Eliciting expert beliefs on the structure of a Bayesian Network

Federico M. Stefanini
Department of Statistics 'G.Parenti'
University of Florence, viale Morgagni 59, 50134 Firenze, Italy

Abstract

The elicitation of prior beliefs about the structure of a Bayesian Network is a formal step of full-Bayesian structural learning which offers the opportunity of exploiting the knowledge accumulated by an expert of the problem domain over years of research in a quantitative way. Motivating applications include molecular biomarkers in gene expression or protein assays, where the use of prior information is often suggested as a promising approach to face the curse of dimensionality. In this paper a general formalization based on propositions describing network features is developed which comprises issues like anchoring and revision. An algorithm is described to estimate the number of structures bearing a-priori relevant features in problem domains characterized by a large number of nodes.

1 Introduction

The structure of a Bayesian Network (BN) and its parameters are in many cases unknown or affected by substantial uncertainty, therefore network learning is performed on the basis of collected data. A prior distribution over the space of structures is a formal ingredient of the Bayesian paradigm. Nevertheless, the elicitation of expert's prior information on network's structure suffers a major limitation due to the super-exponential increase of structures to be considered which becomes critical for five or more random variables. Despite the above mentioned difficulties, there is a wide agreement on the possibility of mitigating the 'curse of dimensionality' occurring in many applied fields by using prior information elicited from experts of the problem domain.

Several approaches have been proposed to define a prior distribution on the set of DAGs for a fixed set of variables. The early work of Buntine (Buntine, 1991) is based on a total ordering of nodes and a full specification of beliefs for each edge which could join pairs of nodes in a DAG. The collection of nodes which precedes a given node v is known given the order relation, therefore the probability of a given parent set Π_v of node v may be calculated as the product of probability for events of type 'there is edge $y \rightarrow v$' or 'there is not an edge $y \rightarrow v$', for each y preceding v. The subjective probability elicited from an expert about structure B_s is defined as the product of probability values for each parent set marginally considered. In the seminal paper of Heckerman (Heckerman et al., 1995) a prior network, $\mathcal{B}_{sc}$, is elicited and compared with candidate networks so that a high degree of belief is assigned to structures closely resembling to the prior network. The number δ_i of different nodes in the parent set of node v_i is calculated for each node to quantify the overall degree of dissimilarity $\delta = \sum_i \delta_i$. Given an elicited hyperparameter $0 < k < 1$, the prior distribution is proportional to k^{δ}. Castelo and Siebes first addressed the issue of partial prior knowledge and they also provided automatic rules to obtain a full prior for a Bayesian network (Castelo and Siebes, 2000). Recent contributions include the development of an informed score function based on the BDe metric (Mascherini and Stefanini, 2007). The use of several types of restrictions to code expert knowledge in structural learning of BNs has been investigated by (de Campos and Castellano, 2007), who also particularized the approach to the local search and to the PC learning algorithms.

This paper is motivated by the need of elicit-

ing beliefs in a more general setup, e.g. avoiding both the a-priori independence among parent sets and the specification of a prior network. A formal approach is developed with the aim of supporting researchers of applied fields in the elicitation and revision of causal and probabilistic beliefs. An algorithm is described which is useful in problem domains characterized by a large number of nodes. A simple case study is presented to illustrate the approach. The key idea of this paper is that in large spaces of structures, elicitation may deal with a limited number of network features.

2 Material and Methods

2.1 Bayesian networks

A graph $\mathcal{G}$ is a pair (V, E) where $V = \{v_1, v_2, \ldots, v_K\}$ is a finite set of nodes and $E \subset V \times V$ is the set of edges. The set E represents the structure of the graph because it defines which nodes are linked by an edge and if such edge is oriented (arrow) or not (undirected). If $(v_i, v_j) \in E$ but $(v_j, v_i) \notin E$ then the ordered pair corresponds to the oriented edge $v_i \rightarrow v_j$. The set of nodes originating oriented edges that enter into node v_j is called parents set, denoted as $pa(v_j)$. In a directed graph all edges are oriented. In a directed graph without cycles a tour following the direction of oriented edges never visits the same node two times. A directed graph without cycles is called Directed Acyclic Graph (DAG). An auxiliary random variable Z is introduced to map the set of DAGs for a fixed V to the set of natural numbers. It follows that a structure E is associated to an arbitrary number z in the set $\Omega_Z = \{1, 2, \ldots, n_z\}$, the sample space of Z.

