Compatible Priors for Casual Bayesian Networks

Valentina Leucari  
(Università di Pavia)

Guido Consonni  
(Università di Pavia)

# 26 (02-03)

Dipartimento di economia politica  
e metodi quantitativi  
Università degli studi di Pavia  
Via San Felice, 5  
I-27100 Pavia

Febbraio 2003
Compatible Priors for Causal Bayesian Networks

VALENTINA LEUCARI
Università di Pavia, Italy
vl@dimat.unipv.it

GUIDO CONSONNI
Università di Pavia, Italy
guido.consonni@unipv.it

SUMMARY
We consider discrete causal DAG-models (or Bayesian Networks) wherein the ordering of the variables is fixed across model structures. Given a prior on the parameter space of a model we describe a method for deriving a compatible prior on the parameter space of a submodel. This allows to generate automatically compatible priors for model parameters starting from a single prior relative to the largest entertained model. Our method makes use of a general procedure for constructing compatible priors for causal DAG-models, named reference conditioning, which is invariant within a suitable class of re-parameterisations and is model intrinsic. We show that if the generating prior satisfies global parameter independence, so does the compatible prior; in addition, prior modularity holds. Further results are obtained when the starting prior is product Dirichlet. A simple illustration of the methodology, and comparisons with alternative methods, are presented.

Keywords: GLOBAL INDEPENDENCE; INVARIANCE; LOCAL INDEPENDENCE; PRIOR MODULARITY; PRODUCT DIRICHLET; REFERENCE CONDITIONING.

1. INTRODUCTION
Bayesian Networks (BNs), based on Directed Acyclic Graphs (DAGs), have become increasingly popular to communicate, model and manage efficiently complex systems. They also represent a major tool in probabilistic expert systems. For an excellent account see Cowell et al. (1999), to which we refer for further background reading.

There are two ways to consider a BN: either as a belief model or as a “causal” model. In the former case, the focus is on the conditional independence properties embodied by the recursive factorisation of the joint distribution, which can be read off from the DAG itself. On the other hand, the causal interpretation is especially appropriate, whenever one can assume the absence of unmeasured confounders, together with an ordering of the variables involved, see e.g. Lauritzen and Richardson (2002). In such models the primitive building blocks are represented by the local conditional distributions.

The structure of the DAG underlying a BN is typically uncertain, so that one may wish to entertain several BNs and eventually perform model determination. If a Bayesian viewpoint is adopted, the issue of specifying a prior distribution on the parameter space of each network (e.g. with the aim of computing a Bayes factor) represents a severe problem, especially when the number of networks is very large. We therefore look for methods capable of automatically generating priors on parameter spaces across model structures. Although not logically necessary, it is also natural to require that these priors be related, in some way.
Specifically, assume that there are two models \( \mathcal{M} \) and \( \mathcal{M}_0 \) for the same observable \( \mathbf{X} \), having parameter space \( \Theta \) and \( \Theta_0 \), respectively, with \( \Theta_0 \subset \Theta \), so that \( \mathcal{M}_0 \) is nested within \( \mathcal{M} \). Write \( \pi(\cdot) \) and \( \pi_0(\cdot) \) for the prior densities on the corresponding parameter spaces. Given \( \pi(\cdot) \), the objective is to design a procedure capable of specifying a “compatible” prior \( \pi_0(\cdot) \). One motivation for this requirement is that the resulting Bayes factor should be least influenced by dissimilarities between the two priors due to differences in the elicitation processes, and should therefore represent more faithfully the strength of the support that the data lend to each model. The issue of prior compatibility was previously discussed, in particular, by Spiegelhalter et al. (1993), Heckerman et al. (1999, Section 11.6) is a very useful review.

Suppose that \( \Theta_0 = \{ \theta \in \Theta : t(\theta) = t_0 \} \), for some function \( t := t(\cdot) \) and fixed value \( t_0 \). Then a “natural” approach is to derive \( \pi_0(\cdot) \) by conditioning \( \pi(\cdot) \) on \( t = t_0 \). The drawback of this approach is that different conditioning functions may identify the same parameter space \( \Theta_0 \) while giving rise to different priors \( \pi_0(\cdot) \). Motivated by this difficulty, Dawid and Lauritzen (2000) suggested a novel approach, called Jeffreys conditioning, to construct \( \pi_0(\cdot) \). Subsequently, Roverato and Consonni (2001) (henceforth R&G) argued that Jeffreys conditioning may be inadequate for causal DAG-models and proposed an alternative method, named reference conditioning, which incorporates explicitly the causal structure of the DAG.

In this paper we apply reference conditioning to find compatible priors for the parameters of causal BNs.

2. BACKGROUND

2.1 Causal Bayesian Networks

A Directed Acyclic Graph (DAG) is a pair \( \mathcal{D} = (V, E) \) where \( V \) is a finite set of vertices and \( E \) is a finite set of directed edges, represented by “\( \rightarrow \)”. If \( u \rightarrow v \) then \( u \) is said to be a parent of \( v \) and the set of parents of \( v \) is denoted by \( \text{pa}(v) \). Random variables or vectors (rvs) \( X_v \) and \( \mathbf{X}_{\text{pa}(v)} \) are associated to \( v \) and \( \text{pa}(v) \), taking values respectively in \( \mathcal{X}_v \) and \( \mathcal{X}_{\text{pa}(v)} = \times_{u \in \text{pa}(v)} \mathcal{X}_u \).

