Biometrical Letters
Vol. 46 (2009), No. 2, 129-152

Learning Bayesian Networks using expert's prior information on structures

Massimiliano Mascherini[1], Alessandro Camussi[2], Federico M. Stefanini[3]

[1]Joint Research Centre of the European Commission, 21027 Ispra (Va), Italy,
e-mail: massimiliano.mascherini@jrc.it
[2]Department of Agricultural Biotechnology, Genetics Unit, University of Florence,
via Maragliano, 75-77 50144 Florence, Italy, e-mail: alessandro.camussi@unifi.it
[3]Department of Statistics, University of Florence, viale Morgagni, 59 50134 Florence, Italy,
e-mail: stefanini@ds.unifi.it

Dedicated to Professor Tadeusz Caliński for his 80th birthday

SUMMARY

Most of the approaches developed in the literature to elicit the *a priori* distribution on Directed Acyclic Graphs (DAGs) require a full specification of graphs. Nevertheless, expert's prior knowledge about conditional independence relations may be weak, making the elicitation task troublesome. This paper presents and evaluates an elicitation procedure for DAGs which exploits prior knowledge on network topology. The elicitation is suited to large Bayesian Networks (BNs) and it accounts for immediate causal link and DAG sparsity. We develop a new quasi-Bayesian score function, the P-metric, to perform structural learning following a score-and-search approach. We tested our score function on two different benchmark BNs by varying sample size and prior belief concerning structures. Our results show the effectiveness of the proposed method and suggest that the use of prior information improves the structural learning process.

Key words: Bayesian Networks; Structural Learning; Prior Information

1. Introduction

Bayesian Networks (BNs) (Jensen, 1996; Pearl, 1988), are a widely used tool in many areas of artificial intelligence and automated reasoning, because they perform probabilistic inference through very efficient algorithms. However, the problem of searching the BN that best depicts the dependence relations entailed in a database of cases is hard to solve.

The Bayesian approach to structural learning exploits algorithms which typically combine expert's knowledge with the information gathered from a database. In particular, it assumes that a space of structures is defined and that one of these structures is the true model of the process which led to the data being observed. Then a prior distribution is defined over the space of structures and for any given structure it represents the expert's belief about such configuration before considering the data. The prior plus the data lead to the posterior distribution over the space of structures, and from a pure Bayesian point of view the posterior distribution is the technical goal of structural learning.

Unfortunately the complete specification of a prior distribution on the topology of a Bayesian Network (BN) is NP-Hard (Chickering, 1995), and most of the approaches in the literature require a complete specification of a prior probability distribution on the space of Directed Acyclic Graphs (DAGs). Nevertheless, there are problem domains in which such complete elicitation is difficult or unfeasible, due to the lack of detailed information about network features. A prior state of partial knowledge about a network's topology may take several forms, like independence relations among subsets of variables or an ordering relation for just a subset of nodes.

In this paper we develop a method to elicit partial beliefs about a network's structure without requiring the *a priori* complete specification of structures. Elicited beliefs are refined by means of dissimilarity measures on the network's topology. In order to perform structural learning in a score-and-search framework, we propose a new score function to evaluate causal Bayesian Networks: the P-metric. It is a quasi-Bayesian score obtained by modifying the Bayesian Dirichlet Equivalent metric (BDe) (Heckerman et al., 1994). The characteristic of a likelihood equivalent metric is that it assigns the same likelihood value to structures entailing the same conditional independence assertions. The P-metric exploits prior information to discriminate among causal structures within equivalence classes, thus it is not likelihood-equivalent.

In section 2 we briefly review some basic concepts about Bayesian Networks. Section 3 contains a description of early approaches to elicit prior information on structures, and in section 4 we detail our approach. A new elicitation procedure using the P-metric is presented in section 5. Numerical results from the analysis of some Machine Learning benchmark datasets are

presented in section 6. Finally, in section 7, we present conclusions and issues to be addressed by further research.

2. Graphs and Bayesian networks

A review of some important definitions in graph theory and of Markov properties is provided. Comprehensive accounts of probabilistic networks may be found in Jensen (1996) and Cowell et al. (1999).

A *graph G* is an ordered pair (*V,E*), with *V* a finite set of nodes $\{v_1, v_2, \ldots\}$ and $E \subset V \times V$ the set of edges. If $(v_i, v_j) \in E$ and $(v_j, v_i) \notin E$ then there is a directed edge from v_i to node v_j , also denoted as $v_i \rightarrow v_j$.

Given $(v_i, v_j) \in E$ we say that v_i and v_j are *adjacent* or *neighbourhoods* of each other: v_i is said to be a *parent* of v_j, and v_j is also called a *child* of v_i. By iterating the two definitions of parent and child recursively, the set of *ancestor* nodes and *descendant* nodes are defined. An ancestral set *A* of node α is a subset of *V* in which for each node in *A* all its parents are in *A* as well. The smallest ancestral set containing a node α is indicated as $An(\alpha)$. A node is called a *root* if it does not have any parent.

For every $v_i \in V$ it holds that $(v_i, v_i) \notin E$ because a node cannot originate an arrow pointing to itself. If $(v_i, v_j) \in E$ and $(v_j, v_i) \in E$ then the edge is said to be undirected.

A *directed graph* G_{DG} contains only directed edges, $(v_i, v_j) \in E \Rightarrow (v_j, v_i) \notin E$.

A path connecting two nodes whatever the direction of edges on the path is called an *adjacency path* or *chain*, to distinguish it from the *directed path, dp,* where edges are all oriented in the same direction, i.e. edges meet head-to-tail for each node. A *Directed Acyclic Graph (DAG)* $G_D = (V, E_D)$ is a directed graph without cycles, i.e. no directed path originated by v_i leads back to the starting node v_i.

A Bayesian Network *B* is a graph-based representation of a joint probability distribution *P* which is Markov with respect to the graph. Random variables are labelled by nodes in the graph, e.g. x_{v_i} with state space χ_{v_i} . For shortness, labels also sometimes indicate random variables. In this paper we will only consider discrete random variables.

