Compatible Prior Distributions for DAG models

Alberto Roverato
University of Modena and Reggio Emilia, Italy

Guido Consonni*
University of Pavia, Italy

July 6, 2001

Abstract

The application of certain Bayesian techniques, such as the Bayes factor and model averaging, requires the specification of prior distributions on the parameters of alternative models. We propose a method for constructing compatible priors on the parameters of models nested in a given DAG (Directed Acyclic Graph) model, using a conditioning approach. We define a class of allowable parameterisations consistent with the well-ordering and the modular structure of the DAG and derive a procedure, invariant within this class, which we name reference conditioning, generalising earlier work based on Jeffreys conditioning. The theory is then applied to Gaussian DAG models.

Keywords: Bayes factor; Conditioning; Directed acyclic graph; Fisher information matrix; Graphical model; Invariance; Jeffreys prior; Ordered group reference prior; Reparameterisation.

*Address for correspondence: Dipartimento di Economia Politica e Metodi Quantitativi, Via S. Felice 5, 27100 Pavia, Italy. E-mail: gconsonni@eco.unipv.it

1 Introduction and summary

Model comparison is an important topic in statistical theory and practice. In particular, Bayesian model comparison, which essentially relies on the Bayes factor, has been an active area of research lately, see Kass and Raftery (1995) for a comprehensive review. The sensitivity of the Bayes factor to the choice of prior distribution is sometimes perceived as a difficulty, and this has stimulated research with the hope of making it a more "objective" tool for scientific investigation, see Berger and Pericchi (2000). Prior specification however involves a more subtle aspect. Indeed the Bayes factor for comparing two models requires the assignment of two prior distributions (one for each model-parameter). Specifically, assume that there are two models $\mathcal{M}$ and $\mathcal{M}^0$, say, for the same observable X, parameterised respectively by θ and θ_0. Write $\pi(\cdot)$ and $\pi_0(\cdot)$ for the corresponding prior densities, and assume, for simplicity, that $\mathcal{M}^0$ is a submodel of $\mathcal{M}$. Given π, it is very often sensible to specify π_0 to be, in a sense to be made precise, as close as possible to π. In this way the resulting Bayes factor should be least influenced by dissimilarities between the two priors due to differences in the elicitation processes, and could thus more faithfully represent the strength of the support that the data lend to each model. Despite being clearly important, this issue has not been adequately dealt with in the literature so far, at least from a foundational perspective. A notable exception is a recent paper by Dawid and Lauritzen (2000), which elucidates the problem through simple and instructive examples, and suggests two strategies to choose, in their terminology, a *compatible* prior π_0, which they name "projection" and "conditioning". In their opinion, the former seems more suitable in the case of *co-existing* models, wherein several models are believed to be simultaneously true, leading to a model-selection approach. The latter, instead, appears more appropriate with *competing* models, corresponding to a situation wherein only one of the specified models is deemed to be true, leading to a model-averaging strategy. The objective of this paper is to elaborate on the "conditioning approach" for the construction of compatible priors.

Specifically we consider statistical models based on Directed Acyclic Graphs, or DAG models for short. These models are very useful in many applied domains, and represent the architecture of probabilistic expert systems and belief networks, see for example Cowell *et al.* (1999). DAG models may be regarded as simply depicting conditional independence relations among the random variables involved, say $X = (X_1, \ldots, X_v)$ or as 'causal' models. In a loose sense the latter interpretation arises, for example, when there exists qualitative prior information that specifies constraints on the ordering of the random variables, as in the context of univariate recursive data generating processes described, among others, in Wermuth and Cox (2000) and Lauritzen and Richardson (2001). In such models, the joint distribution of the observables is not the primitive notion, but rather the end result of the specification of a collection of local conditional distributions. If X_1 denotes the most recent response, while X_v the last, purely explanatory variable, then a generating process starts with the marginal distribution of X_v and progressively generates observations on a response variable X_i, $i = 1, \ldots, v-1$, conditionally on a subset of the potential ancestors $X_{i+1}, \ldots, X_v$, which are called the parents of X_i. A more stringent interpretation of causal DAG models is explicated in Pearl (1995), and more recently in Lauritzen (2001). In this context one distinguishes between conditioning by observation or conditioning by intervention and the causal DAG provides the relevant intervention formulae.

We claim that the causal interpretation of DAG models, which we shall adhere to in this paper, has crucial consequences in terms of allowable model reparameterisations. In the sequel we shall carefully articulate this point, since it represents the cornerstone of our strategy for the construction of compatible priors across parameters of DAG models.

The main idea of the paper is briefly outlined below. The conditioning approach of Dawid and Lauritzen (2000) rests upon the idea of choosing suitable *baseline* (reference, in their terminology) measures ν and ν_0 for models $\mathcal{M}$ and $\mathcal{M}^0$. Given a prior law Π with density π, the prior law Π_0 with density π_0 is then derived by imposing that the Radon-Nykodim derivatives of Π and Π_0, relative

to ν and ν_0 respectively, be proportional. A crucial requirement of the above program is that the baseline measures be, in some sense, *intrinsic* to the models and *independent* of specific ways of parameterising them. Dawid and Lauritzen (2000) suggest to take the appropriate Jeffreys measure as the baseline measure under each model, and accordingly their method is named Jeffreys conditioning. Clearly the Jeffreys measure is model-specific and invariant to reparameterisations. Notice the unconventional use of Jeffreys measure in this context, namely *not* as a possibly uninformative prior, but rather as a half-way house towards obtaining a compatible prior. While we concur that a baseline measure should be intrinsic to the model, we emphasise that invariance to *any* reparameterisation is not appropriate for (causal) DAG models, precisely because such models incorporate an ordering of the variables and a modular structure which would otherwise be lost. Instead we argue in favour of a more restrictive notion of invariance, namely that dictated by the class of allowable DAG parameterisations we alluded to above. A baseline measure consistent with the above requirement corresponds to the group reference prior of Berger and Bernardo (1992), and accordingly we name our method *reference conditioning*. We emphasise that our use of reference priors is unconventional, being unrelated to issues of non-informativity of prior specification. On the other hand the suggestion to employ reference priors stems from (restricted) invariance considerations, an idea already present in Jeffreys (1961, § 3.10). The reference prior measure is at the heart of our procedure for constructing compatible priors on parameters across nested DAG models. Our method is general and automatic and requires a single specification of the prior corresponding to the largest entertained model. Through the notion of 'parameterisation implicitly leading to reference conditioning' the implementation of our method becomes even more straightforward as exemplified in the Gaussian case.

The structure of the paper is as follows. Section 2 describes the conditioning procedure to find compatible priors. Section 3 is devoted to the crucial issue of reparameterisation for DAG models, motivating the definition of the class of allowable reparameterisation. The main result is a characterization of such a class.

An illustration to the Gaussian case is then presented in detail. Section 4 introduces the concept of reference conditioning which is then elaborated for Gaussian DAG models in Section 5. Building on previous work by Roverato (2000) and Consonni and Veronese (2001), see also Consonni, Veronese and Gutiérrez-Peña (2000), the reference measure for Gaussian DAG models is found, relative to an allowable parameterisation, which is especially convenient to work with. Subsequently the reference conditioning approach to find compatible priors for the parameters of DAG models is detailed. Fainally, Section 6 offers some concluding remarks with a special view to future directions.

2 Compatible priors by means of conditioning

Consider the model $\mathcal{M}$ parameterised by θ and let π be a prior density representing uncertainty about θ conditional on $\mathcal{M}$. If $\mathcal{M}^0$ is a submodel obtained from $\mathcal{M}$ by imposing a constraint on θ, say $\theta = \theta_0$, then one way to derive a prior distribution for θ_0 is by conditioning the distribution for θ on the event $\{\theta = \theta_0\}$. Thus, the resulting distribution for θ_0 has density $\pi_0(\theta_0) \propto \pi(\theta_0)$.

This conditioning procedure may appear a natural way to construct "compatible" distributions; however, it is not invariant with respect to the specific parameterisation chosen to describe the model. As a consequence, if the same prior information provided by π is expressed with respect to a different parameterisation, say $\psi = \psi(\theta)$, then the distribution obtained by conditioning on $\{\psi = \psi(\theta_0)\}$ may provide prior information different from that conveyed by π_0. Dawid and Lauritzen (2000) gave an example of this phenomenon, known as the *Borel-Kolmogorov paradox*, with respect to a bivariate normal model. Before presenting such an example, we need some notation.

Let $V = \{1, \ldots, v\}$; for $A \subseteq V$, let X_A denote the random vector having components X_i with $i \in A$. Suppose now that X_V has a multivariate normal distribution with expectation equal to zero and covariance matrix $\Sigma = \{\sigma_{ij}\}$. Instances of alternative parameterisations of this model are: the concentration matrix $\Sigma^{-1} = \{\sigma^{ij}\}$; the matrix P -to be interpreted as an upper-case Greek

"rho"- with variances σ_{ii} in the main diagonal and correlations $\rho_{ij} = \sigma_{ij}/\sqrt{\sigma_{ii}\sigma_{jj}}$ in the off-diagonal entries; the matrix R with elements σ^{ii} in the main diagonal and partial correlations $\rho_{ij \cdot V\setminus\{i,j\}} = -\sigma^{ij}/\sqrt{\sigma^{ii}\sigma^{jj}}$ in the off-diagonal entries. Recall that $\sigma^{ii} = 1/\sigma_{ii \cdot V\setminus\{i\}}$, where $\sigma_{ii \cdot V\setminus\{i\}}$ is the variance of the conditional distribution of X_i given the remaining variables $X_{V\setminus\{i\}}$ (see Whittaker, 1990 p.143).