Let \( \mathcal{M}^\mathcal{D} \) represent a DAG-model, i.e. a family of probability distributions for \( \mathbf{X} = \{X_v, v \in V\} \) parameterised by \( \theta = \{\theta_{v|\text{pa}(v)}, v \in V\} \in \Theta \), where \( \theta_{v|\text{pa}(v)} \) indexes the conditional distribution of \( X_v \) given \( \mathbf{X}_{\text{pa}(v)} \). The joint density of \( \mathbf{X} \), relative to a suitable product measure, admits the following factorisation according to \( \mathcal{D} \)

\[
p(x | \theta) = \prod_{v \in V} p_v(x_v | x_{\text{pa}(v)}, \theta_{v|\text{pa}(v)}).
\]  

(1)

We shall only consider DAG-models with a fixed vertex ordering wherein each \( X_v \) is a discrete rv, and refer to them as causal Bayesian Networks. If \( X_v \) is a \( d_v \)-state rv we let \( \mathcal{X}_v = \{0, \ldots, d_v - 1\} \), \( \mathcal{X}^0_{\text{v}} = \{1, \ldots, d_v - 1\} \) and define

\[
\theta_{v|\text{pa}(v)}^{y_v|y_{\text{pa}(v)}} = \Pr_{\theta_v} \{X_v = y_v | \mathbf{X}_{\text{pa}(v)} = y_{\text{pa}(v)}\}; \quad \theta_{v|\text{pa}(v)}^{y_{\text{pa}(v)}} = (\theta_{v|\text{pa}(v)}^{1|y_{\text{pa}(v)}}, \ldots, \theta_{v|\text{pa}(v)}^{d_v-1|y_{\text{pa}(v)}}),
\]

so that one can write

\[
\theta_{v|\text{pa}(v)} = \{\theta_{v|\text{pa}(v)}^{y_{\text{pa}(v)}}, y_{\text{pa}(v)} \in \mathcal{X}_{\text{pa}(v)}\}.
\]

(2)
We call the collection $\theta = \{\theta_v|\text{pa}(v), v \in V\}$ defined in (2) the conditional-probability-parameter, while the components $\theta_v|\text{pa}(v)$ are named local parameters. The latter are variation-independent under multinomial sampling.

Two structural properties that are often imposed on the prior $\pi(\cdot)$ for $\theta$ are: i) global parameter independence (GPI) and ii) local parameter independence (LPI) defined by

\[
\text{GPI}: \pi(\theta) = \prod_{v \in V} \pi_v(\theta_v|\text{pa}(v)) \quad \text{LPI}: \pi_v(\theta_v|\text{pa}(v)) = \prod_{y \in \text{pa}(v)} \pi_v(y|\theta_v|\text{pa}(v)).
\]

2.2 Compatible priors by means of conditioning

Consider the setup described in the Introduction. Let model $\mathcal{M}$ be parameterised by $\theta \in \Theta$, where $\Theta$ is an open subset of $\mathbb{R}^q$, and let $\mathcal{M}_0$ be a submodel of $\mathcal{M}$ identified by $\Theta_0 = \{\theta \in \Theta : t(\theta) = t_0\}$, where $t := t(\cdot)$ is a measurable function on $\Theta$ and $t_0$ a suitable (vector-valued) constant. Let $r < q$ be the dimension of $\Theta_0$. Assume that $\theta \sim \pi(\cdot)$, with $\pi(\cdot)$ absolutely continuous with respect to Lebesgue measure. Suppose now that $\Theta_0$ can be equivalently expressed as $\Theta_0 = \{\theta \in \Theta : s(\theta) = s_0\}$ using an alternative function $s := s(\cdot)$. As previously remarked, conditioning $\pi(\cdot)$ on $t = t_0$ may give a result different from conditioning it on $s = s_0$. Since the choice of the conditioning function is typically driven by the parameterisation adopted to describe model $\mathcal{M}$ we can conclude, somewhat imprecisely but expressively, that a conditioning procedure is not invariant to model re-parameterisations.

Under Jeffreys conditioning, see Dawid and Lauritzen (2000), the compatible prior distribution on $\Theta_0$ has density, w.r.t. Lebesgue measure (on $\mathbb{R}^r$), given by $\pi_0(\theta) \propto \pi(\theta) \frac{j_0(\theta)}{j(\theta)}$, $\theta \in \Theta_0$, where $j(\cdot)$, $j_0(\cdot)$ represent the densities of the Jeffreys measures under $\mathcal{M}$, respectively $\mathcal{M}_0$. Jeffreys conditioning is invariant to re-parameterisation and is model intrinsic, since the procedure only relies on the information matrix of the model.

Assume now that two causal DAG-models $\mathcal{M}_0^D$ and $\mathcal{M}_0^{D_0}$ are under investigation. They share the same vertex ordering and the latter is nested within the former. R&C argue that a causal DAG-model cannot be arbitrarily re-parameterised, and define the class of $D$-modular parameterisations. Loosely speaking, an element in this class can be unambiguously associated to the original local parameterisation and in this sense it preserves the modular structure of the DAG-model itself. This leads them to recommend a conditioning procedure invariant within this class. While Jeffreys conditioning is clearly invariant within the class of $D$-modular parameterisations, it is not the appropriate measure to be used, as exemplified by R&C in the Gaussian case, since it fails to satisfy prior modularity. If $\pi(\cdot)$ and $\pi_0(\cdot)$ satisfy GPI, prior modularity (defined in Heckerman et al. (1995)) states that $\pi_v(\cdot) = \pi_{0v}(\cdot)$, whenever $\text{pa}_0(v) = \text{pa}(v)$, where $\text{pa}_0(v)$ denotes the set of parents of $v$ under $\mathcal{D}_0$. In other words, parameters corresponding to vertices whose parents are the same under the two models should have the same prior. Arguably Jeffreys conditioning fails to satisfy prior modularity because the corresponding measure is “universally” invariant, while only invariance within the class of $D$-modular parameterisations is called for in causal DAG-models. One measure which satisfies this requirement, and is model intrinsic, is that based on the reference prior, see Bernardo (1979) and Berger and Bernardo (1992) for the general theory. Invariance of reference priors is dealt with in Yang (1995) and Datta and Ghosh (1996).
If \( r(\cdot), r_0(\cdot) \) denote the densities of the reference (prior) measure under \( \mathcal{M}^D \), respectively \( \mathcal{M}^{D_0} \), w.r.t. the local-parameter-grouping, then the compatible prior \( \pi_0(\cdot) \) obtained via reference conditioning is defined by

