DIMACS Technical Report 2002-17
22nd April 2002

Compatible Prior Distributions for DAG models

by

Alberto Roverato[1,2]
Dipartimento di Economia Politica
University of Modena and Reggio Emilia
Modena, Italy 41100

Guido Consonni[3,4]
Dipartimento di Economia Politica e Metodi Quantitativi
University of Pavia
Pavia, Italy 27100

[1]Visitor

[2]We acknowledge support from DIMACS, Rutgers University, during the revision of this work. Partial funding was also provided by the European Union (EU-TMR project on Spatial and Computational Statistics -ERB-FMRX-CT96-0096); MIUR, Rome (COFIN01).

[3]Visitor

[4]We acknowledge support from DIMACS, Rutgers University, during the revision of this work. Partial funding was also provided by the European Union (EU-TMR project on Spatial and Computational Statistics -ERB-FMRX-CT96-0096); MIUR, Rome (COFIN01) and the University of Pavia (Research Project on the Statistical Analysis of Complex Systems).

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 new 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 parameterisations consistent with the modular structure of the DAG and derive a procedure, invariant within this class, which we name reference conditioning.

Keywords: Bayes factor; Directed acyclic graph; Fisher information matrix; Graphical model; Invariance; Jeffreys conditioning; Group reference prior; Reference conditioning; Reparameterisation.

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 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 $\mathcal{M}$. If nothing was elicited to indicate that the two priors should be different, then it is 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 construction 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 suggests two strategies, which they name "projection" and "conditioning", to choose a prior π_0 *compatible* with the given π.

The objective of this paper is to elaborate on the "conditioning approach". 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 alternatively 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 (2002). 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 $X_{i+1}, \ldots, X_v$. A related interpretation of causal DAG models is explicated in Spirtes, Glymour and Scheines (1993), Pearl (2000), and in Lauritzen (2001). In this context one distinguishes between conditioning by observation and conditioning by intervention, with the causal DAG providing the relevant formulae for the latter.

The conditioning approach of Dawid and Lauritzen (2000) rests upon the idea that a rule to construct compatible priors should be invariant to reparameterisations. This is achieved by means of a procedure which they name *Jeffreys conditioning*, since it relies on the Jeffreys measure. We stress 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. We argue, however, that invariance to *any* reparameterisation is not appropriate for (causal) DAG models, precisely because such models incorporate a modular structure which should be preserved. We therefore favour a more restrictive notion of invariance. A baseline measure consistent with our requirement of invariance is shown to be the group reference prior of Berger and Bernardo (1992), and accordingly we name our method *reference conditioning*. Our method is general and automatic and requires a single specification of the prior corresponding to the largest entertained model. In particular we show that reference conditioning is especially straightforward for 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 and characterising the class of modular parameterisations. Section 4 introduces the concept of reference conditioning which is then elaborated for Gaussian DAG models in Section 5. Finally, Section 6 offers a brief discussion of some of the issues involved.

2 Compatible priors by means of conditioning

Consider the model $\mathcal{M}$ parameterised by $\theta \in \Theta$ 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 $\theta = \theta_0 \in \Theta_0$ with $\Theta_0 \subseteq \Theta$, 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 examples of this phenomenon, known as the *Borel-Kolmogorov paradox*, in particular 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 $X_V \equiv X$ 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 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 independent of X_2, shortly $X_1 \perp\!\!\!\perp X_2$. 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) $\quad \sigma_{11} \sim a_{11}/\chi^2_{\delta+2}, \quad \sigma_{22} \sim a_{22}/\chi^2_{\delta+2}, \quad$ *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. For the theory of graphs and graphical models used in this paper we refer to Lauritzen (1996), Cowell *et. al* (1999) and Lauritzen (2001), where the definitions of undirected graph, decomposable graph, Directed Acyclic Graph (DAG), perfect DAG, undirected version of a DAG, parent, family, conditioning by intervention and conditioning by observation are provided.