The Markov property allows the factorization of the joint probability distribution following the child-to-parents structure:

$$p(x) = \prod_{v_i \in V} p\left(x_{v_i} \mid x_{pa(v_i)}\right) \tag{1}$$

It follows that the joint probability distribution may be represented by a collection of conditional probability tables (CPTs) one for every pair (v_i, $pa(v_i)$) in the graph, with $pa(v_i)$ the parent nodes of v_i. To every pair v_i, $pa(v_i)$ of a given network B_s is associated a CPT whose parameters are here indicated as $\theta_{s,v_i,pa(v_i)}$, Given the structure s, the vector of all parameters is $\theta_s = \left\{\theta_{s,v_i,pa(v_i)}\right\}$.

A graph G_D does not always represent all the conditional independence relations entailed by the probability distribution P. If it does, we say that P and G_D are *faithful* to each other. The conditional independence relations which are not determined by "numerical accident" may be represented by a DAG. In a faithful DAG all the conditional independence relations by a BN are revealed by assessing the *direction-dependent separation* property (Geiger and Pearl, 1988), also called *d-separation* (Pearl, 1988).

Given a DAG $G_D = (V,E)$, with $\left(v_i, v_j\right) \in V$ and $v_j \neq v_i$, let C be a subset of V, $C \subset V \setminus \left\{v_i, v_j\right\}$. We say that v_i and v_j are *d-separated* in G_D given C, if and only if there exists no adjacency path *ap* between v_i and v_j such that: (1) every collider on *ap* is in C or has a descendant in C; (2) no other node on path *ap* is in C.

The subset C is the so-called *cut-set*. If v_i and v_j are not separated given C, we say that v_i and v_j are *d-connected* given C. The definition of d-separation of two nodes can be easily extended to the d-separation of two disjoint sets of nodes $X \subset V$ and $Y \subset V$ by iterating the above definition for each pair $\left(v_i, v_j\right)$, with $v_i \in X$ and $v_j \in Y$.

3. Earlier approaches to using prior information on structures in learning Bayesian Networks

In this section we provide a brief overview of the existing approaches to including prior information in BN structural learning. It should be noted that the elicitation problem for prior beliefs on a network's structure has been not much

considered in the literature, where relatively little attention has been paid to the elicitation of beliefs about structures (Friedman and Koller, 2003).

A straightforward elicitation of prior beliefs on complex structures is performed element-by-element, assigning (subjective) probability values to graphs defined on a given set *V* of nodes. The enumerative approach is unfeasible except in networks with a very small set of nodes, because the space of DAGs has superexponential cardinality as the number of nodes in *V* increases.

A simpler approach puts a uniform prior distribution on a subset *H* of all possible DAGs (Heckerman et al., 1994; Srinivas et al., 1990), therefore some structures are *a priori* excluded from the scoring procedure. Bounds on some structural features are established to set hard constraints on elements in *H*. For example a variable can be declared to be a root/leaf node or the parent of another. In addition constraints on the number of parents/children or on partial order between variables can be set. This approach has been applied in both Bayesian and non-Bayesian learning approaches.

Two more elaborate approaches have been proposed by Buntine (1991), Chickering (1995), and Madingan and Raftery (1994) to define a prior distribution on the space of BN structures. Both of them require a complete specification of beliefs over the network, making their implementation not very practical in large networks.

In the so-called Buntine approach (Buntine, 1991), an initial partial theory provided by the expert is transformed into a prior probability over the space of theories. The partial theory consists of: (1) a total ordering $\prec$ on variables, such that if node *y* is in the set of parents of node *x* then $y \prec x$ in the relation set; (2) a full specification of beliefs for each edge in the directed graph, measured in units of subjective probability.

The joint prior distribution conditioned on the total ordering of variables is defined by assuming the independence of parents sets. The joint prior probability distribution is factorized as:

$$p(B_S \mid \prec, \xi) = \prod_{i=1}^{n} p(\pi_i \mid \prec, \xi) \tag{2}$$

By expanding the generic term $p(\pi_i \mid \prec, \xi)$, we have:

$$p(\pi_i \mid \prec, \xi) = \prod_{y \in \pi_i} p(y \rightarrow x_i \mid \prec, \xi) \cdot \left(\prod_{y \notin \pi_i} 1 - p(\pi_i \mid \prec, \xi) \right) \tag{3}$$

In the approach proposed by Heckerman et al. (1994) the expert builds a complete *a priori* network, B_{sc} (*s* for structure and *c* for complete), and the conditional probability of the next case to be seen (observation on a statistical unit) is defined. The joint probability distribution on the domain *U* of random variables is obtained at this purpose, $p(U \mid B_{SC}, \xi)$ where B_{sc} is the complete network. Informative prior distributions for model parameters are built in a peculiar way to obtain the so called Bayesian Dirichlet Equivalent metric (BDe metric).

The prior distribution on BN structures is independent from the prior network, B_{sc}, but in their approach, structures closely resembling the prior network receive a high prior probability, while others are penalized. The number of nodes in the symmetric difference of $\pi_i(B_S)$ and $\pi_i(B_{SC})$ is:

$$\delta_i = \left| \{\pi_i(B_S) \cup \pi_i(B_{SC})\} \setminus \{\pi_i(B_S) \cap \pi_i(B_{SC})\} \right| \tag{4}$$

It follows that the number of different arcs δ between the prior network *Bsc* and a network *Bs* is $\delta = \sum_{i=1}^{n} \delta_i$. By introducing the constant $0 \leq \cdot\ k \leq \cdot\ 1$, the prior distribution penalizing networks not much close to the *a priori* network is

$$p(B_S \mid \xi, B_{SC}) = c \cdot k^{\delta} \tag{5}$$

where *c* is a normalization constant.

Finally, the method proposed by Madigan and Raftery (1994) is similar to the approach used by Heckerman et al. (1994), but it is coarser in avoiding the elicitation of a large number of arc probability values. An arc elicited in one or more DAGs is associated to a constant probability value which is higher than the value for arcs that do not belong to any elicited DAG. Let $\varepsilon = \varepsilon_p \cup \varepsilon_a$ denote the set of all possible links, where ε_p is the set of links which are present in the model and ε_a is the set of absent links; they assume that the evidence in favour of an included link corresponds to a prior link probability for all $e \in \varepsilon_p$ of

$$p(e) = \left(1 + \exp\left(\frac{O_L + O_R}{2} \right) \right)^{-1}$$

and, similarly, the prior link probabilities for $e \in \varepsilon_a$ are given by:

$$p(e) = \exp\left(\frac{O_L + O_R}{2}\right) \cdot \left(1 + \exp\left(\frac{O_L + O_R}{2}\right)\right)^{-1}$$

where the parameters O_L and O_R are set by the users and their effect is to determine the prior bias in favour of arcs included in the model provided by the users.