\[
\pi_0(\theta) \propto \pi(\theta) \frac{r_0(\theta)}{r(\theta)}, \quad \theta \in \Theta_0.
\]  

(3)

We now briefly summarise some construction aspects of reference priors, under simplified conditions, in order to let the reader appreciate our results. Consider a statistical model parameterised by \( \phi \in \Phi \). Berger and Bernardo (1992) motivate and describe a general algorithm to find reference priors for the parameter \( \phi \). Such an algorithm is greatly simplified if the posterior distribution of \( \phi \) is asymptotically normal (the so-called regular case). Let \( H(\phi) \) denote the Fisher information matrix in the parameterisation \( \phi \). We assume that \( \phi \) is grouped into \( s \) components \( (\phi_{(1)}, \ldots, \phi_{(s)}) \). Define \( \phi_{\sim k} \) to be \( \phi \) grouped as above without the component \( \phi_{(k)} \). The elements of \( \phi \) are usually ordered according to inferential importance; in particular, the parameters of interest should come first. For simplicity we assume that the components \( \phi_{(k)} \)'s are variation-independent. Under the regular case, if \( H(\phi) = \text{diag}\{H_{11}(\phi), \ldots, H_{ss}(\phi)\} \) and \( \text{det}\{H_{kk}(\phi)\} = a_k(\phi_{(k)})b_k(\phi_{\sim k}), \forall k \in \{1, \ldots, s\} \), for some positive functions \( a_k(\cdot) \) and \( b_k(\cdot) \), then the density—with respect to Lebesgue measure—of the \( s \)-group reference prior for \( \phi \) is given by

\[
r(\phi_{(1)}, \ldots, \phi_{(s)}) \propto \prod_{k=1}^{s} a_k(\phi_{(k)})^{1/2}.
\]  

(4)

We remark that the prior in (4) does not depend on the ordering of the \( s \) groups (in general the reference prior depends both on the grouping and the ordering). Strictly speaking the above result holds under a further assumption about the structure of the sequence of compact subsets expanding to \( \Phi \), which in our case is most naturally satisfied. We omit details and refer for a proof to Datta and Ghosh (1995); a further useful reference is Gutiérrez-Peña and Rueda (2001).

### 3. REFERENCE CONDITIONING FOR BAYESIAN NETWORKS

Let \( \mathcal{M}^D \) be a causal BN indexed by the conditional-probability-parameter \( \theta \) introduced above. Set \( D_0 = (V, E_0 \subset E) \), so that \( D_0 \) is obtained from \( D \) by deleting some of the edges while preserving the causal ordering for the vertices, and let \( \mathcal{M}^{D_0} \) denote the corresponding BN. In the following subsections we shall describe the steps required to compute the compatible prior \( \pi_0(\cdot) \) given a prior \( \pi(\cdot) \) for \( \theta \), using reference conditioning.

#### 3.1 Information matrix

We detail computations for \( \mathcal{M}^D \) only, since similar calculations hold for \( \mathcal{M}^{D_0} \).

Recall the factorisation in (1) and consider a single vertex \( v \). To each configuration \( y_{\text{pa}(v)} \) of \( X_{\text{pa}(v)} \) there corresponds a local parameter \( \theta_{v|\text{pa}(v)}^{y_{\text{pa}(v)}} \), so that the conditional distribution of \( X_v \) given \( X_{\text{pa}(v)} = y_{\text{pa}(v)} \) can be written as

\[
p_v(x_v \mid y_{\text{pa}(v)}, \theta_{v|\text{pa}(v)}^{y_{\text{pa}(v)}}) = (1 - \sum_{y_v \in \mathcal{X}_v^0} \theta_{v|\text{pa}(v)}^{y_v|y_{\text{pa}(v)}}) 1_{\{y_v\}}(x_v) \prod_{y_v \in \mathcal{X}_v^0} (\theta_{v|\text{pa}(v)}^{y_v|y_{\text{pa}(v)}}) 1_{\{y_v\}}(x_v)
\]
where \(1_A(\cdot)\) is the indicator function of the set \(A\).

