Electronic Journal of Statistics
Vol. 14 (2020) 4110–4132
ISSN: 1935-7524
https://doi.org/10.1214/20-EJS1768

# Compatible priors for model selection of high-dimensional Gaussian DAGs

**Stefano Peluso**

*Università degli Studi di Milano-Bicocca*
*Department of Statistics and Quantitative Methods*
*Via Bicocca degli Arcimboldi 8, 20126 Milan (Italy)*
*e-mail:* stefano.peluso@unimib.it

**Guido Consonni**

*Università Cattolica del Sacro Cuore*
*Department of Statistics*
*Largo Gemelli 1, 20123 Milan (Italy)*
*e-mail:* guido.consonni@unicatt.it

**Abstract:** Graphical models represent a powerful framework to incorporate conditional independence structure for the statistical analysis of high-dimensional data. In this paper we focus on Directed Acyclic Graphs (DAGs). In the Gaussian setting, a prior recently introduced for the parameters associated to the (modified) Cholesky decomposition of the precision matrix is the DAG-Wishart. The flexibility introduced through a rich choice of shape hyperparameters coupled with conjugacy are two desirable assets of this prior which are especially welcome for estimation and prediction. In this paper we look at the DAG-Wishart prior from the perspective of model selection, with special reference to its consistency properties in high dimensional settings. We show that Bayes factor consistency only holds when comparing two DAGs which do not belong to the same Markov equivalence class, equivalently they encode distinct conditional independencies; a similar result holds for posterior ratio consistency. We also prove that DAG-Wishart distributions with arbitrarily chosen hyperparameters will lead to incompatible priors for model selection, because they assign different marginal likelihoods to Markov equivalent graphs. To overcome this difficulty, we propose a constructive method to specify DAG-Wishart priors whose suitably constrained shape hyperparameters ensure compatibility for DAG model selection.

**MSC2020 subject classifications:** Primary 62F15; secondary 62H22.
**Keywords and phrases:** DAG-Wishart prior, graphical models, Markov equivalence class, structural learning.

Received November 2019.

## Contents

## 1. Introduction

The analysis of multivariate data often aims at understanding the dependence structures among variables, as in networks of protein-protein interactions. In this context, graphical models represents a powerful modeling tool (Lauritzen, 1996). In particular graphs based on Directed Acyclic Graphs (DAGs) are suitable to study complex dependencies in a variety of scientific domains; see for instance Friedman (2004), Sachs et al. (2003), Shojaie and Michailidis (2009), Nagarajan and Scutari (2013). A joint distribution that embodies the conditional independence structure represented by a DAG $\mathcal{D}$ is said to be Markov with respect to $\mathcal{D}$. Additionally, if the distribution is Gaussian, its covariance matrix $\boldsymbol{\Sigma}$, as well as its precision matrix $\boldsymbol{\Omega} = \boldsymbol{\Sigma}^{-1}$, will live in a subspace of the set of symmetric and positive definite matrices because of the constraints imposed by $\mathcal{D}$.

Typically, the DAG which supposedly underlies the generating mechanism of the observations is unknown, and the process of inferring it is named *structural learning*. In the Bayesian framework this is cast as a model selection problem: a prior distribution is assigned to the parameter space of each DAG (parameter prior) producing a marginal likelihood which, coupled with a prior distribution on the space of all DAGs, leads to the posterior distribution on model space. In view of model selection, Geiger and Heckerman (2002) proposed a set of assumptions to assign priors on the parameter space of DAGs. Their method is constructive and enjoys desirable properties in terms of the ensuing marginal likelihood; furthermore it reduces to a hyper-Markov law (Dawid and Lauritzen, 1993) when the graph is undirected and decomposable. Letac and Massam (2007) deal with Gaussian decomposable graphs, and generalize the hyper-inverse Wishart distribution on the covariance matrix to a richer family using a representation of the graph in terms of a perfect DAG. Their work is extended to arbitrary DAGs in Ben-David et al. (2015) leading to the DAG-Wishart prior which exhibits the strong directed Markov property when expressed as a prior on the modified Cholesky decomposition parameters of the precision matrix. Theoretical properties of DAG-Wishart priors with regard to DAG selection consistency have been investigated in Cao et al. (2019), also in a high-dimensional setting where the number of nodes is allowed to grow with the sample size.

In this paper we focus on priors for parameters when the goal is model selection. This task presents specific challenges, the main one being *compatibility* of priors across models (Consonni and Veronese, 2008); for a review see Consonni et al. (2018). An additional complication arises when dealing with DAG models because different DAGs may encode the same set of conditional independencies (*Markov equivalent DAGs*). Markov equivalence induces a partition on the space of DAGs into *Markov equivalence classes* (Andersson, Madigan and Perlman, 1997). If no causal interpretation is attributed to DAGs (Lauritzen, 2001; Dawid, 2003), all DAGs in the same Markov equivalence class are indistinguishable based on obervational data, and therefore should have the same marginal likelihood. This represents a basic compatibility requirement for priors for DAG model selection, and is satisfied by the method of Geiger and Heckerman (2002).

Our contribution is threefold. First we show that posterior consistency for model selection under DAG-Wishart priors holds only up to Markov equivalence, that is only when the comparison is between the true graph and one that is not in the same equivalence class. Next we highlight that DAG-Wishart priors without suitably specified hyperparameters will fail to assign the same marginal likelihood to Markov equivalent DAGs, and therefore are not compatible for DAG model selection. This shortcoming is overcome in our third contribution where we present a constructive method to obtain compatible DAG-Wishart priors for model selection of high-dimensional DAGs, and show how this severely constrains the shape hyperparameters of the DAG-Wishart.

The rest of this paper is organized as follows. In Section 2 we provide background material on the DAG-Wishart distribution and its properties for model selection. In Section 3.1 we start from a simple example that illustrates the behavior of the DAG-Wishart prior within a Markov equivalence class, then show in Section 3.2 its inability to correctly select the true DAG among Markov equivalent DAGs, and finally propose a compatible DAG-Wishart prior in Section 3.3. An empirical illustration of the results is provided in Section 4. Section 5 presents a brief summary, together with future directions of investigation on objective compatible DAG-Wishart priors.

## 2. DAG-Wishart prior and equivalence classes

A Directed Acyclic Graph (DAG) $\mathcal{D} = (V, E)$ with vertex set $V$ and edge set $E \subseteq V \times V$, is a graph with no cycles, that is with no paths starting and ending with the same node. The element $e = (i, j) \in E$, also expressed as $i \to j$, denotes the presence in the graph of an edge directed from vertex $i$ to vertex $j$. For a given DAG $\mathcal{D}$, we assume a *parent ordering* of the vertices of $\mathcal{D}$, whereby edges can only be directed from vertices with larger order to those with lower order. For a given $\mathcal{D}$ a parent ordering always exists, although it may not be unique. Additionally the ordering depends on the DAG so that two distinct DAGs will usually have distinct parent ordering of the vertices. The ordering allows to specify a suitable parametrization of the DAG model and the ensuing prior. We denote with $o_i(\mathcal{D})$ the order of node $i \in V$ in graph $\mathcal{D}$, with $pa_i(\mathcal{D})$ the set of parents of node $i$ in $\mathcal{D}$, and with $fa_i(\mathcal{D}) = pa_i(\mathcal{D}) \cup \{i\}$ the *family* of $i$ in $\mathcal{D}$. Let $\boldsymbol{y} = (y_i)_{i \in V}$ be a $p$-variate Gaussian random vector with zero $p$-dimensional vector mean and $p \times p$ positive definite covariance matrix $\boldsymbol{\Sigma} \in P_{\mathcal{D}}$, where $P_{\mathcal{D}}$

is the space of covariance matrices Markov with respect to $\mathcal{D}$. In this case the distribution of $\boldsymbol{y}$ obeys the directed *Markov property* with respect to $\mathcal{D}$ i.e. $y_i \perp\!\!\!\perp \boldsymbol{y}_{\{j\in V: o_j(\mathcal{D})>o_i(\mathcal{D})\}\setminus pa_i(\mathcal{D})} | \boldsymbol{y}_{pa_i(\mathcal{D})}$ (Lauritzen, 1996) where, for any $A \subseteq V$, $\boldsymbol{y}_A$ is the column vector $\boldsymbol{y}_A = (y_i)_{i\in A}$. For the generic square matrix $\boldsymbol{M}$ we will write $\boldsymbol{M}_A$ for $(\boldsymbol{M}_{ij})_{i,j\in A}$, the submatrix of $\boldsymbol{M}$ whose rows and column indexes belong to $A$; by convention $\boldsymbol{y}_\emptyset = \boldsymbol{M}_\emptyset = 1$. For two disjoint sets $A, B \subseteq V$, we write $\boldsymbol{M}_{A|B}$ for the Schur complement of $\boldsymbol{M}_A$ in the matrix $\boldsymbol{M}_{A\cup B}$. We finally write $\boldsymbol{M}_{i\times A}$ and $\boldsymbol{M}_{A\times i}$, $i \in V$, for the column vectors $(\boldsymbol{M}_{ij})_{j\in A}$ and $(\boldsymbol{M}_{ji})_{j\in A}$, respectively.

The positive definite precision matrix $\boldsymbol{\Omega} = \boldsymbol{\Sigma}^{-1}$ has a unique *modified Cholesky decomposition* $\boldsymbol{\Omega} = \boldsymbol{L}\boldsymbol{D}^{-1}\boldsymbol{L}^\top$, where $\boldsymbol{L}$ is a lower-triangular $p \times p$ matrix with $\boldsymbol{L}_{ii} = 1$, $i \in V$, and $\boldsymbol{D}$ is a diagonal $p \times p$ matrix with positive entries on the main diagonal; see e.g. Pourahmadi (2007). There is a relationship between the Markov property of the distribution of $\boldsymbol{y}$ and the structure of $\boldsymbol{L}$: $\boldsymbol{L}_{ij} = 0$ for $o_j(\mathcal{D}) < o_i(\mathcal{D})$ if and only if $i \notin pa_j(\mathcal{D})$. We call $\mathcal{L}_\mathcal{D}$ the set of lower triangular matrices with unit main diagonal entries and with off-diagonal terms coherent with the parent ordering of $\mathcal{D}$, and we call $\mathcal{D}_+^p$ the set of $p \times p$ diagonal matrices with positive entries. Then $\Theta_\mathcal{D} = \mathcal{D}_+^p \times \mathcal{L}_\mathcal{D}$ is the *Cholesky space* corresponding to $\mathcal{D}$, and $(\boldsymbol{D}, \boldsymbol{L}) \in \Theta_\mathcal{D}$ is the *Cholesky parameter*.

