Hindawi Publishing Corporation
The Scientific World Journal
Volume 2014, Article ID 749150, 12 pages
http://dx.doi.org/10.1155/2014/749150

Research Article

Chain Graph Models to Elicit the Structure of a Bayesian Network

Federico M. Stefanini

Dipartimento di Statistica, Informatica, Applicazioni "G. Parenti", Università degli Studi di Firenze, Viale Morgagni 59, 50134 Firenze, Italy

Correspondence should be addressed to Federico M. Stefanini; stefanini@disia.unifi.it

Received 31 August 2013; Accepted 5 November 2013; Published 5 February 2014

Academic Editors: R. M. Rodríguez-Dagnino and M. Saberi

Bayesian networks are possibly the most successful graphical models to build decision support systems. Building the structure of large networks is still a challenging task, but Bayesian methods are particularly suited to exploit experts' degree of belief in a quantitative way while learning the network structure from data. In this paper details are provided about how to build a prior distribution on the space of network structures by eliciting a chain graph model on structural reference features. Several structural features expected to be often useful during the elicitation are described. The statistical background needed to effectively use this approach is summarized, and some potential pitfalls are illustrated. Finally, a few seminal contributions from the literature are reformulated in terms of structural features.

1. Introduction

Bayesian networks (BNs) are possibly the most successful graphical models to represent probabilistic and causal relationships [1, 2]. BNs are used in very different fields including medical domains [3], engineering [4], ecology [5], bioinformatics [6], and many others.

The core of this class of models is made by a directed acyclic graph (DAG) $\mathcal{G}$, where nodes in the graph are labels of modeled variables (elements of vector $\mathbf{X}$), and oriented edges (arrows) capture probabilistic and/or causal relationships. The joint distribution of $\mathbf{X}$ is represented by the product of conditional distribution functions following from the structure of $\mathcal{G}$. If substantial prior information is available on a given problem domain, it is possible that an expert defines the structure of $\mathcal{G}$ and even the parameters inside the conditional distribution functions at a reasonable extent. Otherwise, structure and parameters have to be estimated from data (structural and parameter learning), for example, from a collection of exchangeable observations $\mathcal{D} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)$.

It is often the case that an expert knows some features of DAG $\mathcal{G}$, but knowledge is not enough to define a DAG because several of its aspects are affected by relevant uncertainty. Following the Bayesian paradigm [7, 8], the expert is invited to state his/her degree of belief about $\mathcal{G}$ by eliciting a prior distribution on the set of DAGs which can be considered given a fixed set of variables (nodes). All other unknown quantities, like missing values and parameters, are considered in the prior distribution [9].

Learning the structure of a BN remains nowadays still challenging for the combinatorial explosion of candidate structures with the increase in the number of considered nodes. Following Robinson [10], the number of possible DAGs on 6 nodes is 3781503; thus the enumeration of all structures while eliciting expert beliefs is unfeasible. Computational difficulties in the full enumeration of DAGs follow from about a dozen of nodes on. For these reasons, several restrictions and simplifications in stating prior beliefs were considered in the past, with the aim of making structural learning tasks affordable in large networks. Widely adopted elicitation techniques are based on restrictions like a total ordering of nodes, the presence of sharp order constraints on nodes, the marginal independence of unknowns, or the existence of a prior network which is a good summary of expert beliefs.

In a recent work [11], graphical models were proposed to elicit beliefs about the structure of a BN. The approach is characterized by the possibility of expressing beliefs about

limited (but relevant) aspects of the problem domain, called structural features [12], that by their own nature are often only indirectly related to oriented edges in the unknown DAG.

Here the most general approach based on chain graph (CG) models is reconsidered with the aim of establishing connections with seminal contributions from the literature on structural learning, of detailing a general parameterization, and of describing one way to perform the refinement of elicited beliefs. Common useful structural features are defined and some issues related to their implementation and revision are examined.

2. Methods

Notation and some background information on graphs and Markov properties are provided below. A comprehensive account may be found in [13–15]. An approach to the elicitation of structural features by CG models is described, together with methods for the revision.

2.1. Graphs. 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 directed (arrow) or not. If $(v_i, v_j) \in E$ and $(v_j, v_i) \in E$ then an undirected edge joins v_i and v_j, indicated as v_i—v_j, and v_j is neighbor of v_i. If $(v_i, v_j) \in E$ but $(v_j, v_i) \notin E$ the ordered pair corresponds to the directed edge $v_i \rightarrow v_j$; v_j is said to be the child of v_i and v_i is a parent of v_j. The set $\mathrm{pa}(v_j)$ includes all parents of node v_i, that is, all nodes originating arrows with end in v_i, while the set $\mathrm{ch}(v_i)$ collects all children of v_i, namely, all nodes in which arrows originated from v_i end.

A path is a sequence of vertices such that there is an edge for each pair of subsequent nodes in the sequence, that is, v_i—v_{i+1} or $v_i \leftarrow v_{i+1}$ or $v_i \rightarrow v_{i+1}$. A directed path is a path in which all edges maintain the head-to-tail orientation, for example, (v_i, v_j, v_k) with $v_i \rightarrow v_j \rightarrow v_k$.

In a directed graph all edges are directed. The ancestors $\mathrm{an}(v_i)$ of node v_i are nodes on a directed path reaching v_i, while descent nodes $\mathrm{de}(v_i)$ are nodes on a directed path starting from node v_i. Note that $v_i \in \mathrm{de}(v_i)$ and that $v_i \in \mathrm{an}(v_i)$. The extension of the above definition to $\mathrm{an}(A)$ with $A \subset V$ is obtained by union of sets $\mathrm{an}(v_i)$ for each $v_i \in A$. A similar extension holds for $\mathrm{de}(A)$.

In a directed graph without cycles it is not possible to visit the same node more than one time by following a directed path, and in this case the graph is called directed acyclic graph. A moralized DAG is an undirected graph obtained by joining pairs of nodes sharing children (if not yet connected) with an undirected edge and by removing the direction of all edges. A subgraph $\mathcal{G}_A$ on $A \subset V$ is obtained by removing all nodes in $V \setminus A$ and all edges in which at least one node is in $V \setminus A$ from the graph $\mathcal{G}$.

A graph without directed edges is called undirected graph (UG). An UG with $E = V \times V$ is said to be complete. A subgraph on a subset $S \subset V$ of nodes is obtained by removing nodes not in S and all edges reaching-leaving nodes not in S. Note that a subset of nodes is indicated by capital letters. A clique C is a maximal complete subgraph of an UG.

A chain graph, also called partially directed acyclic graph (PDAG), is made by an ordered collection of chain components $\tau = (\tau_1, \tau_2, \ldots)$ which are undirected graphs and by directed edges between nodes located in different chain components, so that the arrow $v_i \rightarrow v_j$ is allowed only if v_i belongs to a chain component preceding the chain component in which v_j is located. Therefore directed edges are forbidden within a chain component and in the direction from τ_j to τ_{j-k}, with $k > 0$.

The moralization of a chain graph mimics the moralization of a DAG. A moralized CG, indicated as $\mathcal{G}^m$, is obtained by the following steps:

(1) let $k = 2$;

(2) join with undirected edges all pairs of nodes in $\mathrm{pa}(\tau_k)$, with $\mathrm{pa}(\tau_k)$ being the union of parents sets for each node in chain component τ_k;

(3) iterate step (2) for $k = 3, 4, \ldots$;

(4) remove directions to all edges.

2.2. Some Markov Properties. Conditional independence [16] is fundamental to reason out highly structured stochastic systems and to simplify the representation of high dimensional distributions.

In this paper the random vector $\mathbf{X}$ refers to random variables included in the BN. The notation used hereafter is based on nodes in V; thus $\mathbf{X}$ is also indicated as $X_V = X_{v_1,\ldots,v_K} = (X_{v_1}, \ldots, X_{v_K})$ and the sample space $\Omega_{X_V} = \otimes_{v_i \in V} \Omega_{X_{v_i}}$ is the Cartesian product of sample spaces of considered variables.

The joint probability distribution of a random vector X_V is Markov with respect to an UG $\mathcal{G}$ if

$$p\left(x_{v_1}, x_{v_2}, \ldots, x_{v_K}\right) = \sigma^{-1} \prod_{C \in \mathscr{C}} \phi_C\left(x_C\right) \tag{1}$$

with $\mathscr{C}$ being a collection of graph cliques in $\mathcal{G}$ and with ϕ_C nonnegative functions called clique potentials; note that x_C is the coordinate projection of vector x_V on the subset of coordinates defined by C; $\sigma = \sum_{\Omega_{X_V}} \prod_{C \in \mathscr{C}} \phi_C(x_C)$ is the partition function that normalizes the product of potentials in (1).