Setting \(S^0_v(\theta v_{pa(v)}) = \sum_{y_v \in \mathcal{X}_v^0} \theta v_{pa(v)}^{y_v|y_{pa(v)}}\), the joint density becomes

\[
p(x | \theta) = \prod_{v \in V} \prod_{y_{pa(v)} \in \mathcal{X}_{pa(v)}} \left[ p_v(x_v | y_{pa(v)}, \theta v_{pa(v)}^{y_{pa(v)}}) \right] 1\{y_{pa(v)}\}(x_{pa(v)}) = \]

\[
= \prod_{v \in V} \prod_{y_{pa(v)} \in \mathcal{X}_{pa(v)}} \left[ (1 - S^0_v(\theta v_{pa(v)}^{y_{pa(v)}})) 1\{0\}(x_v) \prod_{y_v \in \mathcal{X}_v^0} (\theta v_{pa(v)}^{y_v|y_{pa(v)}}) 1\{y_v\}(x_v) \right] 1\{y_{pa(v)}\}(x_{pa(v)}). \tag{5}
\]

From the first line of (5) it appears that the local parameters \(\theta v_{pa(v)}^{y_{pa(v)}}\)'s are orthogonal, yielding the following result.

**Lemma 1** Let \(\mathcal{M}^D\) be a BN with conditional-probability-parameter \(\theta\). Then

i) The Fisher information matrix relative to \(\theta\) is

\[
H(\theta) = \text{diag}\{H v_{|pa(v)}^{y_{pa(v)}}, v \in V, y_{pa(v)} \in \mathcal{X}_{pa(v)}\}
\]

where \(H v_{|pa(v)}^{y_{pa(v)}} = \left( \text{Pr}_{\theta v}(X_v = y_{pa(v)}) \right) G v_{|pa(v)}^{y_{pa(v)}}\) is a square matrix of order \(\text{dim}(\theta v_{|pa(v)}^{y_{pa(v)}}) = d_v - 1\) and \(G v_{|pa(v)}^{y_{pa(v)}}\) is the Fisher information matrix for a Multinomial family with one trial and cell probabilities \(\theta v_{|pa(v)}^{y_{pa(v)}}\).

ii) The determinant of each matrix block is

\[
\det(H v_{|pa(v)}^{y_{pa(v)}}) = \left( \text{Pr}_{\theta v}(X_v = y_{pa(v)}) \right)^{d_v - 1} \left( 1 - S^0_v(\theta v_{|pa(v)}^{y_{pa(v)}}) \right) F v_{|pa(v)}^{y_{pa(v)}} \]

where \(F v_{|pa(v)}^{y_{pa(v)}} = \prod_{y_v \in \mathcal{X}_v^0} \theta v_{|pa(v)}^{y_v|y_{pa(v)}}\).

**Proof.** i) To each vector \(\theta v_{|pa(v)}^{y_{pa(v)}}\) there corresponds a single term in the double product in (5) and thus a square block of order \(d_v - 1\) in the information matrix

\[
H v_{|pa(v)}^{y_{pa(v)}} = -E_{\theta v}\left\{ 1\{y_{pa(v)}\}(X_v) \log \theta v_{|pa(v)}^{y_{pa(v)}} \frac{\partial^2}{\partial \theta v_{|pa(v)}^{y_{pa(v)}} \partial (\theta v_{|pa(v)}^{y_{pa(v)}})\}} \left( 1\{0\}(X_v) \log (1 - S^0_v(\theta v_{|pa(v)}^{y_{pa(v)}})) + \sum_{y_v \in \mathcal{X}_v^0} 1\{y_v\}(X_v) \log \theta v_{|pa(v)}^{y_v|y_{pa(v)}} \right) \right\} \]

\[
\times \left( 1\{0\}(X_v) \log (1 - S^0_v(\theta v_{|pa(v)}^{y_{pa(v)}})) + \sum_{y_v \in \mathcal{X}_v^0} 1\{y_v\}(X_v) \log \theta v_{|pa(v)}^{y_v|y_{pa(v)}} \right) \}
\]

The expression in square brackets can be regarded as the log-likelihood function relative to the random vector \((1_{d_v-1}(X_v), \ldots, 1_{d_v-1}(X_v)) \sim \text{Mu}(1, \theta v_{|pa(v)})\).

ii) The determinant of each block is easily recovered recalling that the determinant of the Fisher information matrix for a Multinomial family with \(n\) trials and probability vector \((\eta_1, \ldots, \eta_m)\), \(\sum_{j=1}^m \eta_j < 1\), is equal to \(n^m \left( 1 - \sum_{j=1}^m \eta_j \right) \prod_{j=1}^m \eta_j \left( 1 - \sum_{j=1}^m \eta_j \right)^{-1}\), see e.g. Bernardo and Smith (1994, p. 336).
We now compute the reference measure for $\theta$ w.r.t the local-parameter-grouping.

**Proposition 2** Let $\mathcal{M}^D$ be a BN with conditional-probability-parameter $\theta$. Then
the reference measure for $\theta$, relative to the grouping $\{\theta_{v|\text{pa}(v)}^y, v \in V, y_{\text{pa}(v)} \in \mathcal{X}_{\text{pa}(v)}\}$,
is order-invariant and has density

$$r(\theta) \propto \prod_{v \in V} \left[ r_v(\theta_{v|\text{pa}(v)}) \right]^{1/2}$$

(6)

where $r_v(\theta_{v|\text{pa}(v)}) \propto \prod_{y\text{pa}(v) \in \mathcal{X}_{\text{pa}(v)}} \left[ (1 - S^0_v(\theta_{v|\text{pa}(v)}) P^0_v(\theta_{v|\text{pa}(v)}) ) \right]^{-1}$.