The *DAG-Wishart distribution* $\pi_{\boldsymbol{U},\boldsymbol{a}}^{\Theta_\mathcal{D}}$ (Ben-David et al., 2015) on the Cholesky space $\Theta_\mathcal{D}$, with hyperparameter $\boldsymbol{U}$ (a $p \times p$ positive definite matrix) and shape hyperparameter $\boldsymbol{a}(\mathcal{D}) = (a_1(\mathcal{D}), \dots, a_p(\mathcal{D}))$ has density

$$\pi_{\boldsymbol{U},\boldsymbol{a}(\mathcal{D})}^{\Theta_\mathcal{D}}(\boldsymbol{D}, \boldsymbol{L}) = \frac{1}{\mathcal{Z}_\mathcal{D}(\boldsymbol{U}, \boldsymbol{a}(\mathcal{D}))} \exp\left\{-\frac{1}{2}\mathrm{tr}\left(\left(\boldsymbol{L}\boldsymbol{D}^{-1}\boldsymbol{L}^\top\right)\boldsymbol{U}\right)\right\} \prod_{i\in V} \boldsymbol{D}_{ii}^{-\frac{a_{o_i}(\mathcal{D})}{2}} \tag{1}$$

for all $(\boldsymbol{D}, \boldsymbol{L}) \in \Theta_\mathcal{D}$, with $a_{o_i}(\mathcal{D}) := a_{o_i(\mathcal{D})}(\mathcal{D})$ and $\mathcal{Z}_\mathcal{D}(\boldsymbol{U}, \boldsymbol{a}(\mathcal{D}))$ being the normalizing constant. The latter is finite, and hence the prior is proper, if $a_{o_i}(\mathcal{D}) - v_i(\mathcal{D}) > 2$ for all $i \in V$ where $v_i(\mathcal{D}) := |pa_i(\mathcal{D})|$ is the cardinality of the parent set of $i$ in $\mathcal{D}$. Its expression is then

$$\mathcal{Z}_\mathcal{D}(\boldsymbol{U}, \boldsymbol{a}(\mathcal{D})) = \prod_{i=1}^p \Gamma\left(\frac{c_i(\mathcal{D})}{2} - 1\right) 2^{\frac{a_{o_i}(\mathcal{D})}{2}-1} \left(\sqrt{\pi}\right)^{v_i(\mathcal{D})} \frac{\left|\boldsymbol{U}_{pa_i(\mathcal{D})}\right|^{\frac{c_i(\mathcal{D})-3}{2}}}{\left|\boldsymbol{U}_{fa_i(\mathcal{D})}\right|^{\frac{c_i(\mathcal{D})-2}{2}}}, \tag{2}$$

with $c_i(\mathcal{D}) := a_{o_i}(\mathcal{D}) - v_i(\mathcal{D})$. Density (2), when regarded as a parametrized family in $\boldsymbol{U}$, is not identifiable if we require $\boldsymbol{U}$ to be only positive definite; however it becomes identifiable if $\boldsymbol{U} \in P_\mathcal{D}$; see Ben-David et al. (2015).

Theoretical properties in terms of graph selection consistency and estimation consistency of the DAG-Wishart prior are shown in Cao et al. (2019), under regularity conditions that we will assume valid throughout the paper.

Let $\boldsymbol{Y}^n$ be the $n \times p_n$ data matrix, obtained by stacking one upon the other the row-vectors $\boldsymbol{y}_1^\top, \dots, \boldsymbol{y}_n^\top$ which are independent and identically distributed according to $N_{p_n}\left(\boldsymbol{0}, \left(\boldsymbol{L}\boldsymbol{D}^{-1}\boldsymbol{L}^\top\right)^{-1}\right)$, a distribution that obeys the directed Markov property with respect to some DAG $\mathcal{D}$, and the number of random variables $p_n$ is increasing in $n$.

If *a priori* $(\boldsymbol{D}, \boldsymbol{L}) \sim \pi^{\Theta_{\mathcal{D}}}_{\boldsymbol{U}^n, \boldsymbol{a}(\mathcal{D})}$, then *a posteriori* $(\boldsymbol{D}, \boldsymbol{L}) \sim \pi^{\Theta_{\mathcal{D}}}_{n\tilde{\boldsymbol{S}}^n, \boldsymbol{a}(\mathcal{D})+n}$, where $\tilde{\boldsymbol{S}}^n = \boldsymbol{U}^n/n + \boldsymbol{S}^n$ and $\boldsymbol{S}^n = \sum_{j=1}^n \boldsymbol{y}_j \boldsymbol{y}_j^\top/n$ is the sample covariance matrix. Therefore the DAG-Wishart prior is conjugate. For convenience, in the sequel the dependence of $\boldsymbol{Y}^n$, $\boldsymbol{U}^n$, $\boldsymbol{S}^n$, $\tilde{\boldsymbol{S}}^n$, and of all other quantities of interest on $n$ will be omitted, whilst we keep explicit the dependence on the graph dimension $p_n$, to highlight that it increases with the sample size.

Following Cao et al. (2019), the marginal likelihood of a DAG $\mathcal{D}$ can be written as

$$\pi(\boldsymbol{Y}|\mathcal{D}) = (2\pi)^{-n/2} \mathcal{Z}_{\mathcal{D}}(n\tilde{\boldsymbol{S}}, \boldsymbol{a}(\mathcal{D}) + n)/\mathcal{Z}_{\mathcal{D}}(\boldsymbol{U}, \boldsymbol{a}(\mathcal{D})),$$

and, accordingly, the Bayes factor between any two DAGs $\mathcal{D}$ and $\mathcal{D}_0$ is

$$BF_{\mathcal{D},\mathcal{D}_0}(\boldsymbol{Y}) = \frac{\mathcal{Z}_{\mathcal{D}}(n\tilde{\boldsymbol{S}}, \boldsymbol{a}(\mathcal{D}) + n)}{\mathcal{Z}_{\mathcal{D}}(\boldsymbol{U}, \boldsymbol{a}(\mathcal{D}))} \frac{\mathcal{Z}_{\mathcal{D}_0}(\boldsymbol{U}, \boldsymbol{a}(\mathcal{D}_0))}{\mathcal{Z}_{\mathcal{D}_0}(n\tilde{\boldsymbol{S}}, \boldsymbol{a}(\mathcal{D}_0) + n)}. \tag{3}$$

We have *Bayes factor consistency* if, for all $\mathcal{D} \neq \mathcal{D}_0$, $BF_{\mathcal{D},\mathcal{D}_0} \overset{\bar{P}}{\to} 0$ whenever $\mathcal{D}_0$ is the true DAG generating $\boldsymbol{Y}$, where $\overset{\bar{P}}{\to}$ denotes convergence in probability and $\bar{P}$ is the probability measure under the true DAG $\mathcal{D}_0$. On the other hand, we have *posterior ratio consistency* if, with $\mathcal{D}_0$ being the true DAG, it holds that, as $n \to \infty$,

$$\max_{\mathcal{D}\neq\mathcal{D}_0} \frac{\pi(\mathcal{D}|\boldsymbol{Y})}{\pi(\mathcal{D}_0|\boldsymbol{Y})} = \max_{\mathcal{D}\neq\mathcal{D}_0} BF_{\mathcal{D},\mathcal{D}_0}(\boldsymbol{Y}) \frac{\pi(\mathcal{D})}{\pi(\mathcal{D}_0)} \overset{\bar{P}}{\to} 0. \tag{4}$$

Notice that while Bayes factor consistency involves only a pairwise comparison, equation (4) is evaluated over the whole DAG-space.

It is well known that two distinct DAGs can represent the same set of conditional independencies in the joint distribution of the random vector $\boldsymbol{y}$ (see the illustrative example in Section 3.1). In this case the two DAGs are observationally indistinguishable, and they are said to be *Markov equivalent* (Verma and Pearl, 1991). The *Markov equivalence class* of a DAG $\mathcal{D}$, denoted by $[\mathcal{D}]$, is the set of all DAGs Markov equivalent to $\mathcal{D}$. All DAGs in the same equivalence class must have (i) the same skeleton (they are identical if edge directions are neglected); (ii) the same immoralities (the induced sub-graphs of the form $i \to j \leftarrow z$, with $i, j, z \in V$). Since Markov equivalent graphs cannot be distinguished on the basis of observational data, it is appropriate to require *score equivalence*, namely that any two Markov equivalent DAGs $\mathcal{D}$ and $\mathcal{D}_0$ should have the same marginal likelihood, equivalently $BF_{\mathcal{D},\mathcal{D}_0}(\boldsymbol{Y}) = 1$, $\forall \mathcal{D} \in [\mathcal{D}_0]$. If a prior realizes score equivalence, we name it *compatible*. It follows that, if compatible parameter priors are used, the posterior probabilities associated with two Markov equivalent DAGs will differ only if their prior probabilities do.

## 3. DAG-Wishart priors for model selection

### 3.1. A simple motivating example

Consider two DAGs $\mathcal{D}_1 = (V, E_1)$ and $\mathcal{D}_2 = (V, E_2)$, where $V = \{A, B\}$, $E_1 = \{(B, A)\}$ and $E_2 = \{(A, B)\}$, that is $\mathcal{D}_1$ is $A \leftarrow B$ and $\mathcal{D}_2$ is $A \rightarrow B$. They are trivially equivalent, since they encode two alternative factorizations for the *same* unconstrained joint density of $A$ and $B$. Note that $o_A(D_1) = o_B(D_2) = 1$ and $o_B(D_1) = o_A(D_2) = 2$ provide a parent ordering. Clearly no observational data (finite or infinite) can detect the orientation of the edge. Since $\mathcal{D}_1$ and $\mathcal{D}_2$ are observationally equivalent, their marginal likelihood should be the same, and consequently the Bayes Factor for the comparison of the two DAGs should be identically equal to one for each $n$.

We now assign to $(\boldsymbol{D}, \boldsymbol{L})$, separately under $\mathcal{D}_1$ and $\mathcal{D}_2$, a DAG-Wishart prior having shape hyperparameter $\boldsymbol{a}(\mathcal{D}) = (a_1(\mathcal{D}), a_2(\mathcal{D}))$, with $a_{o_i}(\mathcal{D}) > 2 + v_i(\mathcal{D})$, $i \in V$. We assume the same $\boldsymbol{U}$ under both $\mathcal{D}_1$ and $\mathcal{D}_2$, although this is not strictly necessary. Only for the present section we assume that $p_n = p$ constant in $n$. Then the normalizing constant of the prior under $\mathcal{D}_1$ is

$$\mathcal{Z}_{\mathcal{D}_1}\left(\boldsymbol{U}, \boldsymbol{a}(\mathcal{D}_1)\right) = \frac{\Gamma\left(\frac{a_1(\mathcal{D}_1)-3}{2}\right)\Gamma\left(\frac{a_2(\mathcal{D}_1)-2}{2}\right) 2^{\frac{a_1(\mathcal{D}_1)+a_2(\mathcal{D}_1)}{2}-2}\sqrt{\pi}}{\boldsymbol{U}_A^{\frac{a_2(\mathcal{D}_1)-a_1(\mathcal{D}_1)+2}{2}} \det\left(\boldsymbol{U}\right)^{\frac{a_1(\mathcal{D}_1)-3}{2}}},$$

and similarly for $\mathcal{Z}_{\mathcal{D}_2}\left(\boldsymbol{U}, \boldsymbol{a}(\mathcal{D}_2)\right)$ with $A$ and $\boldsymbol{a}(\mathcal{D}_1)$ replaced respectively by $B$ and $\boldsymbol{a}(\mathcal{D}_2)$. The Bayes Factor of $\mathcal{D}_1$ against $\mathcal{D}_2$ is

$$\begin{aligned} BF_{\mathcal{D}_1,\mathcal{D}_2}(\boldsymbol{Y}) &= \frac{m(\boldsymbol{Y}|\mathcal{D}_1)}{m(\boldsymbol{Y}|\mathcal{D}_2)} = \frac{\mathcal{Z}_{\mathcal{D}_1}\left(n\tilde{\boldsymbol{S}}, n + \boldsymbol{a}(\mathcal{D}_1)\right)}{\mathcal{Z}_{\mathcal{D}_1}\left(\boldsymbol{U}, \boldsymbol{a}(\mathcal{D}_1)\right)} \frac{\mathcal{Z}_{\mathcal{D}_2}\left(\boldsymbol{U}, \boldsymbol{a}(\mathcal{D}_2)\right)}{\mathcal{Z}_{\mathcal{D}_2}\left(n\tilde{\boldsymbol{S}}, n + \boldsymbol{a}(\mathcal{D}_2)\right)} \\ &= K(\boldsymbol{a}(\mathcal{D}_2), \boldsymbol{a}(\mathcal{D}_1)) \left[\frac{\det(n\tilde{\boldsymbol{S}})}{\det \boldsymbol{U}}\right]^{\frac{a_1(\mathcal{D}_2)-a_1(\mathcal{D}_1)}{2}} \\ &\quad \left[\frac{\boldsymbol{U}_B + n\boldsymbol{S}_B}{\boldsymbol{U}_B}\right]^{\frac{a_1(\mathcal{D}_1)-a_2(\mathcal{D}_1)-2}{2}} \left[\frac{\boldsymbol{U}_A}{\boldsymbol{U}_A + n\boldsymbol{S}_A}\right]^{\frac{a_1(\mathcal{D}_2)-a_2(\mathcal{D}_2)-2}{2}}, \end{aligned}$$

where

$$K(\boldsymbol{a}(\mathcal{D}_1), \boldsymbol{a}(\mathcal{D}_2)) = \frac{\Gamma\left(\frac{a_1(\mathcal{D}_1)+n-3}{2}\right)\Gamma\left(\frac{a_2(\mathcal{D}_1)+n-2}{2}\right)}{\Gamma\left(\frac{a_1(\mathcal{D}_1)-3}{2}\right)\Gamma\left(\frac{a_2(\mathcal{D}_1)-2}{2}\right)} \frac{\Gamma\left(\frac{a_1(\mathcal{D}_2)-3}{2}\right)\Gamma\left(\frac{a_2(\mathcal{D}_2)-2}{2}\right)}{\Gamma\left(\frac{a_1(\mathcal{D}_2)+n-3}{2}\right)\Gamma\left(\frac{a_2(\mathcal{D}_2)+n-2}{2}\right)}.$$