Markov properties for positive distributions with respect to an undirected graph may be read using the separation theorem [14]. Let $A \subset V$, $B \subset V$, $S \subset V$ be disjoint subsets of nodes. The separation theorem states that subvectors X_A and X_B are conditionally independent given the subvector X_S if and only if all paths from a node in A to a node in B include nodes located in S; thus nodes in S separate nodes in A from nodes in B.

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

$$p\left(x_{v_1}, x_{v_2}, \ldots, x_{v_K}\right) = \prod_{v_i \in V} p\left(x_{v_i} \mid x_{\mathrm{pa}(v_i)}\right), \tag{2}$$

where $x_{\mathrm{pa}(v_i)}$ is the random vector made by variables whose labels belong to the parents set of v_i.

Markov properties on a DAG may be read using the separation theorem for directed graphs [14]. Given three disjoint sets of nodes A, B, S, consider the subgraph $\mathcal{G}$ defined on $\mathrm{an}(A \cup B \cup S)$ after moralization, say $\mathcal{G}^m_{\mathrm{an}(A,B,S)}$. Random subvectors X_A and X_B are conditionally independent given the subvector X_S if and only if nodes in S separate nodes in A from nodes in B in $\mathcal{G}^m_{\mathrm{an}(A,B,S)}$.

A joint probability distribution is Markov with respect to a CG $\mathcal{G}$ if

$$p\left(x_{v_1}, x_{v_2}, \ldots, x_{v_K}\right) = \prod_{\tau_i \in \tau} p\left(x_{\tau_i} \mid x_{\mathrm{pa}(\tau_i)}\right) \tag{3}$$

with $\tau_i \in \tau$, the chain components of $\mathcal{G}$. Furthermore factors on r.h.s. of (3) may be factorized by considering the subgraph on nodes defined by $\tau_i \cup \mathrm{pa}(\tau_i)$:

$$p\left(x_{\tau_i} \mid x_{\mathrm{pa}(\tau_i)}\right) = \sigma_t^{-1} \prod_{C \in \mathcal{C}} \phi_C\left(x_C\right), \tag{4}$$

where σ_t^{-1}, $t \in \Omega_{\mathrm{pa}(\tau_i)}$ are normalization constants, one for each conditioning value of the random subvector $X_{\mathrm{pa}(\tau_i)}$. The working UG in (4) is obtained by removing the orientation of edges from parents of nodes in τ_i and by joining them into a complete undirected subgraph.

Conditional independence relationships in CGs are also obtained from an extension of the separation theorem in UGs and DAGs for positive distributions [17]. Let $\mathcal{G}$ be a chain graph and $A \subset V$, $B \subset V$, $S \subset V$ be three disjoint subsets of V. The separation theorem states that subvectors X_A and X_B are conditionally independent given the subvector X_S if and only if all paths from a node in A to a node in B in $\mathcal{G}^m_{\mathrm{an}(A \cup B \cup S)}$ include nodes of S; thus nodes in S separate nodes in A from nodes in B. Note that $\mathcal{G}^m_{\mathrm{an}(A \cup B \cup S)}$ is the moral graph of the smallest ancestral set for $A \cup B \cup S$, that is, a subgraph of $\mathcal{G}^m$ described in Section 2.1.

2.3. Causal DAGs. A DAG may represent causal relations among variables. According to the causal semantic, an arrow $v_i \rightarrow v_j$ indicates that v_i is a direct cause of v_j with respect to nodes included in V, that is, at the considered model granularity. In principle the intervention on variable x_{v_i} may bear an effect on x_{v_j}. The intervention on a subset of variables $D \subset V$ indicates the external setting of variables in X_D to prescribed values; thus the system or process is perturbed, not merely observed.

Pearl [2, pp. 27–32] starts with the definition of functional causal models, which are deterministic in nature, and he demonstrates [2, theorem 1.4.1] that such formulation induces the Markov factorization in (2), the so-called Markov causal assumption. An equivalent representation embeds exogenous variables into the node of interest and transforms the deterministic relationships into probabilistic conditioning, thus leading to Bayesian networks.

A key property of a casual DAG $\mathcal{G}$ is the stability under external intervention: if a variable x_{v_i} is manipulated all the other variables maintain their relationships as represented by $\mathcal{G}$. In other terms the intervention is local on manipulated variables and it does not break all the other relationships represented in a causal DAG. The intervention regime is in contrast with the plain observation of values taken by the random vector X_V.

The granularity of a causal DAG depends on the variables included in the model. A variable X_L not included in a causal DAG may eventually affect just one variable X_v, with $v \in V$; otherwise if several variables are affected then a more general class of models is needed, called semi-Markovian networks (not considered further in this work).

For an updated presentation of the approach see [18] while [19] warns against the blind definition of DAGs within the causal semantic in observational studies. He also reconsiders the foundations of causal inference by anchoring them to the extended conditional independence (C.I.), both at algebraic level and with a graphical counterpart based on influence diagrams.

2.4. Inference about the Structure of Bayesian Networks. In all cases in which strong prior information is absent, structural learning is performed by means of a database $\mathcal{D} = (x_1, \ldots, x_n)$ of n exchangeable observations of the random vector X_V.

Several algorithms have been proposed to infer causal and probabilistic relations, but in Bayesian inference key quantities enter into the joint probability distribution of $\mathcal{D}$ and network's unknowns given the context ξ:

$$p\left(\mathcal{D}, \theta, z \mid \xi\right) = p\left(\mathcal{D} \mid \theta, z, \xi\right) \cdot p\left(\theta \mid z, \xi\right) \cdot p\left(z \mid \xi\right), \tag{5}$$

where $\theta = (\theta_{v_1,\mathrm{pa}(v_1)}, \ldots, \theta_{v_K,\mathrm{pa}(v_K)})$ are vectors of parameters characterizing the conditional probability distributions of X_{v_i} given $X_{\mathrm{pa}(v_i)}$; variable Z indicates the unknown DAG, and it is built as a bijection from the set of DAGs (fixed V) to a subset Ω_Z of natural numbers.

The likelihood function $p(\mathcal{D} \mid \theta, z, \xi)$ is typically expressed as a product of multinomials by using sufficient statistics. Whenever expert beliefs are reasonably captured by Dirichlet prior distributions for elements of θ, and they are elicited as marginally independent, closed-form integration marginalizes out thetas. The resulting marginal distribution $p(\mathcal{D} \mid z, \xi)$ has a reduced dimensionality and may be optimized with respect to z while looking for optimal structures characterized by the highest posterior probability values [9]:

$$\begin{aligned} p\left(\mathcal{D}, z \mid \xi\right) &= \int p\left(\mathcal{D} \mid \theta, z, \xi\right) \cdot p\left(\theta \mid z, \xi\right) \cdot p\left(z \mid \xi\right) \cdot d\theta \\ &= p\left(z \mid \xi\right) \cdot \prod_{i=1}^{K} \prod_{j=1}^{q_i} \frac{\Gamma\left(\alpha_{i,j}\right)}{\Gamma\left(\alpha_{i,j} + N_{i,j}\right)} \\ &\quad \cdot \prod_{s=1}^{r_i} \frac{\Gamma\left(\alpha_{i,j,s} + N_{i,j,s}\right)}{\Gamma\left(\alpha_{i,j,s}\right)} \end{aligned} \tag{6}$$

with r_i being the number of states taken by X_i and q_i of $X_{\mathrm{pa}(v_i)}$; sufficient statistics are $N_{i,j,s}$ for the ith variable taking

the sth state while its parent configuration is in the jth state; $\alpha_{i,j} = \sum_{s=1}^{r_i} \alpha_{i,j,s}$ and $N_{i,j} = \sum_{s=1}^{r_i} N_{i,j,s}$. A clever choice of hyperparameters guarantees the likelihood equivalence; that is, all BNs equivalent as regards the set of encoded conditional independence relationships are equally scored by (6).

Equation (6) shows that the elicitation of the prior distribution $p(z \mid \xi)$ is a key step to perform Bayesian structural learning using $\mathscr{D}$.

2.5. Plausible Network Features. A candidate structure $z \in \Omega_Z$ on a fixed set of nodes V is plausible if it has got structural features (SFs) believed to be relevant by the expert. Following Stefanini [11, Definition 1], we recall the formal definition.