Example 1 (Dawid and Lauritzen, 2000) *Let $\mathcal{M}$ be a bivariate normal model with zero mean and let $\mathcal{M}^0$ be the submodel with $X_1 \perp\!\!\!\perp X_2$. This constraint can be expressed in several alternative ways, depending on the parameterisation adopted for the model. Dawid and Lauritzen (2000) considered the following cases (the corresponding parameterisation is in square brackets):*

(i) $\sigma_{12} = 0$ $[\Sigma]$
(ii) $\sigma^{12} = 0$ $[\Sigma^{-1}]$
(iii) $\rho_{12} = 0$ $[R]$
(iv) $\beta_{12} = 0$ $[(\sigma^{11}, \beta_{12}, \sigma_{22})]$

where $\beta_{12} = \sigma_{12}/\sigma_{22}$. Let Σ have an inverse Wishart distribution, $\Sigma \sim IW(\delta, A)$ where δ is a positive constant and $A = \{a_{ij}\}$ is a positive definite matrix. Here we use the notation of Dawid (1981) so that $\Sigma^{-1} \sim W(\delta+1, A^{-1})$. In the parameterisation i), the parameter of the submodel is $(\sigma_{11}, \sigma_{22})$ and, by conditioning the distribution of Σ on $\{\sigma_{12} = 0\}$ one obtains

(i) $\sigma_{11} \sim a_{11}/\chi^2_{\delta+2}$, $\sigma_{22} \sim a_{22}/\chi^2_{\delta+2}$, *independently*

where χ^2_g denotes a chi-square random variable with g degrees of freedom. However, different answers result from the application of the same procedure with respect to the alternative parameterisations ii)-iv):

(ii) $\sigma_{11} \sim a_{11}/\chi^2_{\delta}$, $\sigma_{22} \sim a_{22}/\chi^2_{\delta}$, *independently;*
(iii) $\sigma_{11} \sim a_{11}/\chi^2_{\delta+1}$, $\sigma_{22} \sim a_{22}/\chi^2_{\delta+1}$, *independently;*
(iv) $\sigma_{11} \sim a_{11}/\chi^2_{\delta+2}$, $\sigma_{22} \sim a_{22}/\chi^2_{\delta}$, *independently.*

Example 1 may be usefully reinterpreted in terms of graphical models. A graphical model, see Lauritzen (1996) and Cowell *et. al*(1999) is a family of distributions for the vector X_V satisfying a set of conditional independence relations encoded by a graph. A graph is a pair (V, E) where V denotes the vertex set and the edge set E is a subset of the $V \times V$ ordered pairs of distinct vertices. We distinguish between undirected, $i - - j$, and directed edges $j \to i$. A graph is *complete* if all vertices are joined by a directed or an undirected edge.

A graph $\mathcal{G} = (V, E)$ with only undirected edges is itself called undirected and is used to represent symmetric conditional associations between variables. The distribution of X_V is said to satisfy the *pairwise Markov property* with respect to $\mathcal{G}$ if every missing edge $(i, j) \notin E$ corresponds to the conditional independence of X_i and X_j given the remaining variables; shortly $X_i \perp\!\!\!\perp X_j | X_{V\setminus\{i,j\}}$ (see Cowell *et al.*, 1999 p.67). In the Gaussian case the absolute value of the partial correlation coefficient $\rho_{ij \cdot V\setminus\{i,j\}}$ provides the strength of the association represented by the edge (i, j) of $\mathcal{G}$.

A directed acyclic graph, DAG, is a graph with only directed edges and no cycles. If $j \to i$, then j is said to be a *parent* of i and the set of all parents of i is denoted as $\text{pa}(i)$. The set $\{i\} \cup \text{pa}(i)$ is called the *family* of i and denoted by $\text{fa}(i)$. To emphasize that a graph is a DAG in the following we write $\mathcal{D} = (V, E)$. It is always possible to identify a *well-ordering* $(\alpha_1, \alpha_2, \ldots, \alpha_v)$ of the vertices V of a DAG such that, if two nodes are joined by an arrow, the edge points from the vertex with lower position to the vertex with higher position in the ordering. A DAG may not have a unique well-ordering (see Cowell *et. al*, 1999 p. 47). The undirected version of a DAG $\mathcal{D} = (V, E)$ is the undirected graph $\mathcal{D}^{\sim} = (V, E^{\sim})$ obtained by substituting arrows with undirected edges. Note that $\mathcal{D}^{\sim} = (V, E^{\sim})$ along with any well-ordering of the vertices V fully identifies $\mathcal{D} = (V, E)$.

A DAG $\mathcal{D} = (V, E)$ is used to represent a recursive response structure between the variables in X_V such that, for $i = 1, \ldots, v$, X_{α_i} is a response to variables $X_{\alpha_1}, \ldots, X_{\alpha_{i-1}}$ but explanatory to variables $X_{\alpha_{i+1}}, \ldots, X_{\alpha_v}$. The distribution of X_V is said to satisfy the *directed Markov property* with respect to $\mathcal{D} = (V, E)$ if,

for $i = 1, \ldots, v$, it holds that $X_{\alpha_i} \perp\!\!\!\perp X_{\{\alpha_1,\ldots,\alpha_{i-1}\}} | X_{\mathrm{pa}(\alpha_i)}$ (see Cowell *et al.*, 1999 p. 74). As a consequence, in the Gaussian case every missing edge in $\mathcal{D}$ corresponds to a regression coefficient equal to zero.

A chain graph admits both undirected and directed edges, but no partially directed cycles. The corresponding chain graph model is a generalisation of the directed and undirected graphical model.

Two graphs are said to be *Markov equivalent* if they encode the same set of conditional independence assertions (Frydenberg, 1990; Andersson *et al.* 1997). Markov equivalence is an equivalence relation and we denote by $[\mathcal{D}]$ the set of all DAGs Markov equivalent to $\mathcal{D}$. Frydenberg (1990) provided necessary and sufficient conditions for the Markov equivalence of chain graphs. Of special interest is the relation existing between *perfect* DAGs and *decomposable* undirected graphs; for the definition of perfect DAG and decomposable graph see Lauritzen (1996, pp.7-8). Any perfect DAG $\mathcal{D}$ is Markov equivalent to its undirected version $\mathcal{D}^{\sim}$; moreover $\mathcal{D}^{\sim}$ is decomposable (see Lauritzen, 1996, p.52). Conversely, if $\mathcal{G} = (V, E)$ is a decomposable undirected graph, then there exists a perfect DAG $\mathcal{D}$ Markov equivalent to $\mathcal{G}$ with $\mathcal{D}^{\sim} = \mathcal{G}$ (see Lauritzen, 1996, p.18).

In Example 1 model $\mathcal{M}$ may be described by any of the following three Markov equivalent complete graphs: the undirected graph $1 \bullet\!\!-\!\!-\!\!\bullet 2$, the DAG with well-ordering $(2, 1)$, $1\bullet \longleftarrow \bullet 2$, and the DAG with well-ordering $(1, 2)$, $1\bullet \longrightarrow \bullet 2$. The submodel $\mathcal{M}^0$ corresponds to the graph $1 \bullet \quad \bullet 2$.

2.1 Jeffreys conditioning

In order to obtain a conditioning procedure with the property of invariance with respect to reparameterisations of the model, Dawid and Lauritzen (2000) introduced the following generalization of the usual conditioning procedure.

Consider two baseline measures, ν and ν_0, on $\mathcal{M}$ and $\mathcal{M}^0$ respectively. For a distribution Π in $\mathcal{M}$, the compatible distribution Π_0 in $\mathcal{M}^0$ is such that the density function of Π with respect to ν, $d\Pi/d\nu$, and that of Π_0 with respect to ν_0, $d\Pi_0/d\nu_0$, are proportional. This method generalizes the usual conditioning

because if ν is a probability measure and ν_0 is obtained from ν by conditioning on $\{\theta = \theta_0\}$, then it gives the same answer as conditioning Π on $\{\theta = \theta_0\}$.

Invariance with respect to reparameterisations is obtained by choosing measures ν and ν_0 that are intrinsic to the models and independent of specific ways of parameterising them. One possible choice is the Jeffreys measure and the resulting procedure was named by Dawid and Lauritzen (2000) *Jeffreys conditioning*.

The density of the Jeffreys measure with respect to Lebesgue measure is given by $j(\theta) = |H(\theta)|^{1/2}$, where $|H(\theta)|$ is the determinant of the Fisher information matrix for θ.

If π is the density function of Π with respect to Lebesgue measure, then the distribution for θ_0 resulting from the application of Jeffreys conditioning has density function with respect to Lebesgue measure

$$\pi_0(\theta_0) \propto \pi(\theta_0)\frac{j_0(\theta_0)}{j(\theta_0)}. \tag{1}$$

It is worth pointing out that if there exists a parameterisation, say $\psi = \psi(\theta)$, such that the ratio of the Jeffreys measures for ψ under $\mathcal{M}$ and $\mathcal{M}^0$ is constant for all $\psi = \psi(\theta_0)$, then Jeffreys conditioning gives the same answer as usual conditioning, with respect to such a parameterisation. In this case we say that ψ is a parameterisation *implicitly leading to Jeffreys conditioning* for $\mathcal{M}$ and $\mathcal{M}^0$. For instance, Dawid and Lauritzen (2000) showed that, for the problem in Example 1, R is a parameterisation implicitly leading to Jeffreys conditioning; such a result is generalised in Section 5.1.