Note that, if (i) $a_i(\mathcal{D}_1) = a_i(\mathcal{D}_2)$, $i = 1, 2$, and (ii) $a_1(\mathcal{D}_j) - a_2(\mathcal{D}_j) = 2$, $j = 1, 2$, then $BF_{\mathcal{D}_1,\mathcal{D}_2}(\boldsymbol{Y}) = 1$ as it ought to be because of the observational equivalence of $\mathcal{D}_1$ and $\mathcal{D}_2$. On the other hand, for a general DAG-Wishart with no constraints on the hyperparameters $a$'s – beyond those required for the prior being proper- there is no guarantee that the Bayes factor between equivalent graphs will be one. For instance, with the choice $a_{o_i}(\mathcal{D}) = v_i(\mathcal{D}) + 10$, as in Cao et al. (2019, Section 8.1), condition (ii) above is not satisfied because $a_1(\mathcal{D}_1) = a_1(\mathcal{D}_2) = 11$

and $a_2(\mathcal{D}_1) = a_2(\mathcal{D}_2) = 10$. With this choice of hyperparameters, the Bayes factor reduces to

$$BF_{\mathcal{D}_1,\mathcal{D}_2}(\boldsymbol{Y}) = \left[\frac{\boldsymbol{U}_B + n\boldsymbol{S}_B}{\boldsymbol{U}_B}\frac{\boldsymbol{U}_A}{\boldsymbol{U}_A + n\boldsymbol{S}_A}\right]^{-1/2}$$

and then

$$BF_{\mathcal{D}_1,\mathcal{D}_2}(\boldsymbol{Y}) = \left[\frac{\sigma_A^2 + O_p\left(\frac{1}{\sqrt{n}}\right)}{\sigma_B^2 + O_p\left(\frac{1}{\sqrt{n}}\right)}\right]^{1/2}\left[\frac{\boldsymbol{U}_B}{\boldsymbol{U}_A}\right]^{1/2} \xrightarrow{\bar{P}} \sqrt{\frac{\boldsymbol{U}_B}{\boldsymbol{U}_A}}\frac{\sigma_A}{\sigma_B},$$

both under $\mathcal{D}_1$ and under $\mathcal{D}_2$. Note that with this choice of $\boldsymbol{a}(\mathcal{D}_1)$ and $\boldsymbol{a}(\mathcal{D}_2)$, Assumption 5(i) of Cao et al. (2019) is satisfied, and our simplified context of $p_n = p$ is also coherent with their Assumptions 1, 2, 4 and 5(ii). The remaining Assumption 3 concerns the prior on $\mathcal{D}$, and therefore does not affect the result on Bayes factor inconsistency.

If the true DAG generating the data is $\mathcal{D}_2$, the ratio of the posterior probabilities is

$$\frac{\pi(\mathcal{D}_1|\boldsymbol{Y})}{\pi(\mathcal{D}_2|\boldsymbol{Y})} = BF_{\mathcal{D}_1,\mathcal{D}_2}(\boldsymbol{Y})\frac{\pi(\mathcal{D}_1)}{\pi(\mathcal{D}_2)} \xrightarrow{\bar{P}} \sqrt{\frac{\boldsymbol{U}_B}{\boldsymbol{U}_A}}\frac{\sigma_A}{\sigma_B}\frac{\pi(\mathcal{D}_1)}{\pi(\mathcal{D}_2)}.$$

Posterior ratio consistency between the equivalent DAGs $\mathcal{D}_1$ and $\mathcal{D}_2$ fails if the ratio above is different from zero. This however is the case for any choice of priors on model space for which $\pi(\mathcal{D}_1)/\pi(\mathcal{D}_2) > 0$. In particular, because the prior on DAG-space of Cao et al. (2019, formula (3.1)) only depends on the skeleton, it follows that $\pi(\mathcal{D}_1)/\pi(\mathcal{D}_2) = 1$ when $\mathcal{D}_1$ and $\mathcal{D}_2$ belong to the same equivalence class as in our example, and this contradicts their Theorem 4.1.

### 3.2. *Posterior ratio consistency for equivalent graphs*

The example in Section 3.1 shows that a DAG-Wishart prior with freely chosen shape hyperparameter $\boldsymbol{a}(\mathcal{D})$ does not produce the same marginal likelihood, and does not reach posterior ratio consistency, within Markov equivalence classes of DAGs. Proposition 3.1 shows that posterior ratio consistency under a general DAG-Wishart prior only holds *outside* the Markov equivalence class of the true generating DAG $\mathcal{D}_0$. On the other hand, the posterior ratio converges, up to a constant, to the prior ratio within the true equivalence class.

**Proposition 3.1.** *Let $\mathcal{D}_0$ be the true DAG and $(\boldsymbol{D}, \boldsymbol{L}) \in \Theta_{\mathcal{D}_0}$. For $(\boldsymbol{D}, \boldsymbol{L}) \sim \pi^{\Theta_{\mathcal{D}_0}}_{\boldsymbol{U},\boldsymbol{a}(\mathcal{D}_0)}$ and $\boldsymbol{Y}|((\boldsymbol{D}, \boldsymbol{L})) \sim N_{p_n}(\boldsymbol{0}, (\boldsymbol{L}\boldsymbol{D}^{-1}\boldsymbol{L}^\top)^{-1})$, as $n \to \infty$*

$$i)\quad \max_{\mathcal{D}\notin[\mathcal{D}_0]} \frac{\pi(\mathcal{D}|\boldsymbol{Y})}{\pi(\mathcal{D}_0|\boldsymbol{Y})} \xrightarrow{\bar{P}} 0,$$

$$ii)\quad \frac{\pi(\mathcal{D}|\boldsymbol{Y})}{\pi(\mathcal{D}_0|\boldsymbol{Y})} \xrightarrow{\bar{P}} C_{\mathcal{D},\mathcal{D}_0}\frac{\pi(\mathcal{D})}{\pi(\mathcal{D}_0)},\ \text{for all } \mathcal{D} \in [\mathcal{D}_0],\ \mathcal{D} \neq \mathcal{D}_0,$$

*where $C_{\mathcal{D},\mathcal{D}_0}$ is a positive constant.*

*Proof.* See the Appendix. □

We underline that part i) of Proposition 3.1 makes use of the regularity Assumptions 1-5 employed by Cao et al. (2019). In particular, the prior on DAG-space $\pi(\mathcal{D})$ is chosen to be

$$\pi(\mathcal{D}) \propto \prod_{i \in V} q_n^{v_i(\mathcal{D})}(1-q_n)^{p_n - o_i - v_i(\mathcal{D})},$$

where $q_n = e^{-\eta_n n}$ and $\eta_n = d_n(\log p_n/n)^{1/(2+k)}$, for some $k > 0$, so that the probability of edge inclusion vanishes to zero exponentially fast in the limit. On the other hand, part ii) makes no specific assumption on the prior on DAG space. Furthermore, if this prior only depends on the skeleton then $\pi(\mathcal{D})/\pi(\mathcal{D}_0)$ is identically one because DAGs belonging to the same equivalence class share the same skeleton.

Proposition 3.1 states that if $\mathcal{D} \in [\mathcal{D}_0]$ then under any specification of the DAG-Wishart prior, posterior ratio consistency does not hold because $0 < C_{\mathcal{D},\mathcal{D}_0} < \infty$. Furthermore, compatibility within the equivalence class (that is equivalent DAGs should have the same marginal likelihood) is not guaranteed because $C_{\mathcal{D},\mathcal{D}_0}$ can be different from one. We therefore elaborate on Theorem 4.1 of Cao et al. (2019), and carefully distinguish between graphs belonging to the same Markov equivalence class, or living in distinct equivalence classes.

In the following corollary we formalize that the posterior mode $\hat{\mathcal{D}}$ will be in the correct equivalence class with probability 1, but asymptotic model selection consistency of the posterior mode is precluded within the equivalence class, as there is no guarantee that the highest posterior mass will be assigned to the correct DAG.

**Corollary 3.2.** *Let $\mathcal{D}_0$ be the true DAG and $(\boldsymbol{D}, \boldsymbol{L}) \in \Theta_{\mathcal{D}_0}$. For $(\boldsymbol{D}, \boldsymbol{L}) \sim \pi^{\Theta_{\mathcal{D}_0}}_{\boldsymbol{U},\boldsymbol{a}(\mathcal{D}_0)}$ and $\boldsymbol{Y}|((\boldsymbol{D}, \boldsymbol{L})) \sim N_{p_n}(\boldsymbol{0}, (\boldsymbol{L}\boldsymbol{D}^{-1}\boldsymbol{L}^\top)^{-1})$, as $n \to \infty$*

$$\begin{aligned} &\bar{P}(\hat{\mathcal{D}} \in [\mathcal{D}_0]) \to 1, \\ &\bar{P}\left(\hat{\mathcal{D}} \in \underset{\mathcal{D}\in[\mathcal{D}_0]}{\arg\max}\, C_{\mathcal{D},\mathcal{D}_0}\frac{\pi(\mathcal{D})}{\pi(\mathcal{D}_0)}\right) \to 1, \end{aligned}$$

*for finite non-null $C_{\mathcal{D},\mathcal{D}_0}$ given in Proposition 3.1.*

### *3.3. Compatible DAG-Wishart prior*

In this section we provide a constructive method to obtain a DAG-Wishart prior suitable for model selection because it satisfies score equivalence, that is it produces a Bayes factor equal to one when comparing DAGs belonging to the same Markov equivalence class. We follow the method of Geiger and Heckerman (2002), which we view as a method for assigning *compatible* parameter priors across a collection of DAGs; see Consonni et al. (2018) for a discussion of issues arising when specifying priors targeted for model selection. Geiger and

Heckerman (2002) state a set of assumptions on the statistical model and the prior which permit the construction of a parameter prior under *any* DAG model, starting from a *single* parameter prior under a *complete* DAG model. An important byproduct of their methodology is that score equivalence is guaranteed to hold.

First consider an *unrestricted* precision matrix $\boldsymbol{\Omega}$, corresponding to a complete DAG $\mathcal{D}^c$, and assign to it a Wishart prior $\boldsymbol{\Omega} \sim W_{p_n}(a, \boldsymbol{U})$ having density

$$p(\boldsymbol{\Omega} \,|\, a, \boldsymbol{U}) \propto |\boldsymbol{\Omega}|^{\frac{a-p_n-1}{2}} \exp\left\{-\frac{1}{2} tr(\boldsymbol{\Omega}\boldsymbol{U})\right\}, \tag{5}$$

where $\boldsymbol{U}$ is a positive definite matrix and $a > p_n - 1$ to guarantee that the prior is proper.