Definition 1 (SF). A structural feature (SF) $\mathscr{R}_j(z, w)$ in a reference set $\mathscr{R}$ for the set of DAGs on V is a predicate describing a plausible probabilistic or causal characteristic of the unknown directed acyclic graph $z \in \Omega_Z$. Argument w is in the partition $\mathscr{W}$ of a given numeric domain Ω_W of variable W. An atomic structural feature (ASF) $\mathscr{R}_j(z)$ does not depend on any auxiliary variable W.

A reference set $\mathscr{R}$ is a collection $\mathscr{R} = \{\mathscr{R}_j : j \in J\}$ of SFs indexed in a set J, with n_f being the number of considered SFs. A proposition might be defined to carry disbelief to a candidate structure, but the feature-rises-belief direction is conveniently adopted here.

An example makes the previous definition operational. Let us define $\mathscr{R}_1(z, w)$ = "The number of immediate causes of X_{v_3} is in w" to capture the expert belief about the number of parents of node v_3 in the unknown DAG. An expert may consider $\Omega_W = \{0, 1, \ldots, 12\}$ and $\mathscr{W} = (\{0\}, \{1, 2, 3\}, \{4, 5\}, \{6, \ldots, 12\})$. Given a candidate structure z, its plausibility will be determined by the element $w \in \mathscr{W}$ that makes the predicate true. A simple representation of configurations taken by reference features is obtained by descriptors [11, Definition 2].

Definition 2 (descriptors). A descriptor R_i for the SF $\mathscr{R}_i$ is a map:

$$R_i : \mathscr{W}_i \times \{\text{false}, \text{true}\} \longrightarrow \{0, 1, 2, \ldots, |\mathscr{W}_i|\} \tag{7}$$

so that $(w, \text{false}) \mapsto 0$ for all $w \in \mathscr{W}_i$ and $(w, \text{true}) \mapsto h_w$ for all $w \in \mathscr{W}_i$; that is, a different integer is associated with each w if true. The vector $R = (R_1, R_2, \ldots, R_{n_f})$ defined on the Cartesian product $\Omega_R = \otimes_{i=1}^{n_f} \Omega_{R_i} = \otimes_{i=1}^{n_f} \{0, 1, \ldots, |\mathscr{W}_i|\}$ is called vector of descriptors. The descriptor of an ASF is defined by false $\mapsto 0$ and true $\mapsto 1$.

The jth configuration of descriptors in R is indicated as $r_j = (r_{1,j}, r_{2,j}, \ldots, r_{n_f,j}) \in \Omega_R$ while a generic configuration is indicated as $r \in \Omega_R$. The jth configuration of a subvector of R defined by indexes in A is $r_{A,j}$; for example, $r_{\{1,3\},j} = (r_{1,j}, r_{3,j})$.

Vector R induces equivalence classes on Ω_Z; that is, $\mathscr{Z} = \{\mathscr{Z}_r : r \in \Omega_R\}$ contains sets of structures, with $\mathscr{Z}_r$ made by all those DAGs sharing the same configuration r.

Note that members of the same equivalence class must be associated with the same degree of belief by the principle of insufficient reason: they differ in irrelevant and unconsidered ways by construction.

Despite the generality of Definition 1 some features are expected to occur more often than others. Below, some of them are described without pretending to be exhaustive.

2.5.1. Indegree and Outdegree. In applications characterized by a small sample size, it is useful to impose a sharp constraint during the greedy search of top-scored candidate structures. The maximum number of arrows entering into a node, the indegree, is set to a small integer, say 2 to 5, to exclude structures with a large number of parents from the consideration. A large number of parents implies a CPT with a huge number of parameters which are affected by large uncertainty after conditioning to observed data because of sampling zeros. A similar constraint may be set on the number of arrows leaving a node.

Two reference features naturally embed this kind of information.

(i) "The maximum number of arrows reaching a node is in n_{id} for all nodes in V," where n_{id} is a small set of integers close to 1 elicited from the expert.

(ii) "The maximum number of arrows leaving a node is in n_{od} for all nodes in V," where n_{od} is a small set of integers close to 1 elicited from the expert.

The above two features may be exploited to increase the plausibility of candidate structures which are sparsely connected. Higher control on connectivity is obtained by considering the fraction of nodes showing a given degree of connectivity, as described below.

2.5.2. Partitioned Connectivity. A different way of characterizing the connectivity is obtained by eliciting the minimum fraction of nodes in DAG showing a given number of parents (children) and by iterating the elicitation from 0 parents (children) up to a small integer s.

Let s be a small integer representing the maximum number of parents (children) to be considered. Let $W = (W_0, \ldots, W_s)$ be a vector of nondecreasing numbers in $[0, 1]$, with

$$\Omega_W = \{(\omega_0, \omega_1, \omega_2, \ldots, \omega_b, \ldots, \omega_s) : 0 \leq \omega_b \leq \omega_{b+1} < \omega_s = 1\} \tag{8}$$

being the sample space. The elicitation of this feature is based on a vector $a = (a_0, a_1, a_2, \ldots, a_b, \ldots, a_s)$, with $0 \leq a_b \leq a_{b+1} < a_s = 1$, and on the induced partition $\mathscr{W} = \{w_0, w_1\}$ where

$$w_1 = \{(\omega_0, \omega_1, \ldots, \omega_s) : \omega_b \geq a_b, b = 0, 1, \ldots, s\} \tag{9}$$

and $w_0 = \Omega_W \setminus w_1$.

Two reference features naturally embed the extended evaluation of connectivity:

(i) "The a-partitioned inconnectivity of degree s is in w_1," with a, s, and w_1 defined above; a candidate structure z shows this feature if the cumulative relative

Table 1: Elicited vector a, with $s = 5$. In the top row, the number b of parents is reported, while in the second row the correspondent minimum fraction of nodes a_b is shown.

Number of parents	0	1	2	3	4	5
Cumulative fraction of nodes	0.01	0.2	0.4	0.6	0.8	1.0

frequency F_b of nodes in z with number of parents equal or less than b is greater than a_b; that is, $F_b \geq a_b$, $b = 1, 2, \ldots, s$.

(ii) "The a-partitioned outconnectivity of degree s is in w_1," with a, s, and w_1 defined above; a candidate structure z shows this feature if the cumulative relative frequency F_b of nodes in z with number of children equal or less than b is greater than a_b; that is, $F_b \geq a_b$, $b = 1, 2, \ldots, s$.

Trained experts could prefer a conventional total number of nodes equal to 100 to elicit cumulative percentages instead of cumulative fractions of nodes.

In Table 1, an example is shown where $s = 5$ and the elicited vector a is defined on fractions. A candidate structure has the partitioned inconnectivity feature if the fraction of root nodes is equal or above 0.01, while the proportion of nodes with at least 1 parent is equal or above 0.2 and so on.

2.5.3. Direct Cause and Direct Effect. The reference feature "The variable x_{v_i} is an immediate cause of variable x_{v_j}" refers to x_{v_i} as a parent of x_{v_j} so that by setting (intervening on) the value of the variable x_{v_i} to a given value, the distribution of X_{v_j} is modified. This relation holds at the level of selected granularity; thus it may change if the collection of variables (nodes in V) is modified.

2.5.4. Causal Ancestors. The "direct cause" feature may be extended by considering a variable x_{v_i} which is on a causal (directed) path reaching node x_{v_j}. In this case the reference feature is "The variable x_{v_i} is an indirect cause of variable x_{v_j}"; thus the expert believes that one or more variables mediate the effect of x_{v_i} on x_{v_j}.

2.5.5. Causal Hubs. A hub node in a network is characterized by a high number of arrows leaving it. A reference feature which captures the local connectivity of node v_i is "Node v_i is a hub node of at least outdegree w," with the outdegree indicating the number of arrows originated in v_i and w a set of integers. Note that the defined feature is a localized version of the outdegree feature.

The expert might believe that a hub node should be present, but without indicating a specific node. In this case the a-partitioned outconnectivity feature (with a large s) can be exploited for this purpose.

2.5.6. Conditional Independence Relationships. A statement about C.I. among three disjoint subsets of random variables may take the following form: "The random vector X_A is conditionally independent from the random vector X_B given vector X_S," with A, B, S being disjoint subsets of nodes in V.

