Properties of Measures for Bayesian Belief Network Learning

R.R. Bouckaert

UU-CS-1994-35
August 1994

Utrecht University

Department of Computer Science

Padualaan 14, P.O. Box 80.089,
3508 TB Utrecht, The Netherlands,
Tel. : ... + 31 - 30 - 531454

Properties of Measures for Bayesian Belief Network Learning

R.R. Bouckaert

Technical Report UU-CS-1994-35
August 1994

Department of Computer Science
Utrecht University
P.O.Box 80.089
3508 TB Utrecht
The Netherlands

ISSN: 0024-3275

Properties of Measures for Bayesian Belief Network Learning

Remco R. Bouckaert

Utrecht University
Department of Computer Science
P.O.Box 80.089 3508 TB Utrecht, The Netherlands
remco@cs.ruu.nl

Abstract

Bayesian belief network learning algorithms involve three basic components: a quality measure of a network structure given a database, a search heuristic aiming at finding the best network structures, and a method for estimating probabilities from the database. This paper addresses quality measures.

The behavior of the Bayesian measure of Cooper and Herskovits and a minimum description length (MDL) measure are compared with respect to their properties for both limiting size and finite size databases. It is shown that both measure behave the same for infinite size databases and that both measures prefer minimum I-maps overwhelmingly over non minimum I-maps.

For finite size databases, it is shown that the MDL measure will not select network structures with parent sets of size $\log N$ where N is the size of the database. However, the Bayesian measure may select parent sets as large as $N/2$.

1 Introduction

The framework of Bayesian belief networks offers a mathematically sound formalism for representing uncertainty in knowledge-based systems. Efficient algorithms are associated with the formalism for making inferences with knowledge represented in a belief network, [7, 10, 11]. In addition, the framework has proved its practical worth over the last few years, [1, 2, 13]. However, constructing belief networks with the help of human experts is a time-consuming and expensive task. Since more and more large databases of cases become available, automated learning algorithms can help shorten the build and test cycle of a belief network by suggesting an initial set-up. Therefore, learning belief networks from data is an important research issue. In fact, a lot of research effort has been spent on the design of methods for learning Bayesian belief networks from different perspectives such as computer science, statistics, and philosophy.

An algorithm for learning Bayesian belief networks involves three components: a quality measure for comparing network structures, a search heuristic for finding the best network structure given the database of cases, and an estimation method for learning the probabilities for the network from the database.

This paper addresses quality measures only. One of the most promising to date is a measure based on a Bayesian method proposed by Cooper and Herskovits, [5]. However, measures based on the minimum description length (MDL) principle are rapidly gaining

popularity, [3, 9, 14, 15]. We investigate the behavior of the measures proposed by Cooper and Herskovits and a measure based in the MDL principle both for databases of infinite size and for databases of finite size.

In the next section, we give a short introduction to terms and notations used in the remainder of this paper. Section 3 is devoted to the properties of quality measures that we just mentioned. We conclude with some final considerations and directions for further research.

2 Preliminaries

In this section, we define some terms concerning Bayesian belief networks and some notational conventions that will be used throughout this paper.

Let U be a set of n discrete variables denoted x_i, $i = 1,\ldots,n$, $n \geq 1$, that is $U = \{x_1,\ldots,x_n\}$. Each variable x_i may take a value in the range $\{x_{i1},\ldots,x_{ir_i}\}$, $r_i \geq 2$, $i = 1,\ldots,n$. We will use capital letters to denote sets of variables and lower case letters to denote single variables. To prevent an abundant usage of braces, we sometimes write x to denote $\{x\}$, XY to denote $X \cup Y$, and xy to denote $\{x,y\}$. In the sequel, we will assume that every variable is an element of U and every set of variables is a subset of U unless stated otherwise.

A *Bayesian belief network* B over a set of variables U is a pair $B = (B_S, B_P)$. The *network structure* B_S is a directed acyclic graph (DAG) with a node for every variable in U. B_P is a set of conditional probability tables; for every variable $x_i \in U$, B_P contains a conditional probability table with probabilities $P(x_i|\pi_i)$ for all values of x_i given all combinations of values for the variables in x_i's *parent set* π_i in the network structure B_S; in the sequel, such a combination of values will be called an *instantiation*. The joint probability distribution represented by this network is $\prod_{x\in U} P(x|\pi_x)$, [11].

