod_{k=1}^{q(C)} \frac{\Gamma(N'_{Ck} + N_{Ck})}{\Gamma(N'_{Ck})} \ , \tag{15}$$

where $N'_C = \sum_{k=1}^{q(C)} N'_{Ck}$ and N_{Ck} are the counts of the contingency table associated with D_C. (Thus $N = \sum_k N_{Ck}$ is the size of the sample.)

The assumption of a strong hyper Markov prior law for θ_G allows us to use properties of such a prior law to proceed from expression (8). First, by (2), (10) and the exchangeability of the records of D, we can further factorize $f(D|G, \theta_G)$ and $\pi(\theta_G)$ to obtain

$$f(D|G) = \int_{\theta_\tau} \cdots \int \prod_{m=1}^{N} \prod_{\tau \in \mathcal{T}} f(x_\tau^{(m)} | x_{pa_G(\tau)}^{(m)}, \theta_\tau) \pi(\theta_\tau) d\theta_\tau \ . \tag{16}$$

By global independence, we can interchange the product and the integral, yielding

$$f(D|G) = \prod_{\tau \in \mathcal{T}} \int_{\theta_\tau} \prod_{m=1}^{N} f(x_\tau^{(m)} | x_{pa_G(\tau)}^{(m)}, \theta_\tau) \pi(\theta_\tau) d\theta_\tau \equiv$$

$$\equiv \prod_{\tau \in \mathcal{T}} f(D_\tau | D_{pa_G(\tau)}) \ . \tag{17}$$

Next, consider the following proposition.

Proposition 01 *(Dawid and Lauritzen, 1993, Prop. 5.6)*
For a decomposable undirected graph, if the prior law of θ is strong hyper Markov, then the marginal distribution of X is Markov.

Now apply this result to the conditional distribution determined by $f(D_\tau | D_{pa_G(\tau)})$. By the assumption of local independence, the integrated distribution $f(D_\tau | D_{pa_G(\tau)})$ of the data is Markov with respect to the graph G:

$$f(D|G) = \prod_{\tau \in \mathcal{T}} \left[\frac{\prod_{C \in \mathcal{C}_\tau} f_C(D_C | D_{pa_G(\tau)})}{\prod_{S \in \mathcal{S}_\tau} f_S(D_S | D_{pa_G(\tau)})} \right] . \tag{18}$$

Since $|\mathcal{C}_\tau| = |\mathcal{S}_\tau| + 1$, this is equivalent to

$$f(D|G) = \prod_{\tau \in \mathcal{T}} \left[\frac{1}{f_{pa_G(\tau)}(D_{pa_G(\tau)})} \cdot \frac{\prod_{C \in \mathcal{C}_\tau} f_{C,pa_G(\tau)}(D_{C,pa_G(\tau)})}{\prod_{S \in \mathcal{S}_\tau} f_{S,pa_G(\tau)}(D_{S,pa_G(\tau)})} \right] , \tag{19}$$

where every $f_A(D_A)$, A being either $pa_G(\tau)$, $\{C, pa_G(\tau)\}$ or $\{S, pa_G(\tau)\}$, is replaced by the marginal probability distribution for a saturated multinomial model, i.e. (15), just as Dawid and Lauritzen (1993) do with their (41) and (6). Since any separator $S \in \mathcal{S}$ is, by definition, included in some clique $C \in \mathcal{C}$ the marginal probability $f_{S,pa(\tau)}$ corresponds to the marginalization of $f_{C,pa(\tau)}$ over $C \backslash S$. The same is true for $f_{pa(\tau)}$ which corresponds to the marginalization of $f_{C,pa(\tau)}$ over C. From this it follows that the term $\Gamma(N'_C)/\Gamma(N'_C + N)$ in (15) will be the same in the marginals $f_{C,pa(\tau)}$, $f_{S,pa(\tau)}$ and $f_{pa(\tau)}$ such that they will cancel in (19). Therefore, we can write the integrated likelihood for an essential graph G as follows:

$$\begin{aligned} f(D|G) = \prod_{\tau \in \mathcal{T}(G)} \Bigg[& \left(\prod_{k=1}^{q(pa(\tau))} \frac{\Gamma(N'_k)}{\Gamma(N'_k + N_k)} \right) \cdot \\ & \cdot \frac{\prod_{C \in \mathcal{C}_\tau} \prod_{k=1}^{q(C \cup pa(\tau))} (\Gamma(N'_k + N_k)/\Gamma(N'_k))}{\prod_{S \in \mathcal{S}_\tau} \prod_{k=1}^{q(S \cup pa(\tau))} (\Gamma(N'_k + N_k)/\Gamma(N'_k))} \Bigg] , \end{aligned} \tag{20}$$

where, for clarity, we have not specified the sets of variables associated with the N'_k and N_k, and they follow from the context of their running indexes, $q(pa(\tau))$, $q(C \cup pa(\tau))$ and $q(S \cup pa(\tau))$.

For the purpose of model comparison, it is necessary to specify a family of compatible prior laws that permit carrying out computations in a local manner. Such compatible families have been discussed in the context of decomposable models (Dawid and Lauritzen, 1993, Section 6.2) and ADG models (Heckerman et al., 1995; Geiger and Heckerman, 1998; Roverato and Consonni, 2001). The development of specific compatible priors for EG models requires an entire discussion on its own about the topic, but a straightforward uninformative approach is the assignment given by

$$N'_{Ck} = \frac{1}{|\mathcal{X}_{C,pa(\tau)}|} . \tag{21}$$

Its compatibility follows from the equivalence of the EG factorization to the ADG factorization (see Section 5) and the results given by Heckerman et al. (1995).

As part of our prior knowledge, the hyperparameters specified by expression (21) imply that we do not have any preference among the levels of each of the marginal contingency tables formed by the variables from the cliques in $\mathcal{C}_\tau$ and the parents $pa(\tau)$. The consequences of such a policy as prior knowledge for the parameters of the model are best understood by examining the expression of the variance of one of the parameters $\theta_i \in \theta_G$. The variance of θ_i indicates how much the mean of θ_i may vary in the light of new data (DeGroot, 1970, Chapter 5, eqn. (7)):

$$\mathrm{Var}(\theta_i) = \frac{N'_i(N' - N'_i)}{(N')^2(N'+1)} \ , \tag{22}$$

where, recall, $N' = \sum_k N'_k$. This expression shows that the larger the values in ϑ are, the smaller the variance, as noted for instance in Castillo et al. (1997). Since the assignment in (21) is the smallest positive hyperconsistent assignment one can give to the hyperparameters in ϑ, it follows immediately that we let the data determine as much as possible the shapes of the parameters. For this reason, this type of assignment is often also known as an *uninformative* assignment.

5 Equivalence with Respect to Other Factorizations

Every ADG or DEC model is Markov equivalent to an EG Markov model and every EG Markov model is Markov equivalent to a ADG model. From this fact it must follow that the factorization in Theorem 02 is, in fact, equivalent to those provided by any ADG or DEC graph in the equivalence class. We can see this by means of the following two cases:

- A given DEC graph U is in the Markov equivalence class represented by the EG G.
  This case is straightforward, as G will have one single chain component, $\mathcal{T} = \{\tau\}$ with $pa_G(\tau) = \emptyset$, which would be an undirected chordal graph. The factorization of the EG model G is identical to the factorization for U (1).
- A given ADG D is in the Markov equivalence class represented by the EG G.
  In this case we will see that for every clique C_τ in a chain component τ, the corresponding term $f(x_C|x_{pa_G(\tau)}, \theta_{C,\rho})$ can be further factorized and this will lead to transforming the whole factorization of G into a factorization for the Markov equivalent ADG D.
  The cliques, and separators, factorize in the following way

$$\prod_{C\in\mathcal{C}_\tau} f(x_C|x_{pa_G(\tau)}, \theta_{C,\rho}) = \prod_{C\in\mathcal{C}_\tau} \prod_{i=1}^{|C|} f(x_i|x_1, \ldots, x_{i-1}, x_{pa_G(\tau)}, \theta_{C,\rho}). \tag{23}$$