**Proof.** Vectors representing the local-parameter-groups are variation-independent as well as orthogonal, implying order-invariance of the corresponding reference measure, provided each subset in the sequence expanding to $\Theta$ has the structure of a Cartesian product. Moreover $\text{Pr}_{\theta}\{X_{\text{pa}(v)} = y_{\text{pa}(v)}\}$ is independent of $\theta_{v|\text{pa}(v)}^y$ for each $v \in V$ and $y_{\text{pa}(v)} \in \mathcal{X}_{\text{pa}(v)}$. The construction in (4) can thus be exploited, w.r.t. such grouping, setting $a_{k}(\phi_{(i)}) = \left[ (1 - S^0_v(\theta_{v|\text{pa}(v)}) P^0_v(\theta_{v|\text{pa}(v)}) ) \right]^{-1}$, yielding the following factorisation

$$r(\theta) \propto \prod_{v \in V} \prod_{y\text{pa}(v) \in \mathcal{X}_{\text{pa}(v)}} \left[ (1 - S^0_v(\theta_{v|\text{pa}(v)}) P^0_v(\theta_{v|\text{pa}(v)}) ) \right]^{-1/2}$$

The density $r_0(\cdot)$ of the reference measure under $\mathcal{M}^{D_0}$ has an analogous expression. We omit details.

### 3.2 Compatible priors for Bayesian Networks

Consider $\mathcal{M}^D$ and $\mathcal{M}^{D_0}$ as defined at the beginning of Section 3, so that for each $v \in V$, $\text{pa}(v) \subseteq \text{pa}(v)$. Set $\text{pa}^*(v) = \text{pa}(v)|\text{pa}_0(v)$ and $V^* = \{v \in V : \text{pa}(v) \neq \text{pa}_0(v)\}$. The parameter space corresponding to $\mathcal{M}^{D_0}$ is identified through the constraints

$$\theta_{v|\text{pa}^*(v)|\text{pa}_0(v)}^y = \theta_{v|\text{pa}_0(v)}^y$$

say, for all $y_{\text{pa}(v)} \in \mathcal{X}_{\text{pa}(v)}$. Let $\theta_{v|\text{pa}_0(v)} = \left\{\theta_{v|\text{pa}_0(v)}^{y_{\text{pa}(v)}}, y_v \in \mathcal{X}_v, y_{\text{pa}(v)} \in \mathcal{X}_{\text{pa}(v)}\right\}$ and set $\theta_0 = \{\theta_{v|\text{pa}_0(v)}, v \in V \setminus V^*\} \cup \{\theta_{v|\text{pa}_0(v)}, v \in V^*\}$, so that $\theta_0$ represents the conditional probability parameter for model $\mathcal{M}^{D_0}$. The following theorem states the main result of the paper.

**Theorem 3** Let $\mathcal{M}^D$ be a BN with conditional-probability-parameter $\theta$ having prior density $\pi(\cdot)$. Let $\mathcal{M}^{D_0}$ be a submodel of $\mathcal{M}^D$ with conditional-probability-parameter $\theta_0$. Then

i) The compatible prior $\pi_0(\cdot)$ obtained via reference conditioning has density

$$\pi_0(\theta_0) \propto \pi(\theta_0) \prod_{v \in V^*} \frac{1}{r_0_v(\theta_{v|\text{pa}_0(v)})^{1/2} \prod_{u \in \text{pa}^*(v)} du^{-1}}$$

(7)

where $r_0_v(\theta_{v|\text{pa}_0(v)}) \propto \prod_{y_{\text{pa}_0(v)} \in \mathcal{X}_{\text{pa}_0(v)}} \left[ (1 - S^0_v(\theta_{v|\text{pa}_0(v)}) P^0_v(\theta_{v|\text{pa}_0(v)}) ) \right]^{-1}$.
ii) If $\pi(\cdot)$ satisfies GPI, so does $\pi_0(\cdot)$. Moreover, $\pi_0(\cdot)$ satisfies prior modularity, i.e. the distribution of the parameters $\theta^{y_{pa}(v)}_{v|pa(v)}, v \in V \setminus V^*$, is the same under $\pi(\cdot)$ and $\pi_0(\cdot)$.

Proof. i) Applying Proposition 2 w.r.t. $M_0$ and $\theta_0$ and recalling the factorization in equation (6) one obtains

$$r_{0v}(\theta_{v|pa(v)}) = r_v(\theta_{v|pa(v)}) \quad \text{if } v \in V \setminus V^*$$

$$r_{0v}(\theta_{v|pa(v)}) \propto \prod_{y_{pa}(v) \in X_{pa}(v)} \left(1 - S^0_v(\theta_{v|pa(v)})\right) P^0_v(\theta_{v|pa(v)})^{-1} \quad \text{if } v \in V^*.$$  

As a consequence, equation (3) becomes

$$\pi_0(\theta_0) \propto \pi(\theta_0) \left[ \prod_{v \in V^*} r_{0v}(\theta_{v|pa(v)}) \right]^{1/2} \left[ \prod_{v \in V \setminus V^*} r_v(\theta_{v|pa(v)}) \right]^{1/2} \prod_{y_{pa}(v) \in X_{pa}(v)} \left(1 - S^0_v(\theta_{v|pa(v)})\right) P^0_v(\theta_{v|pa(v)})^{-1/2} \prod_{y_{pa}(v) \in X_{pa}(v)} \left(1 - S^0_v(\theta_{v|pa(v)})\right) P^0_v(\theta_{v|pa(v)})^{-1/2}$$

$$\propto \pi(\theta_0) \prod_{v \in V^*} \prod_{y_{pa}(v) \in X_{pa}(v)} \left(1 - S^0_v(\theta_{v|pa(v)})\right) P^0_v(\theta_{v|pa(v)})^{1/2} \prod_{u \in pa^*(v)} du^{-1}.$$  