3 DAG models and reparameterisations

Consider a DAG $\mathcal{D} = (V, E)$ with $V = \{1, \ldots, v\}$ and assume, without loss of generality, that $(v, v-1, \ldots, 1)$ is a well-ordering of V. For each $i \in V$ we specify a family $\mathcal{M}_i^{\mathcal{D}}$ of local conditional densities

$$p(x_i|x_{\,\mathrm{pa}(i)}, \eta_i), \quad \eta_i \in H_i, \tag{2}$$

with the local parameters η_i variation independent. Multiplying the conditional densities (2) one obtains the family $\mathcal{M}^{\mathcal{D}} := \mathcal{M}_1^{\mathcal{D}} \times \cdots \times \mathcal{M}_v^{\mathcal{D}}$ of distributions for

$X_V = (X_1, \ldots, X_v)$ with joint density

$$p(x|\eta) = \prod_{i=1}^{v} p(x_i|x_{\,\mathrm{pa}(i)}, \eta_i), \tag{3}$$

where $\eta := (\eta_1, \ldots, \eta_v)$, with $\eta \in H := H_1 \times \cdots \times H_v$.

While (3) specifies a distribution which is directed Markov with respect to any $\mathcal{D}^* \in [\mathcal{D}]$, we take the view, as discussed in the Introduction, that the vertices of $\mathcal{D}$ are arranged in a causal order, see also Cowell *et* al. (1999, p. 259), and assume the absence of unmeasured confounders that alter the causal interpretation of arrows; see Lauritzen and Richardson (2001) for a detailed discussion of this point. In other words, we are assuming that the ordering of the variables can be used reliably to make inference on the graphical structure. An important special case of this situation is when (3) is the result of a structural assignment system, as described in Lauritzen and Richardson (2001, eqn. 4). It is worth pointing out that in this case the distribution of X_V is causally Markov with respect to $\mathcal{D}$ (Lauritzen, 2001, Theorem 2.20) and so the orientation of the arrows is crucial for the interpretation of the model.

In the causal interpretation of DAG models, the local conditional families $\mathcal{M}_i^{\mathcal{D}}$s represent the primitive modules of the overall model $\mathcal{M}^{\mathcal{D}}$ and the parameter η acquires its meaning from its constituents η_is. Specifically, the partition of η into v blocks, each being associated to a local family, is an integral part of the parametric structure which must be preserved across reparameterisations. The previous remark implies that a DAG model cannot be arbitrarily reparameterised; indeed, only those transformations of η capable of preserving its modular structure will be deemed acceptable. To clarify the above point and eventually reach a definition of allowable reparameterisation we proceed in steps. Consider first a (smooth) one-to-one transformation of η, θ say, such that $\theta = (\theta_1, \ldots, \theta_v)$ and θ_i is a bijection of η_i, for each i. Clearly this transformation is allowable, since it preserves the modular structure of the DAG. Specifically, each component θ_i is uniquely associated to the i-th conditional family, or, equivalently, to η_i. On the other hand, a more general situation could be envisaged, namely one in which θ_i is no longer a bijection of η_i, and yet θ_i can still be uniquely linked to η_i. This

leads us to the notion of unambiguous association.

Definition 1 *Consider the family $\mathcal{M}^{\mathcal{D}}$ with density (3). Let $\theta = \theta(\eta)$ be a bijection of $\eta = (\eta_1, \ldots, \eta_v)$ and let θ_i be a subvector of θ with* $\dim(\eta_i)$=$\dim(\theta_i)$. *We say that θ_i is Unambiguously Associated (UA) to $\mathcal{M}_i^{\mathcal{D}}$ if either*

$$\theta_i = \theta_i(\eta_i)$$

or, for an integer $r = 1, \ldots, v-1$ and distinct indexes $j_1, \ldots, j_r$,

$$\theta_i = \theta_i(\eta_i, \eta_{j_1}, \ldots, \eta_{j_r}),$$

and there exist distinct subvectors $\theta_{j_1}, \ldots, \theta_{j_r}$ of θ with θ_{j_k} UA to $\mathcal{M}_{j_k}^{\mathcal{D}}$, $k = 1, \ldots, r$.

The above is a recursive but effective definition.

Remark In Definition 1 the function $\theta_i = \theta_i(\eta_i)$ is assumed to be one-to-one; furthermore, for fixed $(\eta_{j_1}, \ldots, \eta_{j_r})$, $\theta_i = \theta_i(\eta_i, \eta_{j_1}, \ldots, \eta_{j_r})$ is a one-to-one function of η_i.

Definition 2 *θ is an allowable parameterisation of $\mathcal{M}^{\mathcal{D}}$ with density in (3) if it admits a grouping $\theta_1, \ldots, \theta_v$ such that θ_i is UA to $\mathcal{M}_i^{\mathcal{D}}$, for all $i = 1, \ldots, v$.*

As an immediate consequence of Definition 2 we have the following.

Proposition 1 *θ is an allowable parameterisation of $\mathcal{M}^{\mathcal{D}}$ with density in (3) if and only if there exist a grouping $\theta = (\theta_1, \ldots, \theta_v)$, with* $\dim(\theta_i) = \dim(\eta_i)$, *and an ordering $(j_1, \ldots, j_v)$ of the indexes $(1, \ldots, v)$ such that*

$$\begin{aligned}
\theta_{j_1} &= \theta_{j_1}(\eta_{j_1}) \\
\theta_{j_2} &= \theta_{j_2}(\eta_{j_2}, \eta_{j_1}) \\
\vdots &\quad \vdots \\
\theta_{j_k} &= \theta_{j_k}(\eta_{j_k}, \ldots, \eta_{j_1}) \\
\vdots &\quad \vdots \\
\theta_{j_v} &= \theta_{j_v}(\eta_{j_v}, \ldots, \eta_{j_1}).
\end{aligned} \tag{4}$$

Remark To illustrate the meaning of Proposition 1, suppose $v = 2$ and assume that for any block grouping (θ_1, θ_2) of θ, with $\dim(\theta_1) = \dim(\eta_1)$, θ_1 and θ_2 are each (nontrivial) function both of η_1 and η_2 Then θ is not an allowable transformation of η because it destroys the intrinsic asymmetry of the underlying DAG model parameterised by η.

Example 2 *Consider two local conditional Gaussian families corresponding to the DAG* $\mathcal{D}$ $1\bullet \longleftarrow \bullet 2$.

$$\begin{aligned} X_1|X_2 = x_2, \eta_1 &\sim N(\beta_{12}x_2, \sigma_{11\cdot 2}), \\ X_2|\eta_2 &\sim N(0, \sigma_{22}), \end{aligned}$$

where $\eta_1 = (\beta_{12}, \sigma_{11\cdot 2})$, $\eta_2 = \sigma_{22}$. *As in Example 1, the joint distribution of* (X_1, X_2) *is bivariate normal with zero mean.*

i) Consider the transformation $\eta \mapsto \phi$ *where* $\phi = (\phi_{11}, \phi_{12}, \phi_{22})$ *with*

$$\phi_{11}(\eta_1) = \frac{1}{\sqrt{\sigma_{11\cdot 2}}}, \quad \phi_{12}(\eta_1) = -\frac{\beta_{12}}{\sqrt{\sigma_{11\cdot 2}}} \quad \textit{and} \quad \phi_{22}(\eta_2) = \frac{1}{\sqrt{\sigma_{22}}}.$$

Clearly $\eta \mapsto \phi$ *is an allowable transformation, since* $\phi_1 = (\phi_{11}, \phi_{12})$ *is a bijection of* η_1 *and* $\phi_2 = \phi_{22}$ *is a bijection of* η_2.

ii) Consider the transformation from η *to the local conditional parameters in the reverse DAG* $\mathcal{D}^*$ $1\bullet \longrightarrow \bullet 2$, *namely* $\eta^* = (\beta_{21}, \sigma_{22\cdot 1}, \sigma_{11})$. *It can be verified that*

$$\sigma_{22\cdot 1} = \frac{\sigma_{11\cdot 2}\sigma_{22}}{\beta_{12}^2\sigma_{22} + \sigma_{11\cdot 2}}, \quad \beta_{21} = \frac{\beta_{12}\sigma_{22}}{\beta_{12}^2\sigma_{22} + \sigma_{11\cdot 2}} \quad \textit{and} \quad \sigma_{11} = \beta_{12}^2\sigma_{22} + \sigma_{11\cdot 2}.$$

Hence, this transformation is not allowable. For example, letting $\theta_1(\eta) = (\beta_{21}, \sigma_{22\cdot 1})$ *and* $\theta_2 = \sigma_{11}$, *one immediately realises that* θ_1 *and* θ_2 *are each simultaneously a function of both* η_1 *and* η_2. *Clearly no other grouping would overcome this difficulty. However this is not surprising, since the reverse DAG precisely destroys the modular structure, which includes the direction of the arrow, of the original DAG model. Notice that* $\mathcal{D}^*$ *is Markov equivalent to* $\mathcal{D}$. *Thus the transformation between the local conditional parameters of two Markov equivalent DAGs is in general not allowable.*

iii) Consider the transformation $\eta \mapsto \Sigma$. *Because of symmetry, only the upper*

diagonal elements, say, including the diagonal itself, need be considered. Accordingly, let $\theta_1(\eta) = (\sigma_{11}, \sigma_{12})$ *and* $\theta_2(\eta) = \sigma_{22}$ *denote, respectively, the first and second row of* Σ *excluding the below-diagonal elements. This transformation is allowable. Indeed* θ_2 *is trivially a function of* η_2*; as a consequence* θ_1*, which is a function of* (η_1, η_2)*, is UA to* η_1*, and thus the transformation* $\eta \mapsto \Sigma$ *is allowable.*
iv) It is straightforward that we still have an allowable transformation if in the previous point Σ *is replaced by* P *so that* $\theta(\eta) = (\sigma_{11}, \rho_{12}, \sigma_{22})$*.*
v) Consider now the transformation $\eta \mapsto \Sigma^{-1}$ *and let* $\theta_1(\eta) = (\sigma^{11}, \sigma^{12})$ *and* $\theta_2(\eta) = \sigma^{22}$*. By recalling that* $\beta_{12} = -\sigma^{12}/\sigma^{11} = \sigma_{12}/\sigma_{22}$ *(see Cox and Wermuth, 1996, p.69) we can write*

$$\theta_1(\eta_1) = (1/\sigma_{11\cdot 2}, -\beta_{12}/\sigma_{11\cdot 2}) \quad \textit{and} \quad \theta_2(\eta) = \frac{\beta_{12}^2\sigma_{22} + \sigma_{11\cdot 2}}{\sigma_{11\cdot 2}\sigma_{22}}. \tag{5}$$