By performing this factorization according to any perfect numbering of the vertices - cf. (Lauritzen, 1996, pg. 15) in G we will find repetitions of some terms $f(x_i|x_j, \ldots, x_{i-1}, x_{pa_G(\tau)}, \theta_{C,\rho})$ because the intersections among the cliques in $\mathcal{C}_\tau$ may be non-empty. Since the set of separators S_τ corresponds to the intersections among the cliques, these repeated terms will cancel in (4). Because $|\mathcal{C}_\tau| = |\mathcal{S}_\tau| + 1$, exactly one of these repeated terms will remain, and therefore it follows that the factorization will consist of one term per random variable.

6 Local Computations and Bayes Factors

In order to apply the scoring criterion for EG model selection, we need a characterization of local computations for essential graph Markov models. This characterization will depend on the concept of *local transformation* that we use.

In decomposable models, a local transformation is the addition or removal of an undirected edge. In ADG models, a local transformation is the addition, removal or reversal of an arc.

For essential graphs, we can find different proposals for local transformations in (Chickering, 1996; Perlman, 2001; Chickering, 2002a,b; Auvray and Wehenkel, 2002). We analyze first the transformations provided in Perlman (2001), which we now sketch for completeness.

First note that if an arrow $a \to t \in \tau$ occurs in an EG G, then by Theorem 01(iii), for every other $t' \in \tau$ the arrow $a \to t'$ must occur in G as well. The collection $\{a \to t \mid t \in \tau\}$ is called an *arrow bundle* in G. The transformations are as follows:

Aα Remove an undirected edge, as long as the chain component remains chordal.

Aβ Add an undirected edge, as long as the chain component remains chordal.

Bα Remove an arrow bundle between two comparable chain components as long as strong protection is preserved.

Bβ Add an arrow bundle between two comparable chain components as long as strong protection is preserved.

Bγ Add an arrow bundle between two non-comparable chain components with different parent sets as long as strong protection is preserved.

Bδ Add an undirected edge between two vertices of two non-comparable chain components with the same parent sets as long as strong protection is preserved.

Cα Remove a collection of immoralities formed from two non-adjacent chain components with a single vertex and a third chain component which contains the collision vertices, as long as strong protection is preserved.

Cβ Add a collection of immoralities formed from two non-adjacent chain components with a single vertex and a third chain component which will contain the collision vertices, as long as strong protection is preserved.

For the transformation Aα, we may identify two cases: the case where the removal of an undirected edge leaves the chain component connected, and the case where the chain component is split into two chain components. In the former case, we only need to compute the Bayes factor provided by Dawid and Lauritzen (1993). In the latter case, the set $\mathcal{T}(G)$ of chain components is enlarged by a new chain component. Therefore, the Bayes factor involves the comparison of the former larger chain component against the product of the two smaller new chain components.

For the transformation Aβ it suffices to use the Bayes factor provided by Dawid and Lauritzen (1993) since the transformation is within a single chain component which remains chordal.

In the transformations in Bα, Bβ and Bγ the set $\mathcal{T}(G)$ of chain components does not change, and only one arrow bundle involving two chain components will be added or removed. Therefore it suffices to compare the corresponding term in (20). The term should be entirely recomputed, though, due to the fact that the terms in (20) involve the current parent vertices which may change by any of these three operations.

In the transformation in Bδ, the set $\mathcal{T}(G)$ of chain components does change, by merging two chain components into a single one. In any case, if we have stored the computations made for the two merging components, still only computations for the new larger chain component will be necessary.

In the transformations in C, the set $\mathcal{T}(G)$ of chain components does not change, and it will suffice to compare the term in (20) that corresponds to the chain component containing the collision vertices.