In a network structure B_S, a *trail* is a path in the underlying undirected graph of B_S, that is, the direction of the arcs is immaterial. A *head-to-head node* in a trail in B_S is a node z such that a sequence $x \rightarrow z \leftarrow y$ is part of the trail. A trail in B_S between two nodes x and y is *blocked* by a set of nodes Z if at least one of the following two conditions holds:
- the trail contains a head-to-head node x such that x nor any descendant of x is in Z;
- the trail contains a node x such that $x \in Z$ and x is not a head-to-head node in the trail.

In a network structure B_S, we say that the set of nodes X is *d-separated* from Y given Z if every trail between any node $x \in X$ and any node $y \in Y$ is blocked by Z.

For a joint probability distribution P over U, we call the sets of variables X and Y *conditionally independent* given a set Z, written $I(X,Z,Y)$, if $P(XY|Z) = P(X|Z)P(Y|Z)$ for all instantiations of XYZ. A statement $I(X,Z,Y)$ is called an *independency statement*. Let P be a joint probability distribution over a set of variables U. Then, an *independency model* M of a distribution P is the set of all independency statements that hold in P.

For an arbitrary probability distribution, the first four of the axioms below apply [6]; for positive distributions, the fifth axiom applies as well.

symmetry	$I(X,Z,Y)$	$\Leftrightarrow I(Y,Z,X)$
decomposition	$I(X,Z,WY)$	$\Rightarrow I(X,Z,Y)$
weak union	$I(X,Z,WY)$	$\Rightarrow I(X,ZW,Y)$
contraction	$I(X,ZW,Y) \wedge I(X,Z,W)$	$\Rightarrow I(X,Z,WY)$
intersection	$I(X,ZW,Y) \wedge I(X,ZY,W)$	$\Rightarrow I(X,Z,WY)$

These axioms can be used to derive new independency statements from a given set of independency statements.

A network structure B_S is an *independency map* or *I-map* of a distribution P if X and Y being d-separated by Z in B_S implies that $I(X, Z, Y)$ holds in P. B_S is a *minimal I-map* of P if B_S is an I-map of P and no proper subgraph of B_S is an I-map of P. B_S is a *perfect map* or *P-map* of P if it is an I-map of P and $I(X, Z, Y)$ holding in P implies that X and Y are d-separated by Z in B_S.

Let $<_U$ be a total ordering on U. Let M be the independency model of a joint probability distribution P over U. A *causal input list* $L_{<_U}$ over M is a set of independency statements such that for each $x_i \in U$, the set $L_{<_U}$ contains one and only one independency statement of the form,

$$I(x_i, \pi_i, U_i \backslash \pi_i)$$

where $U_i = \{y | y \in U,\ y <_U x_i\}$ such that $I(x_i, \pi_i, U_i \backslash \pi_i)$ holds in M and for any proper subset S of π_i, $I(x_i, S, U_i \backslash S)$ does not hold in M. A causal input list can be constructed from M in $O(|U|^2)$ consultations of M. Corresponding to a causal input list $L_{<_U}$ over M, $B_{S_<}$ is a network stucture such that each node x_i has parent set π_i, $i = 1, \ldots, n$. This network stucture $B_{S_<}$ is a minimal I-map of P, [12].

We say that a network structure B_S *obeys* a total ordering $<_U$ on U if for each arc $x_i \rightarrow x_j$ in B_S we have that $x_i <_U x_j$. So, $<_U$ is a topological ordering of B_S.

A *case* over U is a value assignment to all variables $x_i \in U$. A *database D of cases* over U is a list of cases over U.

3 Quality Measures

The task of learning a Bayesian belief networks B is twofold: learning the network structure B_S and learning the set of probability tables B_P. The latter task will be addressed in another paper. Here, we investigate learning network structures.

In learning a network stucture, network stuctures have to be compared to decide the 'best' structures given the database. To this end, we use a *quality measure* indicating the fitness of the network structure to the database: the better the network describes the database, the higher the quality. The basic idea is that a network structure that has a higher quality according to the measure employed is preferred over a network structure with lower quality. In the sequel, we will see that a heuristic search procedure aims at finding a network structure that has the highest quality.