Let $(\boldsymbol{D}, \boldsymbol{L})$ denote the modified Cholesky decomposition of $\boldsymbol{\Omega}$, so that $\boldsymbol{\Omega} = \boldsymbol{L}\boldsymbol{D}^{-1}\boldsymbol{L}^\top$. Since $|\boldsymbol{\Omega}| = \prod_{i \in V} \boldsymbol{D}_{ii}^{-1}$, the prior on $(\boldsymbol{D}, \boldsymbol{L})$ induced from (5) is given by

$$p(\boldsymbol{D}, \boldsymbol{L} \,|\, a, \boldsymbol{U}) \propto \left(\prod_{i \in V} \boldsymbol{D}_{ii}^{-\frac{a-p-1}{2}}\right) \exp\left\{-\frac{1}{2} tr((\boldsymbol{L}\boldsymbol{D}^{-1}\boldsymbol{L}^\top)\boldsymbol{U})\right\} \times J(\boldsymbol{D}, \boldsymbol{L})$$

where $J(\boldsymbol{D}, \boldsymbol{L})$ is the Jacobian of the transformation $\boldsymbol{\Omega} \mapsto (\boldsymbol{D}, \boldsymbol{L})$. The following Lemma gives the expression for the Jacobian in the setting of a general DAG.

**Lemma 3.3** (Ben-David et al. 2015, Supplemental Section B)**.** *Let* $\boldsymbol{\Omega}$ *be a precision matrix, Markov with respect to a DAG* $\mathcal{D}$*. The Jacobian of the mapping* $\boldsymbol{\Omega} \mapsto (\boldsymbol{D}, \boldsymbol{L})$ *is*

$$J(\boldsymbol{D}, \boldsymbol{L}) = \prod_{i \in V} \boldsymbol{D}_{ii}^{-(v_i(\mathcal{D})+2)}. \tag{6}$$

If $\mathcal{D} = \mathcal{D}^c$ is complete, so that $\boldsymbol{\Omega}$ is unrestricted, the number of parents of $i$ is $p_n - o_i(\mathcal{D}^c)$ and the Jacobian (6) becomes $\prod_{i \in V} \boldsymbol{D}_{ii}^{-(p_n - o_i(\mathcal{D}^c)+2)}$. Accordingly, the prior on $(\boldsymbol{D}, \boldsymbol{L})$ induced by the Wishart prior (5) is

$$p(\boldsymbol{D}, \boldsymbol{L} \,|\, a, \boldsymbol{U}) \propto \left(\prod_{i \in V} \boldsymbol{D}_{ii}^{-\frac{a+p_n-2o_i(\mathcal{D}^c)+3}{2}}\right) \exp\left\{-\frac{1}{2} tr((\boldsymbol{L}\boldsymbol{D}^{-1}\boldsymbol{L}^\top)\boldsymbol{U})\right\}, \tag{7}$$

because $|\boldsymbol{\Omega}|^{\frac{a}{2}-\frac{p_n+1}{2}} \times J(L, D)$ yields $\prod_{i \in V} \boldsymbol{D}_{ii}^{-\frac{\alpha_i}{2}}$ where $\frac{\alpha_i}{2} = (\frac{a}{2} - \frac{p_n+1}{2}) + (p_n - o_i(\mathcal{D}^c) + 2) = \frac{a+p_n-2o_i(\mathcal{D}^c)+3}{2}$. In other words, the prior distribution on the Cholesky space $\Theta_{\mathcal{D}^c}$ induced by the Wishart distribution on $\mathcal{P}_{\mathcal{D}^c}$ (the space of positive definite matrices) is a DAG-Wishart whose shape hyperparameters have the fixed configuration $a_{o_i}(\mathcal{D}^c) = a - p_n + 2(p_n - o_i(\mathcal{D}^c)) + 3 = a + p_n - 2o_i(\mathcal{D}^c) + 3$.

For a given DAG $\mathcal{D}$, a DAG-Wishart prior with symmetric positive definite matrix $\boldsymbol{U}$ and vector $\boldsymbol{a}(\mathcal{D})$ satisfies the following conditions (Ben-David et al., 2015)

(I) $\underset{i \in V}{\perp\!\!\!\perp} \left(\boldsymbol{D}_{ii}, \boldsymbol{L}_{pa_i(\mathcal{D}) \times i}\right)$ (global parameter independence),

(II) $\boldsymbol{D}_{ii} \sim InvGamma\left(\frac{a_{o_i}(\mathcal{D})-v_i(\mathcal{D})-2}{2}, \frac{1}{2}\boldsymbol{U}_{i|pa_i(\mathcal{D})}\right)$

(III) $\boldsymbol{L}_{pa_i(\mathcal{D})\times i}|\boldsymbol{D}_{ii} \sim N_{v_i(\mathcal{D})}\left(-(\boldsymbol{U}_{pa_i(\mathcal{D})})^{-1}\boldsymbol{U}_{pa_i(\mathcal{D})\times i}, \boldsymbol{D}_{ii}(\boldsymbol{U}_{pa_i(\mathcal{D})})^{-1}\right)$.

Then, given an arbitrary DAG $\mathcal{D}$ and an allied parent ordering, a compatible prior for $(\boldsymbol{D}, \boldsymbol{L})$ can be constructed as follows. For each node $i \in V$, (i) identify a complete DAG $\mathcal{D}^{c(i)}$ s.t. $pa_i(\mathcal{D}^{c(i)}) = pa_i(\mathcal{D})$, (ii) obtain $(\boldsymbol{D}^{c(i)}, \boldsymbol{L}^{c(i)})$, the Cholesky decomposition parameter for the unconstrained $\boldsymbol{\Omega}$ under $\mathcal{D}^{c(i)}$ and (iii) assign to $(\boldsymbol{D}_{ii}, \boldsymbol{L}_{pa_i(\mathcal{D})\times i})$ the same prior of $(\boldsymbol{D}_{ii}^{c(i)}, \boldsymbol{L}_{pa_i(\mathcal{D}^{c(i)})\times i})$. Figure 1 provides an illustration.

Although this procedure looks somewhat convoluted it is actually straightforward to apply. First of all it works separately for each node of the graph. Next, from conditions (II) and (III) above, it appears that it depends on the matrix $\boldsymbol{U}$ only through blocks corresponding to the node under consideration and its parents in each specific DAG we entertain. It remains to work out the shape hyperparameter of the Inverse Gamma distribution for $D_{ii}$. Consider node $i$ under $\mathcal{D}$ with parent set $pa_i(\mathcal{D})$ whose cardinality is $v_i(\mathcal{D})$. By construction node $i$ in the corresponding complete DAG $\mathcal{D}^{c(i)}$ will have the same parent set. Hence $v_i(\mathcal{D}^{c(i)})$, the cardinality of the parent set of $i$ in $\mathcal{D}^{c(i)}$, will be equal to $v_i(\mathcal{D})$. From condition (II) we get

$$D_{ii} \sim InvGamma\left(\frac{a_{o_i}(\mathcal{D}^{c(i)}) - v_i(\mathcal{D}^{c(i)}) - 2}{2}, \frac{1}{2}\boldsymbol{U}_{i|pa_i(\mathcal{D}^{c(i)})}\right),$$

and then $a_{o_i}(\mathcal{D}^{c(i)}) - v_i(\mathcal{D}^{c(i)}) - 2 = a + p_n - 2(p_n - v_i(\mathcal{D}^{c(i)})) + 3 - v_i(\mathcal{D}^{c(i)}) - 2 = a - p_n + v_i(\mathcal{D}) + 1$. In conclusion, for any given DAG, the specification of the compatible prior for the Cholesky parameter associated to each node $i \in V$, $(D_{ii}, \boldsymbol{L}_{pa_i(\mathcal{D})\times i})$, requires only knowledge of the cardinality of its set of parents, as we summarize in Algorithm 1.

**Algorithm 1** Construction of compatible prior for $(\boldsymbol{D}, \boldsymbol{L})$

```
Require: $\mathcal{D}$, parent ordering, $\boldsymbol{a}(\mathcal{D})$ and $\boldsymbol{U}$
  for each node $i \in V$ do
    $\boldsymbol{D}_{ii} \sim InvGamma\left(\frac{a-p_n+v_i(\mathcal{D})+1}{2}, \frac{1}{2}\boldsymbol{U}_{i|pa_i(\mathcal{D})}\right)$
    $\boldsymbol{L}_{pa_i(\mathcal{D})\times i}|\boldsymbol{D}_{ii} \sim N_{v_i(\mathcal{D})}\left(-(\boldsymbol{U}_{pa_i(\mathcal{D})})^{-1}\boldsymbol{U}_{pa_i(\mathcal{D})\times i}, \boldsymbol{D}_{ii}(\boldsymbol{U}_{pa_i(\mathcal{D})})^{-1}\right)$
  end for
  $p(\boldsymbol{D}, \boldsymbol{L}) = \prod_{i\in V} p(\boldsymbol{D}_{ii}, \boldsymbol{L}_{pa_i(\mathcal{D})\times i})$
```

The compatible DAG-Wishart prior has $a_{o_i}(\mathcal{D}) = a - p_n + 2v_i(\mathcal{D}) + 3 = 2v_i(\mathcal{D}) + (a - p_n + 3)$. Interestingly, this has the structure $a_{o_i}(\mathcal{D}) = cv_i(\mathcal{D}) + b$ recommended in the experimental section of Ben-David et al. (2015). Since $a > p_n - 1$ but otherwise is a free hyperparameter, it follows that $a_{o_i}(\mathcal{D}) = 2v_i(\mathcal{D}) + b$ $(b > 2)$ ensures compatibility of priors.

Furthermore, the regularity Assumption 5 of Cao et al. (2019) requires, for all $i \in V$ and $\mathcal{D}$, that $2 < a_{o_i}(\mathcal{D}) - v_i(\mathcal{D}) < d$ for some constant $d$. When the shape hyperparameters are constrained by the compatibility requirements, the left-hand-side inequality $a_{o_i}(\mathcal{D}) - v_i(\mathcal{D}) > 2$ is satisfied; on the other hand, since $a_{o_i}(\mathcal{D}) - v_i(\mathcal{D}) < v_i(\mathcal{D}) + b$, the right-hand-side inequality is satisfied only if the cardinality of the parent sets has an upper bound as $n$ grows (a sparsity condition); see also Proposition 3.5 below.