2.6. The Degree of Belief. The prior distribution on the space of structures on V is obtained by "extending the argument;" that is,

$$P\left[Z = z \mid \xi\right] = \sum_{r \in \Omega_R} P\left[Z = z \mid R = r, \xi\right] \cdot P\left[R = r \mid \xi\right]. \tag{10}$$

By recognizing that R induces the partition $\mathcal{Z}$, it follows that

$$p\left(z \mid \xi\right) = \frac{1}{n_{r^{[z]}}} \cdot P\left[R = r^{[z]} \mid \xi\right] = \frac{p\left(r_1^{[z]}, r_2^{[z]}, \ldots, r_{n_f}^{[z]} \mid \xi\right)}{n_{r^{[z]}}}, \tag{11}$$

where $n_{r^{[z]}}$ is the cardinality of the equivalence class in which z is located and where the structural configuration of DAG z is $r^{[z]}$. In [12] the size of each equivalence class $\mathcal{Z}_r$ was estimated by Monte Carlo simulation to face the combinatorial explosion in the number of DAGs to be enumerated with the increase of the number of nodes:

$$\left|\widehat{\mathcal{Z}}_r\right| = \frac{N_V\left(N_r + 1\right)}{N_T + 1}, \tag{12}$$

where N_T is the total number of DAGs uniformly sampled from the space of all DAGs on V, N_V is the size of such space (see [10]), and $N_r \leq N_T$ is number of sampled DAGs showing configuration r.

The numerator on the right of (11) represents the joint belief on each configuration of descriptors. While the elicitation in full generality becomes cognitively and numerically overwhelming around 7 descriptors on, some parsimony is achieved if a small number of descriptors may be considered at one time, so that conditional independence relationships among descriptors may be exploited. This is a choice available to the expert through the definition of an order relation on descriptors.

Definition 3 (ordered partition). The ordered partition

$$\mathcal{O} = \left(\mathcal{O}_1, \mathcal{O}_2, \ldots\right) \tag{13}$$

of descriptors is defined by the expert to indicate disjoints subsets of SFs to be jointly considered during the elicitation, from the first subset $\mathcal{O}_1$ to the last.

Otherwise stated, the expert decomposes the whole elicitation problem following an order taken from the substantive content of the specific problem domain: features are grouped according to the priority in the elicitation.

The elicitation of the ordered partition is performed by formulating questions in the language typical of a given problem domain. It is indeed difficult to define those questions in general terms, because they result to be quite abstract and they are likely to be obscure for the domain expert (see [11]). We assume here that questions are properly phrased and that an ordered partition is defined.

If a strict order relation is elicited, each element of $\mathcal{O}$ contains one descriptor: this is a special case addressed in [20]. Another special case is represented by a trivial partition of just one subset that contains all descriptors [21]. In the two special cases above it is possible to define, respectively, a Bayesian network and a Markov network on descriptors. The general case made by several subsets, each one of cardinality two or more, was addressed by using CG models in [11] for ASFs. Below details are provided about a reference set of nonatomic features.

The joint probability distribution of descriptors $p(r_1, r_2, \ldots, r_{n_f} \mid \mathcal{O}, \xi)$ is assumed to be Markov with respect to the elicited partition where descriptors within group $\mathcal{O}_j$ define the chain component τ_j of the CG model (see (3)). The elicitation keeps on by asking in the language of the domain expert which descriptors in subset $\mathcal{O}_j$ should be jointly considered while defining the degree of belief. Nodes corresponding to related descriptors are joined by undirected edges and the collection of cliques defining the chain component are found. The elicitation is iterated for all elements in the ordered partition $\mathcal{O}$.

The resulting CG may support the elicitation if model parameters are cognitively suited for the quantitative step, that is, if they are interpretable and easy to assess for the expert, at least after some training.

The first chain component is elicited as a marginal distribution which is not conditioned on other descriptors. An undirected graphical model on descriptors in $\mathcal{O}_1$ is defined through the multiplicative model in (4), under empty conditioning. Nevertheless some care is needed because potential functions on cliques are not uniquely defined. Here a log-linear parameterization is suggested, following [13]. For example, if the first chain component has just one clique made by three descriptors, say R_1, R_2, R_3, we define the multiplicative model as below, after exponentiation and rearrangement of log-linear terms:

$$\begin{aligned}\frac{p(r_1, r_2, r_3 \mid \xi)}{p(0,0,0 \mid \xi)} &= \phi_1(r_1)\,\phi_2(r_2)\,\phi_3(r_3)\,\phi_{1,2}(r_1 r_2) \\ &\quad \times \phi_{1,3}(r_1 r_3)\,\phi_{2,3}(r_2 r_3)\,\phi_{1,2,3}(r_1 r_2 r_3).\end{aligned} \tag{14}$$

Thus the odds value with respect to the no-feature configuration $r_{1,2,3} = (0,0,0)$ is explained by a multiplicative model where each factor, for example, $\phi_{2,3}(r_2 r_3)$, is equal to one if one or more descriptors are null, otherwise they are positive (the so-called treatment parameterization).

The elicitation in (14) is performed on the odds scale by asking the domain expert how many times the configuration (r_1, r_2, r_3) is more plausible than $(0,0,0)$ for descriptors R_1, R_2, R_3. This question is iterated for all configurations in the Cartesian product $\Omega_{R_1} \times \Omega_{R_2} \times \Omega_{R_3}$ after exclusion of $(0,0,0)$. Questions are posed from single main effects towards higher order interactions terms, so that one factor at a time is considered (see algorithm below).

The procedure is iterated for all cliques in the first CG component τ_1, and indeed factors already elicited in previously considered cliques are not reconsidered anymore. For example, the chain component R_1—R_2—R_3 is made by two cliques, R_1, R_2 and R_2, R_3; thus the shared factor $\phi_2(r_2)$ is elicited just one time.

The general algorithm for the first chain component is summarized below.

(1) Consider the undirected graph on descriptors elicited as the first chain component. Control questions include the following: "Are all the relevant features included in the elicitation?", "Are all pairs of features jointly affecting the probability of a structure linked by an undirected edge?". If needed, revise the order relation and (or) links within chain component.

(2) Check out that each descriptor R_{v_i} given its neighbors $R_{\mathrm{ne}(v_i)}$ is independent of all other descriptors in the first CG component.

(3) Find the cliques of such UG (model generator) [13].

(4) For each clique, elicit parameter values by using odds.

(i) Elicit main effects $\{\phi_i(r_{i,j})\}$, one at a time for all configurations, by assigning an odds value with respect to the baseline with no features; that is,

$$\phi_i\left(r_{i,j}\right) = \frac{p\left(0, \ldots, 0, r_{i,j}, 0, \ldots, 0 \mid \xi\right)}{p(0, \ldots, 0 \mid \xi)}. \tag{15}$$

A control question for this step is: "How much above one is the odds value for the sole presence of feature $R_i = r_{i,j} > 0$?"

(ii) Elicit first order interactions $\{\phi_{i,l}(r_{i,j'}, r_{l,j''})\}$, one at a time, by assigning a multiplicative term for each configuration under the question: "Which is the value of the multiplicative term $\{\phi_{i,l}(r_{i,j'}, r_{l,j''})\}$ needed to account for the interaction of feature $R_i = r_{i,j'} > 0$ with feature $R_l = r_{l,j''} > 0$?" The expression helping in this step is

$$\begin{aligned}&\frac{p\left(0, \ldots, r_{i,j'}, 0, \ldots, r_{l,j''}, 0, \ldots, 0 \mid \xi\right)}{p(0, \ldots, 0 \mid \xi)} \\ &\quad = \phi_i\left(r_{i,j'}\right)\phi_l\left(r_{l,j''}\right)\phi_{i,l}\left(r_{i,j'}, r_{l,j''}\right),\end{aligned} \tag{16}$$

where $\phi_i(r_{i,j'})$, $\phi_l(r_{l,j''})$ are already elicited; this is the cross product ratio of features R_i, R_l, after dividing by $\phi_i(r_{i,j'})$, $\phi_l(r_{l,j''})$. It is clear that if the interaction is absent $\phi_{i,l}(r_{i,j'}, r_{l,j''}) = 1$, while $\phi_{i,l}(r_{i,j'}, r_{l,j''}) \in (0,1)$ means that the interaction reduces the plausibility and $\phi_{i,l}(r_{i,j'}, r_{l,j''}) > 1$ raises the plausibility of the considered configuration.