Here, we denote a graph by $\mathcal{G} = (V, E)$ where V is the set of vertices and E is the set of edges and to emphasise that the graph is a DAG we write $\mathcal{D} = (V, E)$. For a DAG $\mathcal{D} = (V, E)$ and a vertex $\alpha \in V$, $\mathrm{pa}(\alpha)$ and $\mathrm{fa}(\alpha)$ denote the parents and the family of α respectively. 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 *higher* position to the vertex with *lower* position in the ordering. Note that the traditional definition of well-ordering (see Cowell *et. al*, 1999 p. 47) is reversed here. A DAG may not have a unique well-ordering.

A graphical model is a family of distributions for the vector X satisfying a set of conditional independence relations encoded by a graph. Two graphs are said to be *Markov equivalent* if they encode the same set of conditional independence assertions (Frydenberg, 1990; Verma and Pearl, 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 $\mathcal{D}$. We remark that all $\mathcal{D}^* \in [D]$ share the same undirected version $\mathcal{D}^\sim$. Furthermore, $\mathcal{D}^\sim = (V, E^\sim)$, together with any well-ordering of the vertices of $\mathcal{D}$ fully identifies $\mathcal{D}$. 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 $(1,2)$, $1\bullet \longleftarrow \bullet 2$, and the DAG with well-ordering $(2,1)$, $1\bullet \longrightarrow \bullet 2$. The submodel $\mathcal{M}^0$ corresponds to the graph $1 \bullet \quad\; \bullet 2$.

2.1 Jeffreys conditioning

Dawid and Lauritzen (2000) introduced the following generalisation of the usual conditioning procedure to construct compatible priors.

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.

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 whose density 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, and the resulting procedure is named Jeffreys conditioning.

If π is the density function of Π with respect to Lebesgue measure, then the distribution for θ_0 resulting from 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 (one-to-one) reparameterisation, 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 $\psi = \psi(\theta_0)$, $\theta_0 \in \Theta_0$, 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 the problem in Example 1, Dawid and Lauritzen (2000) showed that R is a parameterisation implicitly leading to Jeffreys conditioning; we generalise this result in Section 5.2.

3 DAG models and reparameterisations

Consider a DAG $\mathcal{D} = (V, E)$ with $V = \{1, \ldots, v\}$ and assume, without loss of generality, that $(1, 2, \ldots, v)$ 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 $X = (X_1, \ldots, X_v)$ 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 Section 1, that the vertices of $\mathcal{D}$ are arranged in a causal order, see also Cowell *e*t al. (1999, p. 259), and assume the absence of unmeasured confounders that alter the causal interpretation of arrows, see Lauritzen and Richardson (2002) for a detailed discussion of this point. An important special case of this situation is when (3) is the result of a structural assignment system, as described in Lauritzen and Richardson (2002, eqn. 4). It is worth pointing out that in this case the distribution of X is causally Markov with respect to $\mathcal{D}$ (Lauritzen, 2001, Theorem 2.20).

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 $\mathcal{M}^{\mathcal{D}}$, in the sense that they are assumed to remain invariant under conditioning by intervention. Consequently, the partition of η into v groups, 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; to clarify the above point and eventually reach a definition of modular parameterisation 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. Since each θ_i is uniquely associated to $\mathcal{M}_i^{\mathcal{D}}$, we say that this transformation is modular with respect to $\mathcal{D}$. 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 unambiguously linked to $\mathcal{M}_i^{\mathcal{D}}$. This is formalised in the following Definition.

Definition 1 *Consider the family $\mathcal{M}^{\mathcal{D}}$ with density (3). Let $\theta = \theta(\eta)$ be a bijection of $\eta = (\eta_1, \ldots, \eta_v)$. We say that θ is $\mathcal{D}$-modular if it admits a grouping $\theta_1, \ldots, \theta_v$, with* $\dim(\theta_i) = \dim(\eta_i)$*, and an ordering $(j_1, \ldots, j_v)$ of the indexes $(1, \ldots, v)$ such that*

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

where for fixed $(\eta_{j_1}, \ldots, \eta_{j_{k-1}})$, θ_{j_k} is a one-to-one function of η_{j_k}, $k = 1, \ldots, v$.