The joint probability distribution of random variables indexed in V, the random vector $X_{v_1,\ldots,v_K}$, is Markov with respect to a DAG $\mathcal{G}$ if the following factorization holds:

$$p(x_{v_1}, x_{v_2}, \ldots, x_{v_K}) = \prod_{v_i \in V} p(x_{v_i} \mid x_{pa(v_i)})$$

where $x_{pa(v_i)}$ is random vector made by variables whose labels belong to the parents set of v_i. The lack of an arrow from v_i to v_j means irrelevance of X_{v_i} in predicting X_{v_j} if all random variables defined by the parent set have been observed, i.e. it is an instance of conditional independence. More general conditional independence statements may be derived by means of the D-separation theorem, or equivalently separation theorems on moralized graphs (Cowell et al., 1999). Under the stronger Markov Causal Assumption a DAG represents relations among variables which are stable under external manipulation (intervention) of a subset of them, so that causal effects may be in principle estimated.

Structural learning of a BN amounts to process a database $\mathcal{D} = \{d_1, d_2, \ldots, d_{n_d}\}$ of n_d conditionally independent realizations of the random vector $X_{v_1,v_2,\ldots,v_K}$ to infer the conditional independence relations existing in the joint distribution of the random vector. Following the Bayesian approach to inference, the joint probability distribution of $\mathcal{D}$ and network's unknowns given the context ξ is

$$p(\mathcal{D}, \theta, z \mid \xi) = p(\mathcal{D} \mid \theta, z, \xi) \cdot p(\theta \mid z, \xi) \cdot p(z \mid \xi),$$

where $\theta = (\theta_{v_1,pa(v_1)}, \ldots, \theta_{v_K,pa(v_K)})$ are vectors of parameters which appear in the conditional probability distributions of each pair X_{v_i}, $X_{pa(v_i)}$. The likelihood function $p(\mathcal{D} \mid \theta, z, \xi)$ is a product of multinomials and the degree of belief about elements of θ is often expressed as a product of Dirichlet probability density functions (Heckerman et al., 1995). The probability mass function $p(z \mid \xi)$ captures the expert's degree of belief about the unknown structure of a BN.

2.2 Expert's degree of belief about network features

In typical problem domains, we expect that an expert is willing to believe more on candidate structures showing some important features which are a-priori plausible.

Definition 1 (Features)**.** Network features $\{\mathcal{P}_1, \mathcal{P}_2, \ldots\}$ are propositions qualifying graphs defined on a fixed set of nodes V. Given a structure z, a proposition $\mathcal{P}_i(z)$ is either true or false.

Among the examples of features we have: $\mathcal{P}_1$ = 'is an ancestor of v_4', $\mathcal{P}_2$ = 'maximum

cardinality of parents $\leq 2\ \forall v \in V$', $\mathcal{P}_3 =$ 'maximum cardinality of children $\leq 2\ \forall v \in V$', $\mathcal{P}_4 =$ 'node v_3 is neighbor of v_7', $\mathcal{P}_5 =$ 'variable X_{v_2} is an immediate cause of X_{v_7}'. Features may accommodate both probabilistic and causal beliefs according to the choice of suitable propositions and context.

Expert's belief is typically elicited through several network features which may or may not hold at once. Nevertheless the straight specification of $p(z \mid \mathcal{P}_1, \mathcal{P}_2, \ldots)$ may be difficult for a general collection of statements due to relations which might exist among propositions. A collection of features may be organized into a basis for the elicitation by defining canonical features.

Definition 2 (Canonical feature)**.** Let $\mathcal{R} = \{\mathcal{P}_i : i = 1, \ldots, n_f\}$ be the set of reference features selected by an expert for the elicitation. A canonical feature $\mathcal{F}_j, j \in \mathcal{J}$, is a conjunction $\bigwedge_{i=1}^{n_f} \widetilde{\mathcal{P}}_i$ where $\widetilde{\mathcal{P}}_i$ is a proposition chosen between $\mathcal{P}_i$ and its negation $\neg\mathcal{P}_i$.

A convenient index set is $\mathcal{J} = \{(\overline{1}, \ldots, \overline{n_f}), \ldots, (1, \ldots, n_f)\}$ so that the configuration of features in $\mathcal{R}$ which generate a canonical feature $\mathcal{F}_j$, $j \in \mathcal{J}$, is self-evident. Note that a canonical feature is defined in a context ξ which includes a fixed collection of random variables.

Definition 3 (Elicitation basis)**.** A canonical reference set $\mathcal{F} = \{\mathcal{F}_j : j \in J\}$ built on reference set $\mathcal{R}$ is the collection of canonical features defined by all the possible conjunctions in Definition 2. It is a basis for the elicitation.