The Bayesian approach is a well-founded and practical method for selecting statistical models for a given list of cases. In the context of Bayesian belief network learning, the statistical model is a network structure. The basic idea of this approach is to start with a prior distribution on all network stuctures. For each structure, the probability of the database given the structure is computed, and, using Bayes' theorem, the posterior probability of the structure given the database is calculated. The network structure with the highest posterior probability is selected. The posterior probability can be regarded as a measure of the quality of the network structure.

Cooper and Herskovits [5] have proposed a quality measure based on the Bayesian approach. They assume that the cases in the database are supposed to occur independently.

Furthermore, they assume that in the database there are no cases where values are missing. Also, they assume that no probability table B_P is preferred for a given network structure before the database has been inspected.

Let U again be the set of n discrete variables as defined in Section 2. Let D be a database of N cases over U. Now, let B_S be a network structure over the variables in U. For each x_i, let π_i be the parent set of x_i in B_S. Let w_{ij} denote the jth instantiation of π_i relative to D, $j = 1, \ldots, q_i$, $q_i \geq 1$; note that $q_i \leq \prod_{x_j \in \pi_i} r_j$. Furthermore, let N_{ijk} be the number of cases in D in which the variable x_i has the value x_{ik} and π_i is instantiated as w_{ij}, and let $N_{ij} = \sum_{k=1}^{r_i} N_{ijk}$. Let $P(B_S)$ be the probability of B_S prior to observation of the database. Then, the probability of B_S and the database D is,

$$P(B_S, D) = P(B_S) \cdot \prod_{i=1}^{n} \prod_{j=1}^{q_i} \frac{(r_i - 1)!}{(N_{ij} + r_i - 1)!} \prod_{k=1}^{r_i} N_{ijk}!. \tag{1}$$

A proof can be found in [5]. The term $P(B_S)$ models prior information, for example existence and direction of arcs. The other terms in (1) do not have a direct intuitive interpretation. In practical implementations, generally the logarithm of Equation (1) is used because even for small databases with N cases numbers like $N!$ tend to give computational problems (notice that $100! \approx 10^{160}$). The logarithm of (1) will be referred to as the *Bayesian measure* for the quality of a network structure. In this paper, all logarithms are to the base two.

In [3], it has been showed that if all possible instantiations of the parent sets in B_S occur at least once in the database, the Bayesian measure can be approximated by the measure

$$L(B_S, D) = \log P(B_S) - N \cdot H(B_S, D) - \frac{1}{2} K \cdot \log N \tag{2}$$

with $O(1)$ error with respect to N, where $H(B_S, D) = \sum_{i=1}^{n} \sum_{j=1}^{q_i} \sum_{k=1}^{r_i} -\frac{N_{ijk}}{N} \log \frac{N_{ijk}}{N_{ij}}$ and $K = \sum_{i=1}^{n} (r_i - 1) \prod_{x_j \in \pi_i} r_j$. Compared to the Bayesian measure, this measure has a more intuitive interpretation. The term $\log P(B_S)$ models prior information, just as in the Bayesian measure. The term $-N \cdot H(B_S, D)$ is $-N$ times the entropy of a network structure B_S given a database D. Generally, this term increases as arcs are added to a network structure. In the last term, K equals the number of independent probabilities that are needed to define all probability tables in B_P for the Bayesian belief network $B = (B_S, B_P)$. The term $-\frac{1}{2} K \cdot \log N$ thus models the cost of estimating these K probabilities. Contrary to the entropy term, this term decreases when arcs are added to a network structure. If no prior information is available, that is, if $P(B_S)$ is equal for all network structures B_S, then this measure assigns high quality to network structures that fit the database with as few arcs as possible. Therefore, Equation (2) will be referred to as the *minimum description length (MDL) measure* for the quality of a network structure.

3.1 Infinite Database Properties of Quality Measures

This section will be devoted to the investigation of the asymptotic behavior of the Bayesian and MDL measures, that is, the behavior for infinite size databases. The results are of a theoretical character. Nonetheless, the results indicate how the measures may be expected to behave for large databases. In the next subsection, we investigate the behavior for finite size databases.