Remark Clearly the last equation in (4) is always satisfied. Furthermore, as suggested by a referee, the conditions in the above definition are equivalent to requiring that $(\theta_{j_1}, \ldots, \theta_{j_k})$ is a one-to-one function of $(\eta_{j_1}, \ldots, \eta_{j_k})$ for all $k = 1, \ldots, v$.

The property of $\mathcal{D}$-modularity ensures that each component θ_{j_k} can be written as a function of $(\eta_{j_k}, \theta_{j_1}, \ldots, \theta_{j_{k-1}})$ where, for fixed $(\theta_{j_1}, \ldots, \theta_{j_{k-1}})$, θ_{j_k} is a one-to-one function of η_{j_k}, whereas for $i \neq k$, θ_{j_i} can depend on η_{j_k} only indirectly through θ_{j_k}. In this sense we an say that θ_{j_k} is unambiguously associated to the local conditional family $\mathcal{M}_{j_k}^{\mathcal{D}}$.

Example 2 *Consider two local conditional Gaussian families corresponding to the DAG $\mathcal{D}$ $1\bullet \longleftarrow \bullet 2$, such that $(X_1|X_2 = x_2, \eta_1) \sim N(\beta_{12}x_2, \sigma_{11\cdot 2})$ and $(X_2|\eta_2) \sim N(0, \sigma_{22})$, 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 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 $\mathcal{D}$-modular, although $\mathcal{D}^*$ is Markov equivalent to $\mathcal{D}$. For example, letting $\theta_1(\eta) = (\beta_{21}, \sigma_{22\cdot 1})$ and $\theta_2(\eta) = \sigma_{11}$, 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.
ii) Consider the transformation $\eta \mapsto \Sigma$. Because of symmetry, only the upper diagonal elements, say, including the diagonal itself, need be considered. If we set $\theta_1(\eta) = (\sigma_{11}, \sigma_{12})$ and $\theta_2(\eta) = \sigma_{22}$, then θ_2 is a function of η_2 only, and no further check is required to conclude that the transformation is $\mathcal{D}$-modular.
iii) It is straightforward to verify that we still have a $\mathcal{D}$-modular parameterisation if in the previous point Σ is replaced by P, so that $\theta(\eta) = (\sigma_{11}, \rho_{12}, \sigma_{22})$.
iv) 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

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

Since θ_1 is a function of η_1 only, then Σ^{-1} is $\mathcal{D}$-modular.
v) In parallel with case iii) above, consider the transformation $\eta \mapsto R = \{\sigma^{11}, \rho_{12}, \sigma^{22}\}$. Using equations (5) one can reexpress R in terms of (η_1, η_2), to show that no grouping of the elements of R can satisfy (4), hence R is not $\mathcal{D}$-modular.

Working through Example 2 one can derive that Σ, P and Σ^{-1} are also $\mathcal{D}^*$-modular, whereas R is neither $\mathcal{D}$- nor $\mathcal{D}^*$-modular. 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 modular 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 modular parameterisations. Jeffreys measure can clearly be used, however it may lead to questionable 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 $\mathcal{M}^{\mathcal{D}_0} := \mathcal{M}_1^{\mathcal{D}_0} \times \mathcal{M}_2^{\mathcal{D}_0}$, wherein X_1 is independent of 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 that under $\mathcal{M}^{\mathcal{D}}$, $\eta_1 \perp\!\!\!\perp \eta_2$, so that $(X_1, \eta_1) \perp\!\!\!\perp \eta_2 | X_2$. It is often the case that prior distributions on η_1 and η_2 may be regarded as posteriors from $\mathcal{M}^{\mathcal{D}}$ with (possibly fictitious) complete data on X_1 and X_2. As a consequence, the prior on η_2 should depend on "prior" data on X_2 only. This is trivially true under $\mathcal{M}^{\mathcal{D}_0}$ too, and thus the prior distributions on η_2 should be the same under the two models, so that the Bayes factor to compare $\mathcal{M}_2^{\mathcal{D}}$ and $\mathcal{M}_2^{\mathcal{D}_0}$ would be identically one. This is an instance of prior modularity, see Heckerman et al., (1995). 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 Example 6-(iii) of Dawid and Lauritzen, 2000); however, the compatible prior distribution for η_2 under $\mathcal{M}^{\mathcal{D}_0}$, obtained through Jeffreys conditioning, is $a_{22}/\chi^2_{\delta+1}$, see point iii) of Example 1.*