4. From prior information to score functions

The specification of a complete prior network with beliefs over all possible edges is unrealistic for large networks. The elicitation of expert's prior information element by element is performed through the assignment of (subjective) probability values to all possible arrows of a Bayesian Network, as in Buntine (1991), but it becomes very difficult due to the superexponential cardinality of the space of structures as the number of nodes increases. In large networks, a coherent and complete specification of a prior distribution on the space of networks (Chickering et al., 1994), would seem to be extremely difficult.

In this section a score function, $S_{prior}(Bs)$, mirroring prior beliefs, is defined to drive score-and-search algorithms for structural learning. It requires far less elicitation of prior beliefs from the expert than in Buntine (1991) and Heckerman et al. (1994).

Expert's prior information on a large problem domain may be strong but partial, for example it may deal with the orientation of some edges over hundreds, or with global network traits like the size of the graph. In gene expression analysis, for example, a small degree of graph connectivity is *a priori* expected and substantial knowledge may concern the partial ordering of ten out of thousands of genes. In order to fully exploit the *a priori* structural information both local and global features have to be taken into account. In our approach the expert is expected to express: (1) beliefs over some, but not all, possible edges of the network; (2) beliefs over some features of the network topology, like the expected number of node parents or the degree of network connectivity.

Given these assumptions, we propose to elicit the *a priori* belief on the structure of a candidate network B_s by means of a score function $S_{prior}(B_s)$ capturing local and global network features:

$$S_{prior}(B_S) = f\left(S_p^{\delta}(B_S), S_p^{T}(B_S)\right)$$

The score component $S_p^{\delta}(B_S)$ refers to edges elicited one at a time. The second score component $S_p^{T}(B_S)$, describes global network features, related to DAG connectivity.

4.1. Encoding local features

The score component $S_p^{\delta}(B_S)$ encodes the expert's belief ξ on the presence of oriented edges, each one marginally considered.

The DAG's structure is specified by the subset $E \subset V \times V$. We conventionally indicate a pair of nodes (v_i, v_j) in the canonical order $i < j$, and we use the deponent $i \cdot j$ to refer to the edge between nodes v_i and v_j . A structure is more parsimoniously represented by a collection M of $F \leq n(n-1)/2$ variables $M = \{m_1, \ldots m_f, \ldots, m_F\}$ each one taking values on $\chi = \{-1, 0, 1\}$ for each pair of nodes (v_i, v_j), $i < j$, in V .

The respective values in χ are indicated by: an arrow $i \leftarrow j$, no arrow between i and j, and an arrow $i \rightarrow j$. Expert's belief takes the form of a set of probability distributions over the collection $M : \left\{p\left(x_{m_f} \mid \xi\right) : m_f \in M\right\}$.

The distributions over the collection M are coded as vectors of probability values $P_{i.j}^{T} = \left(p_{i.j,-1}, p_{i.j,0}, p_{i.j,+1}\right)$, so that $1^T P_{i \cdot j} = 1$, where the value $p_{i \cdot j,-1}$ represents, for instance, the probability assigned by the expert to the presence of the arrow $i \leftarrow j$.

For each pairs i, j, connectivity vectors $C_{i \cdot j} = (I_{i \cdot j,-1}, I_{i \cdot j,0}, I_{i \cdot j,+1})$ are introduced to indicate the value taken by variables, where $I_{i \cdot j,-1} = 1$ if the arrow $i \leftarrow j$ exists and 0 otherwise. It follows that $1^T C_{i \cdot j} = 1$. The probability value associated to the oriented edge for a pair $i \cdot j$ is $C_{i.j}^{T} P_{i.j}$.

The above construction leads to the specification of a probability distribution on the set of directed graphs G_{DG} in which the candidate directed graph B_D has a prior probability value equal to:

$$P(B_D \mid \xi) = \prod_{\{i.j\}} C_{i.j}^{T} P_{i.j}$$

The above factorization refers to our prior judgment about the existence of a link between v_i and v_j without considering other nodes.

The space of DAGs is contained in the space of Directed Graphs, $G_D \subseteq G_{DG}$, therefore the above construction also induces a probability distribution over DAGs contained in the space of directed graphs, $B_S \in G_{DG}$:

$$P(B_S \mid \xi) \propto I_{DAG}(B_S) \cdot \prod_{\{i.j\}} C_{i.j}^T P_{i.j} \tag{6}$$

with $I_{DAG}(B_s)$ taking the value one if B_s is a DAG, zero otherwise. The proportionality is due to an omitted constant depending on directed graphs which are not DAGs because of cycles. We remark that from a theoretical point of view there is no difficulty in calculating the value of the normalization constant, but given the huge cardinality of spaces of Directed Graphs the computation may not be practical.

We define the score $S_\delta(B_s)$ of a candidate Bayesian Networks using (6):

$$S_\delta(B_S) = \log\left(\frac{P(B_S \mid \xi)}{P(\{0\} \mid \xi)}\right) \tag{7}$$

with $P(\{0\} \mid \delta)$ the prior probability assigned to the Bayesian Network in which E is empty (graphs without edges). By straightforward algebra it may be shown that computation of the normalization constant c is not needed in order to search in the space of networks. In fact, omitting ξ for simplicity, let $P_{GD}(B_s)$ be the probability distribution over DAGs and $P_{GDG}(B^s)$ be the probability distribution over the space of Directed Graphs. By straightforward algebra, and deleting δ from the notation for simplicity, we have:

$$\begin{aligned}
&S_\delta(B_S) = \log\left(\frac{P_{G_D}(B_S)}{P_{G_D}(\{0\})}\right) = \log\left(P_{G_D}(B_S)\right) - \log\left(P_{G_D}(\{0\})\right) = \\
&= \log\left(cP_{DG}(B_S)\right) - \log\left(cP_{DG}(\{0\})\right) = \\
&= \log(c) + \log\left(P_{DG}(B_S)\right) - \log(c) - \log\left(cP_{DG}(\{0\})\right) = \\
&= \log\left(\frac{P_{DG}(B_S)}{P_{DG}(\{0\})}\right) = \log\left(\frac{\prod^{B_S} C_{i.j}^T P_{i.j}}{\prod^{\{0\}} C_{i.j}^T P_{i.j}}\right)
\end{aligned}$$

where $\prod^{B_S} C_{i.j}^T P_{i.j}$ and $\prod^{\{0\}} C_{i.j}^T P_{i.j}$ refer to factorization of the prior judgment respectively over the candidate network *Bs* and to the empty structure.