Theorem 3.1 *Let U be a set of variables and let $<_U$ a total ordering on U. Let the prior distribution over all network stuctures be uniform. Let P_D be a joint probability distribution over U such that P_D has a unique minimal I-map that obeys $<_U$. Furthermore, let D be a database with N cases generated from P_D where N approximates infinity. Let B_S and $B_{S'}$ be network structures obeying $<_U$ where B_S is the minimal I-map of P_D. Let $Q(B_S, D)$ be either the Bayesian or the MDL measure. Then,*

$$\lim_{N \to \infty} (Q(B_{S'}, D) - Q(B_S, D)) = -\infty,$$

if and only if $B_{S'}$ is not a minimal I-map of P_D.

Proof: For the Bayesian measure, the property stated above follows from Theorem 6.3 in [8]. For the MDL measure the theorem was proved in [4]. □

The theorem states that for databases large enough, a network stucture that is a minimal I-map obeying a particular ordering $<_U$ is overwhelmingly preferred over any other network stucture obeying the same ordering. Note that as positive probability distributions P_D have a unique minimal I-map for a given ordering $<_U$ the theorem applies to any positive distribution. Also note that if a distribution has a P-map, then for every ordering its minimal I-map is unique since the intersection rule for conditional independence applies. Therefore, the theorem applies as well for distributions for which a P-map exists. If a P-map exists for the given ordering $<_U$, then this P-map will be overwhelmingly preferred over any other structure obeying $<_U$ since P-maps are unique for a given ordering. In general a minimal I-map need not necessarily have a higher quality than other structures.

When learning a network structure it is desirable that no ordering on the variables is required by the learning algorithm because finding a 'good' ordering may be difficult. Therefore, it is interesting to investigate the properties of the quality measures in case no ordering on the variables is provided. When no ordering needs to be obeyed, minimal I-maps of positive distributions need not be unique. Minimal I-maps need not even be equivalent, that is, represent the same set of independency statements. Consider for example Figure 1. Suppose that the structure on the left is a P-map of some distribution P. Both structures on the right are minimal I-maps of the distribution obtained from P by marginalising over b. However, the upper structure represents $I(a, \emptyset, e)$ whereas the lower structure does not. Note that for the upper structure $e <_U d$ in any ordering $<_U$ obeyed by this structure while for the lower structure it would be $d <_U e$ Uniqueness of optimal structures is a desirable property since it helps deriving theoretical results. Furthermore, when one is not interested in the represented distribution but in the causal structure underlying the domain, it is necessary that the network stucture is unique since the causal structure is unique.

The number of probabilities that need to be estimated for minimal I-maps need not be the same for every such I-map. Since in every estimate of a probability a small error is introduced, it is desirable that as few as possible probabilities need be assessed. This gives reason to distinguish between minimum and non-minimum structures.

Definition 3.1 *Let P be a joint probability distribution on U. A* minimum I-map *B_S of P is a minimal I-map of P such that for any minimal I-map $B_{S'}$ of P in the belief networks $B = (B_S, B_P)$ and $B' = (B_{S'}, B_{P'})$ the set B_P specifies at most as many probabilities as $B_{P'}$.*

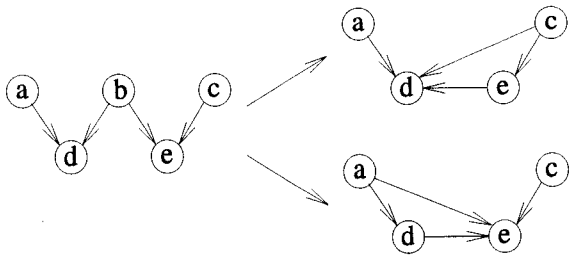

Figure 1: Example of different minimal I-maps.

Consider once more Figure 1. Let all variables except c be binary and let c be ternary. Then for the upper structure, eighteen probabilities need be specified to arrive at a belief network: one for a, two for c, twelve for d, and three for e. However, for the lower structure, only seventeen probabilities need be specified: one for a, two for c, two for d, and twelve for e. If all variables including c would be binary, for both structures twelve probabilities would need to be specified.

Unfortunately, minimum I-maps are not unique, so we cannot deduce causal relations based on a minimum I-map. We have the following theorem for minimum I-maps.

Theorem 3.2 *Let U be a set of variables and let the prior distribution over all network structures over U be positive. Let P_D be a positive distribution over U. Let D be a database with N cases generated by P_D where N approximates infinity. Let B_S be a minimum I-map of P_D. Then, for any network stucture $B_{S'}$ that is not a minimum I-map of P_D we have that,*

$$\lim_{N\to\infty} Q(B_S, D) - Q(B_{S'}, D) = -\infty,$$

where Q is either the Bayesian or the MDL measure.

