Dipartimento di Statistica
"Giuseppe Parenti"

WORKING PAPER 2005/13

Encoding structural prior information to learn large Bayesian Networks

Massimiliano Mascherini,
Federico M. Stefanini

Università degli Studi
di Firenze

Dipartimento di Statistica "G. Parenti" – Viale Morgagni 59 – 50134 Firenze - www.ds.unifi.it

Encoding structural prior information to learn large Bayesian Networks

Massimiliano Mascherini and Federico M. Stefanini
Dipartimento di Statistica 'G.Parenti'
v.le Morgagni 59, 50134 Firenze, Italy
phone: ++39 055 4237211 fax: ++39 055 4223560
mascherini@ds.unifi.it stefanini@ds.unifi.it
http://www.ds.unifi.it/

Abstract

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. Moreover, the detailed specification of prior distributions for structural learning is NP-Hard, [7], making the elicitation of large networks impractical. This is the case, for example, of gene expression analysis, in which a small degree of graph connectivity is a priori plausible and where substantial information may regard dozens against thousands of nodes.

We propose an elicitation procedure for DAGs which exploits prior knowledge on network topology, and that is suited to large Bayesian Networks. Then, we develop a new quasi-Bayesian score function, the P-metric, to perform structural learning following a score-and-search approach.

Contents

1 Introduction

The complete specification of a prior distribution on the topology of a Bayesian Network (BN) is NP-Hard [5]. 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 infeasible, due to the lack of enough information to completely specify one network. A prior state of partial knowledge about network's topology may take several forms, like a 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 network structure without requiring the a-priori complete specification of structures. Elicited beliefs are refined by means of dissimilarity measures on 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, [11]. The peculiarity of a likelihood equivalent metric is to assign 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 shortly review some basic concepts about Bayesian Networks. Section 3 contains the 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 preliminary 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 are provided. Comprehensive accounts on probabilistic networks may be found in [14] and [9].

A *graph* $\mathcal{G}$ 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 indicated as $v_i \rightarrow v_j$.

Given $(v_i, v_j) \in E$ we say that v_i and v_j are *adjacent* or *neighborhoods* of each other: v_i is said *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 *descendent* 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 *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 undirected. A *directed graph* $\mathcal{G}_{DG}$ 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 *adjacency path* or *chains*, 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)* $\mathcal{G}_D = (V, E_D)$ 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 $\mathcal{B}$ is a graph-based representation of a joint probability distribution $\mathcal{P}$ 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, sometimes labels also 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:

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

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 $\mathcal{B}_s$ 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 $\theta_s = \{\theta_{s,v_i,pa(v_i)}\}_{v_i}$.

A graph $\mathcal{G}_D$ does not always represent all the conditional independence relations entailed by the probability distribution $\mathcal{P}$. If it does, we say that $\mathcal{P}$ and $\mathcal{G}_D$ 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 encoded by a BN are revealed by assessing the *direction dependent separation* property [10], also called *d-separation*, [19].

Given a DAG $\mathcal{G}_D = (V, E)$, with $v_i, v_j \in V$, and $v_j \neq v_i$, let C be a subset of V, $C \subset V \backslash \{v_i, v_j\}$. We say that v_i and v_j are *d-separated* in $\mathcal{G}_D$ given C, if and only if there exists no adjacency path *ap* between v_i and v_j such that: (i) every collider on *ap* is in C or has a descendent in C; (ii) 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 d-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 set of nodes $X \subset V$ and $Y \subset V$ by iterating the above definition for each pair (v_i, v_j), with $v_i \in X$ and $v_j \in Y$.

3 Earlier Approaches

The elicitation problem for prior beliefs on network's structure has been not much considered in the literature. 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 infeasible out of networks with a very small set of nodes because the space of DAGs has superexponential cardinality while increasing the number of nodes in V.

A simpler approach puts a uniform prior distribution on a subset H of all possible DAGs, [12], therefore some structures are a-priori excluded from the scoring procedure. Bounds on the number of parents/children are established to set hard constraints on elements in H.

Two more elaborated approaches have been proposed by [4, 6] 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, [4], 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:

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

In the approach proposed by Heckerman, [6], the expert builds a complete a-priori network, $\mathcal{B}_{sc}$, (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, $p(U \mid \mathcal{B}_{sc}, \xi)$, where $\mathcal{B}_{sc}$ 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, $\mathcal{B}_{sc}$, but, in their approach, structures closely resembling to the prior network receive a high prior probability, otherwise they are penalized. The number of nodes in the symmetric difference of $\pi_i(B_s)$ and $\pi_i(\mathcal{B}_{sc})$ is:

$$\delta_i = |\{\pi_i(B_s) \cup \pi_i(\mathcal{B}_{sc}))\} \setminus \{\pi_i(B_s) \cap \pi_i(\mathcal{B}_{sc})\}| \tag{4}$$

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

$$p(B_s \mid \xi, \mathcal{B}_{sc}) = c \cdot k^{\delta} \tag{5}$$

where c is a normalization constant.

4 From prior information to score functions

The specification of a complete prior network with beliefs over all possible edges is quite infeasible 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 [4], but it becomes very difficult due to the superexponential cardinality of the space of structures for an increasing number of nodes. In large networks, a coherent and complete specification of a prior distribution on the space of networks [6] seems very difficult.

In this section a score function, $S_{prior}(B_s)$, 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 [4, 6].

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 (thousands), 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 regard the partial order of ten against thousands 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 a score function $S_{prior}(B_s)$ capturing local and global network features. The score component $S_p^{\delta}(B_s)$ refers to edges elicited one at a time. The second score component, $S_p^{\tau}(B_s)$, describes global network features, related to DAG connectivity.

4.1 Encoding local features

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

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 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 $\mathcal{M}$ of $F \leq n(n-1)/2$ variables $\mathcal{M} = \{m_1, \ldots m_f, \ldots, m_F\}$ each one taking values on $\chi = \{-1, 0, 1\}$ for each pair of nodes $(v_i, v_j), i < j$, in V. Values in the range χ respectively indicate: an arrow $i \leftarrow j$, no arrow, an arrow $i \rightarrow j$. Expert's belief takes the form of a set of probability distributions $\{p(x_{m_f} \mid \xi) : m_f \in \mathcal{M}\}$.