A remarkable property of the score $S_\delta(B_s)$ in equation (7) concerns the possibility of calculating scores by just considering the pair of nodes for which the expert defined a distribution. Let F be the number of pairs of nodes for which the belief has been elicited by the assignment of a distribution $\{p(x_{m_f} \mid \xi): f = 1,...,F\}$ and let k be the constant values assigned to the $F - n(n-1)/2$ cases for which no belief has been elicited. For any given structure B_s we have:

$$P(B_S) = \prod C_{i.j}^T P_{i.j} = \prod_{\{i.j\} \in F} p(x_{m_f}) \cdot \prod_{\{i.j\} \notin F} k$$

and by straightforward algebra we have:

$$S_\delta(B_S) = \log\left(\frac{P(B_S)}{P(\{0\})}\right) = \log\left(\frac{\prod_{\{i.j\} \in F} p\left(x_{m_f}^{B_S}\right) \cdot \prod_{\{i.j\} \notin F} k}{\prod_{\{i.j\} \in F} p\left(x_{m_f}^{\{0\}}\right) \cdot \prod_{\{i.j\} \notin F} k}\right) =$$

$$= \log\left(\frac{\prod_{\{i.j\} \in F} p\left(x_{m_f}^{B_S}\right)}{\prod_{\{i.j\} \in F} p\left(x_{m_f}^{\{0\}}\right)}\right)$$

with constants k cancelled out. It follows that the number of operations to calculate $S_\delta(B_s)$ is equal to $2F + 2$.

4.2. Encoding global features

Partial prior beliefs regarding network topology may take the form of an expected degree of connectivity, for example if the expert has clues about the expected number of parents/children per node. In gene expression analysis, the regulation of one gene is expected to depend on a few other genes, although cases of regulation over many different metabolic pathways are known. The score component $S_p^\tau(B_S)$ captures this class of beliefs about the topology of a candidate network.

In a constructional approach the topology of an n-node network B_s is encoded into a $n \times n$ connectivity matrix C_s, (Larrañaga and Poza, 1996), whose element i, j is 1 if $v_i \in pa(v_i)$, zero otherwise. The matrix C_s is one-to-one with E, therefore it contains the entire structural information. Variables x_{gf} (B_s),

$f = 1, 2, \ldots$ are built to capture global network features, such as the mean cardinality of parent sets, the DAG size, the number of v-structures appearing on a directed path, and the size of a directed path *dp* ending in a node which belongs to the maximal directed path dp_{max}.

We consider here variables $\{x_{g1}, \ldots, x_{gn}\}$ defined to count the number of parents for each $v_i \in V$:

$$x_{g_i} = \sum_j C_{i.j} = \sum_{v_i \in V} | pa(v_i) |. \tag{8}$$

Further variables $x_{gn+1}, \ldots, x_{g2n}$ count the number of children in ch_{vi} for each $v_i \in V$:

$$x_{g_{n+1}} = \sum_i C_{i.j} = \sum_{v_i \in V} | ch(v_i) | \tag{9}$$

The approach adopted here to depict prior beliefs about network topology is based on a reference distribution Q_{pa} representing expert's belief about the fraction of total nodes bearing a given number of parents, (0, 1, . . .) and on the distribution $P_{pa,s}$ of relative frequencies calculated on the candidate network. The support of P_{pa} is $\chi = \{0, 1, 2, \ldots, n-1\}$. Whenever elicitation of the probability distribution on the canonical sample space of the auxiliary variable x_{gf} is beyond the expert's ability, a partitioning of χ into a coarser grid of values is performed before elicitation.

The distribution $P_{pa,s}$ is compared to Q_{pa} and the degree of dissimilarity enters in the score function. The *Kullback-Leibler* divergence is here adopted to assess the degree of dissimilarity among the above distributions:

$$KL(P_{pa} \| Q_{pa}) = \sum_x P_{pa}(x) \log\left(\frac{P_{pa}(x)}{Q_{pa}(x)}\right) \tag{10}$$

Note that the *Kullback-Leiber* divergence is not symmetrical and is equal to 0 if and only if $Q_{pa} \equiv P_{pa}$. A small value of the distance KL means that the candidate network has a structure close to the *a priori* belief as regards the connectivity.

The score component $S_\tau(B_s)$ is defined as a function of the Kullback-Leibler divergence:

$$S_\tau(B_S) = \left(-KL(P_{pa} || Q_{pa})\right) \tag{11}$$

Given P_{pa} and Q_{pa} and j being the number of elements in the partition, the computation of $S_\tau(B_s)$ takes $3j + 1$ operations.

4.3. Score function and calibration

We propose a score function defined by the convex combination of equations (7) and (11):

$$S_{prior}(B_S) = \alpha S_p^{\delta}(B_S) + (1-\alpha) S_p^{\tau}(B_S) \tag{12}$$

with $0 \leq \alpha \leq 1$. By substitution, we have:

$$S_{prior}(B_S) = \alpha \, log\left(\frac{P(B_S)}{P\{0\}}\right) + (1-\alpha)\left(-KL\left(P_{pa} \,||\, Q_{pa}\right)\right) \tag{13}$$

The role of α is to balance the strength of the components due to edge orientation and the strength due to network topology. A value $\alpha = 1$ is suited to the lack of specific prior beliefs on network topology. Without data the best *a priori* structure maximizes (13), which is conveniently reformulated as:

$$S_{prior}(B_S) = log\left[\left(\frac{P(B_S)}{P\{0\}}\right)^{\alpha} \cdot e^{(1-\alpha)\left(-KL\left(P_{pa}||Q_{pa}\right)\right)}\right] \tag{14}$$

5. The P-metric

Structural learning of BNs may be performed using the score function (14) in a Bayesian-inspired metric, called *P-metric*, which mixes prior beliefs and experimental information following Heckerman et al. (1994). The BDe metric is peculiar in assigning the same likelihood value to structures which are likelihood equivalent, i.e. DAGs encoding the same assertions on conditional independence relations. The equivalence is obtained by estimating the parameters through a prior procedure in which Dirichlet hyperparameters are defined using the notion of equivalent sample size.