A canonical reference set induces a partition on the set of structures, as stated in the proposition below:

Proposition 1 (Canonical partition)**.** *Let $\mathcal{F}$ be an elicitation basis on a fixed $\mathcal{R}$. Let $\mathcal{C}_j = \{z : \mathcal{F}_j(z)\}$ be the set of structures satisfying the assertion $\mathcal{F}_j \in \mathcal{F}$. The set $\mathcal{C} = \{\mathcal{C}_j : j \in \mathcal{J}\}$ is a partition of the set of graphs built on a fixed collection of nodes V.*

Proof. By definition, canonical features are incompatible propositions. □

We are now in the position of eliciting a quantitative preference relation on such a partition.

Definition 4 (Preference on $\mathcal{F}$)**.** Let $\mathcal{F}$ be a canonical reference set. A preference relation $\mathcal{U}$ induces an order on $\mathcal{F}$ and the resulting 'precede' and 'succeed' relations are respectively indicated as $\prec$ and $\succ$. If $\mathcal{U}$ determines a partial ordering of features then $\mathcal{E} = \{\mathcal{E}_1, \ldots, \mathcal{E}_e, \ldots, \mathcal{E}_{n_e}\}$ is the induced partition of $\mathcal{F}$ into equivalence classes, with $\mathcal{E}_e$ a generic member of partition $\mathcal{E}$, n_e the total number of equivalence classes and $\mathcal{F}_{[e]}$ a generic member of $\mathcal{E}_e$.

The preference relation $\mathcal{U}$ is not necessarily a strict ordering because different canonical features may be equally plausible for the expert. A non-trivial elicitation basis $\mathcal{F}$, contains at least two distinct elements $\mathcal{F}_L$ and $\mathcal{F}_U$, with $\mathcal{F}_L \prec \mathcal{F}_U$, that respectively precedes and succeeds other canonical features. Therefore degrees of belief satisfy the inequality: $P[\mathcal{F}_L \mid \mathcal{U}, \xi] < P[\mathcal{F}_U \mid \mathcal{U}, \xi]$. A generic canonical feature $\mathcal{F}_j, j \in \mathcal{J}$, does not succeed to $\mathcal{F}_U$ and it does not precedes $\mathcal{F}_L$, that is $\mathcal{F}_L \preceq \mathcal{F}_j \preceq \mathcal{F}_U$, therefore the degree of belief satisfies:

$$P[\mathcal{F}_L \mid \mathcal{U}, \xi] \leq P[\mathcal{F}_j \mid \mathcal{U}, \xi] \leq P[\mathcal{F}_U \mid \mathcal{U}, \xi].$$

Note that if $\mathcal{F}_{j'} \preceq \mathcal{F}_{j''}$ and $\mathcal{F}_{j'} \succeq \mathcal{F}_{j''}$ both holds for two canonical features $\mathcal{F}_{j'}$ and $\mathcal{F}_{j''}$, then they belong to the same equivalence class, namely $\mathcal{F}_{j'} \sim \mathcal{F}_{j''}$ induced by $\mathcal{U}$.

The numerical assignment of degrees of belief is here performed using conditional odds.

Definition 5 (Conditional odds)**.** Let $\mathcal{F}_a, \mathcal{F}_b$ two canonical features and $\mathcal{U}$ a preference relation on $\mathcal{F}$. Conditional odds of $\mathcal{F}_a$ against $\mathcal{F}_b$ given $\mathcal{U}, \xi$ are:

$$\omega_{a,b} = \frac{P[\mathcal{F}_a \mid \mathcal{U}, \xi]}{P[\mathcal{F}_b \mid \mathcal{U}, \xi]} \tag{1}$$

with $\omega_{a,b} \geq 0$.

It follows from (1) that the numerical assignment for two features $\mathcal{F}_{j'} \sim \mathcal{F}_{j''}$ belonging to the same equivalence class is $\omega_{j',j''} = 1.0$. The direct numerical assignment of the degree of belief for pairs of features belonging to distinct equivalence classes $\mathcal{E}_a$ and $\mathcal{E}_b$ exploits an auxiliary experiment, here a hypothetical random

draw of one ball from an urn which contains α_r red balls and α_w white balls, with $\alpha_r + \alpha_w$ conveniently set to 100 or more. Given two features $\mathcal{F}_{[a]}$ and $\mathcal{F}_{[b]}$, the number of white and of red balls in the urn has to be changed by the expert up to the point in which the odds associated to the proposition 'the ball drawn from the urn is white' are equal to conditional odds of $\mathcal{F}_{[a]}$ against $\mathcal{F}_{[b]}$: $\omega_{a,b} = \frac{\alpha_w}{\alpha_r}$.