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 the modular structure, 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. We next show that this can be obtained by means of the *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 $\theta \in \Theta$. If the group-components of θ are variation-independent 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, see the Remark below. 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 satisfying (4), that is the class of $\mathcal{D}$-modular parameterisations.

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 $\eta \in H$ and let $\mathcal{M}^{\mathcal{D}_0}$ be a submodel with $\mathcal{D}_0 = (V, E_0)$, $E_0 \subset E$, and parameter $\eta_0 \in H_0 \subset H$. 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 modular 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 $(\theta_{j_1}, \dots \theta_{j_v})$ of Definition 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 the ordered groups $\theta_1 = \sigma_{11}$ and $\theta_2 = (\sigma_{22}, \sigma_{12})$ for $\mathcal{M}^{\mathcal{D}^*}$.

Suppose there exists a $\mathcal{D}$-modular 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 $\theta = \theta(\eta_0)$, with $\eta_0 \in H_0$; 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 $\mathcal{M}^{\mathcal{D}_0}$.

5 Gaussian DAG models

A Gaussian DAG model is specified through (3), wherein each local conditional -or regression-family is Gaussian. The i-th family, $i = 1, \dots, v$, 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 $\mathcal{D}$-modular parameterisation $\theta_i = (\theta_{ij}; j \in \mathrm{fa}(i))$, $i = 1, \dots, v$, with the θ_is 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 $\phi = (\phi_1, \dots, \phi_v)$ where

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

with the understanding that $(1, 2, \dots, v)$ is a well-ordering of the vertices so that $\mathrm{pa}(i) \subseteq \{i + 1, \dots, v\}$. If Φ denotes the upper triangular matrix with entries ϕ_{ij} and zero elsewhere, then $\Phi^T \Phi = \Sigma^{-1}$ represents the Cholesky decomposition of Σ^{-1} (see Wermuth, 1980; Wermuth and Cox, 2000 and Roverato, 2000, 2002).

With reference to the bivariate case, Example 2 has already provided instances of modular, and non-modular, parameterisations. These results are extended and generalised below. Consider the class $[\mathcal{D}]$ of DAGs Markov equivalent to $\mathcal{D}$. For all $\mathcal{D}^* \in [\mathcal{D}]$, Σ, Σ^{-1}, P and R are all one-to-one transformations of the parameterisation η^* of $\mathcal{M}^{\mathcal{D}^*}$. We remark that, 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 1 *Let $\mathcal{M}^{\mathcal{D}}$ be a Gaussian DAG model. Σ, Σ^{-1}, and P belong to the class of $\mathcal{D}^*$-modular parameterisations for all $\mathcal{D}^* \in [\mathcal{D}]$.*

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 $\mathcal{D}^*$-modular parameterisations for all $\mathcal{D}^* \in [\mathcal{D}]$, and that for all $\mathcal{D}^* \in [\mathcal{D}]\backslash\{\mathcal{D}\}$, the standard parameterisation η^* of $\mathcal{M}^{\mathcal{D}^*}$ is not $\mathcal{D}^*$-modular.