(iii) Iterate the step above with higher order interaction terms (for all configurations) with two constraints: before moving to higher interaction terms all the terms of the same degree must have been already elicited; moreover the maximum degree of interaction among a subset of features

is defined by the size of the clique they belong to (model generator). For example, with the interaction of order two $\{\phi_{i,l,k}(r_{i,j'}, r_{l,j''}, r_{k,j'''})\}$, the control question might be: "Having already elicited values of $\phi_i(r_{i,j'})$, $\phi_l(r_{l,j''})$, $\phi_k(r_{k,j'''})$, $\phi_{i,l}(r_{i,j'}, r_{l,j''})$, $\phi_{i,k}(r_{i,j'}, r_{k,j'''})$, and $\phi_{l,k}(r_{l,j''}, r_{k,j'''})$, which is the value of the multiplicative term $\{\phi_{i,l,k}(r_{i,j'}, r_{l,j''}, r_{k,j'''})\}$ needed to adjust the odds value after considering the interaction among features whose configurations are $r_{i,j'} > 0$, $r_{l,j''} > 0$, $r_{k,j'''} > 0$?"; the helper expression is

$$\frac{p\left(0, \ldots, r_{i,j'}, \ldots, r_{l,j''}, \ldots, r_{k,j'''}, \ldots, 0 \mid \xi\right)}{p\left(0, \ldots, 0 \mid \xi\right)} = Q\,\phi_{i,l,k}\left(r_{i,j'}, r_{l,j''}, r_{k,j'''}\right), \tag{17}$$

where factor

$$Q = \phi_i\left(r_{i,j'}\right)\phi_l\left(r_{l,j''}\right)\phi_k\left(r_{k,j'''}\right)\phi_{i,l}\left(r_{i,j'}, r_{l,j''}\right) \times \phi_{i,k}\left(r_{i,j'}, r_{k,j'''}\right)\phi_{l,k}\left(r_{l,j''}, r_{k,j'''}\right) \tag{18}$$

is already elicited. Similar expressions follow straightforwardly for higher order interactions.

(5) Calculate the baseline probability $p(0, \ldots, 0 \mid \xi)$ of a structure without features by exploiting $\sum_{r \in \Omega_R} p(r \mid \xi) = 1$.

(6) Revise the elicited values following the guidelines of Section 2.7.

(7) Move to the next chain component.

The algorithm presented above may be also applied to chain components $\tau_2, \tau_3, \ldots$, that is, with a nonempty set of parents. The elicitation of parameters in a conditional Markov network is performed by iterating the algorithm for the first chain component over each configuration of conditioning parents. It follows that the elicitation burden depends on the number of parents, more precisely, on the cardinality of the Cartesian product where factors are samples spaces of parents.

The algorithms defined above lead to a prior distribution on configurations, but the elicitation does not end before the revision of elicited values takes place.

2.7. Implementation and Revision Issues. The flexibility achieved by defining predicates entails potential pitfalls that should be considered during an elicitation run with structural features.

The revision of elicited beliefs is always needed because the expert is not expected to provide unbiased elicited values, especially with limited training or in complex problem domains. Revision and elaboration of the prior distribution [22, and references therein] are applied to control elicitation bias and other causes of poor elicitation.

Overelicitation is an important practice to check for the presence of elicited quantities that do not reflect the expert degree of belief. Nevertheless, in large networks elaboration of elicited probability values is the main resource for checking the quality of elicitation. In the proposed framework, the elicitation through reference features produces proper probability distributions; therefore one elaboration consists of inspecting one or more margins of the probability distribution $p(r_1, \ldots, r_{n_f} \mid \xi)$, for example, the bivariate margin $p(r_i, r_j \mid \xi)$, to look for configurations whose plausibility causes surprise or disbelief in the expert. This operation is particularly meaningful if inspected margins do not straightforwardly relate to elicited odds, for example, taking descriptors belonging to different cliques. Surprise and disbelief ask for the revision of elicited values. If an association between selected margins and bias is suspected, random selection of margins is an option.

The stability of elicited values against different order relations on descriptors (reference features) should not be assumed. Nevertheless, in complex problem domains overelicitation made by the repetition of the interview with other ordered partitions not only seems unacceptable as regards the work load, but it could even be cognitively unfeasible, for example, if the original ordered partition selected by the expert is induced by a scientific hypothesis.

The core of the proposed approach is based on predicates representing structural features. The expert might believe that some configurations of features are plausible although they are incompatible with DAGs, for example, because they imply the presence of cycles. A positive probability value would be assigned to such a configuration r^*, but this is not an instance of elicitation bias if the elicited distribution properly matches expert beliefs. If this is the case, $p(z \mid \xi) = 0$ because $P[Z = z \mid R = r^*] = 0$.

The way a predicate is specified determines the granularity of the elicitation and the cardinality of equivalence classes in $\mathscr{Z}$. For example, let us consider two nodes $v_i \in V$ and $v_j \in V$ and the reference feature $\mathscr{R}_1$ = "Nodes v_i, v_j are not descendent of other nodes in V_R".

Reworking the original proposition we have two simpler predicates:

(i) $\mathscr{R}_{1'}$ = "Node v_i is not a descendent of other nodes in V_R";

(ii) $\mathscr{R}_{1''}$ = "Node v_j is not a descendent of other nodes in V_R."

and they can be considered by conjunction. In Table 2, the relation between the original reference feature (right) and the conjoint components (left) is shown: $\neg\mathscr{R}_1$ collects three configurations generated by the conjunction of simpler predicates. It follows that simpler predicates are needed if the expert degree of belief changes over collapsed configurations in $\mathscr{R}_1$, that is, $\neg\mathscr{R}_{1'} \cap \mathscr{R}_{1''}$, $\mathscr{R}_{1'} \cap \neg\mathscr{R}_{1''}$, $\neg\mathscr{R}_{1'} \cap \neg\mathscr{R}_{1''}$. While nothing prevents the expert from defining rich predicates, care should be taken to select a granularity suited to properly represent expert beliefs.

There are indeed several different ways of formulating a predicate. If two reference sets of features induce the same set of equivalence classes $\mathscr{Z}$ then they are operationally equivalent. Nevertheless, from a cognitive standpoint, substantial

Table 2: Relation between the original feature (right) and the conjunction of two simpler structural features (left).

Simpler feature 1	Simpler feature 2	Original feature
$\mathscr{R}_{1'}$	$\mathscr{R}_{1''}$	$\mathscr{R}_1$
$\neg\mathscr{R}_{1'}$	$\mathscr{R}_{1''}$	$\neg\mathscr{R}_1$
$\mathscr{R}_{1'}$	$\neg\mathscr{R}_{1''}$	$\neg\mathscr{R}_1$
$\neg\mathscr{R}_{1'}$	$\neg\mathscr{R}_{1''}$	$\neg\mathscr{R}_1$

differences in the ease of elicitation might depend on the way propositions are formulated [22]. Let us consider the reference feature $\mathscr{R}_1$ = "Nodes v_i, v_j precede all other nodes." Does the expert use "precede" as "come before" all other nodes in the order relationship of nodes in V? Given a DAG with v_j disconnected from other nodes how to answer? Does the expert use "precede" meaning "are ancestors of all other nodes in V" or to indicate that those nodes do not receive arrows coming from other nodes? This example makes clear that some training is mandatory in order to make an expert effective in the elicitation: a trained expert is expected to choose the right granularity in the elicitation and to define meaningful predicates, that is, statements straightforwardly true/false when applied to any DAG defined on V.

Several reference features are jointly considered in actual applications, a number typically far beyond what the expert may simultaneously consider with success. Thus there is the possibility that exclusion and implication relations are not recognized. For example, let us consider two features: $\mathscr{R}_3$ = "The indegree is three or less" and $\mathscr{R}_{25}$ = "v_i is a sink node." After reworking the last feature, the expert reformulates the statement as "Node v_i has indegree ten or more." Clearly the plausibility of $\mathscr{R}_3 \wedge \mathscr{R}_{25}$ should be null; otherwise the actual interpretation of $\mathscr{R}_3$ is "The indegree of all but v_i is three or less."

Implication must be recognized to maintain the probabilistic coherence; for example, let $\mathscr{R}_3$ = "Variable X_{v_i} is a direct cause of X_{v_j}" and $\mathscr{R}_5$ = "Variable X_{v_i} is an ancestral cause of X_{v_j}" be two reference features. Clearly $\mathscr{R}_3$ implies $\mathscr{R}_5$ and the joint plausibility of both features is bound to the plausibility of $\mathscr{R}_3$. In a normalized reference set logical relationships among features are properly handled.

3. Results and Discussion

A few seminal approaches to the elicitation of beliefs on structures are reconsidered as special cases in the proposed framework based on structural features. A published case study on breast cancer (BC case-study) [23] will be (partially) exploited at illustrative purposes. The whole set of nodes V includes: age (AGE), the proliferative index-marker Ki67/MIB-1 (PROLN), oestrogen receptors (ER), progesterone receptors (PR), the receptor tyrosine kinase HER2/neu (NEU), and the P53 protein.