The distributions are now coded as vectors of probability values

$P_{i\cdot j}^T = (p_{i\cdot j,-1}, p_{i\cdot j,0}, p_{i\cdot j,+1})$ so that $\mathbf{1}^T P_{i\cdot j} = 1$. Connectivity vectors $C_{i\cdot j}$ are introduced to indicate the value taken by variables. It follows that $\mathbf{1}^T C_{i\cdot j} = 1$. The probability value associated to the oriented edge for a pair $i \cdot j$ is $C_{i\cdot j}^T P_{i\cdot j}$.

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

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

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

The space of DAGs is contained in the space of Directed Graphs, $\mathcal{G}_D \subseteq \mathcal{G}_{DG}$, therefore the above construction also induces a probability distribution over DAGs contained in the space of directed graphs, $B_s \in \mathcal{G}_{DG}$:

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

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

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

$$S_\delta(B_s) = \log\left(\frac{P(B_s)}{P(\{\emptyset\})}\right) \tag{7}$$

with $P(\{\emptyset\})$ the probability assigned to the Bayesian Network in which E is empty (graphs without edges). By straightforward algebra it may be shown that the computation of the normalization constant c is not needed in order to search in the space of networks, (Appendix 1). If the expert's belief leaves some edges unspecified, than the elicitation is completed using uniform distributions.

A remarkable property of the score $S_\delta(B_s)$ in equation 7 regards the possibility of calculating scores by just considering the pair of nodes for which the expert defined a distribution:

$$S_\delta(B_s) = \log\left(\frac{P(B_s)}{P(\{\emptyset\})}\right) = \log\left(\frac{\prod_f p(x_{m_f}^{B_s})}{\prod_f p(x_{m_f}^{\{\emptyset\}})}\right) \tag{8}$$

(see Appendix 2 for details).

It follows that the number of operations to calculate $S_\delta(B_s)$ is equal to $2 \cdot F + 2$.

4.2 Encoding global features

Partial prior beliefs on 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 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 a n-nodes network B_s is encoded into a $n \times n$ connectivity matrix C_s, [15], whose element i, j is one iff $v_i \in pa(v_j)$, zero otherwise. Matrix C_s is one-to-one with E, therefore it contains the whole structural information. Variables $x_{g_f}(B_s), f = 1, 2, \ldots$ are built to capture global network features like: the mean cardinality of parent sets, the DAG size, the number of v-structures appearing on a directed path, the size of a directed path dp ending into a node which belongs to the maximal directed path dp_{max}.

We consider here variables $\{x_{g_1}, \ldots, x_{g_n}\}$ defined to count the number of parents for each $v_i \in V$:

$$x_{g_i} = \sum_j C_{i,j} = \sum_{v_i \in V} \mid pa(v_i) \mid \tag{9}$$

Further variables $x_{g_{n+1}}, \ldots, x_{g_{2n}}$ count the number of children in ch_{v_i} for each $v_i \in V$:

$$x_{g_{n+i}} = \sum_i C_{i,j} = \sum_{v_i \in V} \mid ch(v_i) \mid \tag{10}$$

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

The distribution $\mathcal{P}_{pa,s}$ is compared to $\mathcal{Q}_{pa}$ 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:

$$KL(\mathcal{P}_{pa} \| \mathcal{Q}_{pa}) = \sum_x \mathcal{P}_{pa}(x) log \left(\frac{\mathcal{P}_{pa}(x)}{\mathcal{Q}_{pa}(x)} \right) \tag{11}$$

Note that the *Kullback-Leiber* divergence is not symmetrical and is equal to 0 if and only if $\mathcal{Q}_{pa} \equiv \mathcal{P}_{pa}$. A small value of KL distance 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:

$$S_\tau(B_s) = (-KL(\mathcal{P}_{pa} \| \mathcal{Q}_{pa})) \tag{12}$$

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

4.3 Score function and calibration

Given the quantities in equations 8 and 12, the proposed score function is a convex combination of two other functions:

$$S_{prior}(B_s) = \alpha S_p^\delta(B_s) + (1 - \alpha) S_p^\tau(B_s) \tag{13}$$

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

$$S_{prior}(B_s) = \alpha \log \left(\frac{P(B_s)}{P(\{\emptyset\})} \right) + (1 - \alpha) \left(-KL(\mathcal{P}_{pa} \| \mathcal{Q}_{pa}) \right) \tag{14}$$

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.

The most a-priori probable structure is the structure that maximizes (14). The logarithmic score is convenient for computational reasons:

$$S_{prior}(B_s) = \log \left[\left(\frac{P(B_s)}{P(\{\emptyset\})} \right)^\alpha \cdot e^{(1-\alpha)(-KL(\mathcal{P}_{pa} \| \mathcal{Q}_{pa}))} \right] \tag{15}$$

The numerical behavior of $S_{prior}(B_s)$ under two different parameter values is shown in Figures 1 and 2.

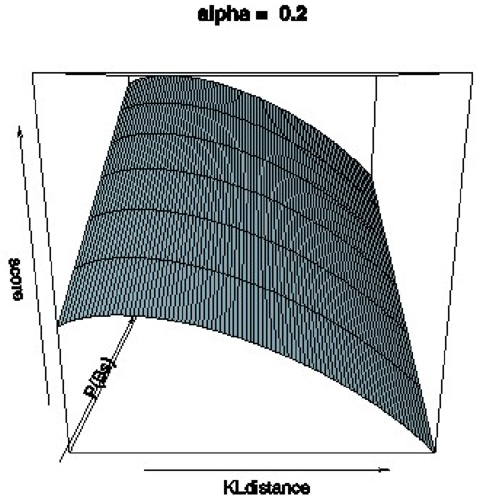

Figure 1: The score prior function $S_{prior}(B_s)$ for $\alpha = 0.2$.