Before turning to more technical matters, we remark that starting from two models with Markov equivalent DAGs $\mathcal{D}$ and $\mathcal{D}^*$ and consistent prior distributions for η and η^* (in the sense that they are induced from the same prior on Σ, say), reference conditioning may lead to different compatible priors for the parameter η_0 of a common submodel. This is not surprising since the prior on η_0 is chosen to be compatible with either the prior on η or that on η^*, depending on whether the assumed DAG is $\mathcal{D}$ or $\mathcal{D}^*$. We have to remark however that our method is not meant to be used when several Markov equivalent models are under consideration, because a fixed ordering of the variables is assumed throughout.

5.1 Reference conditioning

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

For a complete DAG model the Fisher information as well as the group reference measure for ϕ can be easily deduced from that of Consonni and Veronese (2000). These results can be extended to an arbitrary DAG model as follows.

Theorem 2 *Let $\mathcal{M}^{\mathcal{D}}$ be a Gaussian DAG model with DAG $\mathcal{D} = (V, E)$. With respect to the parameterisation ϕ described in (8):*

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

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

where Ψ^i is the unique upper triangular matrix with positive diagonal entries 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 $r(\phi) \propto \prod_{i=1}^{v} 1/\phi_{ii}$.

Proof. See the Appendix. □

As an immediate consequence of Theorem 2 we obtain.

Corollary 3 *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 $\mathcal{M}^{\mathcal{D}_0}$ so that reference conditioning is equivalent to conditioning the distribution of ϕ under $\mathcal{M}^{\mathcal{D}}$ on the event $\{\phi_{ij} = 0; (i, j) \in E\backslash E_0\}$.*

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 both $r_0(\phi_0)$ and $r(\phi_0)$ are proportional to $\prod_{i=1}^{v} 1/\phi_{0\,ii}$. □

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 groups $\phi_1, \ldots, \phi_v$ of ϕ are mutually stochastically independent, so that reference conditioning can be performed separately for each component 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 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 can be shown to satisfy global parameter independence. Furthermore reference conditioning can be routinely applied using Corollary 3, thereby providing a practical method for assigning parameter priors to candidate DAG models *via* a small number of direct assessments.

Example 4 *For the problem in Example 2 let $\Sigma \sim IW(\delta, A)$ as in Example 1. It can be checked, using a result cited in Roverato (2002, R.7), that $(\phi_{11}, \phi_{12}) \perp\!\!\!\perp \phi_{22}$, with $\phi_{22}^2 \sim \chi_\delta^2/a_{22}$ and $\phi_{12}|\phi_{11} \sim N\left(-\phi_{11}a_{12}/a_{22},\ a_{22}^{-1}\right)$, $\phi_{22}^2 \sim \chi_\delta^2/a_{22}$, where $a_{11\cdot 2} = a_{11} - a_{12}^2/a_{22}$. Applying reference conditioning, i.e. conditioning on $\{\phi_{12} = 0\}$, one obtains $\sigma_{11} \sim a_{11}/\chi_{\delta+1}^2$ and $\sigma_{22} \sim a_{22}/\chi_\delta^2$ with $\sigma_{11} \perp\!\!\!\perp \sigma_{22}$ 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 (see Lauritzen, 1996, p.52). Furthermore $\mathcal{G}$ and $\mathcal{D}^{\sim}$, as well as $\mathcal{G}_0$ and $\mathcal{D}_0^{\sim}$, are Markov equivalent. In the Gaussian case this implies $\mathcal{M}^{\mathcal{D}} \equiv \mathcal{M}^{\mathcal{G}}$ 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 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 4 *Let $\mathcal{G}$ be a complete undirected graph and $\mathcal{G}_0$ the graph obtained from $\mathcal{G}$ by removing exactly one edge. Then R is a parameterisation implicitly leading to Jeffreys conditioning for the comparison of $\mathcal{M}^{\mathcal{G}}$ and $\mathcal{M}^{\mathcal{G}_0}$.*