3.1. Buntine 1991. In the seminal paper of Buntine [24], a prior distribution on the set of DAGs for a fixed set V is defined by assuming a total ordering of nodes in the context ξ. The probability of the parent set for node v is defined by the product of probability for events like "There is an edge $y \to v$," shortly $P["y \to v"]$, extended to each y preceding v in the order relation. The subjective probability value elicited for a network structure z is calculated by marginal independence of parent sets. The original formalization defines a total ordering $\prec$, so that if $y \prec x_i$ it may belong to the parent's set Π_i of x_i; then a full specification of beliefs for each edge in the directed graph is needed and measured in units of subjective probability; finally the independence of parent sets $(\Pi_1, \ldots, \Pi_n)$ is assumed. The distribution on the set of structures is ([24], modified)

$$p(z \mid \prec, \xi) = p(\Pi_1, \ldots, \Pi_n \mid \prec, \xi) = \prod_{i=1}^{n} p(\Pi_i \mid \prec, \xi) \tag{19}$$

$$p(\Pi_i \mid \prec, \xi) = \prod_{y \in \Pi_i \wedge y \prec x} P("y \to x_i" \mid \prec, \xi) \cdot \prod_{\neg\{y \in \Pi_i\} \wedge y \prec x} [1 - P("y \to x_i" \mid \prec, \xi)]. \tag{20}$$

Given the order relation (v_3, v_4, v_1, v_2) on four nodes, $n = 4$, parent sets are $\Pi_3 = \emptyset$, $\Pi_4 \subseteq \{y_{v_3}\}$, $\Pi_1 \subseteq \{y_{v_3}, y_{v_4}\}$, $\Pi_2 \subseteq \{y_{v_3}, y_{v_4}, y_{v_1}\}$.

Despite the huge importance of Buntine's seminal work [24], some limitations should be underlined. Under the causal semantic of BNs, the expert might fail in stating such node order which defines the "causal flow" along nodes. In large regulatory networks of system biology a lot of assignments are expected to be 0.5 because expert beliefs may involve a small subset of arrows. We also expect that some plausibility assignments depend on what is already assigned, for example, due to biological substantive laws, but the need of such conditioning is not accounted for. Finally, we remark that several node orders are compatible with the same sparse DAG on V; therefore the specification of a strict order should not be enforced.

The above example may be straightforwardly cast in terms of reference features. The order relation is part of the context ξ, and we define the set of reference features $R_{i,j}$ = "An arrow is from v_i to v_j," with all possible pairs (v_i, v_j) in which v_i precedes v_j in the total ordering $\prec$ equal to (v_3, v_4, v_1, v_2). The reference set is $\mathscr{R} = \{\mathscr{R}_{3,4}, \mathscr{R}_{3,1}, \mathscr{R}_{3,2}, \mathscr{R}_{4,1}, \mathscr{R}_{4,2}, \mathscr{R}_{1,2}\}$. In this case a trivial Bayesian network without arrows is equivalent to the prior distribution in (19) and (20) if conditional probability tables are defined by the same values specified in (20) for each feature; that is, features are marginally independent. To see this, note that the context puts a sharp constraint on the space of structures and that the cardinality of the equivalence classes for each configuration r of reference features is equal to one; that is,

$$p(z \mid \xi) = \prod_{i=1}^{n_f} P\left[R_i = r_i^{[z]} \mid \xi\right]. \tag{21}$$

The formal approach proposed in this paper allows much more flexibility, for example, by restricting the set of nodes to be ordered and by introducing dependence among arrows.

For example, let $V_B = (v_3, v_4, v_1, v_2) \subset V$ be an ordered subset of nodes taken from BC case study. The reference feature $\mathscr{R}_o$ = "V_B is the order on the relevant subset of nodes" induces two equivalence classes, the first made by DAGs on V which do not satisfy $\mathscr{R}_o$, the second one is made by structures without arrows violating the left-to-right order defined in V_B. Besides the sharp constraint obtained by setting $P[R_o = 1 \mid \xi] = 1$, the expert might consider such feature uncertain, thus preferring a degree of belief in the set $(0, 1)$. Further features could be defined to build the reference set:

$$\mathscr{R} = \mathscr{R}_o \cup \{\mathscr{R}_{3,4}, \mathscr{R}_{3,1}, \mathscr{R}_{3,2}, \mathscr{R}_{4,1}, \mathscr{R}_{4,2}, \mathscr{R}_{1,2}\} \quad (22)$$

and the ordered partition

$$\mathscr{O} = (\{R_o\}, \{R_{3,4}, \ldots, R_{1,2}\}) \quad (23)$$

captures weaker causal relationships by features like $\mathscr{R}_{i,j}$ = "Variable X_{v_i} is a causal ancestor of X_{v_j}." A CG model made by two components makes possible to elicit conditional beliefs about $v_4 \to v_2$ given the presence of an arrow $v_3 \to v_4$ and given the lack of $v_3 \to v_1$, without assuming a strict order relation on nodes in V.

3.2. Heckerman et al. (1995). In Heckerman et al. [9], a prior network $z_{\mathscr{P}}$ was elicited and compared to a candidate network z by counting the number of different edges, δ, with a high degree of belief assigned to structures closely resembling the prior network. The authors suggested to elicit the hyperparameter $0 < k < 1$ and to define the prior distribution to be proportional to k^{δ}.

Among the limitations penalizing the use of this prior we found the following:

(i) The impossibility of specifying the degree of belief if it depends not only on the number of different edges δ but also on their position and type; the presence/absence/direction of an arrow may have an impact on the belief about other edges.

(ii) The elicitation about a subset of V is not addressed.

(iii) The causal semantic is natural for this approach because each arrow represents an immediate cause; it seems difficult to mix probabilistic and causal beliefs by counting differences in arrows because a DAG in the probabilistic semantic is just a member of an equivalence class of DAGs representing the same collection of conditional independence relationships.

A simple reformulation is oriented to computation and it involves three operators: arrow deletion, insertion, and change of direction. The reference set of atomic features is $\mathscr{R} = \{\mathscr{R}_j(z) : j = 0, 1, 2, \ldots, J\} \cup \{R_a(z) : a = J + 1\}$ with $\mathscr{R}_j(z)$ = "The application of j operations produces $z_{\mathscr{P}}$," and where $R_a(z)$ = "The application of a or more operations produces $z_{\mathscr{P}}$." An undirected graph made by just one clique is associated with potentials represented in Table 3, where $J = 3$. On the right of Table 3, values of the potential function are shown. The normalization constant is $\sigma = (\theta_0 + \theta_1 + \theta_2 + \theta_3 + \theta_a)^{-1}$. It is obviously possible to make the two approaches as close as desired by setting $\theta_i \propto k^i$ with $i \leq 3$.

Table 3: Potential function for the Heckerman et al. [9] prior.

R_0	R_1	R_2	R_3	R_a	Potential
0	0	0	1	0	θ_3
0	0	1	0	0	θ_2
0	1	0	0	0	θ_1
1	0	0	0	0	θ_0
0	0	0	0	1	θ_a
Otherwise					0

An even simpler reformulation exploits the incompatibility of the above features and it is based on just one reference feature: $\mathscr{R}_1(z, a)$ = "The application of a operations produces $z_{\mathscr{P}}$," with $a \in \mathscr{A}_1 = (\{0\}, \{1\}, \ldots, \{4, 5, \ldots\})$, with arrow manipulations (insert-delete-change) defined as above. Feature $\mathscr{R}_1$ essentially defines a plausible neighborhood with respect to a prior network. Further reference features may be introduced to refine such plausibility, for example, by also considering one causal, $\mathscr{R}_2$, and one C.I., $\mathscr{R}_3$, relationships. In other words, a candidate network "close" to the prior network could be associated with a high prior probability which is then tuned according to the presence/absence of two other relevant features. A natural ordered partition on descriptors could be $\mathscr{O} = (\{R_1\}, \{R_2, R_3\})$; thus a two-component chain graph model may support the elicitation.

A different kind of reformulation is based on the full-probabilistic semantic in which a prior DAG $z_{\mathscr{P}}$ is just a way to define a collection of C.I. relationships. In this case, it is natural to define a structural feature for each conditional independence relationship if it is relevant according to the expert among those represented by $z_{\mathscr{P}}$. The reference set $\mathscr{R} = \{\mathscr{R}_1, \ldots, \mathscr{R}_{n_f}\}$ in this case is a collection of plausible C.I. statements taken from the prior DAG. Note that, although it is not essential to draw such prior DAG, it may be useful because a general collection of C.I. statements does not necessarily imply the existence of a compatible DAG.

If none among the C.I. relations is preeminent the ordered partition of descriptors contains just one element, say $\mathscr{O} = (\{R_1, \ldots, R_{n_f}\})$, and a CG model made by one component (undirected graphical model) is suited for the elicitation.