Proof: In [3], we have shown that for N approximating infinity the MDL measure is an approximation of the Bayesian measure with $O(1)$ error if all instantiations of parent sets are observed in the database. The condition that P_D is a positive distribution assures that all these instantiations are in the database. It therefore suffices to prove the theorem for the MDL measure.

For B_S, let K, r_i, x_i, π_i, x_{ij}, w_{ij} be as before. Let q_i be the number of all possible instantiations of π_i. Likewise, let K', r_i', etc. be defined for $B_{S'}$. We consider the expression $\lim_{N\to\infty}(L(B_{S'}, D) - L(B_S, D))$, which by definition is equal to $\lim_{N\to\infty}(\log P(B_{S'}) - N \cdot H(B_{S'}, D) - \frac{1}{2}K' \cdot \log N - \log P(B_S) + N \cdot H(B_S, D) + \frac{1}{2}K \cdot \log N)$ which can be written as,

$$\lim_{N\to\infty}\left(\frac{\log P(B_{S'})}{P(B_S)} - N \cdot H(B_{S'}, D) + N \cdot H(B_S, D) - \frac{1}{2}(K' - K) \cdot \log N\right). \qquad (3)$$

First, consider the entropy term $N \cdot H(B_{S'}, D)$. Note that by the strong law of large numbers, we have $\lim_{N\to\infty} \frac{N'_{ijk}}{N} = P_D(x_i = x'_{ik}, \pi'_i = w'_{ij})$ and $\lim_{N\to\infty} \frac{N'_{ijk}}{N'_{ij}} = P_D(x_i = x'_{ik} | \pi'_i = w'_{ij})$. Therefore for very large N, $N \cdot H(B_{S'}, D)$ can be written as,

$$N \cdot \sum_{i=1}^{n}\sum_{j=1}^{q_i}\sum_{k=1}^{r_i} -P_D(x_i = x'_{ik}, \pi'_i = w'_{ij}) \cdot \log P_D(x_i = x'_{ik} | \pi'_i = w'_{ij}).$$

Let $<_U$ be an ordering for $B_{S'}$ and let $U_i = \{y | y <_U x_i\}$ for $i = 1, \ldots, n$, let w''_{ij} be the jth instantiation of U_i. Then we can write the above formula as,

$$N \cdot \sum_{i=1}^{n} \sum_{j=1}^{q_i} \sum_{k=1}^{r_i} - \left(\sum_{U_i \backslash \pi_i} P_D(x_i = x'_{ik}, U_i = w'_{ij} \wedge c_{U_i \backslash \pi_i}) \right) \log P_D(x_i = x'_{ik} | \pi'_i = w'_{ij}),$$

where $\sum_{U_i \backslash \pi_i}$ represents the sum over all instantiations of the nodes from $U_i \backslash \pi_i$ and $U_i = w'_{ij} \wedge c_{U_i \backslash \pi_i}$ denotes that of the variables in U_i the variables in π_i take values according to w_{ij} and the variables in $U_i \backslash \pi_i$ take values in the summation. By carefully grouping terms, this can be shown to equal,

$$N \cdot \sum_{U} - \left(\prod_{i=1}^{n} P_D(x_i | U_i) \right) \log \left(\prod_{i=1}^{n} P_D(x_i | \pi'_i) \right),$$

where $\sum_U$ denotes the summation over all instantiations of the variables in U and x_i, π'_i, and U_i take values conform the instantiation of U. Now, let $P'_D(U) = \prod_{i=1}^{n} P_D(x_i | \pi'_i)$, that is, let $P'_D(U)$ be the distribution defined the belief network wit network stucture $B_{S'}$. Then, we can write the above formula as

$$N \cdot \sum_{U} -P_D(U) \cdot \log P'_D(U).$$

For the entropy term $N \cdot H(B_S, D)$, we find that this term equals

$$N \cdot \sum_{U} -P_D(U) \cdot \log P_D(U).$$

So, the Formula (3) can be written as,

$$\lim_{N \to \infty} \left(\log \frac{P(B_{S'})}{P(B_S)} + N \cdot \sum_{U} (P_D(U) \log P'_D(U) - P_D(U) \log P_D(U)) - \frac{1}{2}(K' - K) \cdot \log N \right). \tag{4}$$

We now distinguish between two cases:

- $B_{S'}$ is not an I-map of P_D;
- $B_{S'}$ is an I-map but not a minimum I-map of P_D.