Proof. See the Appendix. □

6 Discussion

In this paper we consider DAG models wherein the order of the variables is fixed; this specifies a collection of local conditional families which represents the modular structure of the DAG model. It is the latter aspect, as opposed to vertex ordering, which is emphasised throughout the paper. The η-parameterisation of equation (3), or alternative one-to-one component-wise transformations such as ϕ described in (7), naturally encapsulates this modular structure, since its components are variation independent, and each uniquely parameterises a local family. Reference conditioning does not depend on the order of the η-components, whereas, when applied to an alternative $\mathcal{D}$-modular parameterisation, the ordered groupings follow automatically from (4) of Definition 1.

Reference conditioning is invariant within the class of $\mathcal{D}$-modular parameterisations. Arguably, this property may be shared by other baseline measures, distinct from the reference measure. We contend however that the reference measure is model intrinsic and appears as a very natural choice when trying to generalise Jeffreys conditioning: indeed the latter becomes a special case when no ordering of the variables is specified, equivalently when no modular structure is entertained, as for example with undirected graphical models.

Reference conditioning 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 induced through conditioning. 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 prior modularity is automatically satisfied. The scope of the latter property is indeed more restrictive in our approach than in the original formulation of Heckerman *et al.* (1995), wherein the whole class of Markov equivalent DAGs is considered. This is however consistent with our starting assumption, namely that the ordering of the variables is fixed, so that only one specific DAG for each equivalence class need be considered. This restriction offers the bonus of extra flexibility when it comes to prior modelling: for example in the Gaussian case, we are not confined to using the inverse Wishart family, as instead one ought to do if prior modularity, together with global parameter independence, were required for the entire class of Markov equivalent DAGs, see Geiger and Heckerman (1999).

Acknowledgements

This work has greatly benefited from the comments of an Associate Editor and two referees. We thank Christian Robert for helpful discussions during our stay at CREST-INSEE, Paris.

Appendix: Proofs

Proof of Proposition 1

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 $\Sigma^{*-1} = \Phi^{*T}\Phi$. 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 $\mathcal{D}^*$ so that also the rows and columns of Σ will follow the same ordering. 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 $\mathcal{D}^*$-modular. 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 this grouping satisfies Definition 1 and we can conclude that Σ^{-1} is $\mathcal{D}^*$-modular.

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 $\Sigma^*_{BB} = \left(\Phi^{*T}_{BB}\Phi^*_{BB}\right)^{-1}$ (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 $\mathcal{D}^*$-modular. To show that also P is $\mathcal{D}^*$-modular 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 this completes the proof.

Proof of Theorem 2

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

Let $\mathcal{D}^i = (\text{fa}(i), E^i)$ be any complete DAG such that i is the first vertex, that is the set of parents of i in $\mathcal{D}^i$ is $\text{pa}(i)$, and assume that the rows and columns of $\Sigma^{-1}_{\text{fa}(i),\text{fa}(i)}$ 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 $H_{ii}(\phi)$.

Hereafter we make use of the theory related to commutation, duplication and elimination matrices. For a $p\times p$ matrix A let $\mathrm{vec}(A)$ be the p^2 column vector made up of the columns of A and $\mathrm{vech}(A)$ the $\{p(p+1)/2\}$ column vector obtained by stacking the columns of A from the principal diagonal downwards. The commutation matrix K_{pp} is defined as the $p^2 \times p^2$ matrix such that $\mathrm{vec}(A^T) = K_{pp}\mathrm{vec}(A)$ for any $p\times p$ matrix A; the duplication matrix D_p is the $p^2 \times \{p(p+1)/2\}$ matrix such that $\mathrm{vec}(A) = D_p\mathrm{vech}(A)$ for any $p\times p$ matrix A, and the elimination matrix L_p is the $\{p(p+1)/2\} \times p^2$ matrix such that $\mathrm{vech}(A) = L_p\mathrm{vec}(A)$ for any $p\times p$ matrix A (see, Lütkepohl, 1996, pp.8-9); we refer to Lütkepohl (1996) for a description of the properties of such matrices.