6.1 Inclusion Friendly Local Computations

The graphical Markov inclusion order, or inclusion order for short, is a partial order among the graphs that belong to a common class of GMMs and have the same number of vertices. This partial order is defined as follows: a graph G precedes a different graph G', denoted $G \subseteq G'$, if and only if all the conditional independence statements that can be read off from G can be also read off from G'. For a more thorough description the reader should consult (Castelo and Kočka, 2002, Section 3). A trivial example is that of the fully connected graph preceeding the fully disconnected graph. From the concept of inclusion order, follows the concept of inclusion boundary:

Definition 02 *(Kočka and Castelo, 2001)*
Let H, K, L be three GMMs. Let $H \prec L$ denote that $H \subset L$ and for no K, $H \subset K \subset L$. The inclusion boundary *of the GMM G, denoted by $\mathcal{IB}(G)$, is*

$$\mathcal{IB}(G) = \{H \,|\, H \prec G\} \cup \{L \,|\, G \prec L\}.$$

Several authors (Kočka and Castelo, 2001; Castelo and Kočka, 2002; Chickering, 2002b) have shown that learning algorithms for ADG models that perform local transformations which can reach any GMM in the inclusion boundary, perform better than those without this feature.

Meek (1997) conjectured a graphical characterization of the inclusion order for ADGs which basically tells us that the inclusion boundary for a given ADG is only one adjacency away. Kočka et al. (2001) proved Meek's conjecture for a the particular case where two ADGs differ in one single adjacency and recently, Chickering (2002b) has proved Meek's conjecture in its general form. This implies we should strive to provide local transformations and local computations for EGs that may differ in one single adjacency.

The recent works by Chickering (2002b) and Auvray and Wehenkel (2002) provide such local transformations for EGs, combined with local computations for ADGs, because they use a scoring metric for ADGs. However, in both works the resulting graph, while being a chain graph, may not be a valid EG representation and therefore further transformations to obtain the corresponding EG may be necessary. The fact that we have (yet) no cheap graphical characterization of the inclusion boundary in terms of EGs, makes it difficult to find efficient local computations for a scoring metric specific for EGs, as the one presented in this paper. We expect, however, that this open problem can be solved in the future and therefore obtain an inclusion-friendly learning algorithm based exclusively upon the EG representation.

To illustrate our approach to this open problem, we shall consider the necessary local computations for a particular example. First, for each chain component $\tau \in \mathcal{T}$, consider a perfect ordering of the cliques - cf. (Lauritzen, 1996, pg. 14) of G_τ as $C_1, \ldots, C_k$. Let $H_j = \bigcup_{i=1}^{j} C_i$, $S_{j+1} = C_{j+1} \cap H_j$, $R_{j+1} = C_{j+1} \backslash H_j$. The expression of the marginal likelihood of the data in (18) may be now written as

$$f(D|G) = \prod_{\tau \in \mathcal{T}} \left[f_{C_1|pa_G(\tau)}(D_{C_1}|D_{pa_G(\tau)}) \prod_{i=2}^{k} f_{R_i|S_i,pa_G(\tau)}(D_{R_i}|D_{S_i,pa_G(\tau)}) \right] . \tag{24}$$

Now consider the EGs in Figure 1.

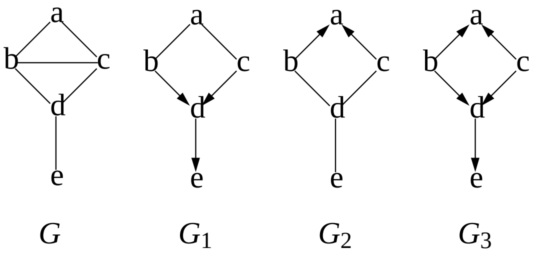


Fig. 1. The EGs G_1, G_2 and G_3 are obtained from G by removing the undirected edge b–c