ii) Suppose GPI holds for $\pi(\cdot)$. Then one can write

$$\pi_0(\theta_0) \propto \left( \prod_{v \in V \setminus V^*} \pi_v(\theta_{v|pa(v)}) \right) \times \prod_{v \in V^*} \left\{ \pi_v(\theta_{v|pa(v)}) \prod_{y_{pa}(v) \in X_{pa}(v)} \left(1 - S^0_v(\theta_{v|pa(v)})\right) P^0_v(\theta_{v|pa(v)})^{1/2} \prod_{u \in pa^*(v)} du^{-1} \right\}$$

from which it appears immediately that GPI is satisfied by $\pi_0(\cdot)$. Furthermore prior modularity holds since priors for vertices $v \in V \setminus V^*$ are unaffected by the conditioning procedure.

We remark that formula (7) is computationally inexpensive. Furthermore notice that $\pi_0(\cdot)$ may be improper even if $\pi(\cdot)$ is proper.

Suppose now that a prior distribution is assigned to $\theta$ satisfying GPI and LPI with local densities

$$\pi_v(y_{pa}(v)_{v|pa(v)}) \propto (1 - S^0_v(\theta_{v|pa(v)})\theta_{v|pa(v)})^{-1} \prod_{y_{pa}(v) \in X_{pa}(v)}^{y_{pa}(v)} \frac{y_{v|pa(v)}}{\alpha_{v|pa(v)}}$$

i.e. a Dirichlet with strictly positive hyperparameters $\alpha_{v|pa(v)}^{y_{pa}(v)}$. Such a prior is called product Dirichlet. The total precision associated to vertex $v$ is denoted by $\alpha_v = \sum_{y_{pa}(v) \in X_{pa}(v)} \sum_{y_{pa}(v) \in X_{pa}(v)} \alpha_{v|pa(v)}^{y_{pa}(v)}$. The following result provides an expression for the compatible prior $\pi_0(\cdot)$.
Proposition 4 Assume that \( \pi(\cdot) \) is product Dirichlet. If the compatible prior \( \pi_0(\cdot) \) is obtained by means of reference conditioning then

i) \( \pi_0(\cdot) \) is product Dirichlet (so that GPI and LPI hold) and satisfies prior modularity.

ii) The marginal distribution of \( \theta_{y|\text{pa}_0(v)} \), \( v \in V^* \), under \( \pi_0(\cdot) \) has hyperparameters

\[
\alpha_{v|\text{pa}_0(v)} = \left( \sum_{y_{\text{pa}^*(v)} \in X_{\text{pa}^*(v)}} \alpha_{v|y_{\text{pa}^*(v)}} \right) - \frac{1}{2} \left( \prod_{u \in \text{pa}^*(v)} d_u - 1 \right).
\]

iii) The total precision associated to vertex \( v \in V^* \) under \( \pi_0(\cdot) \) is

\[
\alpha_{v|\text{pa}_0(v)}^{++} = \alpha_{v|\text{pa}(v)}^{++} - \frac{1}{2} \left( \prod_{u \in \text{pa}^*(v)} d_u - 1 \right) \left( \prod_{u \in \text{pa}_0(v)} d_u \right) d_v.
\]

Proof. i), ii) Compute equation (7) for the product Dirichlet prior. 

iii) Use (ii) and the definition of total precision.

We remark that the total precision associated to each vertex \( v \in V^* \) is lower under \( \pi_0(\cdot) \) than under \( \pi(\cdot) \). Moreover we stress the fact that a product Dirichlet structure for \( \pi_0(\cdot) \) is implied under reference conditioning, unlike in other approaches, see Section 4 below.

4. EXAMPLES

We illustrate our methodology using two simple BNs and compare our results in Proposition 4 with the method proposed by Spiegelhalter et al. (1993) and that suggested by Cowell (1996). Both papers assume that the prior under either model is product Dirichlet. Spiegelhalter et al. (1993) use the expansion and contraction (EC) algorithm to find a matching prior which preserves total precisions for all vertices. Cowell (1996) considers a fixed ordering, as we do, and finds the closest prior based on a criterion of minimum prior and expected posterior Kullback-Leibler divergences (KL). His method admits a few variants depending on the distribution used to compute expectations over all possible future observations, and thus does not lead to a unique answer. In the examples this distribution corresponds to that under the larger model.

Example 1 Let \( M^{D_i} \) be a bivariate BN where \( V = \{a, b\} \) and \( \text{pa}(b) = \{a\} \), see Figure 1. \( X_a \) and \( X_b \) are binary “0−1” rvs, so that \( \theta = (\theta^1_1, \theta^1_{1|a}, \theta^0_{1|a}) \), each component being the probability of value 1. Suppose \( \theta \) is assigned a product Beta distribution with \( \theta^1_a \sim \text{Be}(1, 8) \), \( \theta^1_{1|a} \sim \text{Be}(1, 9) \), \( \theta^0_{1|a} \sim \text{Be}(7, 21) \). If \( M^{D_0} \) is the independence model

\begin{figure}[h]
\centering
\begin{tikzpicture}
\node (a) at (0,0) {a};
\node (b) at (1,0) {b};
\node (c) at (1.5,0) {c};
\draw (a) -- (b);
\end{tikzpicture}
\caption{DAGs corresponding to \( M^{D_i} \), \( i = 1, 2 \) (\( M^{D_0} \) obtained by dropping all the arrows).}
\end{figure}
Table 1. $\mathcal{M}^{D1}$, $\mathcal{M}^{D0}$: hyperparameters under reference conditioning, EC and KL.