Proposition 2 (Complete minimal ensemble)**.** *Let $\mathcal{U}$ be an order relation on $\mathcal{F}$ and $\mathcal{E}$ the induced partition into equivalence classes. An ensemble is a collection of conditional odds $\{\omega_{a,b}\}$ elicited from the expert. The ensemble is complete and minimal if it contains $n_e - 1$ odds values between pairs of features belonging to distinct equivalence classes, so that at least one feature is taken from each equivalence class in $\mathcal{E}$.*

Proof. The ensemble is complete because a probability distribution on $\mathcal{F}$ is obtained by transformation of elicited odds, that is $\sum_{j\in\mathcal{J}} P[\mathcal{F}_j \mid \mathcal{U}, \xi] = 1$ and $P[\mathcal{F}_j \mid \mathcal{U}, \xi] \geq 0$. The ensemble is minimal because its size can not be further reduced without compromising the full specification of a probability distribution on $\mathcal{F}$. □

The assignment of conditional odds has to be performed according to the order induced by $\mathcal{U}$ in units of subjective probability.

Two structures z_1 and z_2 may belong to the same equivalence class $\mathcal{C}_j$ and in this case they are on equal footing for what concerns expert's prior information. The probability $P[Z = z \mid \mathcal{F}_j, \xi]$ represents the expert degree of belief about the proposition: 'the unknown structure z is one of those structures characterized by $\mathcal{F}_j$'.

Proposition 3 (Beliefs on Z)**.** *Given the canonical partition $\mathcal{C}$ induced by an elicitation basis $\mathcal{F}$, the probability mass function $p(z \mid \xi)$ is given by:*

$$p(z \mid \mathcal{U}, \xi) = \frac{1}{n_{j(z)}} \cdot P[\mathcal{F}_{j(z)} \mid \mathcal{U}, \xi] \qquad (2)$$

where $j(z)$ is the element of the canonical partition in which z is located, and with $n_{j(z)}$ the cardinality of such subset.

Proof. The starting factorization is:

$$p(z \mid \mathcal{U}, \xi) = \sum_{j\in\mathcal{J}} P[Z = z \mid \mathcal{F}_j, \xi] \cdot P[\mathcal{F}_j \mid \mathcal{U}, \xi]$$

but $P[Z = z \mid \mathcal{F}_j, \xi]$ is null for all but one conditioning feature, say $\mathcal{F}_{j(z)}$. Moreover, under indifference among members within class $\mathcal{C}_{j(z)}$ the probability $P[Z = z \mid \mathcal{F}_{j(z)}, \xi]$ is one over $n_{j(z)}$, the cardinality of such equivalence class. □

2.3 The elicitation of $p(z \mid \mathcal{U}, \xi)$

An elicitation basis is a general object, nevertheless it is convenient to describe some practical details both to support algorithms formulation and to prepare the expert to variations which also depend on the amount of information being elicited. The conjunction of two or more incompatible features, like $\mathcal{P}_{i'}$ = 'has $v_i \to v_j$' and $\mathcal{P}_{i''}$ = 'has $v_j \to v_i$', determines a canonical feature which is indeed false for DAGs. Therefore the probability of $\mathcal{P}_{i'} \wedge \mathcal{P}_{i''}$ is null. Similar remarks hold if a feature implies another feature, say $\mathcal{P}_2 \Rightarrow \mathcal{P}_1$. In such case the degree of belief in the conjunction $\neg\mathcal{P}_1 \wedge \mathcal{P}_2$ should be zero.

Substantial prior information in the problem domain may result in a narrow partition, $n_{j(z)} = 1, \forall j \in \mathcal{J}$, and the burden of assessment is equivalent to the one-by-one elicitation of beliefs on structures. Nevertheless, in large spaces of structures it is likely that the elicitation brings to coarse partitions in which $n_j \gg 1$, and the number of DAGs belonging to an equivalence class may be hard to assess.

An approximated solution to the counting problem may be obtained by simulation. The core of our algorithm is defined in (Ide et al., 2002) who build a Markov Chain (MC) that at convergence provides a DAG uniformly sampled from the space of all DAGs on a fixed set of nodes V. We extended the algorithm described in (Ide et al., 2002, Algorithm 1) by adding steps `00, 09, 10`, so that the auxiliary experiment made by M runs of such a MC results in a sample of M DAGs:

```
INPUT: number of nodes n, number of
iterations N, number of DAGs M.
```

```
OUTPUT: a vector of counts.
00.Repeat M times:
01.Initialize a simple tree in which each
   node has just one parent, except the
   root node without parents;
02.Repeat the next loop N times:
03    Generate uniformly  a pair of
      distinct nodes i,j;
04    If arc(i,j) exists in the graph,
         delete the arc providing the
         graph remains connected;
05    else
06    Add the arc, provided that the
      graph  remains a DAG;
07    Otherwise keep the same state;
08.Return the current graph after N
   iterations;
09.Assign the returned DAG to an
   element of the partition;
10.Return the vector of counts after
   M iterations;
```