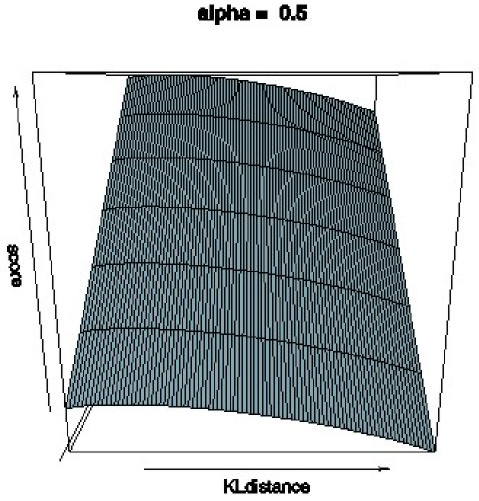

Figure 2: The score function $S_{prior}(B_s)$ for $\alpha = 0.5$.

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 [6]. 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 in [6] may be used both in 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 [6] nor $S_p(B_s)$ are prior equivalent functions, therefore the proposed P-metric can be only used for causal Bayesian networks.

Using the BDe function, the P-metric inherits all the assumptions described in [6]: (1) the database of cases $\mathcal{D}$ 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 [6], these five assumptions imply that the prior distribution over parameters of each CPT is Dirichlet [8].

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

$$S_{P\text{-}metric}(B_s) = S_p(B_s)^{\beta_z} \cdot P_{BDe}(D \mid B_s, \theta) \tag{16}$$

that may be rewritten as:

$$log\left(S_{P\text{-}metric}(B_s)\right) = \beta_z \cdot \log(S_p(B_s)) + ll_{BDe}(D \mid B_s, \theta) \tag{17}$$

The role of the parameter β_z is to calibrate the strength of the prior score with respect to the likelihood function. The value of β_z depends on the size of the problem domain and on the sample size of cases as well as on the elicited belief. Even if heuristics to set β_z are still under investigation, here we propose to set β_z as a function of the score prior and the likelihood computed for the empty structure:

$$\beta_z = z \cdot \frac{ll_{BDe}(D \mid \{\emptyset\}, \theta)}{\log\left(S_p(\{\emptyset\})\right)}$$

with $0 \leq z \leq 1$. Clearly when $z = 0$ then $\beta_z = 0$ and the P-metric is equal to the BDe metric when uniform prior distribution over structures is assumed.

The P-metric makes easy to quantify beliefs taking the form of both global network features and (marginal) causal assertions on pairs of variables. The joint use of the score prior $S_p(B_s)$ 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, [4, 6], $S_p(B_s)$ makes the elicitation easy even in large networks.

Numerical explorations on 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 biased elicitation.

6 Results

We implement the P-metric on top of package DEAL, [3], coded in the R environment, [13]. DEAL is a software package which includes several methods for analyzing data using Bayesian Networks and conditionally Gaussian networks (CG-BNs), [2]. We numerically investigated the proposed metric by means of two benchmark datasets which are often referred to in the machine learning literature. One is the famous ASIA network, proposed by [16] and the other is a subnetwork from the Hepatic Glucose Homeostasis network proposed by [17]. These are two discrete networks, which handle respectively 8 and 20 variables. We run the learning algorithm over three different sample of: 500, 1500, 3000 observations and we test the P-metric for different combinations of parameters $z \in \beta_z$ and α.

6.1 The ASIA network

Asia is a small fictitious Bayesian network, [16], 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 one if the patient recently visited Asia. All variables in this network are binary. The ASIA network is implemented in the software HUGIN, [1], which is also used to generate the database of cases. The problem

$P(A \rightarrow [S, L, X]) = 0.01$	$P([all] \rightarrow A) = 0.01$
$P(S \rightarrow [B, L]) = 0.6$	$P(S \cdots [E, X]) = 0.98$
$P(B \rightarrow D) = 0.6$	$P(B \rightarrow [L, X]) = 0.01$
$P(D \rightarrow [E, L, T, X]) = 0.01$	$P(X \rightarrow [L, T, D]) = 0.01$
$P([L, T] \rightarrow E) = 0.98$	$P(E \rightarrow S) = 0.01$

Table 1: Expert's domain for the ASIA network.

domain is here 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), as neither does the presence or absence of dyspnoea.

The above prior information is supposed to be partially quantified by experts as listed in the expert domain of Table 1.

In the adopted expert domain, the node ”Visiting Asia” (A) is defined as root and is not reputed to change Smoking habits; Lung Cancer and X-ray, as well as Smoking, are believed to have an effect on Bronchitis and Lung Cancer. Bronchitis (B) is supposed to have an effect on Dyspnoea (D) and no effect on Lung Cancer and X-ray. Dyspnoea (D) is believed to have no effect on variables E,L,T,X; X can not provoke Lung Cancer, Tuberculoses and Dyspnoea. Variables L and T have an effect on E by construction; E does not have any effect on Smoking. As regards the network topology, we believe that 80% of network nodes has at most one parent.

We repeated the learning process under three different sample sizes, respectively of 500, 1500 and 3000 cases. We also evaluated the algorithm's behavior with different combinations of parameter values for z and α. The comparison among the actual network and those learned by means of the P-metric and the DEAL score has been performed in terms of number of correctly/incorrectly learned arcs. Results of the learned network using the BDe metric implemented in DEAL are shown on Table 2, furthermore results of the performance of the P-metric are shown in Tables 7,8,9. The P-metric seems to improve the overall performance of the BDe metric implemented in DEAL. In all the samples, the best network found by the P-metric correctly identifies all the arcs of the ASIA networks and one incorrect arc is added; in the best case with DEAL, just two arcs are correctly identified, six arcs are identified but with wrong orientation, and nineteen incorrect

sample	Tot. Arcs	Cor.Arc	Wr.Dir.	I.Ad.	I.Mis.
500	27	2/8	6	19	0
1500	26	1/8	7	18	0
3000	26	1/8	7	18	0

Table 2: The ASIA network, [16], learned by DEAL.

arcs are added.

Results about the calibrating parameters suggest that by increasing the sample size the best network is obtained even with smaller values of z. Small values of α seems to improve the overall performance of the search.