<table>
<thead>
<tr>
<th></th>
<th>$\pi_0$</th>
<th>$\pi_0^{EC}$</th>
<th>$\pi_0^{KL}$</th>
<th>$\pi_0$</th>
<th>$\pi_0^{EC}$</th>
<th>$\pi_0^{KL}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\alpha_a^1$</td>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
</tr>
<tr>
<td>$\alpha_a^0$</td>
<td>8.0</td>
<td>8.0</td>
<td>8.0</td>
<td>8.0</td>
<td>8.0</td>
<td>8.0</td>
</tr>
<tr>
<td>$\alpha_b^0$</td>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
</tr>
<tr>
<td>$\alpha_b^1$</td>
<td>7.0</td>
<td>7.0</td>
<td>7.0</td>
<td>7.0</td>
<td>7.0</td>
<td>7.0</td>
</tr>
<tr>
<td>$\alpha_b^2$</td>
<td>7.5</td>
<td>7.5</td>
<td>7.5</td>
<td>7.5</td>
<td>7.5</td>
<td>7.5</td>
</tr>
<tr>
<td>$\alpha_b^3$</td>
<td>8.9</td>
<td>8.9</td>
<td>8.9</td>
<td>8.9</td>
<td>8.9</td>
<td>8.9</td>
</tr>
<tr>
<td>$\alpha_c^0$</td>
<td>7.2</td>
<td>7.2</td>
<td>7.2</td>
<td>7.2</td>
<td>7.2</td>
<td>7.2</td>
</tr>
<tr>
<td>$\alpha_c^1$</td>
<td>9.0</td>
<td>9.0</td>
<td>9.0</td>
<td>9.0</td>
<td>9.0</td>
<td>9.0</td>
</tr>
<tr>
<td>$\alpha_c^2$</td>
<td>21.0</td>
<td>21.0</td>
<td>21.0</td>
<td>21.0</td>
<td>21.0</td>
<td>21.0</td>
</tr>
<tr>
<td>$\alpha_c^3$</td>
<td>29.5</td>
<td>29.5</td>
<td>29.5</td>
<td>29.5</td>
<td>29.5</td>
<td>29.5</td>
</tr>
<tr>
<td>$\alpha_c^4$</td>
<td>23.8</td>
<td>23.8</td>
<td>23.8</td>
<td>23.8</td>
<td>23.8</td>
<td>23.8</td>
</tr>
</tbody>
</table>

Table 2. $\mathcal{M}^{D2}$: hyperparameters and precisions under reference conditioning and EC.

<table>
<thead>
<tr>
<th></th>
<th>$\alpha_b^0$</th>
<th>$\alpha_b^1$</th>
<th>$\alpha_b^2$</th>
<th>$\alpha_b^3$</th>
<th>$\alpha_b^4$</th>
<th>$\alpha_b^5$</th>
<th>$\alpha_c^0$</th>
<th>$\alpha_c^1$</th>
<th>$\alpha_c^2$</th>
<th>$\alpha_c^3$</th>
<th>$\alpha_c^4$</th>
<th>$\alpha_c^5$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\pi_0^{EC}$</td>
<td>66.0</td>
<td>28.0</td>
<td>14.0</td>
<td>17.0</td>
<td>125.0</td>
<td>91.5</td>
<td>107.5</td>
<td>61.5</td>
<td>153.5</td>
<td>145.5</td>
<td>155.4</td>
<td>168.0</td>
</tr>
<tr>
<td>$\pi_0^{KL}$</td>
<td>75.0</td>
<td>17.2</td>
<td>11.8</td>
<td>25.0</td>
<td>129.0</td>
<td>151.9</td>
<td>56.6</td>
<td>55.1</td>
<td>155.4</td>
<td>168.0</td>
<td>168.0</td>
<td>155.4</td>
</tr>
</tbody>
</table>

wherein $X_a \perp X_b$ parameterised with $\theta_0 = (\theta_0^1, \theta_0^2)$, reference conditioning implies that $\pi_0(\cdot)$ is still product Beta with $\theta_0^1 \sim \text{Be}(1, 8)$, $\theta_0^2 \sim \text{Be}(7.5, 29.5)$. Table 1 reports the hyperparameters for six different prior specifications of $\pi(\cdot)$ under $\mathcal{M}^{D1}$ (including the one above, indicated by $\pi_1$), together with the corresponding hyperparameters for three types of compatible priors, namely those computed via reference conditioning ($\pi_0^{EC}$), expansion-contraction ($\pi_0^{KL}$) and Kullback-Leibler minimizations ($\pi_0^{KL}$). Clearly only the hyperparameters corresponding to vertex $b$ are shown, since those pertaining to vertex $a$ do not change by prior modularity, which is satisfied by each of the three compatible priors. On the whole it appears that reference conditioning and EC are in closer agreement between themselves than with KL.