Consequently, $\theta_1(\eta_1)$ *and* $\theta_2(\eta)$ *are UA to* η_1 *and* η_2 *respectively and the transformation* $\eta \mapsto \Sigma^{-1}$ *is allowable.*
vi) In parallel with case iv) above, we now consider the transformation $\eta \mapsto R = \{\sigma^{11}, \rho_{12}, \sigma^{22}\}$*. Applying equations (5) to* $\rho_{12} = -\sigma^{12}/\sqrt{\sigma^{11}\sigma^{22}}$*, it can be easily checked that no grouping of the elements of* R *can satisfy (4), so that the transformation is not allowable because of Proposition 1.*

Example 2 is interesting because it clarifies the nature of an allowable parameterisation. Case i) and ii) are instances of "asymmetric" reparameterisations, in the sense that they depend on some well-ordering of the vertices, - i) being allowable, because it is faithful to the original structure, and ii) not allowable for the opposite reason. Cases iii) to vi) are instances of "symmetric" parameterisations. Σ is itself allowable, possibly contrary to our intuition. Upon reflection, however, this appears to be sensible, because η_i is a function of σ_{ij}, $j \in \text{fa}(i)$ only. Thus, rearranging the elements of Σ in a way consistent with a well-ordering of the vertices of $\mathcal{D}$ allows Σ to incorporate the required structural information to become allowable. Note that, because this is a "symmetric" parameterisation of the model, it is allowable to both $\mathcal{M}^{\mathcal{D}}$ and $\mathcal{M}^{\mathcal{D}^*}$. This structure is not destroyed

when σ_{12} is replaced by ρ_{12}. Of special interest are cases v) and vi). Σ^{-1} and R play a key role in parameterising undirected Gaussian graphical models: the former is the canonical parameter of the model and the latter provides a direct measure of the interactions between variables. However Σ^{-1} is allowable to both $\mathcal{M}^{\mathcal{D}}$ and $\mathcal{M}^{\mathcal{D}^*}$ but R is neither allowable for $\mathcal{M}^{\mathcal{D}}$ nor for $\mathcal{M}^{\mathcal{D}^*}$. Interestingly R is the parameterisation implicitly leading to Jeffreys conditioning for this problem.

4 Reference conditioning

According to the theory developed in the previous section, a procedure for the construction of compatible priors for DAG models should be invariant with respect to the class of allowable parameterisations for the model. This can be obtained by applying Dawid and Lauritzen's (2000) conditioning method with respect to two measures, ν and ν_0, each being intrinsic to the model and invariant within the class of allowable parameterisations. Jeffreys measure can clearly be used, however it may lead to inconsistent results as illustrated in the following example.

Example 3 *Let $\mathcal{M}^{\mathcal{D}} := \mathcal{M}_1^{\mathcal{D}} \times \mathcal{M}_2^{\mathcal{D}}$ be as in Example 2 so that $\eta = (\eta_1, \eta_2)$ with $\eta_1 = (\sigma_{11\cdot 2}, \beta_{12})$ and $\eta_2 = \sigma_{22}$. The submodel $\mathcal{M}^{\mathcal{D}_0} := \mathcal{M}_1^{\mathcal{D}_0} \times \mathcal{M}_2^{\mathcal{D}_0}$ with $X_1 \perp\!\!\!\perp X_2$ differs from $\mathcal{M}^{\mathcal{D}}$ only with respect to the first conditional density, that is $\mathcal{M}_2^{\mathcal{D}} \equiv \mathcal{M}_2^{\mathcal{D}_0}$. Assume, for simplicity, $\eta_1 \perp\!\!\!\perp \eta_2$ under $\mathcal{M}^{\mathcal{D}}$. In the causal interpretation of the model, the association represented by the arrow $1 \longleftarrow 2$ is truly asymmetric in the sense that no feedback relationship is present. It follows that prior beliefs on η_2 conditional on $\mathcal{M}^{\mathcal{D}}$ must be the same as prior beliefs on η_2 conditional on $\mathcal{M}_2^{\mathcal{D}}$ as well as on $\mathcal{M}_2^{\mathcal{D}_0}$. Accordingly, the prior distribution for η_2 should be the same under $\mathcal{M}^{\mathcal{D}}$ and $\mathcal{M}^{\mathcal{D}_0}$ (this property is named prior modularity by Heckerman et al., 1995) and the Bayes factor to compare $\mathcal{M}_2^{\mathcal{D}}$ and $\mathcal{M}_2^{\mathcal{D}_0}$ should be identically one. In Example 1 the prior distribution for η under $\mathcal{M}^{\mathcal{D}}$ is such that $\eta_1 \perp\!\!\!\perp \eta_2$ with $\eta_2 \sim a_{22}/\chi^2_\delta$ (see Dawid and Lauritzen, 2000); nevertheless the application of Jeffreys conditioning to $\mathcal{M}^{\mathcal{D}}$ and $\mathcal{M}^{\mathcal{D}_0}$ modifies the distribution of*

η_2 to $a_{22}/\chi^2_{\delta+1}$, see point iii) of the same example. In this case the Bayes factor to compare $\mathcal{M}_2^{\mathcal{D}}$ and $\mathcal{M}_2^{\mathcal{D}_0}$ would be in favour of one of the two models, thereby reflecting prior discrepancies merely due to the elicitation procedure rather than actual prior belief.

The unsatisfactory results deriving from the application of Jeffreys conditioning in DAG models can be regarded as a consequence of the fact that the prior information on variable well-ordering, which in turn gives rise to the grouping of η into $(\eta_1, \ldots, \eta_v)$, does not enter in the definition of the procedure. This suggests to revert to a conditioning approach whose baseline measure takes into explicit consideration the above parameter grouping, and this naturally leads to the notion of *ordered group reference prior*.

Reference priors were introduced by Bernardo (1979) and generalised to multiparameter problems by Berger and Bernardo (1992), see also Bernardo and Smith (1994) for the definition of reference priors and a description of the procedure to derive them. Here we wish to emphasise that a reference prior for a parameter θ is derived with respect to an *ordered grouping* of the elements of $\theta \in \Theta$. The algorithm for constructing reference priors typically requires to specify a nested sequence of compact subsets of Θ whose union is Θ itself. Such an algorithm is greatly simplified if the posterior distribution of θ is asymptotically normal, the so-called "regular" case. Furthermore if the group-components of θ are variation-independent (suggesting to choose cartesian products of subspaces for the nested sequence above) and the Fisher information matrix is block diagonal, as in the case of the modular components of the η parameterisation for DAG models, then the ordering of the grouping is irrelevant. Accordingly we shall employ the term *group reference measure* for η and denote its density with respect to Lebesgue measure by $r(\eta)$. Datta and Ghosh (1996) showed that such a measure is invariant with respect to the class of reparameterisations in (4); that is the group reference measure for η is invariant within the class of allowable parameterisations for $\mathcal{M}^{\mathcal{D}}$.

Along with Jeffreys conditioning, we can now define a new approach, named *reference conditioning*, by imposing that ν and ν_0 be group reference measures.

Let $\mathcal{M}^{\mathcal{D}}$ be a DAG model with $\mathcal{D} = (V, E)$ and parameter η and let $\mathcal{M}^{\mathcal{D}_0}$ be a submodel with $\mathcal{D}_0 = (V, E_0)$, $E_0 \subset E$, and parameter η_0. If π is the density with respect to Lebesgue measure of the prior distribution Π under $\mathcal{M}^{\mathcal{D}}$ for η, then the distribution for η_0 resulting from the application of reference conditioning has density function with respect to Lebesgue measure

$$\pi_0(\eta_0) \propto \pi(\eta_0) \frac{r_0(\eta_0)}{r(\eta_0)}. \tag{6}$$

Remark Reference conditioning may be defined with respect to any allowable parameterisation θ of $\mathcal{M}^{\mathcal{D}}$. However, if the blocks of θ are not variation independent the reference measure has to be derived with respect to the ordered grouping $(\theta_{j_1}, \ldots \theta_{j_v})$ of Proposition 1. For instance, if in Example 2 the parameterisation Σ is used, then the reference measure has to be computed with respect to the ordered groups $\theta_1 = \sigma_{22}$ and $\theta_2 = (\sigma_{11}, \sigma_{12})$ for $\mathcal{M}^{\mathcal{D}}$, but with respect to ordered groups $\theta_1 = \sigma_{11}$ and $\theta_2 = (\sigma_{22}, \sigma_{12})$ for $\mathcal{M}^{\mathcal{D}^*}$.