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

We test the P-metric with the Hepatic Glucose Homeostasis network (HGH) in [17]. The HGH depicts a model for the genetic network controlling glucose metabolism in perinatal hepatocytes, where specific focus is placed on the effects of insuline, 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 signaling 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 here modelled 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 into 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 [17], the data are randomly generated using the HGH network, as it would be obtained from experiments involving microarrays.

In order to re-construct the HGH genetic network using the proposed P-metric, we considered a sparse structure in which the cardinality of $pa(v_i)$ is small for each $v_i \in V$.

For computational reasons we consider here a reduced version of the HGH network composed by 20 genes and 33 regulatory interactions. Prior information take the from of a partial order among few variables and high structural sparsity.

Formally, we assume that insuline, glucagon and glucocorticoid hor-

sample	Tot. Arcs	Cor.Arc	Wr.Dir.	I.Ad.	I.Mis.
500*	48	1/33	18	29	14
1500**	38	1/33	18	19	14
3000***	19	0/33	7	12	26

Table 3: The HGH network, [17], learned by DEAL (out of memory error invoked after 49(*),40(**) and 19(***) iterations).

mones, (respectively IPA, CPA and GPA) precede AC3, G6P, IP1, TAT, PEP, G6T, IP1 and that the 80% of nodes have less than 2 incoming arrows.

The adoption of a simplified version of the HGH network is justified by the computational problems arisen with the R environment, [13]. The hardware running R is an IBM e-Server Type 325 8835-51X, a dual processors computer equipped with 2xAMD Opteron 2.0GHz (1MB L2 Cache) with 5giga RAMs and the operative system is Red Hat Enterpriser Linux AS Ver.4. Two reasons forced towards the reduction of the original network: first of all, the way in which R manages multidimensional arrays and the way in which networks are coded by DEAL, which limits the number of nodes up to 27. Secondarily, the way DEAL and R manage memory, which cause "Out of memory" messages during the learning process. Running the learning process under the DEAL package, without implementing the P-metric, the out of memory error appeared after few iterations using 27 variables. We reduced the number of variables to 20 and the sample size was limited to 3000 cases. Despite the above limitations, we were able to test the P-metric.

We tested the P-metric with 3 different samples of cardinality 500, 150 and 3000 using different combinations of parameters z and α. Here data were generated using the software HUGIN, [1], as well as in [17] data were simulated using BNet toolbox,[18]. Results were compared to those from the BDe metric implemented in DEAL, where a uniform distribution over structures is assumed.

Results of the learned process under the BDe implemented in DEAL are shown on Table 3. Even if the search of the best BN using DEAL is stopped after almost 50 iterations due to the "Out of Memory" message, it is clear that our metric performed quite well. The use of prior information indeed improved the performance of structural learning. The small number of correct arcs found by DEAL in the best case is probably due to the imaginary sample size that here was automatically set high according to the large number of variables in

		sample = 500				
z	α	Tot. Arcs	Cor.Arc	Wr.Dir.	I.Ad.	I.Mis.
0.05	0.2	37	23/33	1	13	9
0.05	0.5	50	23/33	1	26	9
0.05	0.8	50	25/33	1	24	7
0.10	0.2	37	23/33	1	13	9
0.10	0.5	37	23/33	1	13	9
0.10	0.8	37	23/33	1	13	9
0.20	0.2	37	23/33	1	13	9
0.20	0.5	37	23/33	1	13	9
0.20	0.8	37	23/33	1	13	9
0.50	0.2	37	23/33	1	13	9
0.50	0.5	37	23/33	1	13	9
0.50	0.8	37	23/33	1	13	9

Table 4: The HGH network, [17], learned by the P-metric and a sample of size 500.

		sample = 1500				
z	α	Tot. Arcs	Cor.Arc	Wr.Dir.	I.Ad.	I.Mis.
0.05	0.2	41	22/33	1	18	10
0.05	0.5	45	22/33	1	22	10
0.05	0.8	40	19/33	1	20	13
0.10	0.2	39	23/33	1	15	9
0.10	0.5	40	21/33	1	18	11
0.10	0.8	40	19/33	1	20	13
0.20	0.2	39	23/33	1	15	9
0.20	0.5	39	23/33	1	15	9
0.20	0.8	40	23/33	1	16	9
0.50	0.2	39	23/33	1	15	9
0.50	0.5	39	23/33	1	15	9
0.50	0.8	39	23/33	1	15	9

Table 5: The HGH network, [17], learned by P-metric and a sample size equal to 1500 observations.

		sample = 3000				
z	α	Tot. Arcs	Cor.Arc	Wr.Dir.	I.Ad.	I.Mis.
0.05	0.2	37	23/33	1	11	9
0.05	0.5	37	21/33	1	13	11
0.05	0.8	35	19/33	1	15	13
0.10	0.2	35	23/33	1	11	9
0.10	0.5	35	23/33	1	11	9
0.10	0.8	35	23/33	1	11	9
0.20	0.2	35	23/33	1	11	9
0.20	0.5	35	23/33	1	11	9
0.20	0.8	35	23/33	1	11	9
0.50	0.2	35	23/33	1	11	9
0.50	0.5	35	23/33	1	11	9
0.50	0.8	35	23/33	1	11	9

Table 6: The HGH network, [17], learned by P-metric with a sample size of 3000 observations.

the network.

This problem may cause an almost constant BDe score even for quite different networks. Unfortunately the use of larger samples of cases was infeasible due to the computational burden. The original HGH network and the learned network are shown in Figures 6 and 7.

7 Conclusion

In this paper we defined a new quasi-bayesian score function, called P-metric, to score networks representing causal relations among variables. The metric component dealing with structural information takes account of marginal causal beliefs on arcs and global network features without requiring the elicitation of a complete network, [4, 11]. The second component is based on the BDe metric, thus it exploits its peculiarities well known in the literature.

The BDe metric does not distinguish structures entailing the same conditional independence assertions, but our score function makes possible to discriminate structures belonging to the same likelihood equivalence class at the price of loosing score equivalence property: the P-metric is suited to learn causal networks, [11].