**Example 3.4.** *Consider DAG $\mathcal{D}$ in Figure 1, whose node labels satisfy a parent ordering, with corresponding complete DAGs $\mathcal{D}^{c(i)}, i = A, B, C, D$. A compatible prior for $(\boldsymbol{D}, \boldsymbol{L})$ is assigned, for any $a > 3$, as follows:*

$$\begin{aligned}
\boldsymbol{D}_{DD} &\sim \mathit{InvGamma}\left(\frac{a-3}{2}, \frac{1}{2}\boldsymbol{U}_{DD}\right)\\
\boldsymbol{D}_{CC} &\sim \mathit{InvGamma}\left(\frac{a-2}{2}, \frac{1}{2}\boldsymbol{U}_{C|D}\right)\\
\boldsymbol{L}_{DC}|\boldsymbol{D}_{CC} &\sim N_1\left(-\boldsymbol{U}_{DC}/\boldsymbol{U}_{DD}, \boldsymbol{D}_{CC}/\boldsymbol{U}_{DD}\right)\\
\boldsymbol{D}_{BB} &\sim \mathit{InvGamma}\left(\frac{a-2}{2}, \frac{1}{2}\boldsymbol{U}_{B|D}\right)\\
\boldsymbol{L}_{DB}|\boldsymbol{D}_{BB} &\sim N_1\left(-\boldsymbol{U}_{DB}/\boldsymbol{U}_{DD}, \boldsymbol{D}_{BB}/\boldsymbol{U}_{DD}\right)\\
\boldsymbol{D}_{AA} &\sim \mathit{InvGamma}\left(\frac{a-1}{2}, \frac{1}{2}\boldsymbol{U}_{A|BC}\right)\\
\boldsymbol{L}_{BC,A}|\boldsymbol{D}_{AA} &\sim N_2\left(-\boldsymbol{U}^{-1}_{BC,BC}\boldsymbol{U}_{BC,A}, \boldsymbol{U}^{-1}_{BC,BC}\boldsymbol{D}_{AA}\right),
\end{aligned}$$

*where the shape hyperparameter for the prior of $\boldsymbol{D}_{DD}$ is given by $a + p_n - 2v_D(\mathcal{D}) + 1$, with $p_n = 2$ and $v_D(\mathcal{D}) = 1$, and similarly for the remaining priors.*

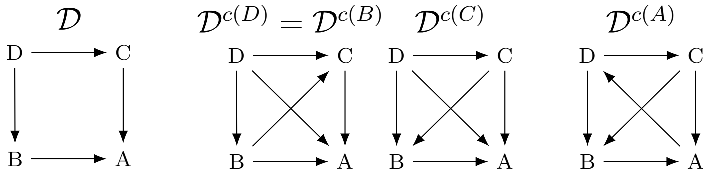

Fig 1. *A DAG $\mathcal{D}$ with corresponding complete DAGs $\mathcal{D}^{c(i)}$, $i = A, B, C, D$ for compatible prior specification.*

The next proposition specializes the result ii) of Proposition 3.1 to the case of compatible DAG-Wishart priors.

**Proposition 3.5.** *Let $\mathcal{D}_0$ be the true DAG and $(\boldsymbol{D}, \boldsymbol{L}) \in \Theta_{\mathcal{D}_0}$. For $(\boldsymbol{D}, \boldsymbol{L}) \sim \pi^{\Theta_{\mathcal{D}_0}}_{\boldsymbol{U},\boldsymbol{a}(\mathcal{D}_0)}$ and $\boldsymbol{Y}|((\boldsymbol{D}, \boldsymbol{L})) \sim N_{p_n}(\boldsymbol{0}, (\boldsymbol{L}\boldsymbol{D}^{-1}\boldsymbol{L}^\top)^{-1})$, if $a_{o_i}(\mathcal{D}) = a - p_n + 2v_i(\mathcal{D}) + 3$, for some $a \geq p_n$, $i \in V$,*

$$\frac{\pi(\mathcal{D}|\boldsymbol{Y})}{\pi(\mathcal{D}_0|\boldsymbol{Y})} = \frac{\pi(\mathcal{D})}{\pi(\mathcal{D}_0)}, \text{ for all } \mathcal{D} \in [\mathcal{D}_0],\ \mathcal{D} \neq \mathcal{D}_0.$$

*Furthermore, if* $v_i(\mathcal{D}) < d$*, for all* $i \in V$*, all* $n$ *and some finite constant* $d$*, then*

$$\max_{\mathcal{D} \notin [\mathcal{D}_0]} \frac{\pi(\mathcal{D}|\boldsymbol{Y})}{\pi(\mathcal{D}_0|\boldsymbol{Y})} \xrightarrow{\bar{P}} 0.$$

*Proof.* See the Appendix. □

We remark that the first statement in Proposition 3.5 is not an asymptotic result but it holds for each finite sample size $n$. Since score equivalence for Markov equivalent DAGs is a natural and important requirement we believe that posterior ratio consistency should only be required outside the Markov equivalence class of the true DAG. Our compatible DAG-Wishart achieves this double aim.

Propositions 3.1 and 3.5 together clarify that in the comparison between two DAGs, posterior ratio consistency and score equivalence of Markov equivalent DAGs are conflicting goals. In general a DAG is not identifiable based on observational data alone. This is still true if observational and interventional data are entertained (Hauser and Bühlmann, 2015). Outside the Gaussian setup identifiability of the true generating DAG is possible (Shimizu et al., 2006; Peters et al., 2011). Interestingly within the Gaussian setting identification is still possible using a structural equation model with noise components all having the same variance (Peters and Bühlmann, 2014).

**Example 3.6** (continued)**.** *Back to the example in Section 3.1, the compatible DAG-Wishart prior requires*

$$\begin{aligned} a_1(\mathcal{D}_1) &= a_1(\mathcal{D}_2) = a + 3, \\ a_2(\mathcal{D}_1) &= a_2(\mathcal{D}_2) = a + 1, \end{aligned}$$

*for some* $a > 1$*. Therefore,*

$$\begin{aligned} \mathcal{Z}_{\mathcal{D}_1}\left(n\tilde{\boldsymbol{S}}, n + \boldsymbol{a}(\mathcal{D}_1)\right) &= \frac{\Gamma\left(\frac{a+n}{2}\right)\Gamma\left(\frac{a+n-1}{2}\right)2^{a+n}\sqrt{\pi}}{\det\left(n\tilde{\boldsymbol{S}}\right)^{\frac{a+n}{2}}} \\ &= \mathcal{Z}_{\mathcal{D}_2}\left(n\tilde{\boldsymbol{S}}, n + \boldsymbol{a}(\mathcal{D}_2)\right). \end{aligned}$$

*and similarly* $\mathcal{Z}_{\mathcal{D}_1}(\boldsymbol{U}, \boldsymbol{a}(\mathcal{D}_1)) = \mathcal{Z}_{\mathcal{D}_2}(\boldsymbol{U}, \boldsymbol{a}(\mathcal{D}_2))$*. Coherently with equivalents graphs, the Bayes Factor is then equal to one, solving the contradiction.*

## 4. Simulation studies

In this section we present two simulation studies to contrast the behavior of compatible and non-compatible, DAG-Wishart priors with regard to model choice. With reference to the Example of Section 3.1, we consider two scenarios: 1) the true DAG model is that of independence between the two variables and the alternative is DAG $\mathcal{D}_1$, so the models are not equivalent; 2) the true DAG

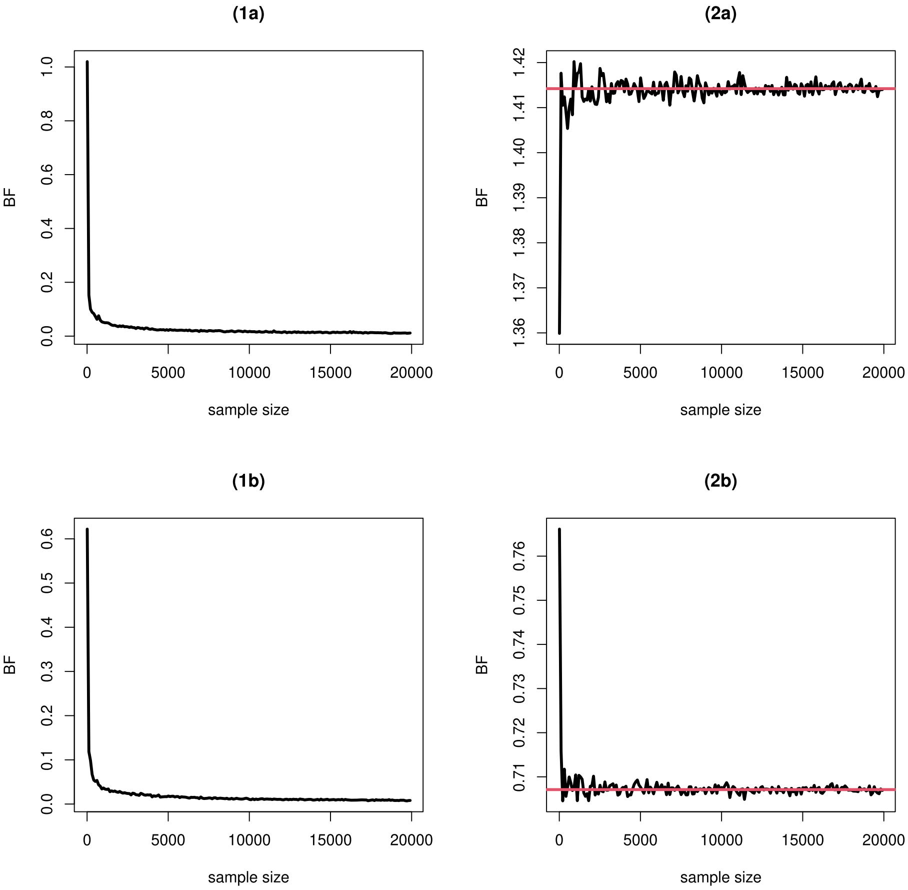

Fig 2. *Bayes factor (BF) between competing and true model using a non-compatible DAG-Wishart prior. Scenario 1): models are not Markov equivalent; scenario 2): models are Markov equivalent. Reported BF is an average over $M = 100$ repetitions. The theoretical asymptotic value is highlighted as a horizontal line.*

is $\mathcal{D}_1$ and the alternative is $\mathcal{D}_2$, so the models are equivalent. In both scenarios, the prior hyperparameter $\boldsymbol{U}$ is set equal to the identity matrix; additionally $\boldsymbol{a}(\mathcal{D}) = (11, 10)$ for all DAG models involved, so that the compatibility requirement explicated in Section 3 is not satisfied. For each scenario we generate $M = 100$ data sets, and we replicate the experiment under two settings: (a) $\sigma_A = \sqrt{2}$, $\sigma_B = 1$ and (b) $\sigma_A = 1$, $\sigma_B = \sqrt{2}$. For each of the four combinations (1a), (1b), (2a) and (2b), Figure 2 reports the average Bayes factor (over the $M$ repetitions) of the competing model versus the true generating model as a function of the sample size. As expected, under (1a) and (1b), wherein DAGs are not Markov equivalent, we observe Bayes factor consistency, regardless of the values of $\sigma_A$ and $\sigma_B$. On the contrary, under (2a) and (2b) the Bayes factor converges to a theoretical value that is finite and, respectively, strictly lower and higher than one. Therefore in the latter scenarios neither Bayes factor consistency nor compatibility is assured, and even if the data are in both cases generated in accordance with $\mathcal{D}_1$, the two scenarios suggest contradicting results, depending on the true underlying variances.

In a second study we extend our analysis to include compatible, as well as non-compatible, DAG Wishart priors for the comparison of Markov equivalent and non-equivalent DAGs. Specifically, we randomly generate three non-equivalent DAGs using the `R` package `pcalg` (Kalisch et al., 2012; Hauser and Bühlmann, 2012), in a sparse setting with edge inclusion probability $3/(2p_n - 2)$ (Peters and Bühlmann, 2014; Castelletti et al., 2018). We also generate three equivalent DAGs: we start from one randomly chosen DAG, and we generate a new DAG by attempting to invert the direction of a randomly chosen edge, until the inversion does not alter the corresponding equivalence class. Shape hyperparameters $\boldsymbol{a}(\mathcal{D})$ for non-compatible priors have been all set equal to $v_i(\mathcal{D})+3$, whilst compatible $\boldsymbol{a}(\mathcal{D})$ are set to $2v_i(\mathcal{D}) + 3$. The results for sample size $n = 100$ and number of nodes $p_n \in \{20, 50, 200\}$ are shown in Tables 1, 2 and 3, respectively. For $M = 100$ repetitions, we compute the log Bayes factors between the DAG in the column against the DAG in the row. For each given DAG, each dataset is obtained with $n$ random extractions from the zero-mean multivariate Gaussian distribution, Markov with respect to the row DAG, $N_{p_n}\left(\mathbf{0}, \left(\boldsymbol{L}\boldsymbol{D}^{-1}\boldsymbol{L}^{\top}\right)^{-1}\right)$, where $\boldsymbol{D}_{ii} = 1$ for all $i \in V$ and each $\boldsymbol{L}_{ij}$, $i, j \in V$, are uniformly chosen in the interval $[0.1, 1]$ if $i \in \mathrm{pa}_j(\mathcal{D})$, and $\boldsymbol{L}_{ij} = 0$ otherwise. We then report the mean and standard deviation (in parenthesis) of each log Bayes factor, and highlight the model with the highest log-BF in bold. We also report the proportion of times that the corresponding model on the column is selected. Across all tables, we note that when comparing non-equivalent DAGs, the correct model is always chosen both with compatible and non-compatible hyperparameter specifications. On the other hand, when we compare equivalent graphs, non-compatible choices for $\boldsymbol{a}(\mathcal{D})$ may lead to wrong conclusions, whilst compatible specifications return, by construction, Bayes factors that do not discriminate among models.