3.3. Imoto et al. (2003). In the seminal paper of Imoto et al. [25], the authors developed a framework for combining microarray data and biological knowledge while learning the structure of a BN representing relationships among genes. The proposed criterion has two components, and the second one is particularly interesting because it captures the a priori biological knowledge.

Following their original notation with minor modifications, $\pi(G)$ is the prior distribution of network G. Then, the interaction energy $U_{i,j}$ of the edge from (gene) v_i to (gene) v_j is defined on a sample space which is categorized into I values, say $H_1, H_2, \ldots, H_I$. For example if gene v_i regulates gene v_j then $U_{i,j} = H_1 > 0$, but if not much is known about such potential regulation then $U_{i,j} = H_2 > H_1$. The total energy of a network G is $E(G) = \sum_{(v_i, v_j) \in G} U_{i,j}$; thus the sum is taken over existing edges in G. The total energy may be

rewritten by collecting the parents of each node; thus $E(G) = \sum_{v_j \in V} \sum_{v_i \in pa(v_j)} U_{i,j} = \sum_{v_j \in V} E_j$. The (prior) probability of network G is modeled by the Gibbs distribution:

$$\pi(G) = \sigma^{-1} \exp(-\zeta E(G)), \tag{24}$$

where $\zeta > 0$ is a hyperparameter and σ is the normalizing constant, also called the partition function $\sigma = \sum_{G \in \mathcal{G}_V} \exp(-\zeta E(G))$, with $\mathcal{G}_V$ being the collection of all DAGs on a fixed set of nodes V.

Operationally, the prior information is coded into a square matrix U of size defined by the number of genes, with each $u_{i,j}$ corresponding to ζH_1 or ζH_2 according to the prior belief. Beliefs in protein-protein interactions are coded by $u_{i,j} = u_{j,i} = \zeta H_1$. Protein-DNA interactions between the transcription regulator v_i and the controlled gene v_j are accounted by setting $u_{i,j} = \zeta H_1$ and $u_{j,i} = \zeta H_2$. Some genes are controlled by a transcription regulator through a consensus motif in their DNA promoter region. If genes $v_{j_1}, v_{j_2}, \ldots, v_{j_n}$ have the consensus motif and they are regulated by gene v_i then $u_{j_1,i} = \cdots = u_{j_n,i} = \zeta H_2$ and $u_{i,j_1} = \cdots = u_{i,j_n} = \zeta H_1$.

The seminal approach of Imoto et al. [25] suffers of two main limitations. They stated that the biological knowledge should suggest the partitioning of the underlining continuous energy function but it is not clear how, even after invoking the metaphor of energy from physics. In Example 3.2, they tried $\zeta H_1 = 0.5$ (but also $\zeta H_1 = 1$) and optimized the selection of $\zeta H_2 = 2.5$, a procedure not much in line with pure preexperimental Bayesian elicitation. Moreover, the sum in the partition function is taken on the set of DAGs on V, and it becomes quite intractable from 5 nodes on. It follows that their approach is substantially a way to build a score function in a spirit similar to [26], but without providing a flexible support for the calibration of the score function.

The prior information considered by these authors can be also expressed in terms of reference features, for example, by considering features like $\mathcal{R}_{i,j}$ = "Gene v_i regulates gene v_j," for all relevant pairs of genes. If preeminent features are absent, an UG model formally resembles the expression based on energy functions but with some major differences:

(1) the parameterization is related to subjective probability through odds values;

(2) the normalization constant is easier to calculate, at least if the number of features is less than the total pair of genes (some genes omitted);

(3) the general calibration constant ζ disappears, although a similar tuning has been considered elsewhere [21] to smooth raw elicited odds values while trying to compensate for the well known human tendency towards overstating odds values.

3.4. Werhli and Husmeier (2007). The authors [27], building on the work of Imoto et al. [25], defined a prior information matrix B whose elements $B_{i,j} \in [0, 1]$, with i, j being a pair of integers for nodes v_i and v_j and the relation $v_i \rightarrow v_j$. If no prior preference about the presence of such arrow is elicited, then $B_{i,j} = 0.5$; if $0 \leq B_{i,j} < 0.5$ elicited beliefs put more plausibility on the lack of arrow $v_i \rightarrow v_j$; if $0.5 < B_{i,j} \leq 1$ higher plausibility favors the presence of the arrow $v_i \rightarrow v_j$. Note that elements $B_{i,j}$ in B are not probabilities.

The calculation of the prior probability for a candidate DAG G is straightforward if it is represented through an adjacency matrix in which the element $G_{i,j}$ in row i and column j is 1 if the DAG has an arrow form v_i to v_j, and it is zero otherwise. At first the energy of the DAG is calculated as

$$\mathcal{E}(G) = \sum_{i,j} \left| B_{i,j} - G_{i,j} \right| \tag{25}$$

with $G_{i,j}$ and $B_{i,j}$ being, respectively, the elements of G and B. The probability elicited for G is

$$\pi(G \mid \beta, B) = \sigma^{-1}(\beta, B) \exp(-\beta \mathcal{E}(G)), \tag{26}$$

where β is a hyperparameter regulating the overall strength of the elicited degree of belief and with σ being the partition function which depends on a sum of energy values over the collection of all DAGs on a fixed set of nodes V. It follows that an increase of energy is related to a larger mismatch with the prior B. In the limit of $\beta \rightarrow 0$ a noninformative prior is obtained, while for β diverging to infinity peaked priors are defined.

The relationship with the seminal approach of Imoto et al. [25] is clear but despite the easier parameterization chosen to define the energy function, several limitations are still present. First of all, the overall strength β is not straightforwardly related to (subjective) probability; therefore at some point the expert has to play numerically with fake examples to get the feeling on reasonable values for such hyperparameter. The calibration of the whole approach is difficult because it depends on the calculation of the normalization constant which is hard to obtain due to the calculation of energy values for all DAGs on a fixed set of nodes V. Despite the shortcut proposed by the authors, the calculation of the normalization constant remains as a bottleneck of the approach.

The authors increased the flexibility of their approach in representing prior beliefs by using two matrices $B^{(1)}$ and $B^{(2)}$:

$$\pi\left(G \mid \beta_1, B^{(1)}, \beta_2, B^{(2)}\right) = \left\{\sigma\left(\beta_1, B^{(1)}, \beta_2, B^{(2)}\right)\right\}^{-1} \times \exp\left(-\left\{\beta_1 \mathcal{E}_1(G) + \beta_2 \mathcal{E}_2(G)\right\}\right) \tag{27}$$

with $\mathcal{E}_1(G)$, $\mathcal{E}_2(G)$ being the energy functions, respectively, depending on $B^{(1)}$ and $B^{(2)}$. In the elicitation based on a widely adopted database of metabolic pathways, values of B are defined as the ratio $m_{i,j}/M_{i,j}$, with $M_{i,j}$ being the number of times two genes appear in a pathway and $m_{i,j}$ the number of times that they are linked inside such pathway. Results of a study on 25 genes measured at 73 time points suggested that the procedures using prior information outperformed those without it. Nevertheless, the above mentioned limitations get even worse under such generalization; for example, the explosion in the number of terms entering the partition function must be taken under control by introducing a

limitation on the number of arrows entering a node, a technical constraint which may lead to biased elicited beliefs.

The expressivity obtained by these authors through the use of two B matrices may be similarly obtained using reference features with some advantages. Back to the BC case study, a major hypothesis under which the elicitation may be performed could be $\mathscr{R}_1$ = "ER is a hub gene," so that given the configuration $R_1 = 1$ the expert has to express further conditional beliefs about the other markers; for example, $\mathscr{R}_2$ = "PR regulates P53," $\mathscr{R}_3$ = "P53 acts on KI67," $\mathscr{R}_4$ = "P53 acts on NEU." In this context it is natural to consider the order relation on features $\mathcal{O} = (\{R_1\}, \{R_2, R_3, R_4\})$; thus a chain graph model is suited for the elicitation with just node R_1 in the first chain component and as many arrows leaving from R_1 as needed to capture changes of conditional beliefs due to a switch in the major hypothesis, from $R_1 = 1$ to $R_1 = 0$. Note that, using a chain graph on features, it is possible to tune the flexibility exactly of the amount needed, without the unnecessary and blind consideration of all pairs of genes.