The P-metric has been tested under two different Machine Learning benchmark datasets and compared against the metric implemented in the DEAL package. Successful numerical findings suggest that the P-metric could be very useful in large problem domains with associated substantial and partial information. Unfortunately, computational constraints forbade wide numerical testing in large networks using the R environment. Further code improvement is needed, especially an implementation under C++ or Java, in order to perform extensive numerical testings including the analysis of calibration parameters with large networks.

Acknowledgements

We are pleased to thank Finn V. Jensen, Thomas D. Nielsen and all the Machine Intelligence Group of the Computer Science Department at Aalborg University, Denmark, for their hospitality and for inspiring discussion with their many valuable comments about the idea underlying this paper.

<table>
<tr><th></th><th></th><th>sample = 500</th><th></th><th></th><th></th><th></th></tr>
<tr><th>z</th><th>α</th><th>Learned Network</th><th>Correct</th><th>Inc. Add.</th><th>Inc. Mis.</th><th>Wr.Dir.</th></tr>
<tr><td>0.05</td><td>0.2</td><td>[B|S:E][D|B:A:S:E][A][S][T|A][L|S][E|T:L][X|T:E]</td><td>8/8</td><td>4</td><td>0</td><td>0</td></tr>
<tr><td>0.05</td><td>0.5</td><td>[B|S:E][D|B:A:S:E][A][S][T|A][L|S][E|T:L][X|E]</td><td>8/8</td><td>3</td><td>0</td><td>0</td></tr>
<tr><td>0.05</td><td>0.8</td><td>[B|S:E][D|B:A:S:E][A][S][T|A][L|S][E|T:L][X|E]</td><td>8/8</td><td>3</td><td>0</td><td>0</td></tr>
<tr><td>0.10</td><td>0.2</td><td>[B|S:E][D|B:A:S:E][A][S][T|A][L|S][E|T:L][X|E]</td><td>8/8</td><td>3</td><td>0</td><td>0</td></tr>
<tr><td>0.10</td><td>0.5</td><td>[B|S:E][D|B:A:S:E][A][S][T|A][L|S][E|T:L][X|E]</td><td>8/8</td><td>3</td><td>0</td><td>0</td></tr>
<tr><td>0.10</td><td>0.8</td><td>[B|S:E][D|B:A:S:E][A][S][T|A][L|S][E|T:L][X|E]</td><td>8/8</td><td>3</td><td>0</td><td>0</td></tr>
<tr><td>0.20</td><td>0.2</td><td>[B|S:E][D|B:A:S:E][A][S][T|A][L|S][E|T:L][X|E]</td><td>8/8</td><td>3</td><td>0</td><td>0</td></tr>
<tr><td>0.20</td><td>0.5</td><td>[B|S:E][D|B:E][A][S][T|A][L|S][E|T:L][X|E]</td><td>8/8</td><td>1</td><td>0</td><td>0</td></tr>
<tr><td>0.20</td><td>0.8</td><td>[B|S:E][D|B:E][A][S][T|A][L|S][E|T:L][X|E]</td><td>8/8</td><td>1</td><td>0</td><td>0</td></tr>
<tr><td>0.50</td><td>0.2</td><td>[B|S:E][D|B:E][A][S][T|A][L|S][E|T:L][X|E]</td><td>8/8</td><td>1</td><td>0</td><td>0</td></tr>
<tr><td>0.50</td><td>0.5</td><td>[B|S:E][D|B:E][A][S][T|A][L|S][E|T:L][X|E]</td><td>8/8</td><td>1</td><td>0</td><td>0</td></tr>
<tr><td>0.50</td><td>0.8</td><td>[B|S:E][D|B:E][A][S][T|A][L|S][E|T:L][X|E]</td><td>8/8</td><td>1</td><td>0</td><td>0</td></tr>
</table>

Table 7: The ASIA network, [16], learned by the P-metric with a sample size of 500 observations.

		sample = 1500				
z	α	Learned Network	Correct	Inc. Add.	Inc. Mis.	Wr. Dir.
0.05	0.2	[B|S:E][D|B:A:S:E][A][S][T|A][L|S][E|T:L][X|E]	8/8	3	0	0
0.05	0.5	[B|S:E][D|B:A:S:E][A][S][T|A][L|S][E|T:L][X|E]	8/8	3	0	0
0.05	0.8	[B|S:E][D|B:A:S:E][A][S][T|A][L|S][E|T:L][X|E]	8/8	3	0	0
0.10	0.2	[B|S:E][D|B:E][A][S][T|A][L|S][E|T:L][|E]	8/8	1	0	0
0.10	0.5	[B|S:E][D|B:A:S:E][A][S][T|A][L|S][E|T:L][X|E]	8/8	3	0	0
0.10	0.8	[B|S:E][D|B:A:S:E][A][S][T|A][L|S][E|T:L][X|E]	8/8	3	0	0
0.20	0.2	[B|S:E][D|B:E][A][S][T|A][L|S][E|T:L][|E]	8/8	1	0	0
0.20	0.5	[B|S:E][D|B:E][A][S][T|A][L|S][E|T:L][|E]	8/8	1	0	0
0.20	0.8	[B|S:E][D|B:A:S:E][A][S][T|A][L|S][E|T:L][X|E]	8/8	3	0	0
0.50	0.2	[B|S:E][D|B:E][A][S][T|A][L|S][E|T:L][|E]	8/8	1	0	0
0.50	0.5	[B|S:E][D|B:E][A][S][T|A][L|S][E|T:L][|E]	8/8	1	0	0
0.50	0.8	[B|S:E][D|B:E][A][S][T|A][L|S][E|T:L][|E]	8/8	1	0	0

Table 8: The ASIA network, [16], learned by the P-metric with a sample size of 1500 observations.