The BDe function defined by Heckerman et al. (1994) may be used in both causal and acausal networks.

In order to work with acausal networks, the score equivalence condition must be fulfilled. Nevertheless, a prior equivalent score is needed to obtain

a score equivalent metric. Neither the prior function proposed in Heckerman et al. (1994) nor $S_p(B_s)$ are prior equivalent functions, therefore the proposed P-metric is better recommended for learning causal Bayesian networks.

The P-metric inherits from the BDe function all the assumptions described in Heckerman et al. (1994): (1) the database of cases D is a multinomial sample from a Bayesian Network with parameters θ; (2) missing data are not allowed; (3) the structure B_s defines the number of CPTs needed, each CPT with its own parameter θ; (4) parameters for each CPT are independent; (5) given two networks B_1 and B_2 with $p(B_1 \mid \xi) > 0$ and $p(B_2 \mid \xi) > 0$, if they are equivalent, then they have the same likelihood value; as shown in (11), these five assumptions imply that the prior distribution over parameters of each CPT is Dirichlet (Cooper and Herskovitz, 1992).

We propose the P-metric below to assess the score of a candidate structure B_s, given a complete database of cases D:

$$S_{P.metric}(B_S) = S_p(B_S)^{\beta_z} \cdot P_{BDe}(D \mid B_S, \theta) \tag{15}$$

and on the log scale it may be rewritten as:

$$\log(S_{P.metric}(B_S)) = \beta_z \cdot \log(S_p(B_S)) + \log(P_{BDe}(D \mid B_S, \theta)) \tag{16}$$

The parameter β_z is introduced to calibrate the score so that the strength of the *a priori* component is balanced against the contribution of the BDe component. The numerical choice of β_z is related to the size of node set V, to the sample size of cases in the database but also to the strength of the elicited beliefs. We propose to define an omnibus-default value for β_z that is based on indirect assessment of the aforementioned relations by making β_z depend on a function of the prior score and data likelihood of the empty DAG through a user-selected value z which sets the relevance of prior belief:

$$\beta_z = z \cdot \frac{\log(P_{BDe}(D \mid \{0\}, \theta))}{\log(S_p(\{0\}))} \tag{17}$$

with $0 \le z \le 1$. Clearly if $z = 0$ then $\beta_z = 0$ and the P-metric is equal to the BDe metric when uniform prior distribution over structures is assumed. The role of z is to set the importance of the prior score with respect to the BDe likelihood function.

Normalized prior would set a probabilistically coherent calibration in which the likelihood function reshapes prior beliefs.

The P-metric makes it easy to quantify beliefs taking the form of both global network features and (marginal) causal assertions concerning pairs of variables. The joint use of the prior score $S_p(B_s)$ and of the BDe likelihood enables the detection of score differences in causally distinct structures, even if they would be collapsed into the same equivalence class by using a uniform prior distribution over structures. As shown in section 3, although several methods are available to define prior distributions on structures (Buntine, 1991; Heckerman et al. ,1994), $S_p(B_s)$ makes the elicitation easy even in large networks.

Numerical investigations in benchmark case studies suggest that the P-metric is a valuable tool for large and structured domains, like gene expression studies. Note that the proposed approach is one step beyond the use of hard constraints, which may cause a loss of information and even a biased elicitation.

6. Results

We implemented the P-metric on top of the MASTINO package (Mascherini, 2006), coded in the *R* environment (Ihaka and Gentleman, 1996), and built on the top of the library DEAL (Bøttcher and Dethlefsen, 2003). MASTINO is a suite of *R* functions, which includes several algorithms to learn Bayesian Networks.

The package MASTINO can be freely downloaded from the website http://statind.jrc.it/mastino.

We numerically investigated the P-metric by means of two benchmark datasets which are often referred to in the machine learning literature. One is the famous ASIA network (Lauritzen and Spiegehalter, 1988) and the other is a subnetwork from the Hepatic Glucose Homeostasis network (Le et a.,2004). They are both discrete networks of respectively 8 and 20 variables.

We used the Iterated Hill Climbing with Random Restarts (IHC) (Chickering et al., 1995) as heuristic search strategy and we ran the learning algorithm over three different samples of 500, 1500, 3000 observations. We tested the P-metric for different combinations of parameters $z \in \beta_z$ and α. Finally, we compared our approach against three other algorithms: the PC and

NPC algorithm (Spirtes et al., 2000; Steck, 2001), implemented in HUGIN, and with the BDe metric (Heckerman et al. ,1994), still using the ICH as search algorithm.

6.1. The ASIA network

Asia is a small fictitious Bayesian network (Lauritzen and Spiegehalter, 1988), to calculate the probability of a patient having tuberculosis, lung cancer or bronchitis given values taken by some other variables, like "visit-to-Asia", which is equal to 1 if the patient recently visited Asia.

All variables in this network are binary. The ASIA network is implemented in the software HUGIN (Andreassen et al., 1989), which is also used to generate the database of cases. The problem domain here is quite rich: shortness-of-breath, dyspnoea (D), may be due to different factors, i.e. tuberculosis (T), lung cancer (L), bronchitis (B). Then a recent visit to Asia (A) increases the risk of tuberculosis, while smoking (S) is known to be a risk factor for both lung cancer and bronchitis. Results of a single chest x-ray (X) do not discriminate between lung cancer and tuberculosis (E), and neither does the presence or absence of dyspnoea.