Suppose there exists an allowable parameterisation $\theta = \theta(\eta)$ such that the ratio of the reference measures for θ under $\mathcal{M}^{\mathcal{D}}$ and $\mathcal{M}^{\mathcal{D}_0}$ is constant for all $\theta = \theta(\eta_0)$; then reference conditioning gives the same answer as usual conditioning. In this case we say that θ is a parameterisation *implicitly leading to reference conditioning* for $\mathcal{M}^{\mathcal{D}}$ and $\mathcal{M}^{\mathcal{D}_0}$.

We close this section by noticing that if no well-ordering of the vertices is specified, as for example in undirected graphical models, then the model parameter can be considered as a single block. Accordingly the group reference measure turns out to be the Jeffreys measure: thus Jeffreys conditioning becomes a special case of reference conditioning.

5 Gaussian DAG models

We now specialise the general theory described above with respect to the important class of Gaussian DAG models. A Gaussian DAG model is specified through

(3), wherein each local conditional -or regression- family is Gaussian. The i-th family, $i = 1, \ldots, v$, is typically parameterised by

$$\eta_i = (\beta_i, \sigma_{ii \cdot \mathrm{pa}(i)}), \tag{7}$$

with $\beta_i = (\beta_{ij}, j \in \mathrm{pa}(i))$ representing the regression coefficients. Clearly the local parameters η_is are variation independent. The rest of this paper is entirely devoted to this case, so that in the following $\mathcal{M}_i^{\mathcal{D}}$ will denote the i-th local regression family, parameterised by (7), and $\mathcal{M}^{\mathcal{D}}$ the overall Gaussian DAG model. We remark that any allowable parameterisation $\theta_i = (\theta_{ij}; j \in \mathrm{fa}(i))$, $i = 1, \ldots, v$, such that θ_i parameterises $\mathcal{M}_i^{\mathcal{D}}$ and the θ_is are variation independent, would be equivalent for the development of the theory in this paper. One such parameterisation, which plays a key role here, is given by $\phi = (\phi_1, \ldots, \phi_v)$ with

$$\phi_{ii} = \frac{1}{\sqrt{\sigma_{ii \cdot \mathrm{pa}(i)}}} \quad \text{and} \quad \phi_{ij} = -\frac{\beta_{ij}}{\sqrt{\sigma_{ii \cdot \mathrm{pa}(i)}}}, \quad j \in \mathrm{pa}(i), \tag{8}$$

with the understanding that $(v, v-1, \ldots, 1)$ is a well-ordering of the variables so that $\mathrm{pa}(i) \subseteq \{i+1, \ldots, v\}$. If Φ denotes the upper triangular matrix with entries ϕ_{ij} and zero elsewhere, then $\Phi^T \Phi = \Sigma^{-1}$, which represents the Cholesky decomposition of Σ^{-1} (see Wermuth, 1980; Wermuth and Cox, 2000 and Roverato, 2000, 2001). For an interpretation of the elements ϕ_{ij} notice that β_{ij} and β_{rj} represent the unit-change effect of X_j on variables X_i and X_r respectively. The standardised versions ϕ_{ij} and ϕ_{rj} allow a direct comparison of these effects since they are expressed on the same scale.

With reference to the bivariate case, Example 2 has already provided instances of allowable, and not allowable, parameterisations. These results are extended and generalised below.

Consider the class $[\mathcal{D}]$ of DAGs Markov equivalent to $\mathcal{D}$. In this case $\mathcal{M}^{\mathcal{D}}$ is distribution equivalent to $\mathcal{M}^{\mathcal{D}^*}$, for all $\mathcal{D}^* \in [\mathcal{D}]$. Furthermore, for all $\mathcal{D}^* \in [\mathcal{D}]$, Σ, Σ^{-1}, P and R are all bijective transformations of the parameterisation η^* of $\mathcal{M}^{\mathcal{D}^*}$.

Remark When $\mathcal{D}$ is not complete, Σ has free entries σ_{ij} with $j \in \mathrm{fa}(i)$, while all the remaining entries are functions of these. The same is true for Σ^{-1}, P

and R.

Proposition 2 *Let $\mathcal{M}^{\mathcal{D}}$ be a Gaussian DAG model with DAG $\mathcal{D} = (V, E)$. Σ, Σ^{-1}, and P belong to the class of allowable parameterisations of every $\mathcal{M}^{\mathcal{D}^*}$ such that $\mathcal{D}^* \in [\mathcal{D}]$.*

Proof. See the Appendix. □

On the other hand, Example 2 makes clear that in general the matrix R does not belong to the class of allowable parameterisations of any $\mathcal{M}^{\mathcal{D}^*}$ with $\mathcal{D}^* \in [\mathcal{D}]$, and that for all $\mathcal{D}^* \in [\mathcal{D}] \backslash \{\mathcal{D}\}$, the standard parameterisation η^* is not allowable for $\mathcal{M}^{\mathcal{D}}$. As a consequence, starting from two Markov equivalent DAG models with the same parameter prior distribution, the reference conditioning approach may lead to different compatible priors for the parameter of a common submodel.

5.1 Reference conditioning

For Gaussian DAG models the method of reference conditioning is easily implementable using the allowable parameterisation ϕ described in (8).

For a complete DAG model the Fisher information as well as ordered group reference prior for ϕ can be easily deduced from that of Consonni and Veronese (2001). These results can be extended to an arbitrary DAG model as follows.

Theorem 3 *Let $\mathcal{M}^{\mathcal{D}}$ be a Gaussian DAG model with DAG $\mathcal{D} = (V, E)$. Relative to the parameterisation ϕ described in (8):*

i) the Fisher information is block-diagonal with the i-th block given by

$$H_{ii}(\phi) \;=\; \Psi^i \begin{pmatrix} 2 & 0 \\ 0 & I \end{pmatrix} \Psi^{iT},$$

where Ψ^i is the unique upper triangular matrix with positive diagonal such that $\Sigma_{\mathrm{fa}(i),\,\mathrm{fa}(i)} = \Psi^i (\Psi^i)^T$ and I is the identity matrix with dimension equal to the cardinality of the set $\mathrm{pa}(i)$.

ii) The group reference measure for the parameterisation ϕ has density with respect to Lebesgue measure

$$r(\phi) = \prod_{i=1}^{v} \frac{1}{\phi_{ii}}.$$

Proof. See the Appendix. □

As an immediate consequence of Theorem 3 we obtain.

Corollary 4 *Let $\mathcal{M}^{\mathcal{D}}$ be a Gaussian DAG model with $\mathcal{D} = (V, E)$ and $\mathcal{M}^{\mathcal{D}_0}$ a submodel with $\mathcal{D}_0 = (V, E_0)$ and $E_0 \subset E$. Then ϕ is a parameterisation implicitly leading to reference conditioning for $\mathcal{M}^{\mathcal{D}}$ and $\mathcal{M}^{\mathcal{D}_0}$.*

Proof. We have to prove that $r_0(\phi_0)/r(\phi_0)$ is constant for all ϕ_0, and this is obviously true because

$$\frac{r_0(\phi_0)}{r(\phi_0)} \propto \frac{\prod_{i=1}^{v} \phi_{0\,ii}}{\prod_{i=1}^{v} \phi_{0\,ii}} = 1.$$

□

By Corollary 4, reference conditioning can be straightforwardly applied within the ϕ-parameterisation by simply conditioning on the event $\{\phi_{ij} = 0; (i, j) \in E \backslash E_0\}$.

In DAG modelling it is common practice to specify prior distributions which satisfy the property of global parameter independence, see Spiegelhalter and Lauritzen (1990). In our context this amounts to assuming that the blocks $\phi_1, \ldots, \phi_v$ of ϕ are mutually stochastically independent, so that reference conditioning can be performed separately for each block ϕ_i. Trivially, the resulting compatible distribution for the parameter of a submodel will also satisfy the global parameter independence property. Moreover, if a local regression $\mathcal{M}_i^{\mathcal{D}}$ of $\mathcal{M}^{\mathcal{D}}$ is unchanged in the submodel, then reference conditioning leaves the corresponding parameter prior also unchanged, so that prior modularity, see Heckerman *et al.* (1995), is automatically satisfied.

The inverse Wishart distribution represents the most used prior distribution for the parameter Σ of a complete Gaussian DAG model. In this case the distribution of ϕ is easily derived and reference conditioning can be routinely applied

using Corollary 4, thereby providing a practical method for assigning parameter priors to candidate DAG models *via* a small number of direct assessments. A related but different procedure is described in Geiger and Heckerman (1999) for DAG models having no causal interpretation.