## 5. Discussion

The DAG-Wishart distribution is a flexible prior for the parameters associated to the Cholesky decomposition of the precision matrix of a Gaussian model Markov with respect to a given DAG. By allowing distinct shape hyperparameters (one for each vertex) it can enhance estimation in high dimensional settings through differential shrinkage (Ben-David et al., 2015). However care must be exercised when using it in a model selection setting. In particular we show that consistency properties proved in Cao et al. (2019) only hold for DAGs that do not belong to the same Markov equivalence class. Furthermore, if the shape hyperparameters of the prior distribution are freely chosen, two Markov equivalent DAGs will be assigned different marginal likelihoods, an unreasonable feature in model selection based on observational data. We present a construc-

Table 1. *Simulation study, $p_n = 20$ and $n = 100$. Average log-Bayes factors between models on columns and rows, for compatible and non-compatible DAG-Wishart priors, for three Markov equivalent, as well as non-equivalent DAGs. For each row 100 datasets were generated according to the corresponding DAG. Highest log-Bayes factor is reported in bold, and standard deviation in parenthesis, together with proportion of times corresponding model on the column is selected.*

| | Non-compatible prior | | | Compatible prior | | |
|---|---|---|---|---|---|---|
| | $\mathcal{D}_1$ | $\mathcal{D}_2$ | $\mathcal{D}_3$ | $\mathcal{D}_1$ | $\mathcal{D}_2$ | $\mathcal{D}_3$ |
| Non-equivalent DAGs | | | | | | |
| $\mathcal{D}_1$ | **0.0000** | -141.8839 | -150.9445 | **0.0000** | -144.9076 | -154.0877 |
| | (0.0000) | (14.8990) | (15.1606) | (0.0000) | (15.1913) | (15.4763) |
| | 100% | 0% | 0% | 100% | 0% | 0% |
| $\mathcal{D}_2$ | -175.7093 | **0.0000** | -135.9295 | -180.3152 | **0.0000** | -138.4078 |
| | (18.0930) | (0.0000) | (17.0051) | (18.2697) | (0.0000) | (17.1789) |
| | 0% | 100% | 0% | 0% | 100% | 0% |
| $\mathcal{D}_3$ | -138.2916 | -100.1602 | **0.0000** | -142.2880 | -101.6536 | **0.0000** |
| | (15.5685) | (11.9462) | (0.0000) | (15.7725) | (12.1668) | (0.0000) |
| | 0% | 0% | 100% | 0% | 0% | 100% |
| Equivalent DAGs | | | | | | |
| $\mathcal{D}_1$ | 0.0000 | -0.0559 | **0.2587** | 0.0000 | 0.0000 | 0.0000 |
| | (0.0000) | (0.0869) | (0.0879) | (0.0000) | (0.0000) | (0.0000) |
| | 1% | 2% | 97% | – | – | – |
| $\mathcal{D}_2$ | **0.0061** | 0.0000 | -0.0778 | 0.0000 | 0.0000 | 0.0000 |
| | (0.0975) | (0.0000) | (0.1229) | (0.0000) | (0.0000) | (0.0000) |
| | 44% | 49% | 7% | – | – | – |
| $\mathcal{D}_3$ | **0.0142** | -0.0467 | 0.0000 | 0.0000 | 0.0000 | 0.0000 |
| | (0.0853) | (0.1184) | (0.0000) | (0.0000) | (0.0000) | (0.0000) |
| | 43% | 17% | 40% | – | – | – |

Table 2. *Simulation study, $p_n = 50$ and $n = 100$. Average log-Bayes factors between models on columns and rows, for compatible and non-compatible DAG-Wishart priors, for three Markov equivalent, as well as non-equivalent DAGs. For each row 100 datasets were generated according to the corresponding DAG. Highest log-Bayes factor is reported in bold, and standard deviation in parenthesis, together with proportion of times corresponding model on the column is selected.*

| | Non-compatible prior | | | Compatible prior | | |
|---|---|---|---|---|---|---|
| | $\mathcal{D}_1$ | $\mathcal{D}_2$ | $\mathcal{D}_3$ | $\mathcal{D}_1$ | $\mathcal{D}_2$ | $\mathcal{D}_3$ |
| Non-equivalent DAGs | | | | | | |
| $\mathcal{D}_1$ | **0.0000** | -413.5615 | -368.6975 | **0.0000** | -426.6263 | -375.3692 |
| | (0.0000) | (28.7591) | (27.5878) | (0.0000) | (29.3854) | (28.0724) |
| | 100% | 0% | 0% | 100% | 0% | 0% |
| $\mathcal{D}_2$ | -573.4748 | **0.0000** | -496.4580 | -583.3696 | **0.0000** | -502.6452 |
| | (31.1688) | (0.0000) | (28.5000) | (31.8421) | (0.0000) | (29.0201) |
| | 0% | 100% | 0% | 0% | 100% | 0% |
| $\mathcal{D}_3$ | -425.8196 | -388.6999 | **0.0000** | -434.5435 | -401.2242 | **0.0000** |
| | (26.2763) | (22.6814) | (0.0000) | (26.6887) | (23.1208) | (0.0000) |
| | 0% | 0% | 100% | 0% | 0% | 100% |
| Equivalent DAGs | | | | | | |
| $\mathcal{D}_1$ | 0.0000 | **0.0903** | -0.2258 | 0.0000 | 0.0000 | 0.0000 |
| | (0.0000) | (0.1028) | (0.1137) | (0.0000) | (0.0000) | (0.0000) |
| | 19% | 79% | 2% | – | – | – |
| $\mathcal{D}_2$ | **0.2385** | 0.0000 | 0.1554 | 0.0000 | 0.0000 | 0.0000 |
| | (0.0967) | (0.0000) | (0.1503) | (0.0000) | (0.0000) | (0.0000) |
| | 81% | 0% | 19% | – | – | – |
| $\mathcal{D}_3$ | 0.0547 | **0.1295** | 0.0000 | 0.0000 | 0.0000 | 0.0000 |
| | (0.1049) | (0.1395) | (0.0000) | (0.0000) | (0.0000) | (0.0000) |
| | 8% | 76% | 16% | – | – | – |

Table 3. *Simulation study, $p_n = 200$ and $n = 100$. Average log-Bayes factors between models on columns and rows, for compatible and non-compatible DAG-Wishart priors, for three Markov equivalent, as well as non-equivalent DAGs. For each row 100 datasets were generated according to the corresponding DAG. Highest log-Bayes factor is reported in bold, and standard deviation in parenthesis, together with proportion of times corresponding model on the column is selected.*

| | Non-compatible prior | | | Compatible prior | | |
|---|---|---|---|---|---|---|
| | $\mathcal{D}_1$ | $\mathcal{D}_2$ | $\mathcal{D}_3$ | $\mathcal{D}_1$ | $\mathcal{D}_2$ | $\mathcal{D}_3$ |
| Non-equivalent DAGs | | | | | | |
| $\mathcal{D}_1$ | **0.0000** | -2125.1471 | -2105.1915 | **0.0000** | -2172.6796 | -2144.7647 |
| | (0.0000) | (61.1288) | (60.3531) | (0.0000) | (62.1403) | (61.1484) |
| | 100% | 0% | 0% | 100% | 0% | 0% |
| $\mathcal{D}_2$ | -1787.5176 | **0.0000** | -1767.1222 | -1822.8475 | **0.0000** | -1806.6065 |
| | (55.6965) | (0.0000) | (56.7089) | (56.7273) | (0.0000) | (57.8676) |
| | 0% | 100% | 0% | 0% | 100% | 0% |
| $\mathcal{D}_3$ | -1899.1510 | -1904.9126 | **0.0000** | -1929.2827 | -1941.2513 | **0.0000** |
| | (55.9835) | (57.5023) | (0.0000) | (56.8492) | (58.3046) | (0.0000) |
| | 0% | 0% | 100% | 0% | 0% | 100% |
| Equivalent DAGs | | | | | | |
| $\mathcal{D}_1$ | **0.0000** | -0.2662 | -0.0570 | 0.0000 | 0.0000 | 0.0000 |
| | (0.0000) | (0.0717) | (0.0705) | (0.0000) | (0.0000) | (0.0000) |
| | 82% | 0% | 18% | – | – | – |
| $\mathcal{D}_2$ | 0.4220 | 0.0000 | **0.8709** | 0.0000 | 0.0000 | 0.0000 |
| | (0.1007) | (0.0000) | (0.1411) | (0.0000) | (0.0000) | (0.0000) |
| | 0% | 0% | 100% | – | – | – |
| $\mathcal{D}_3$ | -0.4416 | -0.7750 | **0.0000** | 0.0000 | 0.0000 | 0.0000 |
| | (0.1051) | (0.1502) | (0.0000) | (0.0000) | (0.0000) | (0.0000) |
| | 0% | 0% | 100% | – | – | – |

tive method for eliciting a DAG-Wishart prior that guarantees score equivalence among Markov equivalent DAGs, by imposing constraints on the shape hyperparameters.

The analysis in Cao et al. (2019) and in this paper is predicated on the choice of a proper prior for the Cholesky decomposition parameter. Often however, either because prior information is lacking or for a more neutral scientific communication of findings, an Objective Bayes (OB) approach is deemed preferable (Berger, 2006). An effective OB approach for model selection is represented by the Fractional Bayes Factor (FBF, O'Hagan 1995). Briefly the idea is to start with a default, typically improper, prior which is subsequently trained, using a *fraction* of the entire likelihood, to produce a prior which is then coupled with the likelihood raised to the complementary fraction to obtain an FBF marginal likelihood for model comparison; see also Consonni and La Rocca (2012) and Consonni, La Rocca and Peluso (2017) for an implementation of the FBF in a graphical model setup. An interesting line of research is the construction of an FBF compatible DAG-Wishart prior for model selection. Proving consistency in high dimensional settings is challenging because the prior is data-dependent. We conjecture that suitable sparsity conditions on the complexity of the graph as $n$ increases might allow the fraction to grow at a sufficiently slow rate to achieve consistency.