<table>
<tr><th></th><th></th><th>sample = 3000</th><th></th><th></th><th></th><th></th></tr>
<tr><th>z</th><th>α</th><th>Learned Network</th><th>Correct</th><th>Inc. Add.</th><th>Inc. Mis.</th><th>Wr. Dir.</th></tr>
<tr><td>0.05</td><td>0.2</td><td>[B|S:E][D|B:E][A][S][T|A][L|S][E|T:L][|E]</td><td>8/8</td><td>1</td><td>0</td><td>0</td></tr>
<tr><td>0.05</td><td>0.5</td><td>[B|S:E][D|B:A:S:E][A][S][T|A][L|S][E|T:L][X|E]</td><td>8/8</td><td>3</td><td>0</td><td>0</td></tr>
<tr><td>0.05</td><td>0.8</td><td>[B|S:E][D|B:A:S:E][A][S][T|A][L|S][E|T:L][X|E]</td><td>8/8</td><td>3</td><td>0</td><td>0</td></tr>
<tr><td>0.10</td><td>0.2</td><td>[B|S:E][D|B:E][A][S][T|A][L|S][E|T:L][|E]</td><td>8/8</td><td>1</td><td>0</td><td>0</td></tr>
<tr><td>0.10</td><td>0.5</td><td>[B|S:E][D|B:E][A][S][T|A][L|S][E|T:L][|E]</td><td>8/8</td><td>1</td><td>0</td><td>0</td></tr>
<tr><td>0.10</td><td>0.8</td><td>[B|S:E][D|B:A:S:E][A][S][T|A][L|S][E|T:L][X|E]</td><td>8/8</td><td>3</td><td>0</td><td>0</td></tr>
<tr><td>0.20</td><td>0.2</td><td>[B|S:E][D|B:E][A][S][T|A][L|S][E|T:L][|E]</td><td>8/8</td><td>1</td><td>0</td><td>0</td></tr>
<tr><td>0.20</td><td>0.5</td><td>[B|S:E][D|B:E][A][S][T|A][L|S][E|T:L][|E]</td><td>8/8</td><td>1</td><td>0</td><td>0</td></tr>
<tr><td>0.20</td><td>0.8</td><td>[B|S:E][D|B:E][A][S][T|A][L|S][E|T:L][|E]</td><td>8/8</td><td>1</td><td>0</td><td>0</td></tr>
<tr><td>0.50</td><td>0.2</td><td>[B|S:E][D|B:E][A][S][T|A][L|S][E|T:L][|E]</td><td>8/8</td><td>1</td><td>0</td><td>0</td></tr>
<tr><td>0.50</td><td>0.5</td><td>[B|S:E][D|B:E][A][S][T|A][L|S][E|T:L][|E]</td><td>8/8</td><td>1</td><td>0</td><td>0</td></tr>
<tr><td>0.50</td><td>0.8</td><td>[B|S:E][D|B:E][A][S][T|A][L|S][E|T:L][|E]</td><td>8/8</td><td>1</td><td>0</td><td>0</td></tr>
</table>

Table 9: The ASIA network, [16], learned by the P-metric with a sample size of 3000 observations.

References

[1] S. K. Andreassen, K. G. Olesen, F. V. Jensen, and F. Jensen. Hugin: a shell for building bayesian belief universes for expert systems. *Proceedings of the Eleventh International Joint Conference on Artificial Intelligence*, 1989.

[2] S. G. Bøttcher. *Learning Conditional Gaussian Networks.* Tecnical Report R2005-22, Aalborg University, Denmark, 2005.

[3] S. G. Bøttcher and C. Dethlefsen. DEAL: A package for learning bayesian networks. *Journal of Statistical Software*, 8(20):1–40, 2003.

[4] W. L. Buntine. Theory of refinement on bayesian networks. *Proceedings of 7th Conference on Uncertainty in Artificial Intelligence*, pages 52–60, 1991.

[5] D. M. Chickering. Learning bayesian networks is NP-complete. *Proceedings on Artificial Intelligence and Statistics*, pages 121–130, 1995.

[6] D. M. Chickering, D. Geiger, and D. Heckerman. Learning bayesian network: A combination of knowledge and statistical data. *Tecnical Report MSR-TR-94-17, Microsoft Research, Advanced Technology Division*, 1994.

[7] D. M. Chickering, D. Geiger, and D. Heckerman. Learning bayesian networks: Search methods and experimental results. *Preliminary papers of the 5th Intl. Workshop on Artificial Intelligence and Statistics*, pages 112–128, 1995.

[8] G. F. Cooper and E. Herskovitz. A bayesian method for the induction of probabilistic networks from data. *Machine Learning*, 9:309–357, 1992.

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

[10] D. Geiger and J. Pearl. Logical and algorithmic properties of conditional independence. *Tecnical Report R97, Cognitive System Laboratory, UCLA*, 1988.

[11] D. Heckerman, D. Geiger, and D. M. Chickering. Learning bayesian network: A combination of knowledge and statistical data. *Proceedings of 10th Conf. Uncertainty in Artificial Intelligence*, pages 293–301, 1994.

[12] D. Heckerman, C. Meek, and G. Cooper. A bayesian approach to causal discovery. *Technical Report MSR-TR-97-05*, 1997.

[13] R. Ihaka and R. Gentleman. R: a language for data analisys and graphics. *Journal of Computational and Graphical Statistics*, 5:299–314, 1996.

[14] F. V. Jensen. *An introduction to Bayesian Networks.* Springer Verlag, New York, N.Y., 1996.

[15] P. Larrañaga and M. Poza. 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, 1996.

[16] S. Lauritzen and D. Spiegehalter. 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, 1988.

[17] P. Le, A. Bahl, and L. Ungar. Using prior knowledge to improve genetic network reconstruction from microarray data. *InSilico Biology*, 27(4), 2004.

[18] K. P. Murphy. The bayes net toolbox for MATLAB. *Computer Science and Statistics*, 33:331–349, 2001.

[19] J. Pearl. *Probabilistic Reasoning in Intelligent Systems.* Morgan Kaufmann, San Mateo, CA, 1988.