3.5. Discussion. The expressivity achieved through reference features is wide whether probabilistic or causal information is elicited. Many limitations found in other approaches depend on the consideration of arrows as the key building block of the elicitation. Further restrictions are due to the use of marginal relationships among arrows, so that severe constraints on the expressivity of the approach follow. The elicitation based on reference features has maximum resolution because, for a suitable set of reference features, it is possible to define a prior distribution characterized by a probability value for each DAG defined on V.

There are useful side effects in the approach based on structural features. First, the elicitation effort does not depend on the size of the space of DAGs on V, but on the size of the collection of features. Another side effect is related to the cardinality of equivalence classes in $\mathscr{Z}$. Two candidate DAGs z' and z'' characterized by two different configurations of structural features may receive a very different prior probability value just because the cardinality of the equivalence classes they belong to is very different, say $n_{r[z']} \lll\ggg n_{r[z'']}$ (see (11)). Note that the cardinality of equivalence classes must be taken into account to preserve probabilistic coherence.

The above description of the elicitation process did not deal with computational issues that are very important for applications. The proposed approach is suited to large networks if the number of DAGs within each equivalence class is available, for example, as a Monte Carlo point estimate. Monte Carlo simulation makes it possible to explore the space of DAGs and it provides evidences about the presence of features which are logically incompatible; thus it may also suggest predicates on which the expert should focus to improve the definition of reference features. Moreover, the elicitation defines a proper prior distribution, so that different MCMC algorithms could be developed to obtain the posterior distribution of Z. New greedy search algorithms to find plausible structures in Ω_Z could be investigated to exploit the elicited reference features, for example, by developing an algorithm that generates candidate DAGs belonging to different equivalence classes in $\mathscr{Z}$ at each iteration of the optimization.

The approach to elicitation based on graphical models does not necessitate very esoteric software, although software libraries for platforms commonly adopted in scientific computing would offer the opportunity of performing extensive testing and of investigating human heuristics specifically relevant in this elicitation framework. It is well known that graphical user interfaces facilitate the elicitation, especially if experts are not much trained in Bayesian statistics, and this is a resource which is anyway almost mandatory to face the elicitation in problem domains involving a large number of structural features.

4. Conclusions

Graphical models may be exploited to elicit beliefs about the structure of an unknown BN from experts. The joint plausibility on configurations of structural features is decomposed according to conditional independence relationships that are considered plausible by an expert. The expert may use CG models to elicit structural prior information in quite complex domains without leaving a full Bayesian framework. From the elicited CG model, and eventually by using an auxiliary Monte Carlo simulation to estimate the cardinality of equivalence classes, a (proper) subjective prior distribution on the space of DAGs is built and ready to be used with the likelihood function in order to find BN structures supported both by expert beliefs and by collected observations.

No surprise that in complex domains the definition of a prior distribution may be costly. A trade-off should be found by considering the goal of the analysis, how much prior information is available, and the cost and importance of collected data. Here system biology and medicine are expected to be fields in which the proposed approach might be useful, because subjective prior information, besides providing the above mentioned benefits, also tempers the curse of dimensionality caused by structures defined on a high number of variables.

Conflict of Interests

The author declares that there is no conflict of interests regarding the publication of this paper.

Acknowledgment

This work was partially funded by a grant from the University of Florence.

References

[1] J. Pearl, *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*, Morgan Kaufmann, San Francisco, Calif, USA, 1988.

[2] J. Pearl, *Causality*, Cambridge University Press, 2000.

[3] D. E. Heckerman, E. J. Horvitz, and B. N. Nathwani, "Toward normative expert systems: part I. The Pathfinder project,"

Methods of Information in Medicine, vol. 31, no. 2, pp. 90–105, 1992.

[4] K. McNaught and A. Chan, "Bayesian networks in manufacturing," *Journal of Manufacturing Technology Management*, vol. 22, no. 6, pp. 734–747, 2011.

[5] B. G. Marcot, R. S. Holthausen, M. G. Raphael, M. M. Rowland, and M. J. Wisdom, "Using Bayesian belief networks to evaluate fish and wildlife population viability under land management alternatives from an environmental impact statement," *Forest Ecology and Management*, vol. 153, no. 1–3, pp. 29–42, 2001.

[6] N. Friedman, M. Linial, I. Nachman, and D. Pe'er, "Using Bayesian networks to analyze expression data," *Journal of Computational Biology*, vol. 7, no. 3-4, pp. 601–620, 2000.

[7] A. O'Hagan and J. Foster, *Bayesian Inference*, vol. 2 of *Kendall's Advanced Theory of Statistics*, John Wiley & Sons, Chichester, UK, 2nd edition, 2004.

[8] D. V. Lindley, "The philosophy of statistics," *Journal of the Royal Statistical Society D*, vol. 49, no. 3, pp. 293–337, 2000.

[9] D. Heckerman, D. Geiger, and D. M. Chickering, "Learning Bayesian networks: the combination of knowledge and statistical data," *Machine Learning*, vol. 20, no. 3, pp. 197–243, 1995.

[10] R. W. Robinson, "Counting labeled acyclic digraphs," in *New Directions in the Theory of Graphs*, pp. 239–273, Academic Press, New York, NY, USA, 1973.

[11] F. M. Stefanini, "Graphical models for eliciting structural information," in *Classification and Data Mining*, A. Giusti, G. Ritter, and M. Vichi, Eds., Springer, Berlin, Germany, 2013.

[12] F. M. Stefanini, "Eliciting expert beliefs on the structure of a Bayesian network," in *Proceedings of the Probabilistic Graphical Models (PGM '08)*, Hirtshals, Denmark, 2008.

[13] J. Whittaker, *Graphical Models in Applied Multivariate Statistics*, John Wiley & Sons, Chichester, UK, 1990.

[14] R. G. Cowell, A. P. Dawid, S. L. Lauritzen, and D. J. Spiegelhalter, *Probabilistic Networks and Expert Systems*, Springer, New York, NY, USA, 1999.

[15] D. Koller and N. Friedman, *Probabilistic Graphical Models: Principles and Techniques*, MIT Press, Cambridge, Mass, USA, 2009.

[16] A. P. Dawid, "Conditional independence in statistical theory," *Journal of the Royal Statistical Society B*, vol. 41, no. 1, pp. 1–31, 1979.

[17] M. Frydenberg, "The chain graph Markov property," *Scandinavian Journal of Statistics*, vol. 17, no. 4, pp. 333–353, 1990.

[18] J. Pearl, "Causal inference in statistics: an overview," *Statistics Surveys*, vol. 3, pp. 96–146, 2009.

[19] A. P. Dawid, "Beware of the DAG," *Journal of Machine Learning Research*, vol. 6, pp. 59–86, 2008.

[20] F. M. Stefanini, "Prior beliefs about the structure of a probabilistic network," in *Proceedings of the Statistical Methods for the Analysis of Large Data-Sets (SIS '09)*, Pescara, Italy, 2009.

[21] F. M. Stefanini, "The revision of elicited beliefs on the structure of a Bayesian network," in *Proceedings of the Complex Data Modeling and Computationally Intensive Statistical Methods for Estimation and Prediction (SCo '09)*, Milano, Italy, 2009.

[22] P. H. Garthwaite, J. B. Kadane, and A. O'Hagan, "Statistical methods for eliciting probability distributions," *Journal of the American Statistical Association*, vol. 100, no. 470, pp. 680–700, 2005.

[23] F. M. Stefanini, D. Coradini, and E. Biganzoli, "Conditional independence relations among biological markers may improve clinical decision as in the case of triple negative breast cancers," *BMC Bioinformatics*, vol. 10, supplement 12, article S13, 2009.

[24] W. Buntine, "Theory of refinement on Bayesian networks," in *Proceedings of the 7th Conference on Uncertainty in Artificial Intelligence*, pp. 52–60, 1991.

[25] S. Imoto, T. Higuchi, T. Goto, K. Tashiro, S. Kuhara, and M. Satoru, "Combining microarrays and biological knowledge for estimating gene networks via Bayesian networks," in *Proceedings of the IEEE Computational Systems Bioinformatics (CSB '03)*, pp. 104–113, IEEE Computer Society, 2003.

[26] M. Mascherini and F. M. Stefanini, "Using weak prior information on structures to learn Bayesian networks," in *Knowledge-Based Intelligent Information and Engineering Systems*, B. Apolloni, R. J. Howlett, and L. Jain, Eds., vol. 4692 of *Lecture Notes in Computer Science*, pp. 413–420, 2007.

[27] A. V. Werhli and D. Husmeier, "Reconstructing gene regulatory networks with Bayesian networks by combining expression data with multiple sources of prior knowledge," *Statistical Applications in Genetics and Molecular Biology*, vol. 6, no. 1, article 15, 2007.