The above prior information was supposed to be partially quantified by experts concerning three pairs of nodes: (*A*, *T*), (*S*, *L*) and (*L*, *T*). In particular, in the adopted expert domain, the node "Tuberculosis" (T) was not reputed to have any effect on "Visiting Asia" (A), so the probability of the event $A \leftarrow T$ was set to be equal to 0.01; then, "Smoking" (S) was believed to have an effect on "Lung Cancer" (L) but "Lung Cancer" did not have any effect on "Smoking". The probability of those events was set to $P(S \rightarrow L) = 0.6$ and $P(S \leftarrow L) = 0.01$ respectively. Finally, no effects between "Lung Cancer" and "Tuberculosis" (T) were believed to exist, so the probability of the event $L \Leftrightarrow T$ was set equal to 0.8. The uniform distribution was used to complete the probability vectors referred to the pairs of nodes listed above. As regards the network topology, it was set that 80% of network nodes have at most one parent.

The elicited prior belief was used to build an instance of the P-metric. The structural learning algorithm was repeatedly run for three different sample sizes, respectively of 500, 1500 and 3000 cases. Reversed arcs entailing the same equivalent structures were considered as correct answers.

In order to assess the sensitivity of the P-metric to the input parameters, the algorithm's behaviour was evaluated with different combinations of parameter values for z and α. The parameter z was set to take values from the grid $\{0.1, 0.25, 0.5\}$, where the higher the value of z, the stronger is the role of the prior information in the proposed metric. Furthermore parameter α takes values from the set $\{0.25, 0.5, 0.75\}$, where the higher the value of α, the stronger is the effect of prior information on local features. When $\alpha = 1$ the effect of prior information on global features is null. For each sample we performed the learning process for all possible combinations of z and α.

Results of sensitivity analysis on the calibrating parameters are shown in Table 1, in which the robustness of the P-metric is evident, and it suggests that when the sample size is increased the algorithm would find the best network even with a smaller contribution of prior information, i.e. smaller values of z. Furthermore, for each value of z better performances are reached for higher values of α, implying that the contribution of prior information on local features is apparently more important than the contribution of global features. Given the small size of the network this behaviour was expected.

The comparison between the ASIA network and those learned by means of the P-metric, the PC algorithm, the NPC algorithm and the BDe score was performed in terms of number of correctly/incorrectly learned arcs. The comparison of P-metric with other algorithms is shown in Table 2. For the P-metric we reported the worst and the best performance obtained under different configurations of the calibrating parameters.

Overall the P-metric performed very well by comparison with other well-known learning algorithms. Our results suggest a general utility of the search for optimal structures based on our P-metric. In fact, it is important to notice that for all the considered samples the overall performances of the P-metric are better than those of all the other algorithms. At sample size of 500 observations, the worst performance of our score function is equal to the performances of other algorithms as regards the number of correct/missing arcs. Nevertheless the P-metric can be anyway considered superior, since it does not add incorrect arcs to the optimal networks. At other sample sizes, the worst performance of the P-metric is also better than the performances of other algorithms which maintained similar performances at each sample size. In particular, for the smallest sample size, in the best case the P-metric discovered 6 arcs out of 8

(4 arcs in the worst case), compared with 4 arcs for the BDe, NPC and PC algorithms. With a sample size of 1500 observations, the P-metric correctly identified 7 arcs in the best case (6 arcs in the worst case) compared with 5 arcs

Table 1. ASIA network: P-metric performances under different combinations of parameters.

Sample	z	α	Correct	Incorrect	Missing
500	0.10	0.25	4	0	4
		0.50	4	0	4
		0.75	5	0	3
		1	5	0	3
	0.25	0.25	4	0	4
		0.50	4	0	4
		0.75	5	0	3
		1	6	0	2
	0.50	0.25	4	0	4
		0.50	5	0	3
		0.75	6	0	2
		1	6	0	2
1500	0.10	0.25	5	0	3
		0.50	5	0	3
		0.75	6	0	2
		1	7	0	1
	0.25	0.25	5	0	3
		0.50	6	0	2
		0.75	6	0	2
		1	7	0	1
	0.50	0.25	6	0	2
		0.50	6	0	2
		0.75	7	0	1
		1	7	0	1
3000	0.10	0.25	5	0	3
		0.50	6	0	2
		0.75	6	0	2
		1	7	0	1
	0.25	0.25	5	0	3
		0.50	6	0	2
		0.75	7	0	1
		1	7	0	1
	0.50	0.25	6	0	2
		0.50	6	0	2
		0.75	7	0	1
		1	7	0	1

Table 2. Comparison of the algorithms' performances for the ASIA network, where P-metric[1] and P-metric[2] represent the worst and best P-metric performance respectively .

Sample	Algorithm	Correct Arcs	Missing	Incorrect Added
500	PC	4	4	2
	NPC	4	4	1
	Bde	4	4	1
	P-metric[1]	4	4	0
	P-metric[2]	6	2	0
1500	PC	5	3	0
	NPC	5	3	1
	Bde	5	3	2
	P-metric[1]	6	2	0
	P-metric[2]	7	1	0
3000	PC	5	3	1
	NPC	5	3	2
	Bde	6	2	2
	P-metric[1]	6	2	0
	P-metric[2]	7	1	0

for the other algorithms. Finally, for the largest sample size, the P-metric again discovered 7 arcs (6 arcs in the worst case) compared with 6 arcs for BDe and 5 arcs for the PC and NPC algorithms.

Although the performances of all the algorithms improved with increasing sample size, it is important to emphasize the robustness shown by the P-metric, which – whatever the sample size and the combination of calibrating parameters – always obtained better networks than those built by the PC, NPC and BDe algorithms.

6.2. The Hepatic Glucose Homeostasis network: A case study in functional genomics

The performances of the P-metric were assessed by learning the structure of the Hepatic Glucose Homeostasis network (HGH) (Le, 2004). The HGH depicts a model for the genetic network controlling glucose metabolism in perinatal hepatocyte, where specific focus is placed on the effects of insulin, glucagon and glucocorticoid hormones. In addition, several transcription factors known to be important in controlling the expression of key genes are also thoroughly incorporated in the model.