Proposition 4 (MC estimate of cardinalities)**.** *Given a Markov Chain algorithm providing DAGs uniformly sampled from the spaces of DAGs on a fixed set V, an estimate of cardinality $n_j, j \in \mathcal{J}$, of elements in the canonical partition $\mathcal{C}$ is:*

$$\widehat{n}_j = \frac{f(n_K)}{M} \cdot \sum_{i=1}^{M} \mathbf{I}_{\mathcal{C}_j}(z_i) \tag{3}$$

with $f(n_K)$ the total number of DAGs on a fixed set V and M the number of simulated chains. The indicating function $\mathbf{I}_{\mathcal{C}_j}(z_i)$ is equal to 1 if the structure z_i belongs to $\mathcal{C}_j$, zero otherwise.

Proof. The above algorithm generates a sample of M DAGs. Each DAG is assigned to the equivalence class in $\mathcal{C}$ which corresponds to the canonical feature $\mathcal{F}_{j(z)}$. The total number of DAGs on a fixed set of nodes V is obtained by recursion (Robinson, 1977):

$$f(n_K) = \sum_{i=1}^{n_K} (-1)^{i+1} \cdot \binom{n_K}{i} \cdot 2^{i \cdot (n_K - i)} \cdot f(n_K - i) \tag{4}$$

where $f(1) = 1, f(0) = 1$ and $n_K \geq 2$. □

2.4 Coherence, stability and revision

The elicitation of expert beliefs is not made by one straight operation. It is closer to a self-untangling adaptive procedure which increase in clarity during its dynamic. In this perspective the need of revision and elaboration of elicited values is pretty understandable and generally accepted in practice. The psychological nature of the elicitation process may lead to poorly elicited quantities, as it has been discussed in the literature (Garthwaite et al., 2005, and references therein). For this reason it is convenient to elicit more quantities than needed, that is a redundant collection of conditional odds is elicited.

Definition 6. (Coherent anchoring) Let $\widetilde{\omega}_{\mathcal{R}}$ be the collection of distinct complete minimal ensembles based on $\mathcal{R}$, that is ensembles in which at least one value among conditional odds is built on features taken from different equivalence classes. Then degree of beliefs are coherently anchored if all complete minimal ensembles provide the same distribution of subjective probability values on $\mathcal{F}$.

Elaboration of elicited quantities is performed to improve the correspondence between expert's belief and numerical assignments. Coherent anchoring leads to the definition of a probability measure on the algebra of features $\mathcal{A}(\mathcal{F})$. Subjective probability values for marginal canonical features may be compared to actual expert beliefs about the same joint statements for the unknown structure.

Definition 7 (Reduced reference set)**.** Let $\mathcal{R}$ be a set of reference features. A reduced reference set $\mathcal{R}_r$ is a proper subset of $\mathcal{R}$.

Proposition 5 (Stability)**.** *Let $\widetilde{\mathcal{R}}$ be the collection of all reduced reference sets obtained from a given reference set: $\widetilde{\mathcal{R}} = \{\mathcal{R}_r : \mathcal{R}_r \subset \mathcal{R}.\}$. Let $\widetilde{\omega}_{\mathcal{R}_r}$ be a collection of distinct complete minimal ensembles, as in Definition 6, based on $\mathcal{R}_r$ and the preference relation $\mathcal{U}_r$. Then elicited degree of beliefs are stable under reduction $\mathcal{R}_r$ if:*

$$P\left[\mathcal{F}_j \mid \mathcal{U}_r, \mathcal{R}_r, \xi\right] = P\left[\bigvee_{s \in S} \mathcal{F}_s \mid \mathcal{U}, \mathcal{R}, \xi\right], \tag{5}$$

where S is the collection of index values denoting canonical features based on $\mathcal{R}$ which appear in the disjunction producing the canonical feature $\mathcal{F}_j$ based on $\mathcal{R}_r$. The elicited degree of beliefs are stable if they are stable for all reductions in $\widetilde{\mathcal{R}}$.

Proof. If one or more features are removed from a reference set then the canonical basis of elicitation partially collapses to one of larger granularity. The associated algebra is given by unions of elements taken from the starting algebra. □

Full stability may be heavy to check and a useful compromise is to limit the number of reductions, for example to the collection of all one-feature reference sets. The revision of elicited beliefs is mandatory if stability or coherent anchoring are violated for some reductions.