The EG G is formed by a single chain component. By removing the edge b–c we may obtain either G_1, G_2 or G_3 which have, respectively, 3, 2 and 5 chain components. Using expression (24), we are going to write down the different Bayes factors for comparing the EGs G_1, G_2, G_3 against the EG G. In G_1, the cliques in the chain component $\tau = \{a, b, c\}$ may follow the perfect ordering $C_1 = \{a, b\}, C_2 = \{a, c\}$. Thus, we obtain $H_1 = C_1, S_2 = C_2 \cap H_1 = \{a\}, R_2 = \{C_2 \backslash H_1\} = \{c\}$. The other chain components, $\tau = \{d\}$ and $\tau = \{e\}$, are singletons and, therefore, each of them has just one clique containing one single vertex. By (24) the Bayes factor of the EG G_1 against G may be written as

$$\frac{f(D|G_1)}{f(D|G)} = \frac{f_{a,b}(D_{a,b}) f_{c|a}(D_c|D_a) f_{d|b,c}(D_d|D_{b,c}) f_{e|d}(D_e|D_d)}{f_{a,b,c}(D_{a,b,c}) f_{d|b,c}(D_d|D_{b,c}) f_{e|d}(D_e|D_d)} ,$$

which after some simplification becomes:

$$\frac{f(D|G_1)}{f(D|G)} = \frac{f_{a,b}(D_{a,b}) f_{a,c}(D_{a,c})}{f_a(D_a) f_{a,b,c}(D_{a,b,c})} . \tag{25}$$

Similar to the case of Bayes factors in decomposable models (Dawid and Lauritzen, 1993, pg. 1300), expression (25) provides a formula to test the conditional independence $X_b \perp\!\!\!\perp X_c | X_a$.

The Bayes factor for G_2 against G is

$$\begin{aligned}\frac{f(D|G_2)}{f(D|G)} &= \frac{f_{a|b,c}(D_a|D_{b,c}) f_{b,d}(D_{b,d}) f_{c|d}(D_c|D_d) f_{e|d}(D_e|D_d)}{f_{a,b,c}(D_{a,b,c}) f_{d|b,c}(D_d|D_{b,c}) f_{e|d}(D_e|D_d)} = \\ &= \frac{f_{a,b,c}(D_{a,b,c}) f_{b,c}(D_{b,c}) f_{b,d}(D_{b,d}) f_{c,d}(D_{c,d})}{f_{a,b,c}(D_{a,b,c}) f_{b,c,d}(D_{b,c,d}) f_{b,c}(D_{b,c}) f_d(D_d)} = \frac{f_{b,d}(D_{b,d}) f_{c,d}(D_{c,d})}{f_{b,c,d}(D_{b,c,d}) f_d(D_d)} ,\end{aligned}$$

which corresponds to a formula to test whether $X_b \perp\!\!\!\perp X_c | X_d$. Finally, the Bayes factor for G_3 against G is

$$\begin{aligned}\frac{f(D|G_3)}{f(D|G)} &= \frac{f_{a|b,c}(D_a|D_{b,c}) f_b(D_b) f_c(D_c) f_{d|b,c}(D_d|D_{b,c}) f_{e|d}(D_e|D_d)}{f_{a,b,c}(D_{a,b,c}) f_{d|b,c}(D_d|D_{b,c}) f_{e|d}(D_e|D_d)} = \\ &= \frac{f_b(D_b) f_c(D_c)}{f_{b,c}(D_{b,c})} ,\end{aligned}$$

which corresponds to a formula to test whether $X_b \perp\!\!\!\perp X_c | \emptyset$.

7 Concluding Remarks

We have presented a scoring metric for EGs with discrete multinomial data. Such a scoring metric opens the way to devise a learning, or selection, procedure that relies only on the EG representation. Another advantage is that the

EG scoring metric automatically assigns the same score to equivalent ADG models and bypasses the need to impose strong constraints on the prior distribution of both graphs and parameters as discussed by Andersson et al. (1997, Section 7.2).

We have investigated possible ways of performing local computations with this scoring metric. We have briefly analized the question of finding the corresponding local computations for local transformations on single adjacencies, which would respect the inclusion order. In that analysis, we can see how this scoring metric in fact provides formulae for investigating the conditional independence restrictions being removed (or added) while performing model comparison. This feature is important if we want to monitor, and understand, each step in the learning process.