Appendix 1

In this note we show that the computation of the normalization constant c is not needed to use the proposed score function. Let $P_{\mathcal{G}_D}(B_s)$ be the probability distribution over DAGs and $P_{\mathcal{G}_{DG}}(B_s)$ be the probability distribution over Digraphs . By straightforward algebra, we have:

$$S_\delta(B_s) = \log\left(\frac{P_{\mathcal{G}_D}(B_s)}{P_{\mathcal{G}_D}(\{\emptyset\})}\right) = \log\left(P_{\mathcal{G}_D}(B_s)\right) - \log\left(P_{\mathcal{G}_D}(\{\emptyset\})\right) =$$

$$= \log\left(cP_{\mathcal{G}_{DG}}(B_s)\right) - \log\left(cP_{\mathcal{G}_{DG}}(\{\emptyset\})\right) =$$

$$= \log(c) + \log\left(P_{\mathcal{G}_{DG}}(B_s)\right) - \log(c) - \log\left(P_{\mathcal{G}_{DG}}(\{\emptyset\})\right) =$$

$$= \log\left(\frac{P_{\mathcal{G}_{DG}}(B_s)}{P_{\mathcal{G}_{DG}}(\{\emptyset\})}\right) = \log\left(\frac{\prod^{B_s} C^T_{i\cdot j}P_{i\cdot j}}{\prod^{\{\emptyset\}} C^T_{i\cdot j}P_{i\cdot j}}\right)$$

where $\prod^{B_s} C^T_{i\cdot j}P_{i\cdot j}$ and $\prod^{\{\emptyset\}} C^T_{i\cdot j}P_{i\cdot j}$ refer to factorization of the prior judgment respectively over the candidate network B_s and to the empty structure.

Appendix 2

The score computation may be limited to pairs of nodes for which the expert explicitly defined a probability distribution.. Let F be the number of pairs of nodes for which the belief has been elicited by the assignment of a distribution $\{p(x_{m_f} \mid \xi) : f = 1, \ldots, F\}$ 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:

$$P(B_s) = \prod C^T_{i\cdot j}P_{i\cdot j} = \prod_{\{i\cdot j\}\in F} p(x_{m_f}) \cdot \prod_{\{i\cdot j\}\notin F} k$$

and by straightforward algebra we have:

$$S_\delta(B_s) = \log\left(\frac{P(B_s)}{P(\{\emptyset\})}\right) = \log\left(\frac{\prod_{\{i\cdot j\}\in F} p(x^{B_s}_{m_f}) \cdot \prod_{\{i\cdot j\}\notin F} k}{\prod_{\{i\cdot j\}\in F} p(x^{\{\emptyset\}}_{m_f}) \cdot \prod_{\{i\cdot j\}\notin F} k}\right) =$$

$$= \log\left(\frac{\prod_{\{i\cdot j\}\in F} p(x^{B_s}_{m_f})}{\prod_{\{i\cdot j\}\in F} p(x^{\{\emptyset\}}_{m_f})}\right)$$

with constants k cancelled out.

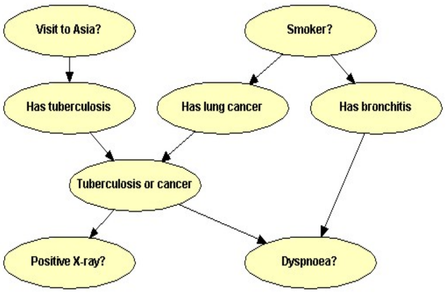

Figure 3: The ASIA network, [16].

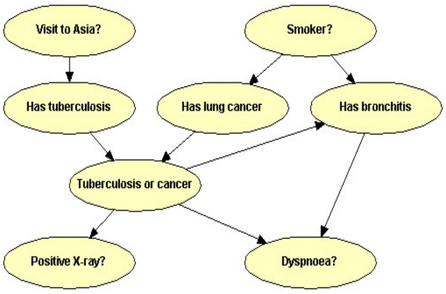

Figure 4: ASIA network learned by the P-metric with sample=500 and parameters $z = 0.2$ and $\alpha = 0.5$.

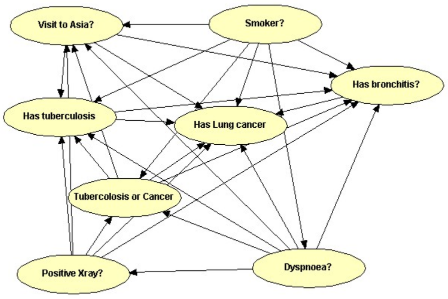

Figure 5: The ASIA network learned with DEAL.