2.5 A case study in breast cancer

Classical biomarkers in breast cancer studies include progesterone receptors (PR), oestrogen receptors (ER), age (AG), menopausal status (MS), number of positive lymph nodes (PL). Variables of interest for patients are tumor grade (TG) and tumor size (TS). The reference set $\mathcal{R}$ contains the propositions below:
$\mathcal{P}_1$ ='Nodes AG, MS precede all other nodes';
$\mathcal{P}_2$ ='TS, TG, NL follow all other nodes';
$\mathcal{P}_3$ ='The parent set is made by three or less nodes for each node in V';
$\mathcal{P}_4$ ='ER is independent on AG given MS'.

An important particularization of the general elicitation scheme is obtained by a preference relation $\mathcal{U}$ which sets the order over canonical features according to the the number of true propositions making each canonical feature. The canonical feature $\neg\mathcal{P}_1 \wedge \neg\mathcal{P}_2 \wedge \neg\mathcal{P}_3 \wedge \neg\mathcal{P}_4$ precedes all the other canonical features, while $\mathcal{P}_1 \wedge \mathcal{P}_2 \wedge \mathcal{P}_3 \wedge \mathcal{P}_4$ succeeds to all the other canonical features. It follows that the first and last equivalence classes are: $\mathcal{E}_0 = \{\neg\mathcal{P}_1 \wedge \neg\mathcal{P}_2 \wedge \neg\mathcal{P}_3 \wedge \neg\mathcal{P}_4\}$ and $\mathcal{E}_4 = \{\mathcal{P}_1 \wedge \mathcal{P}_2 \wedge \mathcal{P}_3 \wedge \mathcal{P}_4\}$. Four canonical features have just one proposition true and they define the equivalence class $\mathcal{E}_1 = \{\neg\mathcal{P}_1 \wedge \mathcal{P}_2 \wedge \mathcal{P}_3 \wedge \mathcal{P}_4, \mathcal{P}_1 \wedge \neg\mathcal{P}_2 \wedge \mathcal{P}_3 \wedge \mathcal{P}_4, \mathcal{P}_1 \wedge \mathcal{P}_2 \wedge \neg\mathcal{P}_3 \wedge \mathcal{P}_4, \mathcal{P}_1 \wedge \mathcal{P}_2 \wedge \mathcal{P}_3 \wedge \neg\mathcal{P}_4\}$ which follows $\mathcal{E}_0$. Equivalence classes $\mathcal{E}_2$ and $\mathcal{E}_3$ are defined in a similar way. The cardinality of an equivalence class $\mathcal{E}_i$ in this case study is equal to $\mid \mathcal{E}_i \mid = \binom{n_f}{i}$, where n_f is the total number of propositions and i the number of true propositions for each canonical feature in the equivalence class $\mathcal{E}_i$: $\mid \mathcal{E}_0 \mid = 1$, $\mid \mathcal{E}_1 \mid = 4$, $\mid \mathcal{E}_2 \mid = 6$, $\mid \mathcal{E}_3 \mid = 4$, $\mid \mathcal{E}_4 \mid = 1$. In this particular preference relation the total number of equivalence classes within the partition $\mathcal{E}$ is $n_e = n_f + 1 = 5$.

In the elicitation, three distinct complete minimal ensembles are considered, say $\widetilde{\omega}_{\mathcal{R}} = \{\widetilde{\omega}_{\mathcal{R},1}, \widetilde{\omega}_{\mathcal{R},2}, \widetilde{\omega}_{\mathcal{R},3}\}$. The first ensemble is: $\widetilde{\omega}_{\mathcal{R},1} = \{\omega_{1,0} = 2.0, \omega_{2,0} = 3.0, \omega_{3,0} = 4.0, \omega_{4,0} = 5.0\}$, with indices $i = 0, 1, 2, 3, 4$ denoting any canonical feature belonging to equivalence class $\mathcal{E}_i$. The second and third ensembles are: $\widetilde{\omega}_{\mathcal{R},2} = \{\omega_{1,0} = 1, \omega_{2,1} = \frac{3}{2}, \omega_{3,2} = \frac{4}{3}, \omega_{4,3} = \frac{5}{4}\}$, $\widetilde{\omega}_{\mathcal{R},3} = \{\omega_{0,4} = \frac{1}{5}, \omega_{1,3} = \frac{1}{2}, \omega_{2,4} = \frac{3}{5}, \omega_{3,2} = \frac{4}{3}\}$. The anchoring is coherent because the three ensembles provide the same probability values: $P[\mathcal{F}_{[0]} \mid \mathcal{U}, \xi] = \frac{1}{48}$, $P[\mathcal{F}_{[1]} \mid \mathcal{U}, \xi] = \frac{1}{24}$, $P[\mathcal{F}_{[2]} \mid \mathcal{U}, \xi] = \frac{1}{16}$, $P[\mathcal{F}_{[3]} \mid \mathcal{U}, \xi] = \frac{1}{12}$, $P[\mathcal{F}_{[4]} \mid \mathcal{U}, \xi] = \frac{5}{48}$.