The interactions between the hormones signalling pathways and liver-specific transcription factors define the genetic network that controls the expression of genes maintaining glucose homeostasis in the liver. Each gene is modelled here as a node, for a total of 35 nodes in the network. In the original HGH network a directed edge from a parent node to a child is added to the network when a published resource indicates that the parent gene has a direct effect on the transcription process of the child gene. In the HGH network a total of 52 modelled regulatory interactions are added. In Le at al. (2004) the data are randomly generated using the HGH network, as it would be obtained from experiments involving microarrays.

The HGH network is formed by 20 genes and 33 regulatory interactions, because this is the size of the problem domain of our major interest, and to keep the computational burden to a reasonable size. The adoption of a simplified version of the HGH network is also justified by the limits imposed by the implementation of multidimensional arrays in R, currently quite limited, on which the package MASTINO is based.

Prior information takes the form of a plausible partial order on a few variables and high levels of network sparsity. Formally, we place high plausibility on the event that insulin, glucagon and glucocorticoid hormones (respectively IPA, CPA and GPA) preceded AC3, G6P, IP1 and TAT. The probability of the event $\{IPA,CPA,GPA\} \rightarrow \{AC3,G6P, IP1, TAT\}$ was set equal to 0.50 for each pair of nodes. As regards the network topology of the HGH network, we quantified our belief about sparsity setting the cardinality of $pa(v_i)$ of each node $v_i \in V$ to a fairly small value, more precisely 80% of nodes are expected to receive less than 3 incoming arrows.

We tested the P-metric at 3 different sample sizes, 500, 150 and 3000, using different combinations of parameters z and α. Data were simulated using the software HUGIN (Andreassen et al., 1989), following the same approach of Le et al. (2004), in which data were simulated using the BNet toolbox (Murphy, 2001). Analysis of the algorithm's sensitivity to choice of parameter values was performed by running the algorithm on the aforementioned grid of values for z and α described in the ASIA case study. For each sample size we performed the learning process for all possible combinations of z and α.

Results of the sensitivity analysis are shown in Table 3, in which the robustness of the proposed approach for the HGH network is also evident at all

sample sizes. The difference between the best and the worst network is always limited to the orientation of just one arc. Results also suggest that the best network is also found if the contribution of the prior information is set to be small, i.e. a smaller value of z, whatever the sample size. It is interesting to notice that, in contrast with the result obtained for the ASIA network, in this case study for each value of z better performances are reached with smaller values of α. This result implies that for the HGH network the contribution of prior information on network topology is more important than the contribution of local features.

The comparison of the P-metric with other learning algorithms again shows the overall good performances of the proposed metric, because at all sample sizes the performances of the P-metric were always equal to or better than those shown by other algorithms. The results in Table 4 reveal that at the smallest sample size, the performance of the proposed score is comparable with that obtained with the NPC and PC algorithms, which found 21 and 19 correct arcs respectively. The P-metric outperformed the BDe metric, which found 12 correct arcs while the P-metric found 22 corrected arcs (21 in the worst case). At a size of 1500 observations, our metric again outperforms all the other algorithms by correctly identifying 25 arcs in the best case (24 arcs in the worst case) compared with 22 and 21 arcs for NPC and PC respectively, and 18 arcs for the BDe metric. Finally, at the largest sample size, the P-metric discovered 28 arcs (27 arcs in the worst case), achieving a similar performance to PC and NPC and again outperforming BDe, which found 24 arcs.

7. Conclusions

In this paper we have defined a new Bayesian-inspired score function, called P-metric, to learn the structure of networks representing causal relations among variables. The metric component dealing with structural information takes account of marginal causal beliefs concerning arcs and global network features without requiring the elicitation of a complete network (Buntine, 1991; Heckerman et al., 1994). The likelihood component is based on the BDe metric, thus it exploits the characteristics of the latter, which are well reported in the literature.

The BDe metric does not distinguish structures entailing the same conditionnal independence assertions, but our score function makes it possible to discriminate structures belonging to the same likelihood equivalence class using

Table 3. HGH network: P-metric performances under different combinations of parameters.

Sample	z	α	Correct	Incorrect	Missing
500	0.10	0.25	22	1	11
		0.50	21	1	12
		0.75	21	1	12
		1	21	1	12
	0.25	0.25	22	1	11
		0.50	22	1	11
		0.75	21	1	12
		1	21	2	12
	0.50	0.25	22	1	11
		0.50	22	1	11
		0.75	21	1	12
		1	21	2	12
1500	0.10	0.25	25	1	8
		0.50	25	1	8
		0.75	25	1	8
		1	24	1	9
	0.25	0.25	25	1	8
		0.50	25	1	8
		0.75	24	1	9
		1	24	2	9
	0.50	0.25	25	1	8
		0.50	25	1	8
		0.75	24	1	9
		1	24	2	9
3000	0.10	0.25	28	1	5
		0.50	28	1	5
		0.75	27	1	6
		1	27	1	6
	0.25	0.25	28	1	5
		0.50	28	1	5
		0.75	27	1	6
		1	27	1	6
	0.50	0.25	28	1	5
		0.50	27	1	6
		0.75	27	1	6
		1	27	1	6

Table 4. Comparison of the algorithms' performance for the HGH network, where P-metric[1] and P-metric[2] represent the worst and best P-metric performance.

Sample	Algorithm	Correct Arcs	Missing	Incorrect Added
500	PC	19	11	3
	NPC	21	10	2
	Bde	12	21	1
	P-metric[1]	21	11	2
	P-metric[2]	22	10	1
1500	PC	21	10	3
	NPC	22	10	2
	Bde	18	15	1
	P-metric[1]	24	8	1
	P-metric[2]	25	8	0
3000	PC	27	2	5
	NPC	27	3	5
	Bde	24	9	5
	P-metric[1]	27	6	1
	P-metric[2]	28	5	1

the elicited causal information. The BDe metric may be used to learn causal networks (Heckerman et al., 1994) and our P-metric extends its flexibility.