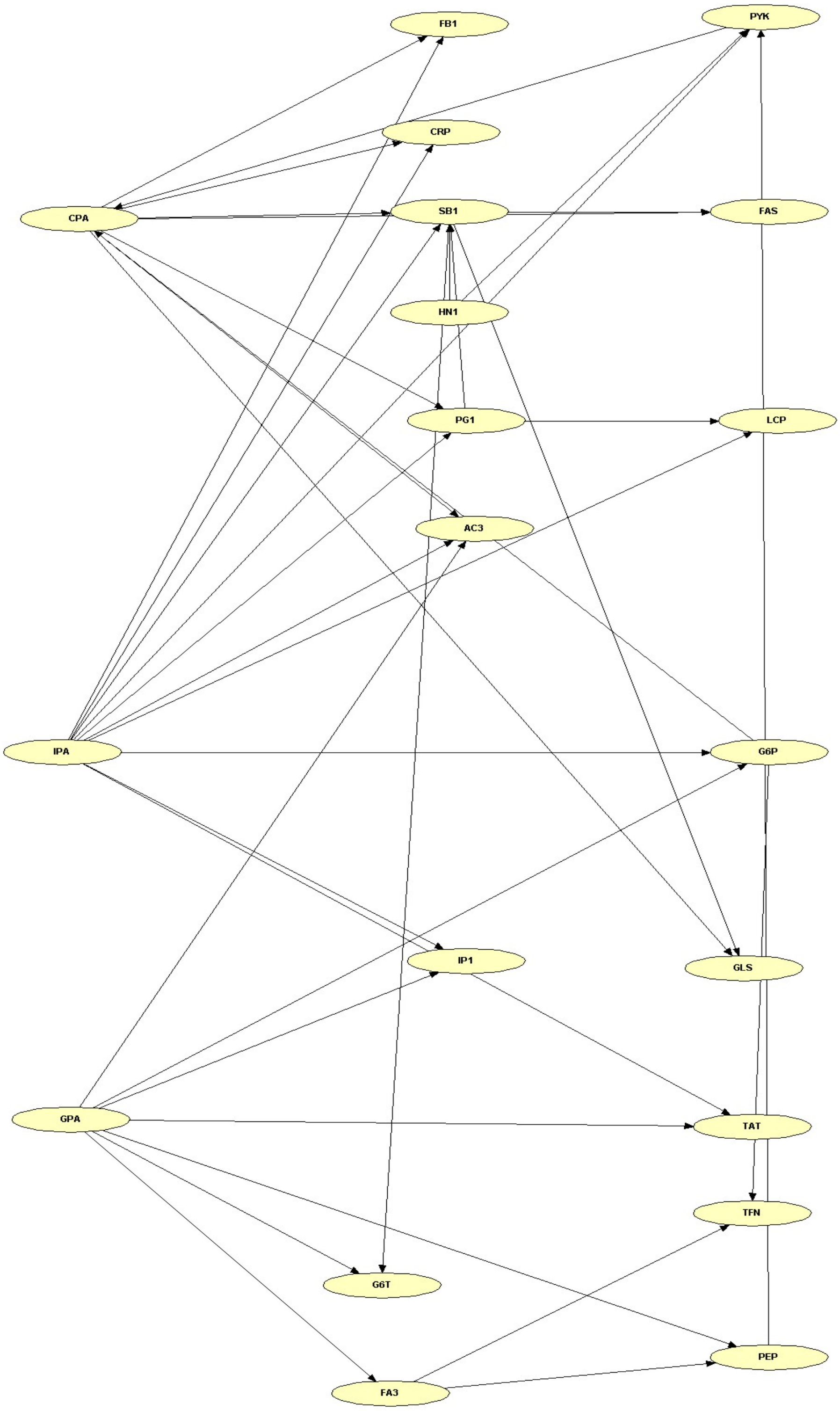

Figure 6: The HGH network learned by the P-metric with parameters $z = 0.5$ and $\alpha = 0.5$.

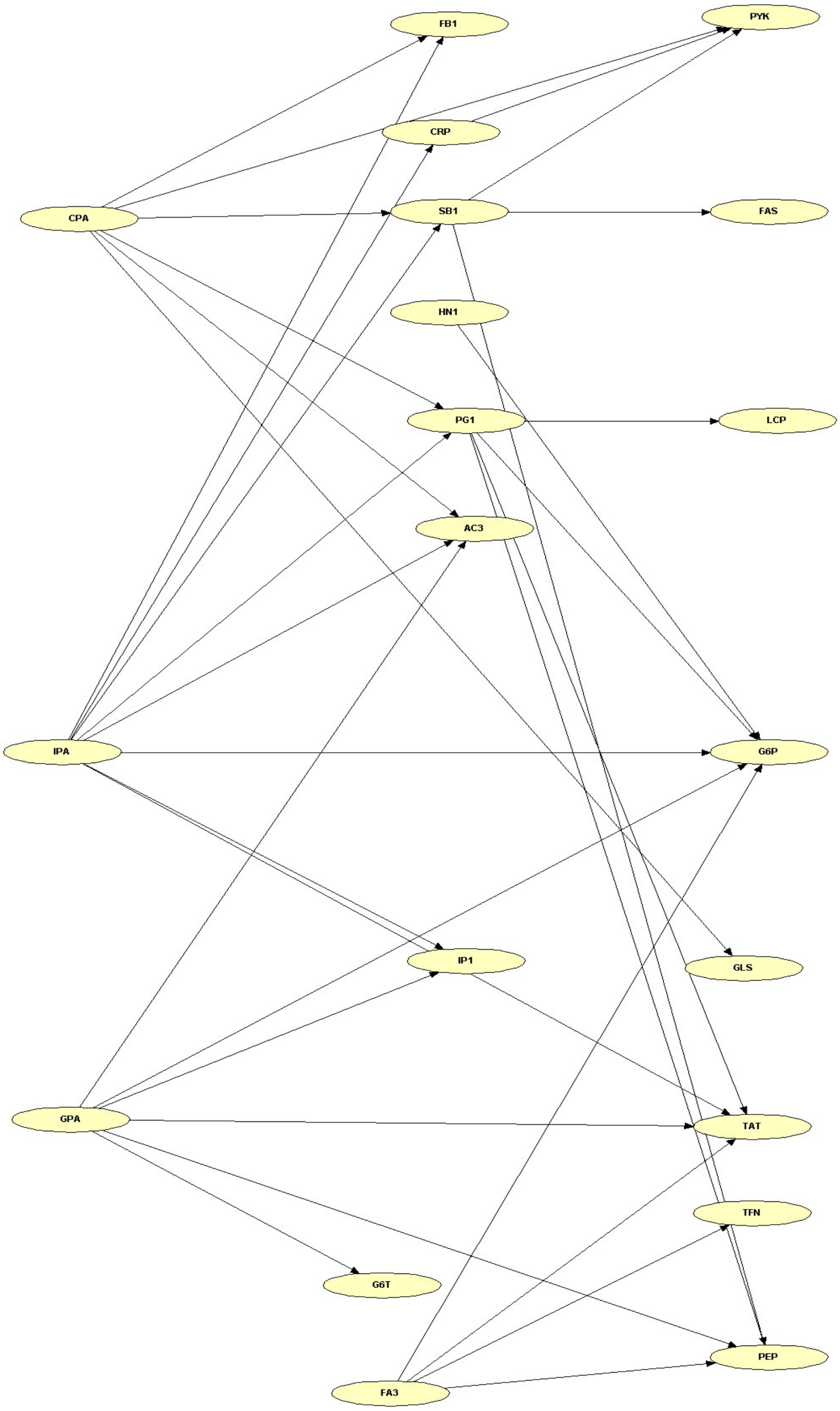

Figure 7: The original HGH network, [17].