The algorithm described in Section (2.3) run with parameters $M = 10000$ and $N = 294$. In (Ide et al., 2002) the authors motivated the choice of $N = K^2 * 6$ through empirically findings. We replicated the above simulation for one hundred times to assess the variability of point estimates of the fractions of DAGs in $\mathcal{C}_j, j \in \mathcal{J}$ (full results not shown): minimum and maximum standard deviations of $\hat{n}_j$ observed for the $2^4 = 16$ elements of the canonical partition $\mathcal{C}$ are, respectively, 0.000475 and 0.004797 in 100 replicated simulations. The total number of DAGs for seven nodes is $f(7) = 1'138'779'265$ (equation 4).

The stability of elicited values has been examined limited to four reduced reference sets: $\widetilde{\mathcal{R}} = \{\mathcal{R}_i : i = 1, 2, 3, 4\}$, so that the reduced reference set $\mathcal{R}_i = \{P_i\}$ contains just one proposition. An explicit expression exploiting the already introduced index set $\mathcal{J}$ is easily obtained for such reductions, for example: $P[\mathcal{F}_{[1]} \mid \mathcal{U}_1, \mathcal{R}_1, \xi] = \sum_{s \in S} P[\mathcal{F}_s \mid \mathcal{R}, \mathcal{U}, \widetilde{\omega}_R, \xi]$, where $S = \{(1, \bar{2}, \bar{3}, \bar{4}), \ldots, (1, 2, 3, 4)\}$. In this way we obtained four marginal probability val-

ues of each proposition $\mathcal{P}_i, i = 1, 2, 3, 4$ and they are all equal to $\frac{7}{12}$. The expert did not reject the above values which follow from the unrestricted elicitation, so the revision did not take place (see the discussion).

3 Discussion

The generality of the approach described in this work is mainly due to the use of propositions describing network features. Nevertheless the usefulness also depends on the amount of work left to the expert in actual problem domains. The good scaling of the proposed elicitation with an increasing number of nodes rests on a number of propositions which is far smaller than the number of DAGs. Moreover structures counting is left to simulation, and reasonable estimates in equation (3) are obtained by sampling a number of DAGs M which is much smaller than the total number of DAGS $f(n_K)$. Even for an increasing number of propositions and very large spaces of structures the computation may remain feasible if the order relation $\mathcal{U}$ induces large equivalence classes of canonical features. Nevertheless, the overall computational burden partially depends on the nature of features. For example, features local to Markov blankets are quickly checked, while features involving the consideration of the whole structure are computationally heavy to assess.

Particularized instances of our approach may serve as starting elicitation from which semi-automatic prior distributions may be quickly obtained. The preference relation $\mathcal{U}$ discussed in the case study induces a partition in which $\mathcal{E}_i$ collects all canonical features made by i true propositions. A computationally efficient assignment of probability values to canonical features is obtained by eliciting a value $0 < k < 1$ and by setting $P[\mathcal{F}_{[i]} \mid \mathcal{U}, \xi] \propto k^{(n_f - i)}$, so that for two features $\mathcal{F}_{[i']}$ and $\mathcal{F}_{[i'']}$ satisfying $i' - i'' = \delta$ we have $\omega_{i',i''} = k^{-\delta}$. Reconciliation of incoherent anchoring may be automatically performed by defining an equally weighted mixture of probability distributions obtained from different ensembles. This reconciliation scheme might be an automatic first step towards a detailed revision and it could be useful in problem domains with a large number of features and weak prior information.

Simpler elicitation schemes may work in practice, and in these setups our formalization may be useful in obtaining the elicitation bias, for example, due to the assumption of a-priori independence among propositions. Sharp prior information has been coded as structural restrictions in (de Campos and Castellano, 2007), but it may be as well coded using canonical features and degree of beliefs very close to 0 or to 1, so that management of restrictions and potential overstatement of beliefs are avoided. A comparison of computational burden and of flexibility between the two approaches in similar case studies is a theme for future research. The elicitation based on network features and conditional odds is more demanding than the approach in (Castelo and Siebes, 2000), where edges are units of elicitation that are combined up to a full prior for a BN in a quite implicit way. The use of oriented graphs and the distribution of the extra-DAGs amount of weight lead to a prior distribution for BNs, but in large sized problem domains the resulting prior distribution may be difficult to submit to further inspection as regards non-local properties. Levels of information including full structures, correlation or causation among nodes, temporal order (O'Donnell et al., 2006) may be also captured through features, while the implications due to the use of a uniform distribution on the space of Totally Ordered Models (O'Donnell et al., 2006, TOM) has still to be investigated. Finally, at the best of our knowledge the approach described in this work seems the only one to reach full generality in considering global network features, therefore numerical explorations of its performances under common learners, like greedy search, deserve research efforts.