Example 4 *For the problem in Example 2 let* $\Sigma \sim IW(\delta, A)$ *as in Example 1. It can be easily checked from standard result for the Wishart distribution (see Muirhead, 1982, Theorem 3.2.10) that* $(\phi_{11}, \phi_{12}) \perp\!\!\!\perp \phi_{22}$ *and*

$$\begin{aligned} \phi_{11}^2 &\sim \chi^2_{\delta+1}/a_{11\cdot 2} \\ \phi_{12}|\phi_{11} &\sim N\left(-\phi_{11}a_{12}/a_{22},\ a_{22}^{-1}\right) \\ \phi_{22}^2 &\sim \chi^2_{\delta}/a_{22} \end{aligned}$$

where $a_{11\cdot 2} = a_{11} - a_{12}^2/a_{22}$. *Applying reference conditioning, i.e. conditioning on* $\{\phi_{12} = 0\}$, *one obtains*

$$\sigma_{11} \sim a_{11}/\chi^2_{\delta+1}, \quad \sigma_{22} \sim a_{22}/\chi^2_{\delta}, \qquad \textit{independently}$$

which is different from any of the results obtained in Example 1. In particular, unlike the result iii) arising from Jeffreys conditioning, the distribution of $\sigma_{22} = 1/\phi_{22}^2$ *is the same under the two models.*

5.2 Undirected Gaussian graphical models

For an undirected graph $\mathcal{G}$ with vertex set V, we denote by $\mathcal{M}^{\mathcal{G}}$ the Gaussian graphical model with graph $\mathcal{G}$, that is the family of v-dimensional Gaussian distributions, with zero mean, satisfying the pairwise Markov property with respect to $\mathcal{G}$ (Dempster, 1972; Wermuth, 1976). If $\mathcal{D}$ and $\mathcal{D}_0$ are perfect DAGs then $\mathcal{G} = \mathcal{D}^{\sim}$ and $\mathcal{G}_0 = \mathcal{D}_0^{\sim}$ are decomposable undirected graphs. Furthermore $\mathcal{G}$ and $\mathcal{D}^{\sim}$, as well as $\mathcal{G}_0$ and $\mathcal{D}_0^{\sim}$, are Markov equivalent, so that $\mathcal{M}^{\mathcal{D}} \equiv \mathcal{M}^{\mathcal{G}}$ and $\mathcal{M}^{\mathcal{D}_0} \equiv \mathcal{M}^{\mathcal{G}_0}$, and thus the only difference between the two types of models is that in the directed case vertex ordering is of relevance. When no well-ordering is specified, as in undirected graphical models, Jeffreys conditioning appears more

natural, since no constraint on the set of allowable parameterisations seems sensible. As a consequence it is of some interest to generalise Example 1 and see which is the corresponding implicit parameterisation. The following Proposition provides a partial answer to this problem.

Proposition 5 *Let $\mathcal{G}$ be a complete graph and $\mathcal{G}_0$ the graph obtained from $\mathcal{G}$ by removing exactly one edge. Consider the comparison of the undirected Gaussian graphical models $\mathcal{M}^{\mathcal{G}}$ and $\mathcal{M}^{\mathcal{G}_0}$. Then R is a parameterisation implicitly leading to Jeffreys conditioning.*

Proof. See the Appendix. □

6 Conclusions

We argue in this paper that (causal) DAG models encapsulate a modular structure given by the local conditional distributions of nodes given parents together with a well-ordering of the variables. As a consequence, unlike other statistical models, DAG models cannot be arbitrarily reparameterised; indeed we define a class of allowable parameter transformations, which essentially preserves the original modular structure. Building on work of Dawid and Lauritzen (2000), we propose a new method to construct compatible prior distributions for the parameters of DAG models, which we name reference conditioning. This differs from Jeffreys conditioning, since the baseline measure is provided by an order-invariant group reference prior. Our method is more general than Jeffreys conditioning and reduces to it when no restriction is imposed on the set of allowable parameter transformations, save for the usual smoothness assumptions, as for example with undirected graphical models. An interesting byproduct of our research is the identification of parameterisations implicitly leading to reference conditioning. We discuss the issue in detail for Gaussian DAG models. Using the parameterisation induced by the Cholesky decomposition of the concentration matrix, we show that reference conditioning can be performed through usual conditioning

simply setting to zero the components of the parameter corresponding to arrows that are absent in the underlying DAG.

Our approach only requires the specification of a prior distribution for the parameter of the most complex (possibly complete) model; prior distributions for the parameters of each submodel are then determined by usual conditioning as described above. In particular, for Gaussian DAG models, we show that if the starting prior satisfies the property of global parameter independence, this property propagates through reference conditioning, thus facilitating prior-to-posterior inference under complete sampling. Furthermore parameter modularity is automatically satisfied.

A natural extension of our methodology is to discrete Bayesian networks, followed by chain graphs, and possibly mixed graphical models, containing both discrete and continuous nodes. We hope to report on these developments in the near future.

Acknowledgements

Guido Consonni's research was partially supported by the University of Pavia research project on the Statistical Analysis of Complex Systems.

Appendix: Proofs

Proof of Proposition 2

Let $\mathcal{D}^*$ be any DAG in the set of Markov equivalent DAGs $[\mathcal{D}]$ and let η^* denote the standard parameterisation of $\mathcal{M}^{\mathcal{D}^*}$. We denote by ϕ^* the alternative parameterisation of $\mathcal{M}^{\mathcal{D}^*}$ defined by (8) and by $\Phi^* = \{\phi^*_{ij}\}$ the matrix obtained from ϕ^* as described in Section 5, so that $\Sigma^{*-1} = \Phi^{*T}\Phi^{*T}$. It is worth pointing out that, although Σ^{-1} is the same for all the DAGs in $[\mathcal{D}]$, by deriving it from Φ^* we need to specify some well-ordering $(\alpha_1, \ldots, \alpha_v)$ of the vertices of $\mathcal{D}^*$. As a consequence, the first row and column of Σ will correspond to X_{α_v}, the second to $X_{\alpha_{v-1}}$ and so on. Two Markov equivalent DAGs differ for the well-ordering

of their vertex sets, and we add an asterisk to Σ^{-1} when row and column well-ordering is of relevance for the operation being performed, and similarly for Σ and P.

We first show that Σ^{-1} is allowable for $\mathcal{M}^{\mathcal{D}^*}$. Consider the group ordering of the unconstrained entries of Σ^{-1} given by $(\sigma^{*1}, \ldots, \sigma^{*v})$ with $\sigma^{*i} = (\sigma^{ij}, j \in \text{fa}^*(i))$ where $\text{fa}^*(i)$ denotes the family of i with respect to $\mathcal{D}^*$. Note that σ^{*i} is made up of the unconstrained entries of the i-th row of Σ^{*-1}. It can be easily checked that the first row of Σ^{*-1} is a function of the first row of Φ^*, namely $\sigma^{*1} = \sigma^{*1}(\phi_1^*)$, the second row of Σ^{*-1} is a function of the first two rows of Φ^*, $\sigma^{*2} = \sigma^{*2}(\phi_1^*, \phi_2^*)$, and so on. Thus, by Proposition 1, Σ^{-1} is allowable for $\mathcal{M}^{\mathcal{D}^*}$.

We proceed in a similar way for Σ. Consider the grouping $(\sigma_1^*, \ldots, \sigma_v^*)$ with $\sigma_i^* = (\sigma_{ij}, j \in \text{fa}^*(i))$ so that σ_i^* is made up of the unconstrained entries of the i-th row of Σ^*. Recalling that, for all $i = 1, \ldots, v$, if $B = \{i, i+1, \ldots, v\}$ then $\Sigma_{BB}^* = \left(\Phi_{BB}^{*T} \Phi_{BB}^*\right)^{-1}$ (see Roverato, 2000) we obtain that σ_i^* is a function of the last i rows of Φ^*, that is $\sigma_i^* = \sigma_i^*(\phi_i^*, \phi_{i+1}^*, \ldots, \phi_v^*)$. As a consequence Σ is allowable for $\mathcal{M}^{\mathcal{D}^*}$. To show that also P is allowable for $\mathcal{M}^{\mathcal{D}^*}$ we put $(\rho_1^*, \ldots, \rho_v^*)$ with $\rho_i^* = (\sigma_{ii}, \rho_{ij}; j \in \text{pa}^*(i))$. It is easy to check that $\rho_i^* = \rho_i^*(\sigma_i^*, \sigma_{i+1}^*, \ldots, \sigma_v^*)$ for $i = 1, \ldots, v$, so that ρ_i^* is itself a function of $(\phi_i^*, \phi_{i+1}^*, \ldots, \phi_v^*)$ and therefore P is allowable for $\mathcal{M}^{\mathcal{D}^*}$ by Proposition 1.

Proof of Theorem 3

i) The parameter ϕ is made up of variation independent groups $(\phi_1, \ldots, \phi_v)$ and it is clear from (3) that the Fisher information matrix for ϕ is block-diagonal with i-th block given by

$$H_{ii}(\phi) = -E\left\{\frac{\partial^2 \log p(X_i | X_{\text{pa}(i)}, \phi_i)}{\partial \phi_i^T \partial \phi_i}\right\}, \tag{9}$$

where $X_{\text{fa}(i)} \sim N(0, \Sigma_{\text{fa}(i),\text{fa}(i)})$ and $p(\cdot|\cdot)$ denotes density function of the conditional distribution of X_i given $X_{\text{pa}(i)}$. Therefore, the computation of the i-th block of $H(\phi)$ only involves the distribution of $X_{\text{fa}(i)}$. More precisely, $p(x_i | x_{\text{pa}(i)}, \phi_i)$ in (9) is the density of the first local regression of any complete

DAG model for $X_{\mathrm{fa}(i)}$ with X_i as pure response variable and $H_{ii}(\phi)$ can be computed locally with respect to such a complete DAG model.

Let $\mathcal{D}^i = (\,\mathrm{fa}(i), E^i)$ be a complete DAG such that i is the first vertex, that is the set of parents of i in $\mathcal{D}^i$ is $\mathrm{pa}(i)$, and assume that the rows and columns of $\Sigma^{-1}_{\mathrm{fa}(i),\,\mathrm{fa}(i)}$ are ordered according to the vertex ordering. In this way, the upper-triangular matrix Φ^i obtained from the Cholesky decomposition $\Sigma^{-1}_{\mathrm{fa}(i),\,\mathrm{fa}(i)} = (\Phi^i)^T\Phi^i$ provides the ϕ-parameterisation of $\mathcal{M}^{\mathcal{D}^i}$ and the first row of Φ^i is ϕ_i as in (9).