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 Gutiérrez-Peña and Smith, 1995, equation (18)) and p is the cardinality of the set $\mathrm{fa}(i)$. One can check that

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

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

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

$$
\begin{aligned}
&= L_p(\Phi^i \otimes I_p)(\Psi^i \otimes \Psi^i)(I_{p^2} + K_{pp}) \\
&= L_p(\Phi^i \Psi^i \otimes I_p \Psi^i)(I_{p^2} + K_{pp}) \\
&= L_p(I_p \otimes \Psi^i)(I_{p^2} + K_{pp}) \\
&= 2L_p(I_p \otimes \Psi^i) D_p D_p^+ .
\end{aligned}
$$

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 to ϕ does not depend on the order of the v groups and has density with respect to Lebesgue measure $r(\phi) \propto \prod_{i=1}^{v} a_i(\phi_i)^{1/2}$. As far uniqueness is concerned see Proposition 1 of Gutiérrez-Peña and Rueda (2002).

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 $a_i(\phi_i) = 1/\phi^2_{ii}$ 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 4

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*, (eds J.M. Bernardo, J.O. Berger, D.V. Lindley, and A.F.M. Smith), pp. 35-60. London: Oxford University Press.

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

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

Consonni, G. and Veronese, P. (2000). Enriched Conjugate and Reference Priors for the Wishart Family on Symmetric Cones. *Technical Report* 123, Dipartimento di Ec. Pol. Met.Quant., University of Pavia, Italy. Submitted. http://economia.unipv.it/PagP/Pagine_Personali/gconsonni/www/papers/Wishart2.pdf

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

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

Datta, G.S., and Ghosh, M. (1995) Some remarks on non-informative priors. *J. Am. Statist. Assoc.*, **90**, 1357-1363.

Datta, G.S., and Ghosh, M. (1996) On the invariance of noninformative priors. *Ann. Statist.*, **24**, 141-159.

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

Dawid, A.P. and Lauritzen, S.L. (2000) Compatible prior distributions. In *Bayesian Methods with Applications to Science Policy and Official Statistics. The sixth world meeting of the International Society for Bayesian Analysis* (ed. E.I. George), pp. 109-118. http://www.stat.cmu.edu/ISBA/index.html.

Dempster, A.P. (1972) Covariance selection. *Biometrics*, **28**, 157-175.

Frydenberg, M. (1990) The chain graph Markov property, *Scand. J. Statist.*, **17**, 333-353.

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

Gutiérrez-Peña, E. and Rueda, R. (2002) Reference priors for exponential families, *J. Statist. Plan. Inf.*, to appear.

Gutiérrez-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.

Kass, R.E. and Raftery, A.E. (1995) Bayes factors. *J. Am. Statist. Assoc.*, **90**, 773-795.

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

Lauritzen, S.L. (2001) Causal inference from graphical models. In *Complex Stochastic Systems*, (eds O. E. Barndorff Nielsen, D. R. Cox and C. Klüppelberg), pp. 163-107. Boca Raton: Chapman and Hall-CRC.

Lauritzen, S.L. and Richardson, T.S. (2002) Chain graphs models and their causal interpretation (with discussion). *J. Roy. Statist. Soc. B*, to appear.

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

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

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

Roverato, A. (2002) 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**, 711-725.

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

Spirtes, P., Glymour, C., and Scheines, R. (1993) *Causality, Prediction and Search.* New York: Springer-Verlag.

Verma, T. and Pearl, J. (1990) Equivalence and synthesis of causal models. In *Proceedings of the 6th Conference on Uncertainty in Artificial Intelligence*, (ed. P. Bonissone, M. Henrion, L.N. Kanal, and J.F. Lemmer), pp. 255-70. North-Holland, Amsterdam.

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**, 963-972.

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.* Chichester: Wiley.