Performances of the P-metric were tested under two different Machine Learning benchmark datasets, varying the sample size and the structural prior. The sensitivity analysis of the P-metric was performed by testing several combinations of calibrating parameters. The results were compared against three well-known learning algorithms: the PC algorithm (Spirtes et al., 2000), the NPC algorithm (Steck, 1990) and the BDe metric (Heckerman et al., 1994). Successful numerical findings prove the effectiveness of the P-metric, which achieved performances always equal to or better than those shown by other algorithms for both the benchmark BNs and at all the sample sizes. In particular, it is important to notice that our score always outperforms the BDe (i.e. the P-metric when the parameter z is set equal to 0), showing the importance of the prior information in the performance of BNs learning algorithms. Moreover, sensitivity analysis also highlights the overall robustness of the proposed metric, demonstrating the limited effect of the input parameters on the score function.

The P-metric is not highly demanding as regards the elicitation of prior information, therefore it could be very useful in large problem domains in

which substantial but partial information is available, like gene expression studies. In this paper results from the case studies have clearly shown the outstanding impact that prior information may have in improving the learning process. For the omnibus calibration we proposed, just one parameter, *z*, besides the elicited quantities needs to be set in order to obtain a working algorithm. More work is needed to improve the algorithm's calibration for specific problem domains; bootstrapping the omnibus setting might be a good start.

Search under P-metric may stop at a local maximum, like other greedy search algorithms, therefore population based algorithms (Larrañaga and Poza, 1996; , Mascherini and Stefanini, 2005, Pelikan et al., 1999), might be considered as a useful alternative reducing the probability of premature convergence. Future work might deal with the development of elicitation aids about local features on a grid of values (Jeffreys, 1961). Sensitivity analysis on the elicited values would also be useful.

REFERENCES

Andreassen S.K., Olesen K.G., Jensen F.V., Jensen F. (1989): HUGIN: a shell for building Bayesian belief universes for expert systems. Proceedings of the Eleventh International Joint Conference on Artificial Intelligence.

Bøttcher S.G., Dethlefsen C. (2003): DEAL: A package for Learning Bayesian Networks. Journal of Statistical Software 8/20: 1–40.

Buntine W.L. (1991): Theory of Refinement on Bayesian Networks. Proceedings of 7th Conference on Uncertainty in Artificial Intelligence: 52–60.

Chickering D.M., Geiger D., Heckerman D. (1994): Learning Bayesian Network: A combination of knowledge and statistical data. Technical Report MSR-TR-94-17, Microsoft Research, Advanced Technology Division.

Chickering D.M., Geiger D., Heckerman D. (1995): Learning Bayesian Networks: Search methods and Experimental Results. Preliminary papers of the 5th Intl. Workshop on Artificial Intelligence and Statistics: 112–128.

Chickering D.M. (1995): Learning Bayesian Networks is NP-Complete. Proceedings on Artificial Intelligence and Statistics: 121–130.

Cooper G.F., Herskovitz E. (1992): A Bayesian method for the induction of probabilistic networks from data. Machine Learning 9: 309–357.

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

Friedman N., Koller D. (2003): Being Bayesian about network structures: A Bayesian approach to structure discovery in Bayesian Networks. Machine Learning 50: 95-126.

Geiger D., Pearl J. (1988): Logical and algorithmic properties of conditional independence. Technical Report R97, Cognitive System Laboratory, UCLA.

Heckerman D., Geiger D., Chickering D.M. (1994): Learning Bayesian Network: A combination of knowledge and statistical data. Proceedings of 10th Conf. Uncertainty in Artificial Intelligence: 293–301.

Heckerman D., Meek C., Cooper G. (1997): A Bayesian Approach to Causal Discovery. Technical Report MSR-TR-97-05. Microsoft Corporation, Redmond, WA.

Ihaka R., Gentleman R. (1996): A Language for Data Analysis and Graphics. Journal of Computational and Graphical Statistics 5/3: 299–314.

Jeffreys H. (1961): Theory of Probability. Oxford University Press Oxford, U.K.

Jensen F.V. (1996): An introduction to Bayesian Networks. Springer-Verlag, Berlin Heidelberg New York.

Kullback S., Leibler R.A.M. (1951): On Information and Sufficiency. Annals of Mathematical Statistics 22: 79–86.

Larrañaga P., Poza M. (1996): Structure Learning of Bayesian Networks by Genetic Algorithms: A Performance Analysis of Control Parameters. IEEE Journal on Pattern Analysis and Machine Intelligence 18/9: 912–926.

Lauritzen S.L., Spiegehalter D.J. (1988): Local Computation with probabilities on graphical structures and their application to expert system. Journal of the Royal Statistical Society, B Series 50/2: 157–192.

Le P.P., Bahl A., Ungar L.H. (2004): Using prior knowledge to improve genetic network reconstruction from microarray data. In Silico Biology 27/4.

Madingan D., Raftery A.E. (1994): Model Selection and accounting for model uncertainty in graphical model using Occam's window. Journal of the American Statistical Association 24: 2271–2292.

Mascherini M., Stefanini F.M. (2005): M-GA: A genetic algorithm to learn Conditional Gaussian Bayesian Networks. Proceedings of the IEEE International Conference on Computational Intelligence for Modelling, Control and Automation, IEEE Computer Society: 61–67.

Mascherini M. (2006): MASTINO: A Suite of R Functions to learn Bayesian Networks from data. UseR! International Conference of R Users, Vienna(Austria).

Murphy K.P. (2001): The Bayes Net Toolbox for MATLAB. Computer Science and Statistics 33: 331–349.

Pearl J. (1988): Probabilistic Reasoning in Intelligent Systems. Morgan Kaufmann, San Mateo, CA.

Pelikan M., Goldberg D.E., Cantu-Paz E. (1999): BOA: The Bayesian Optimization Algorithm. Proceedings of the Genetic and Evolutionary Computation Conference: 525–532.

Spirtes P., Glymour C., Scheines R. (2000): Causation, Prediction, and Search, 2nd ed. MIT Press, New York, N.Y.

Srinivas S., Russel S., Agogino A.M. (1990): Automated construction of sparse Bayesian Networks from unstructured probabilistic models and domain information. Uncertainty in Artificial Intelligence: Proceedings of the Fifth Conference, Elsevier Science Publishing Company, New York, NY: 295–308.

Steck H. (2001): Constraint-Based structural learning in Bayesian networks using finite data sets. PhD thesis, University of Munich. Munich, Germany.