**Example 2** Consider the complete DAG $D^2$ of Figure 1 and a model $\mathcal{M}^{D2}$ where the corresponding rvs $X_a$, $X_b$ and $X_c$ take values respectively in $\{0, 1, 2\}$, $\{0, 1, 2, 3\}$ and $\{0, 1, 2, 3, 4\}$. A product Dirichlet is specified for the conditional-probability-parameter of such a model with total precisions: $\alpha_a^1 = 18$, $\alpha_b^1 = 129$, $\alpha_c^1 = 587$. Specifically, defining $\alpha_{y_{pa}(v)} = (\alpha_{y_{pa}(v)}^0, \ldots, \alpha_{y_{pa}(v)}^{d_y-1})$, we fix $\alpha_a = (9, 2, 7)$, $\alpha_b = (12, 1, 1, 5)$, $\alpha_c = (4, 23, 3, 1)$. $\alpha_{y_{pa}(v)}^{0} = (51, 5, 11, 12)$ (to save space we do not report the hyperparameters for the prior on $\theta_{c_{lab}}$). Compatible priors for the independence model $\mathcal{M}^{D0}$ are computed using both reference conditioning and the expansion-contraction method, yielding the results in Table 2. While the total precisions $\alpha_b^1$ and $\alpha_c^1$ are broadly similar under the two approaches, there appear to be some marked differences
between specific pairs of hyperparameters.

5. DISCUSSION

Under the product Dirichlet assumption our method allows for varying degrees of total precisions of priors across vertices (for a given prior) and across model structures for each given vertex. This is in contrast to previous works such as Heckerman et al. (1995) and Spiegelhalter et al. (1993), and similar, in this respect, to Cowell (1996). A further similarity with the latter paper is the assumption of a causal ordering of the variables. However Cowell’s method is rooted in minimizing divergence, while our approach stems from invariance considerations. We also remark that reference conditioning works from the prior on the parameter space of a larger model to that of a nested model. This is not required in the alternative approaches mentioned above. On the other hand, while the latter impose a priori the structural properties of the compatible prior, such as GPI or product Dirichlet, such features become consequences of the structure of the generating prior within the reference conditioning framework.

REFERENCES


ACKNOWLEDGEMENTS

Research partially supported by MIUR, Italy, grant 2001138888.
# List of the lately published Technical Reports
(available at the web site: "http://economia.unipv.it/Eco-Pol/quaderni.htm").

## Quaderni di Dipartimento

<table>
<thead>
<tr>
<th>#</th>
<th>Date</th>
<th>Author(s)</th>
<th>Title</th>
</tr>
</thead>
<tbody>
<tr>
<td>136</td>
<td>11-01</td>
<td>F.Chelli L.Rosti</td>
<td>Gender Discrimination, Entrepreneurial Talent and Self-Employment in Italy ‘If you think you’re so discriminated against, why don’t you set up on your own?’</td>
</tr>
<tr>
<td>137</td>
<td>12-01</td>
<td>L. E. Nieto-Barajas</td>
<td>Normalized Random Measures driven by Increasing Additive Processes</td>
</tr>
<tr>
<td></td>
<td></td>
<td>I. Pruenster</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>S.G. Walker</td>
<td></td>
</tr>
<tr>
<td>138</td>
<td>02-02</td>
<td>A.Lijoi</td>
<td>Means of a Dirichlet process and multiple hypergeometric functions</td>
</tr>
<tr>
<td></td>
<td></td>
<td>E.Regazzini</td>
<td></td>
</tr>
<tr>
<td>139</td>
<td>02-02</td>
<td>M.Maggi</td>
<td>Immunization in an affine term structure framework</td>
</tr>
<tr>
<td>140</td>
<td>02-02</td>
<td>F.Menoncin</td>
<td>Optimal Portfolio With Benchmark for Fund Managers</td>
</tr>
<tr>
<td>141</td>
<td>04-02</td>
<td>M.Balconi</td>
<td>Il trasferimento di conoscenze tecnologiche dall’università all’industria in Italia: nuova evidenza sui brevetti di paternità dei docenti</td>
</tr>
<tr>
<td></td>
<td></td>
<td>S.Breschi</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>F.Lissoni</td>
<td></td>
</tr>
<tr>
<td>142</td>
<td>04-02</td>
<td>M.Balconi</td>
<td>Ivory Tower vs Spanning University: il caso dell’Università di Pavia</td>
</tr>
<tr>
<td></td>
<td></td>
<td>S.Borghini</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>A.Moisello</td>
<td></td>
</tr>
<tr>
<td>143</td>
<td>05-02</td>
<td>L.Bottolo</td>
<td>Bayesian analysis of extreme values by mixture modelling</td>
</tr>
<tr>
<td></td>
<td></td>
<td>G.Consonni</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>P.Dellaportas</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>A.Lijoi</td>
<td></td>
</tr>
<tr>
<td>144</td>
<td>06-02</td>
<td>I.Epifani</td>
<td>Exponential functionals and means of neutral to the right priors</td>
</tr>
<tr>
<td></td>
<td></td>
<td>A.Lijoi</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>I.Pruenster</td>
<td></td>
</tr>
<tr>
<td>145</td>
<td>06-02</td>
<td>P.Bertoletti</td>
<td>A note on third-degree price discrimination and output</td>
</tr>
<tr>
<td>146</td>
<td>12-02</td>
<td>P.Berti</td>
<td>Limit theorems for predictive sequences of random variables</td>
</tr>
<tr>
<td></td>
<td></td>
<td>L.Fratelli</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>P.Rigo</td>
<td></td>
</tr>
<tr>
<td>147</td>
<td>01-03</td>
<td>A. Lijoi</td>
<td>Practicable alternatives to the Dirichlet process</td>
</tr>
<tr>
<td></td>
<td></td>
<td>I. Pruenster</td>
<td></td>
</tr>
<tr>
<td>148</td>
<td>02-03</td>
<td>A. Lijoi</td>
<td>Extending Doob’s consistency theorem to nonparametric densities</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Pruenster</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>S.G.Walker</td>
<td></td>
</tr>
</tbody>
</table>