A key step of our scheme is the auxiliary Monte Carlo experiment which provides the cardinality of equivalence classes when straight counting DAGs is too heavy. The algorithm discussed in (Ide et al., 2004) may be used to count DAGs in restricted spaces, for example for a sharply believed feature like "no more then

three nodes in each parent set" with a probability equal to 1. The counting problem has been also considered by (Peña, 2007), who provides MCMC algorithms to approximately calculate the ratio of DAG models to DAGs up to 20 nodes and the fraction of chain graph models that are neither DAG models nor DAG models up to 13 nodes. The extension of the approach to elicitation described in this work for more general classes of graphs deserves attention in future work.

4 Conclusions

In this paper we proposed a formal procedure to elicit expert beliefs on the structure of a Bayesian network by means of propositions which capture relevant features. While a detailed elicitation may be overwhelming for the expert in large problem domains, particularizations of the general approach offer automatic completion and limited expert efforts.

Further work on elicitation is needed both on theoretical and applied sides. An inferential engine could suggest sure-false and sure-true canonical features given a reference set. Moreover a graphical user interface could make elicitation and revision easier to perform for applied scientists. It is has been found that different propositions embedding the same meaning may lead to different elicited values. Finally, human cognitive peculiarities related to the elicitation of beliefs on plausible structures for a BN are still largely unexplored.

Acknowledgments

This work is funded by Italian MIUR (PRIN). The author thanks anonymous referees for their comments.

References

Wray Buntine. 1991. Theory of Refinement on Bayesian Networks, In *Proceedings of 7th Conference on Uncertainty in Artificial Intelligence*, pages 52–60.

Roberto Castelo and Arno Siebes. 2000. Priors on network structures. Biasing the search for Bayesian networks, *International Journal of Approximate Reasoning*, 24:39–57.

Robert G. Cowell, A. Philip Dawid, Steffen L. Lauritzen and David J. Spiegelhalter. 1999. *Probabilistic Networks and expert systems*, Springer Verlag.

Luis M. de Campos and Javier G. Castellano. 2007. Bayesian network learning algorithms using structural restrictions, *International Journal of Approximate Reasoning*, 45:233–254.

Paul H. Garthwaite, Joseph B. Kadane and Antony O'Hagan. 2005. Statistical methods for eliciting probability distributions, *Journal of the American Statistical Association*, 100:680–701.

David Heckerman, Dan Geiger, David M. Chickering. 1995. Learning Bayesian Networks: The Combination of Knowledge and Statistical Data, *Machine Learning*, 20:197–243.

Jaime S. Ide and Fabio Gagliardi Cozman. 2002. Random generation of Bayesian Networks. In *Proceedings of the Brazilian Symposium on Artificial Intelligence*, Brazil, pages 366–375.

Jaime S. Ide, Fabio Gagliardi Cozman and Fabio T. Ramos. 2004. Generating Random Bayesian Networks with Constraints on Induced Width. In *Proceedings of the 16th European Conference on Artificial Intelligence (ECAI-04)*, IOS Press, Amsterdam, pages 323-327.

Massimiliano Mascherini and Federico M. Stefanini. 2007. Using weak prior information on structures to learn Bayesian Networks. In *B. Apolloni et al. (Eds.): KES 2007/WIRN 2007,Part I, LNAI 4692*, Springer Verlag, Berlin, pages 413-420.

Rodney T. O'Donnell, Ann E. Nicholson, Bin Han, Kevin B. Korb, Jahangir M. Alam and Lucas R. Hope. 2006. Causal discovery with prior information. In *A: Sattar and B. H. Kang (Eds): AI 2006, LNAI 4304*, Springer Verlag, Berlin, pages 1162-1167.

Jose M. Peña. 2007. Approximate counting of Graphical Models via MCMC, In *Proceedings of 11th International Conference on Artificial Intelligence and Statistics (AISTATS 2007)*, pages 352–359.

Robert W. Robinson. 1997. Counting unlabeled acyclic digraphs. In *Combinatorial Mathematics V* (C.H.C. Little, ed.) Springer Lectures Notes in Mathematics 622, pages 28–43.