We first give the Fisher information matrix for the inverse variance, $H(\Sigma^{-1}_{\mathrm{fa}(i),\,\mathrm{fa}(i)})$, and derive the Fisher information for Φ^i by using the relation $H(\Phi^i) = J^T H(\Sigma^{-1})J$ where J is the Jacobian of the transformation from $\Sigma^{-1}_{\mathrm{fa}(i),\,\mathrm{fa}(i)}$ to Φ^i. In $H(\Sigma^{-1}_{\mathrm{fa}(i),\,\mathrm{fa}(i)})$ we take the distinct elements of $\Sigma^{-1}_{\mathrm{fa}(i),\,\mathrm{fa}(i)}$ ordered according to the rows of $\Sigma^{-1}_{\mathrm{fa}(i),\,\mathrm{fa}(i)}$ and similarly for $H(\Phi^i)$. Thereby, $H(\Phi^i)$ is block-diagonal and its first block is $H_{ii}(\phi)$. We make use of the theory related to the duplication matrix D_p, the commutation matrix K_{pp} and the elimination matrix L_p, where $p = |\,\mathrm{fa}(i)|$. We refer to Lütkepohl (1996) for the definition of such matrices and a description of their properties.

The Fisher information matrix for $\Sigma^{-1}_{\mathrm{fa}(i),\,\mathrm{fa}(i)}$ has form

$$H\left(\Sigma^{-1}_{\mathrm{fa}(i),\,\mathrm{fa}(i)}\right) = \frac{1}{2}D_p^T\left(\Sigma_{\mathrm{fa}(i),\,\mathrm{fa}(i)} \otimes \Sigma_{\mathrm{fa}(i),\,\mathrm{fa}(i)}\right)D_p$$

where $\otimes$ denotes the Kronecker product (see Gutierrez-Peña and Smith, 1995, equation (18)), and can be factorised as

$$\begin{aligned} H\left(\Sigma^{-1}_{\mathrm{fa}(i),\,\mathrm{fa}(i)}\right) &= \frac{1}{2}D_p^T(\Psi^i\Psi^{iT} \otimes \Psi^i\Psi^{iT})D_p \\ &= \frac{1}{2}D_p^T\left[(\Psi^i \otimes \Psi^i)(\Psi^{iT} \otimes \Psi^{iT})\right]D_p \\ &= \frac{1}{2}\left[D_p^T(\Psi^i \otimes \Psi^i)\right]\left[D_p^T(\Psi^i \otimes \Psi^i)\right]^T \end{aligned}$$

where $\Psi^i = (\Phi^i)^{-1}$.

The Jacobian J of the transformation $\Sigma^{-1}_{\mathrm{fa}(i),\,\mathrm{fa}(i)} \to \Phi^i$ is given in Lütkepohl (1996, §10.5.4 (3)) and has form

$$J = 2D_p^+(\Phi^{iT} \otimes I_p)L_p^T$$

where I_p is the $p \times p$ identity matrix and $D_p^+ = (D_p^T D_p)^{-1} D_p^T$. Making use of some known matrix algebra results (see Lütkepohl, 1996, §2.4 (5), §9.2.1 (16), §9.2.2 (2) and §9.5.2 (1)) we have that

$$\begin{aligned} J^T \left[D_p^T (\Psi^i \otimes \Psi^i) \right] &= \left[2D_p^+ (\Phi^{iT} \otimes I_p) L_p^T \right]^T \left[D_p^T (\Psi^i \otimes \Psi^i) \right] \\ &= L_p (\Phi^i \otimes I_p)(I_{p^2} + K_{pp})(\Psi^i \otimes \Psi^i) \\ &= L_p (\Phi^i \otimes I_p)(\Psi^i \otimes \Psi^i)(I_{p^2} + K_{pp}) \\ &= L_p (\Phi^i \Psi^i \otimes I_p \Psi^i)(I_{p^2} + K_{pp}) \\ &= L_p (I_p \otimes \Psi^i)(I_{p^2} + K_{pp}) \\ &= 2L_p (I_p \otimes \Psi^i) D_p D_p^+. \end{aligned}$$

Hence, we can write

$$\begin{aligned} H(\Phi^i) &= J^T H \left(\Sigma^{-1}_{\mathrm{fa}(i),\,\mathrm{fa}(i)} \right) J \\ &= M \left(2D_p^+ D_p^{+T} \right) M^T \end{aligned}$$

where $M = L_p (I_p \otimes \Psi^i) D_p$. It can be checked that M is block-diagonal and that its first block, corresponding to ϕ_i, is Ψ^i. The matrix $D_p^+ D_p^{+T}$ is described in Lütkepohl (1996 §9.5.1 (11)) and it is such that the submatrix of $H(\Phi^i)$ corresponding to ϕ_i can be written as

$$H_{ii}(\phi) = \Psi^i \begin{pmatrix} 2 & 0 \\ 0 & I_{p-1} \end{pmatrix} \Psi^{iT}. \tag{10}$$

ii) To construct the reference prior on $\phi = (\phi_1, \ldots, \phi_v)$ we make use of a result in Datta and Ghosh (1995). Assume that the Fisher information matrix for ϕ is block-diagonal, $H(\phi) = \mathrm{diag}\{H_{11}(\phi), H_{22}(\phi), \ldots, H_{vv}(\phi)\}$, with $H_{ii}(\phi)$ a square submatrix of dimension $\dim(\phi_i)$. If the determinant of $H_{ii}(\phi)$ can be factorised as

$$|H_{ii}(\phi)| = a_i(\phi_i) b_i(\phi_1, \ldots, \phi_{i-1}, \phi_{i+1}, \ldots, \phi_v) \qquad i = 1, \ldots, v \tag{11}$$

for some positive functions $a_i(\cdot)$ and $b_i(\cdot)$, then the ordered group reference prior relative for ϕ does not depend on the order of the v groups and has density with respect to Lebesgue measure $r(\phi) \propto \prod_{i=1}^{v} a_i(\phi_i)^{1/2}$.

We now prove that the determinant of $H_{ii}(\phi)$ in (10) can be factorised as in (11) with $a_i(\phi_i) = 1/\phi_{ii}^2$ so that

$$r(\phi) \propto \prod_{i=1}^{v} a_i(\phi_i)^{1/2} = \prod_{i=1}^{v} \frac{1}{\phi_{ii}}. \tag{12}$$

Now partition Ψ^i and Φ^i as

$$\Psi^i = \begin{pmatrix} \psi^i_{11} & \Psi^i_{\mathrm{pa}(i)} \\ 0 & \Psi^i_{\mathrm{pa}(i),\,\mathrm{pa}(i)} \end{pmatrix} \quad \text{and} \quad \Phi^i = \begin{pmatrix} \phi^i_{11} & \Phi^i_{\mathrm{pa}(i)} \\ 0 & \Phi^i_{\mathrm{pa}(i),\,\mathrm{pa}(i)}. \end{pmatrix}$$

The determinant of $H_{ii}(\phi)$ can now be written as

$$|H_{ii}(\phi)| \;=\; 2(\psi^i_{11})^2 |\Psi^i_{\mathrm{pa}(i),\,\mathrm{pa}(i)}|^2,$$

and we show that $\Psi^i_{\mathrm{pa}(i),\,\mathrm{pa}(i)}$ is a function of $(\phi_{i+1}, \ldots, \phi_v)$ whereas ψ^i_{11} is a function of ϕ_i.

If we put $\mathrm{pr}(i) = \{i+1, \ldots, v\}$, then $\mathrm{pa}(i) \subseteq \mathrm{pr}(i)$ and $\Sigma_{\mathrm{pa}(i),\,\mathrm{pa}(i)} = \Psi^i_{\mathrm{pa}(i),\,\mathrm{pa}(i)} (\Psi^i_{\mathrm{pa}(i),\,\mathrm{pa}(i)})^T$ is a submatrix of $\Sigma_{\mathrm{pr}(i),\,\mathrm{pr}(i)}$. Hence $\Psi^i_{\mathrm{pa}(i),\,\mathrm{pa}(i)}$ is a function of $\Sigma_{\mathrm{pr}(i),\,\mathrm{pr}(i)}$. Let $\Phi = \{\phi_{ij}\}$ denote the matrix, defined in Section 5, made up of the elements of ϕ so that $\Sigma^{-1} = \Phi^T \Phi$. It is not difficult to check that, because of the upper-triangular form of Φ, $\Sigma^{-1}_{\mathrm{pr}(i),\,\mathrm{pr}(i)} = \Phi^T_{\mathrm{pr}(i),\,\mathrm{pr}(i)} \Phi_{\mathrm{pr}(i),\,\mathrm{pr}(i)}$ for all $i = 1, \ldots, v$ (see also Roverato, 2000). Thus $\Sigma_{\mathrm{pr}(i),\,\mathrm{pr}(i)}$ is a function of $(\phi_{i+1}, \ldots, \phi_v)$ and the same is true for $\Psi^i_{\mathrm{pa}(i),\,\mathrm{pa}(i)}$.

Because of the upper-triangular form of $\Psi^i = (\Phi^i)^{-1}$ it holds that $\psi^i_{11} = 1/\phi^i_{11}$. By standard results on the multivariate normal distribution (see Whittaker, 1990 p.143) $(\phi^i_{11})^2 = \left\{\Sigma^{-1}_{\mathrm{fa}(i),\,\mathrm{fa}(i)}\right\}_{11} = 1/\sigma_{ii\cdot\,\mathrm{pa}(i)}$. Since, by (8), $\phi_{ii} = 1/\sqrt{\sigma_{ii\cdot\,\mathrm{pa}(i)}}$, it turns out that $\phi^i_{11} = \phi_{ii}$ and $\psi^i_{11} = 1/\phi_{ii}$.