## Appendix

*Proof of Proposition 3.1.* We first prove ii). Assume that $\mathcal{D} \in [\mathcal{D}_0]$, that is $\mathcal{D}$ is Markov equivalent to the true DAG $\mathcal{D}_0$. By Chickering (1995, Theorem 2), there exists a sequence of Markov equivalent DAGs $\mathcal{D}_0 =: \mathcal{D}^0, \mathcal{D}^1, \dots, \mathcal{D}^{k-1}, \mathcal{D}^k := \mathcal{D}$ from $\mathcal{D}_0$ to $\mathcal{D}$, such that graphs $\mathcal{D}^{i-1}$ and $\mathcal{D}^i$ adjacent in the sequence differ only for the reversal of a *covered edge*, that is for an edge $(i,j)$ of $\mathcal{D}$ for which $pa_j(\mathcal{D}) = pa_i(\mathcal{D}) \cup \{i\}$. Therefore it is sufficient to compare $\mathcal{D}$ and $\mathcal{D}_0$ differing only by one covered edge reversal. This means that, for some nodes $i_1$ and $i_2$,

$$\begin{aligned} pa_{i_1}(\mathcal{D}) &= pa_{i_2}(\mathcal{D}_0) =: pa_*, \\ pa_{i_1}(\mathcal{D}_0) &= pa_* \cup i_2, \\ pa_{i_2}(\mathcal{D}) &= pa_* \cup i_1 \\ pa_i(\mathcal{D}) &= pa_i(\mathcal{D}_0) \text{ for all } i \in V \setminus \{i_1, i_2\}. \end{aligned}$$

We first write the posterior ratio as

$$\begin{aligned} \frac{\pi(\mathcal{D}|\boldsymbol{Y})}{\pi(\mathcal{D}_0|\boldsymbol{Y})} &= \frac{m(\boldsymbol{Y}|\mathcal{D})}{m(\boldsymbol{Y}|\mathcal{D}_0)} \frac{\pi(\mathcal{D})}{\pi(\mathcal{D}_0)} \\ &= \frac{\mathcal{Z}_{\mathcal{D}}\left(n\tilde{\boldsymbol{S}}, n + \boldsymbol{a}(\mathcal{D})\right) \mathcal{Z}_{\mathcal{D}_0}\left(\boldsymbol{U}, \boldsymbol{a}(\mathcal{D}_0)\right)}{\mathcal{Z}_{\mathcal{D}_0}\left(n\tilde{\boldsymbol{S}}, n + \boldsymbol{a}(\mathcal{D}_0)\right) \mathcal{Z}_{\mathcal{D}}\left(\boldsymbol{U}, \boldsymbol{a}(\mathcal{D})\right)} \frac{\pi(\mathcal{D})}{\pi(\mathcal{D}_0)}. \end{aligned}$$

Define the useful quantities $v_* := |pa_*|$, $c_i(\mathcal{D}) := a_{o_i}(\mathcal{D}) - v_i(\mathcal{D})$, $\tilde{c}_i(\mathcal{D}) := a_{o_i}(\mathcal{D}) - v_i(\mathcal{D}_0)$ and $c_i^*(\mathcal{D}) := a_{o_i}(\mathcal{D}) - v_*$. Note that $c_i(\mathcal{D}) - c_i(\mathcal{D}_0) =$

$\tilde{c}_i(\mathcal{D}) - \tilde{c}_i(\mathcal{D}_0) = c_i^*(\mathcal{D}) - c_i^*(\mathcal{D}_0) = a_{o_i}(\mathcal{D}) - a_{o_i}(\mathcal{D}_0)$. We can write the ratio $\mathcal{Z}_{\mathcal{D}}\left(\boldsymbol{U}, \boldsymbol{a}(\mathcal{D})\right) / \mathcal{Z}_{\mathcal{D}_0}\left(\boldsymbol{U}, \boldsymbol{a}(\mathcal{D}_0)\right)$ as

$$\prod_{i\in\{i_1,i_2\}} \left\{ \frac{\Gamma\left(\frac{c_i^*(\mathcal{D})}{2} - 1\right)}{\Gamma\left(\frac{c_i^*(\mathcal{D}_0)-1}{2} - 1\right)} \frac{\left|\boldsymbol{U}_{pa_i(\mathcal{D})}\right|^{\frac{c_i^*(\mathcal{D})-1}{2}-1}}{\left|\boldsymbol{U}_{pa_i(\mathcal{D}_0)}\right|^{\frac{c_i^*(\mathcal{D}_0)-2}{2}-1}} \frac{\left|\boldsymbol{U}_{fa_i(\mathcal{D}_0)}\right|^{\frac{c_i^*(\mathcal{D}_0)-1}{2}-1}}{\left|\boldsymbol{U}_{fa_i(\mathcal{D})}\right|^{\frac{c_i^*(\mathcal{D})}{2}-1}} \right\} \cdot$$
$$\prod_{i\in V\setminus\{i_1,i_2\}} \left\{ \frac{\Gamma\left(\frac{\tilde{c}_i(\mathcal{D})}{2} - 1\right)}{\Gamma\left(\frac{\tilde{c}_i(\mathcal{D}_0)}{2} - 1\right)} \left(\frac{\left|\boldsymbol{U}_{pa_i(\mathcal{D}_0)}\right|}{\left|\boldsymbol{U}_{fa_i(\mathcal{D}_0)}\right|}\right)^{\frac{a_{o_i}(\mathcal{D})-a_{o_i}(\mathcal{D}_0)}{2}} \right\} \left(\sqrt{2}\right)^{\sum_{i=1}^{pn}(a_i(\mathcal{D})-a_i(\mathcal{D}_0))}$$

Using the following relationships:

$$\left|\boldsymbol{U}_{pa_{i_1}(\mathcal{D})}\right| = \left|\boldsymbol{U}_{pa_{i_2}(\mathcal{D}_0)}\right| = \left|\boldsymbol{U}_{pa_*}\right|, \tag{8}$$
$$\left|\boldsymbol{U}_{pa_{i_1}(\mathcal{D}_0)}\right| = \left|\boldsymbol{U}_{fa_{i_2}(\mathcal{D}_0)}\right| = \left|\boldsymbol{U}_{pa_*}\right| \boldsymbol{U}_{i_2|pa_*}, \tag{9}$$
$$\left|\boldsymbol{U}_{fa_{i_1}(\mathcal{D})}\right| = \left|\boldsymbol{U}_{pa_{i_2}(\mathcal{D})}\right| = \left|\boldsymbol{U}_{pa_*}\right| \boldsymbol{U}_{i_1|pa_*}, \tag{10}$$
$$\left|\boldsymbol{U}_{fa_{i_1}(\mathcal{D}_0)}\right| = \left|\boldsymbol{U}_{fa_{i_2}(\mathcal{D})}\right| = \left|\boldsymbol{U}_{pa_*}\right| \left|\boldsymbol{U}_{i_1,i_2|pa_*}\right|, \tag{11}$$

we can then write $\mathcal{Z}_{\mathcal{D}}\left(\boldsymbol{U}, \boldsymbol{a}(\mathcal{D})\right) / \mathcal{Z}_{\mathcal{D}_0}\left(\boldsymbol{U}, \boldsymbol{a}(\mathcal{D}_0)\right)$ as

$$\prod_{i\in\{i_1,i_2\}} \left\{ \frac{\Gamma\left(\frac{c_i^*(\mathcal{D})}{2} - 1\right)}{\Gamma\left(\frac{c_i^*(\mathcal{D}_0)-1}{2} - 1\right)} \right\} \frac{\boldsymbol{U}_{i_1|pa_*}^{\frac{a_{o_{i_2}}(\mathcal{D})-a_{o_{i_1}}(\mathcal{D})}{2}-1}}{\boldsymbol{U}_{i_2|pa_*}^{\frac{a_{o_{i_1}}(\mathcal{D}_0)-a_{o_{i_2}}(\mathcal{D}_0)}{2}-1}} \left|\boldsymbol{U}_{i_1,i_2|pa_*}\right|^{\frac{a_{o_{i_1}}(\mathcal{D}_0)-a_{o_{i_2}}(\mathcal{D})}{2}}$$
$$\cdot \prod_{i\in V\setminus\{i_1,i_2\}} \left\{ \frac{\Gamma\left(\frac{\tilde{c}_i(\mathcal{D})}{2} - 1\right)}{\Gamma\left(\frac{\tilde{c}_i(\mathcal{D}_0)}{2} - 1\right) \boldsymbol{U}_{i|pa_i(\mathcal{D}_0)}^{\frac{a_{o_i}(\mathcal{D})-a_{o_i}(\mathcal{D}_0)}{2}}} \right\} \left(\sqrt{2}\right)^{\sum_{i=1}^{pn}(a_i(\mathcal{D})-a_i(\mathcal{D}_0))}.$$

The terms $\tilde{c}_i(\mathcal{D})$ and $c_i^*(\mathcal{D})$ are bounded for all $i$ and $\mathcal{D}$ from Assumption 5 of Cao et al. (2019), and therefore $C^1_{\mathcal{D},\mathcal{D}_0} := \mathcal{Z}_{\mathcal{D}_0}\left(\boldsymbol{U}, \boldsymbol{a}(\mathcal{D}_0)\right) / \mathcal{Z}_{\mathcal{D}}\left(\boldsymbol{U}, \boldsymbol{a}(\mathcal{D})\right) \in (0, \infty)$.

Define $c_i^n(\mathcal{D}) := a_{o_i}(\mathcal{D}) - v_i(\mathcal{D}) + n$ and note that

$$\frac{\Gamma\left(\frac{c_i^n(\mathcal{D})}{2} - 1\right)}{\Gamma\left(\frac{c_i^n(\mathcal{D}_0)}{2} - 1\right)} = \frac{\Gamma\left(\frac{c_i^n(\mathcal{D}_0)+(c_i^n(\mathcal{D})-c_i^n(\mathcal{D}_0))}{2} - 1\right)}{\Gamma\left(\frac{c_i^n(\mathcal{D}_0)}{2} - 1\right)} \sim n^{\frac{c_i^n(\mathcal{D})-c_i^n(\mathcal{D}_0)}{2}}. \tag{12}$$

We can state

$$\frac{\pi(\mathcal{D}|\boldsymbol{Y})}{\pi(\mathcal{D}_0|\boldsymbol{Y})} = C^1_{\mathcal{D},\mathcal{D}_0} \frac{\pi(\mathcal{D})}{\pi(\mathcal{D}_0)} \frac{\mathcal{Z}_{\mathcal{D}}\left(n\tilde{\boldsymbol{S}}, \boldsymbol{a}(\mathcal{D}) + n\right)}{\mathcal{Z}_{\mathcal{D}_0}\left(n\tilde{\boldsymbol{S}}, \boldsymbol{a}(\mathcal{D}_0) + n\right)}$$

$$\asymp_{\bar{P}} \quad C^2_{\mathcal{D},\mathcal{D}_0} \frac{\pi(\mathcal{D})}{\pi(\mathcal{D}_0)} \frac{\tilde{\boldsymbol{S}}_{i_1|pa_*}^{\frac{a_{o_{i_2}}(\mathcal{D})-a_{o_{i_1}}(\mathcal{D})}{2}-1}}{\tilde{\boldsymbol{S}}_{i_2|pa_*}^{\frac{a_{o_{i_1}}(\mathcal{D}_0)-a_{o_{i_2}}(\mathcal{D}_0)}{2}-1}} \frac{\left|\tilde{\boldsymbol{S}}_{i_2|pa_*}\right|^{\frac{a_{o_{i_1}}(\mathcal{D}_0)-a_{o_{i_2}}(\mathcal{D})}{2}}}{\prod_{i \in pa_*} \tilde{\boldsymbol{S}}_{i|pa_i(\boldsymbol{D}_0)}^{\frac{a_{o_i}(\mathcal{D}^j)-a_{o_i}(\mathcal{D}_0)}{2}}}$$

with $C^2_{\mathcal{D},\mathcal{D}_0} := C^1_{\mathcal{D},\mathcal{D}_0}\left(\sqrt{2}\right)^{\sum_{i=1}^{p_n}(a_i(\mathcal{D})-a_i(\mathcal{D}_0))}$, and where $x_n \asymp_{\bar{P}} y_n$ for two generic sequences $x_n$ and $y_n$ means that $x_n$ is of the same order as $y_n$ in probability $\bar{P}$. Note that the sub-matrices of $\tilde{\boldsymbol{S}}$ in the formula above are of fixed dimension, and we can then rely on $\tilde{\boldsymbol{S}}_{i_1|pa_*} = \boldsymbol{\Sigma}_{i_1|pa_*} + O_p\left(1/\sqrt{n}\right)$, and similarly for the other quantities. Therefore we can further simplify the posterior ratio to

$$\frac{\pi(\mathcal{D}|\boldsymbol{Y})}{\pi(\mathcal{D}_0|\boldsymbol{Y})} \asymp_{\bar{P}} C^3_{\mathcal{D},\mathcal{D}_0} \frac{\pi(\mathcal{D})}{\pi(\mathcal{D}_0)},$$

with

$$C^3_{\mathcal{D},\mathcal{D}_0} := C^2_{\mathcal{D},\mathcal{D}_0} \frac{\boldsymbol{\Sigma}_{i_1|pa_*}^{\frac{a_{o_{i_2}}(\mathcal{D})-a_{o_{i_1}}(\mathcal{D})}{2}-1}}{\boldsymbol{\Sigma}_{i_2|pa_*}^{\frac{a_{o_{i_1}}(\mathcal{D}_0)-a_{o_{i_2}}(\mathcal{D}_0)}{2}-1}} \frac{\left|\boldsymbol{\Sigma}_{i_1,i_2|pa_*}\right|^{\frac{a_{o_{i_1}}(\mathcal{D}_0)-a_{o_{i_2}}(\mathcal{D})}{2}}}{\prod_{i \in pa_*} \boldsymbol{\Sigma}_{i|pa_i(\mathcal{D}_0)}^{\frac{a_{o_i}(\mathcal{D})-a_{o_i}(\mathcal{D}_0)}{2}}}.$$