Bibliography

Andersson, S., Madigan, D., and Perlman, M. (1997). A characterization of Markov equivalence classes for acyclic digraphs. Annals of Statistics, 25:505–541.

Andersson, S., Madigan, D., and Perlman, M. (2001). Alternative Markov properties for chain graphs. Scandinavian Journal of Statistics, 28:33–85.

Auvray, V. and Wehenkel L. (2002). On the construction of the inclusion boundary neighborhood for Markov equivalence classes of Bayesian network structures. In Proc. of the Eighteenth Conf. on Uncertainty in Art. Int.. Morgan Kaufmann.

Castelo, R. and Kočka, T. (2002). Towards an inclusion driven learning of Bayesian networks. Tech. Rep. CS-2002-05, Jan. 2002, Institute for Computing and Information Sciences, University of Utrecht, The Netherlands.

Castillo, E., Hadi, A., and Solares, C. (1997). Learning and updating of uncertainty in Dirichlet models. Machine Learning, 26:43–63.

Chickering, D. (1996). Learning equivalence classes of Bayesian network structures. In Proc. of the Twelfth Conf. on Uncertainty in Art. Int., pg. 150–157. Morgan Kaufmann.

Chickering, D. (2002a). Learning equivalence classes of Bayesian-network structures. Journal of Machine Learning Research, 2:445-498.

Chickering, D. (2002b). Optimal Structure Identification with Greedy Search. Journal of Machine Learning Research, 3:507–554

Cowell, R., Dawid, A., Lauritzen, S., and Spiegelhalter, D. (1999). Probabilistic Networks and Expert Systems. Springer-Verlag, New York.

Dawid, P. and Lauritzen, S. (1993). Hyper-Markov laws in the statistical analysis of decomposable graphical models. Annals of Statistics, 21(3):1272–1317.

DeGroot, M. H. (1970). Optimal Statistical Decisions. McGraw-Hill.

Frydenberg, M. (1990). The chain graph Markov property. Scandinavian Journal of Statistics, 17:333–353.

Geiger, D. and Heckerman, D. (1998). Parameter priors for directed acyclic graphical models and the characterization of several probability distributions. Tech. Rep. MSR-TR-98-67, Oct. 1998, Microsoft Research.

Heckerman, D., Geiger, D., and Chickering, D. (1995). Learning Bayesian networks: The combination of knowledge and statistical data. Machine Learning, 20:194–243.

Kočka, T., Bouckaert, R., and Studený, M. (2001). On characterizing inclusion of Bayesian networks. In Proc. of the Seventeenth Conf. on Uncertainty in Art. Int., pg. 261–268. Morgan Kaufmann.

Kočka, T. and Castelo, R. (2001). Improved learning of Bayesian networks. In Proc. of the Seventeenth Conf. on Uncertainty in Art. Int., pg. 269–276. Morgan Kaufmann.

Lauritzen, S. (1996). Graphical Models. Oxford University Press, Oxford.

Meek, C. (1997). Graphical models, selecting causal and statistical models. PhD Thesis, Carnegie Mellon University.

Perlman, M. (2001). Graphical model search via essential graphs. In *Algebraic Methods in Statistics and Probability*, V. 287, American Math. Soc, Providence, Rhode Island.

Roverato, A. and Consonni, G. (2001). Compatible Prior Distributions for DAG models. Tech. Rep. 134, Sept. 2001, University of Pavia.

Spiegelhalter, D. and Lauritzen, S. (1990). Sequential updating of conditional probabilities on directed graphical structures. Networks, 20:579–605.

Verma, T. and Pearl, J. (1990). Equivalence and synthesis of causal models. In Proc. of the Sixth Conf. on Uncertainty in Art. Int., pg. 255–268. Morgan Kaufmann.

Wilks, S.S. (1962). Mathematical Statistics. Wiley, New York.