We can conclude that, with respect to (11), we can put $a_i(\phi_i) = 1/\phi_{ii}^2$ and $b_i(\phi_{i+1}, \ldots, \phi_v) = |\Psi^i_{\mathrm{pa}(i),\,\mathrm{pa}(i)}|^2$ from which (12) follows.

Proof of Proposition 5

Let $\mathcal{G}_1 = (V, E_1)$ be a complete graph and $\mathcal{G}_0 = (V, E_0)$ the graph obtained from $\mathcal{G}_1$ by removing edge (r, s). Moreover, let $j_1(R_1) = |H_1(R_1)|^{1/2}$ and $j_0(R_0) =$

$|H_0(R_0)|^{1/2}$ be the densities of the Jeffrey measures for the parameters R_1 of $\mathcal{M}^{\mathcal{G}_1}$ and R_0 of $\mathcal{M}^{\mathcal{G}_0}$ respectively. We have to show that $j_0(R_0)/j_1(R_0)$, or equivalently $|H_0(R_0)|/|H_1(R_0)|$, is constant for all R_0.

The Fisher information matrix for the parameter R of the Gaussian graphical model $\mathcal{M}^{\mathcal{G}}$ with arbitrary undirected graph $\mathcal{G} = (V, E)$, can be derived from the Fisher information matrix for the parameter $\Sigma^{-1} = \{\sigma^{ij}\}$ of the model by using the relation $H(R) = J^T H(\Sigma^{-1})J$ where J is the Jacobian matrix of the transformation from Σ^{-1} to R. Consequently, $|H(R)| \propto |J|^2|H(\Sigma^{-1})|$.

We first consider $|J|$. If we order the distinct nonzero entries of Σ^{-1} by taking first the diagonal elements and then the off-diagonal elements, and similarly for R, the Jacobian matrix $J = \frac{\partial}{\partial R}\Sigma^{-1}$ is triangular and its determinant is the product of the diagonal elements. Since for $i = 1, \ldots, v$ $\{\Sigma^{-1}\}_{ii} = \{R\}_{ii}$ and for $(i, j) \in E$, $\{\Sigma^{-1}\}_{ij} = -\rho_{ij\cdot V\setminus\{i,j\}}\sqrt{\sigma^{ii}\sigma^{jj}}$, the diagonal elements of J are $\frac{\partial}{\partial\{R\}_{ii}}\sigma^{ii} = 1$ for $i = 1, \ldots, v$ and $\frac{\partial}{\partial\{R\}_{ij}}\sigma^{ij} = -\sqrt{\sigma^{ii}\sigma^{jj}}$ for $(i, j) \in E$. Therefore, $|J|^2 = \prod_{(i,j)\in E}\sigma^{ii}\sigma^{jj}$.

The graph $\mathcal{G}_0$ has two cliques, $C_r = V\setminus\{r\}$ and $C_s = V\setminus\{s\}$, and one separator, $S = V\setminus\{r, s\}$, and the determinants of $H_1(\Sigma_1^{-1})$ and $H_0(\Sigma_0^{-1})$ are

$$|H_1(\Sigma_1^{-1})| \propto |\Sigma_1|^{v+1} \quad \text{and} \quad |H_0(\Sigma_0^{-1})| \propto \frac{|\Sigma_{C_rC_r}|^v|\Sigma_{C_sC_s}|^v}{|\Sigma_{SS}|^{v-1}}$$

respectively (Roverato and Whittaker, 1998).

Recalling that $|\Sigma_0| = |\Sigma_{C_rC_r}||\Sigma_{C_sC_s}||\Sigma_{SS}|^{-1}$ (see Lauritzen, 1996 p.145), $|\Sigma_{C_rC_r}| = \sigma_{rr\cdot S}|\Sigma_{SS}|$ and $|\Sigma_{C_sC_s}| = \sigma_{ss\cdot S}|\Sigma_{SS}|$ (see Lauritzen, 1996 equation (B.1)) and that $\sigma_{rr\cdot S} = 1/\sigma_0^{rr}$ and $\sigma_{ss\cdot S} = 1/\sigma_0^{ss}$ we obtain

$$\begin{aligned}\frac{|H_0(R_0)|}{|H_1(R_0)|} &\propto \frac{\prod_{(i,j)\in E_0}\sigma_0^{ii}\sigma_0^{jj}}{\prod_{(i,j)\in E_1}\sigma_0^{ii}\sigma_0^{jj}} \times \frac{|\Sigma_{SS}|^2}{|\Sigma_{C_rC_r}||\Sigma_{C_sC_s}|} \\ &= \frac{1}{\sigma_0^{rr}\sigma_0^{ss}} \times \frac{1}{\sigma_{rr\cdot S}\sigma_{ss\cdot S}} \\ &= 1\end{aligned}$$

and the proof is complete.

References

Andersson, S.A., Madigan, D. and Perlman, M.D. (1997). On the Markov equivalence of chain graphs, undirected graphs, and acyclic digraphs. *Scand. J. Statist.*, **24**, 81-102.

Berger, J.O. and Bernardo, J.M. (1992). On the development of the reference prior method. In *Bayesian Statistics 4*, J.M. Bernardo, J.O. Berger, D.V. Lindley, and A.F.M. Smith (eds). Oxford University Press, London.

Berger, J.O. and Pericchi, L. (2000). Objective Bayesian methods for model selection: introduction and comparison (with discussion). In *Model selection*, P. Lahiri (ed.). IMS Lecture Notes. To appear

Bernardo, J.M. (1979). Reference posterior distributions for Bayesian inference. *J. Roy. Statist. Soc. Ser. B*, 41, 113-147 (with discussion).

Bernardo, J.M. and Smith, A.F.M. (1994). *Bayesian Theory.* Chichester: Wiley.

Consonni, G. and Veronese, P. (2001). Conditionally reducible natural exponential families and enriched conjugate priors, *Scand. J. Statist.*, **28**, 377-406.

Consonni, G., Veronese, P. and Gutiérrez-Peña E. (2000). Order-invariant group reference priors for natural exponential families having a simple quadratic variance function. Technical Report, University of Pavia, Italy. Submitted.

Cox, D.R. and Wermuth, N. (1996). *Multivariate Dependencies - Models, analysis and interpretation*, Chapman and Hall, London.

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

Datta, G.S., and Ghosh, M. (1995). Some remarks on non-informative priors. *J. Am. Statist. Assoc.*, **90**, 1357-63.

Datta, G.S., and Ghosh, M. (1996). On the invariance of noninformative priors. *Ann. Statist.*, **24**, 141-59.

Dawid, A.P. (1981). Some matrix-variate distribution theory: notational considerations and a Bayesian application, *Biometrika*, **68**, 265-74.

Dawid, A.P. and Lauritzen, S.L. (2000). Compatible prior distributions. ISBA2000 Proceedings, ISBA and Eurostat, 2001, to appear.

Dempster, A.P. (1972). Covariance selection, *Biometrics* **28**, 157-75.

Frydenberg, M. (1990). The chain graph Markov property, *Scand. J. Statist.*, **17**, 333-53.

Geiger, D. and Heckerman, D. (1999). Parameters priors for directed acyclic graphical models and the characterization of several probability distributions. Technical Report MSR-TR-98-67, Microsoft Research (Revised version).

Gutierrez-Peña, E. and Smith, A.F.M. (1995). Conjugate parameterizations for natural exponential families, *J. Am. Statist. Assoc.*, **90**, 1347-56.

Heckerman, D, Geiger, D. and Chickering, D. (1995). Learning Bayesian networks: the combination of knowledge and statistical data. *Machine Learning*, **20**, 197-243.

Jeffreys, H. (1961). *The Theory of Probability - Third Edition.* Oxford University Press, Oxford.

Kass, R.E. and Raftery, A.E. (1995). Bayes factors. *J. Am. Statist. Assoc.*, **90**, 773-95.

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

Lauritzen, S.L. (2001). Causal inference from graphical models. In *Complex Stochastic Systems*, (ed. O. E. Barndorff Nielsen, D. R. Cox and C Klüppelberg), pp. 63-107. Chapman and Hall/CRC Press, London/Boca Raton.

Lauritzen, S.L. and Richardson, T.S. (2001). Chain graphs models and their causal interpretation. Research Report R-01-2003, Department of Mathematical Science, Aalborg University.

Lütkepohl, H. (1996). *Handbook of Matrices*, Wiley, Chichester.

Muirhead, R.J. (1982). *Aspects of Multivariate Statistical Theory*, Wiley, New York.

Pearl, J. (2000). *Causality: Models, Reasoning and Inference.* Cambridge University Press, Cambridge.

Roverato, A. (2000). Cholesky decomposition of a hyper inverse Wishart matrix, *Biometrika*, **87**, 99-112.

Roverato, A. (2001). Hyper inverse Wishart distribution for non-decomposable graphs and its application to Bayesian inference for Gaussian graphical models. *Scand. J. Statist.*, to appear.

Roverato, A. and Whittaker, J. (1998). The Isserlis matrix and its application to non-decomposable graphical Gaussian models, *Biometrika*, **85**, 711-25.

Spiegelhalter, D. and Lauritzen, S.L. (1995). Sequential updating of conditional probabilities on directed graphical structures. *Networks*, **20**, 579-605.

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. Am. Statist. Assoc.*, **75**, 963-72.

Wermuth, N., and Cox, D.R. (2000). A sweep operator for triangular matrices and its statistical applications. Research report, ZUMA 00-04.

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