With $k$ covered edges to be inverted between $\mathcal{D}$ and $\mathcal{D}_0$, we finally have

$$\frac{\pi(\mathcal{D}|\boldsymbol{Y})}{\pi(\mathcal{D}_0|\boldsymbol{Y})} = \prod_{j=1}^{k} \frac{\pi(\mathcal{D}^j|\boldsymbol{Y})}{\pi(\mathcal{D}^{j-1}|\boldsymbol{Y})} = C_{\mathcal{D},\mathcal{D}_0} \frac{\pi(\mathcal{D})}{\pi(\mathcal{D}_0)} \tag{13}$$

$$C_{\mathcal{D},\mathcal{D}_0} := \prod_{j=1}^{k} C^3_{\mathcal{D}^j,\mathcal{D}^{j-1}} \in (0,\infty).$$

This completes the proof of part ii). The proof of part i) which deals with posterior ratio consistency between $\mathcal{D}_0$ and a DAG $D \notin [\mathcal{D}_0]$ (not Markov equivalent to $\mathcal{D}_0$) follows from Cao et al. (2019, Theorem 4.1). □

*Proof of Proposition 3.5.* As outlined in the proof of Proposition 3.1, it is sufficient to compare equivalent $\mathcal{D}$ and $\mathcal{D}_0$ differing by one covered edge reversal, where the covered edge nodes are $i_1$ and $i_2$. Since $\mathcal{D}$ and $\mathcal{D}_0$ have the same skeleton, $\sum_i v_i(\mathcal{D}) = \sum_i v_i(\mathcal{D}_0)$, and then $\sum_i a_i(\mathcal{D}) = \sum_i a_i(\mathcal{D}_0)$. Using this fact, together with $a_i(\mathcal{D}) = a_i(\mathcal{D}_0)$ for all $i \in V \setminus \{i_1, i_2\}$, we can write

$$\frac{\mathcal{Z}_{\mathcal{D}}(\boldsymbol{U}, \boldsymbol{a}(\mathcal{D}))}{\mathcal{Z}_{\mathcal{D}_0}(\boldsymbol{U}, \boldsymbol{a}(\mathcal{D}_0))} = \prod_{i \in \{i_1,i_2\}} \left\{ \frac{\Gamma\left(\frac{c_i(\mathcal{D})-2}{2}\right)}{\Gamma\left(\frac{c_i(\mathcal{D}_0)-2}{2}\right)} \frac{|\boldsymbol{U}_{pa_i(\mathcal{D})}|^{\frac{c_i(\mathcal{D})-3}{2}}}{|\boldsymbol{U}_{pa_i(\mathcal{D}_0)}|^{\frac{c_i(\mathcal{D}_0)-3}{2}}} \frac{|\boldsymbol{U}_{fa_i(\mathcal{D}_0)}|^{\frac{c_i(\mathcal{D}_0)-2}{2}}}{|\boldsymbol{U}_{fa_i(\mathcal{D})}|^{\frac{c_i(\mathcal{D})-2}{2}}} \right\},$$

where $c_i(\mathcal{D}) = a_{o_i}(\mathcal{D}) - v_i(\mathcal{D})$. Since $v_{i_1}(\mathcal{D}) = v_{i_2}(\mathcal{D}_0)$ and $v_{i_2}(\mathcal{D}) = v_{i_1}(\mathcal{D}_0)$, also $c_{i_1}(\mathcal{D}) = c_{i_2}(\mathcal{D}_0)$ and $c_{i_2}(\mathcal{D}) = c_{i_1}(\mathcal{D}_0)$, and then the terms with Gamma functions cancel. Using also Equations (8)-(11), all other terms cancel and $\mathcal{Z}_{\mathcal{D}}\left(\boldsymbol{U}, \boldsymbol{a}(\mathcal{D})\right) / \mathcal{Z}_{\mathcal{D}_0}\left(\boldsymbol{U}, \boldsymbol{a}(\mathcal{D}_0)\right) = 1$. Similarly, the ratio between $\mathcal{Z}_{\mathcal{D}}(n\tilde{\boldsymbol{S}}, \boldsymbol{a}(\mathcal{D}) + n)$ and $\mathcal{Z}_{\mathcal{D}_0}(n\tilde{\boldsymbol{S}}, \boldsymbol{a}(\mathcal{D}_0) + n)$ is equal to one, proving the first statement.

If furthermore $v_i(\mathcal{D}) < d$, some finite $d$, for all $i \in V$ and for all $n$, then Assumption 5 of Cao et al. (2019) is respected and their Theorem 4.1 implies that $\max_{\mathcal{D} \notin [\mathcal{D}_0]} \pi(\mathcal{D}|\boldsymbol{Y})/\pi(\mathcal{D}_0|\boldsymbol{Y}) \xrightarrow{\bar{P}} 0$. □

## Acknowledgments

The authors thank the Editor, the Associate Editor and the Reviewer for useful suggestions on a first version of the paper. The research of the authors has been partially supported by a grant from UCSC (track D1) and by the EU COSTNET project (CA15109).

## References

Andersson, S. A., Madigan, D. and Perlman, M. D. (1997). A characterization of Markov equivalence classes for acyclic digraphs. *The Annals of Statistics* **25** 505–541. MR1439312 (99a:62076)

Ben-David, E., Li, T., Massam, H. and Rajaratnam, B. (2015). High dimensional Bayesian inference for Gaussian directed acyclic graph models. *arXiv preprint* arXiv:1109.4371.

Berger, J. (2006). The case for objective Bayesian analysis. *Bayesian Anal.* **1** 385–402. MR2221271

Cao, X., Khare, K., Ghosh, M. et al. (2019). Posterior graph selection and estimation consistency for high-dimensional Bayesian DAG models. *The Annals of Statistics* **47** 319–348. MR3909935

Castelletti, F., Consonni, G., Della Vedova, M. and Peluso, S. (2018). Learning Markov Equivalence Classes of Directed Acyclic Graphs: an Objective Bayes Approach. *Bayesian Analysis* **13** 1231–1256. MR3855370

Chickering, D. M. (1995). A transformational characterization of equivalent Bayesian network structures. In *Proceedings of the Eleventh conference on Uncertainty in artificial intelligence* 87–98. Morgan Kaufmann Publishers Inc.

Consonni, G. and La Rocca, L. (2012). Objective Bayes Factors for Gaussian Directed Acyclic Graphical Models. *Scandinavian Journal of Statistics* **39** 743–756. MR3000846

Consonni, G., La Rocca, L. and Peluso, S. (2017). Objective Bayes Covariate-Adjusted Sparse Graphical Model Selection. *Scandinavian Journal of Statistics* **3** 741–764. MR3687971

Consonni, G. and Veronese, P. (2008). Compatibility of Prior Specifications Across Linear Models. *Statistical Science* **23** 332–353. MR2483907

Consonni, G., Fouskakis, D., Liseo, B., Ntzoufras, I. et al. (2018). Prior distributions for objective Bayesian analysis. *Bayesian Analysis* **13** 627–679. MR3807861

Dawid, A. P. (2003). Causal inference using influence diagrams: the problem of partial compliance. In *Highly structured stochastic systems* (P. J. Green, N. L. Hjort and S. Richardson, eds.) 45–81. Oxford Univ. Press, Oxford. MR2082406

Dawid, A. P. and Lauritzen, S. L. (1993). Hyper Markov Laws in the Statistical Analysis of Decomposable Graphical Models. *The Annals of Statistics* **21** 1272–1317. MR1241267

Friedman, N. (2004). Inferring Cellular Networks Using Probabilistic Graphical Models. *Science* **303** 799–805.

Geiger, D. and Heckerman, D. (2002). Parameter priors for directed acyclic graphical models and the characterization of several probability distributions. *The Annals of Statistics* **30** 1412–1440. MR1936324 (2003h:62019)

Hauser, A. and Bühlmann, P. (2012). Characterization and greedy learning of interventional Markov equivalence classes of directed acyclic graphs. *Journal of Machine Learning Research* **13** 2409–2464. MR2973606

Hauser, A. and Bühlmann, P. (2015). Jointly interventional and observational data: estimation of interventional Markov equivalence classes of directed acyclic graphs. *Journal of the Royal Statistical Society. Series B (Methodology)* **77** 291–318. MR3299409

Kalisch, M., Mächler, M., Colombo, D., Maathuis, M. H. and Bühlmann, P. (2012). Causal Inference Using Graphical Models with the R Package pcalg. *Journal of Statistical Software* **47** 1–26.

Lauritzen, S. L. (1996). *Graphical Models*. Oxford University Press. MR1419991

Lauritzen, S. L. (2001). Causal inference from graphical models. In *Complex stochastic systems (Eindhoven, 1999). Monogr. Statist. Appl. Probab.* **87** 63–107. Chapman & Hall/CRC, Boca Raton, FL. MR1893411

Letac, G. and Massam, H. (2007). Wishart distributions for decomposable graphs. *Ann. Statist.* **35** 1278–1323. MR2341706 (2009j:62170)

Nagarajan, R. and Scutari, M. (2013). *Bayesian Networks in R with Applications in Systems Biology*. Springer, New York. ISBN 978-1-4614-6445-7, 978-1-4614-6446-4. MR3059206

O'Hagan, A. (1995). Fractional Bayes Factors for Model Comparison. *Journal of the Royal Statistical Society. Series B (Methodological)* **57** 99–138. MR1325379

Peters, J. and Bühlmann, P. (2014). Identifiability of Gaussian structural equation models with equal error variances. *Biometrika* **101** 219–228. MR3180667

Peters, J., Mooij, J., Janzing, D. and Schölkopf, B. (2011). Identifiability of causal graphs using functional models. 589-598. AUAI Press, Corvallis, OR, USA.

Pourahmadi, M. (2007). Cholesky decompositions and estimation of a covariance matrix: orthogonality of variance–correlation parameters. *Biometrika* **94** 1006–1013. MR2376812

Sachs, K., Perez, O., Pe'er, D., Lauffenburger, D. and Nolan, G. (2003). Causal protein-signaling networks derived from multiparameter single-cell data. *Science* **308** 504–506.

Shimizu, S., Hoyer, P. O., Hyvärinen, A. and Kerminen, A. (2006). A linear non-Gaussian acyclic model for causal discovery. *Journal of Machine Learning Research* **7** 2003–2030. MR2274431

Shojaie, A. and Michailidis, G. (2009). Analysis of gene sets based on the underlying regulatory network. *Journal of Computational Biology* **16** 407–426. MR2487566

Verma, T. and Pearl, J. (1991). Equivalence and Synthesis of Causal Models. In *Proceedings of the Sixth Annual Conference on Uncertainty in Artificial Intelligence. UAI 90* 255–270. Elsevier Science Inc., New York, NY, USA.