We consider these cases separately. First, suppose that $B_{S'}$ is not an I-map of P_D. Now observe that since $B_{S'}$ is not an I-map of P_D, there is an instantiation of U such that $P'_D(U) \neq P_D(U)$; assume that no such instantiation could be found, then the belief networks with network stuctures $B_{S'}$ and B_S would define the same probability distribution and thus satisfy the same set of independency statements.

By Shannon's inequality which states $\sum_i a_i \log b_i \leq \sum_i a_i \log a_i$, for all $a_i, b_i \geq 0$, $\sum_i a_i = \sum_i b_i = 1$ (with equality only if $a_i = b_i$ for all i), we have that in (4) the summation $\sum_U (P_D(U) \log P'_D(U) - P_D(U) \log P_D(U))$ is smaller than 0. The result is multiplied by N yielding an $O(N)$ term that goes to minus infinity. Since the prior distribution on network structures is positive, the probabilities $P(B_S)$ and $P(B_{S'})$ are positive; so, $\log(P(B_{S'})/P(B_S))$ is a constant that is negligible when $N \to \infty$. Since an $O(N)$ term dominates an $O(\log N)$ term for $N \to \infty$ we have that the $\frac{1}{2}(K' - K) \cdot \log N$ term does not influence the result. So, Formula (4) is $-\infty$.

Now suppose that $B_{S'}$ is an I-map of P_D, yet not a minimum I-map. Let $<_U$ be an ordering obeyed by $B_{S'}$. Without loss of generality, we take that $x_i <_U x_j$ if $i < j$. Then, because $B_{S'}$

is an I-map of P_D, we have that $P_D(x_i|\pi_i') = P_D(x_i|U_i)$ where U_i is the set of all variables lower ordered according to $<_U$ than x_i. So, $P_D'(U) = \prod_{i=1}^{n} P_D(x_i|\pi_i') = \prod_{i=1}^{n} P_D(x_i|U_i)$, which by the chain rule, equals $P_D(U)$. We conclude that the entropy terms in Formula (4) cancel out. However, since $B_{S'}$ is not a minimum I-map of P_D, we know that compared to B_S at least one extra probability need to be estimated for $B_{S'}$. Therefore, $K' - K > 0$. So, the term $-\frac{1}{2}(K' - K) \cdot \log N$ will go to minus infinity for $N \to \infty$. The term $\log(P(B_S)/P(B_{S'}))$ can once more be neglected. So, Formula (4) is $-\infty$ as was to be shown. □

Note that if $B_{S'}$ is a minimum I-map of P_D in Theorem 3.2 then $H(B_{S'}, D) = H(B_S, D)$ and $K = K'$. So $L(B_{S'}, D) - L(B_S, D)$ depends solely on the term $\log P(B_{S'})/P(B_{S'})$ which for positive priors is a constant.

The above theorem suggests that for databases large enough, network structures that are minimum I-maps are preferred over network structures that are non-minimum I-maps. Note that contrary to Theorem 3.1, Theorem 3.2 is not restricted to network stuctures obeying a certain topological ordering but for all possible network stuctures.

From the theorem, it is easily seen that if a P-map exists for a given distribution then a P-map is preferred over any non P-map for large databases. This property follows from the observation that all P-maps require the same number of probabilities and every I-map that is not a P-map requires more probabilities. Note that a joint probability distribution need not have a unique P-map. Consider for example the simple network $a \to b$. If this network structure is a P-map of some distribution P, then $a \leftarrow b$ is also a P-map of P. However, for every P-map the same number of probabilities needs to be specified, [4].

Corollary 3.1 *Let U be a set of variables and let the prior distribution over all network structures over U be positive. Let P_D be a positive distribution over U such that a P-map exists for P_D. Let D be a database with N cases generated by P_D where N approximates infinity. Let B_S be a P-map of P_D. Then, for any non P-map $B_{S'}$ we have that,*

$$\lim_{N \to \infty} Q(B_S, D) - Q(B_{S'}, D) = -\infty,$$

where Q is either the Bayesian or the MDL measure.

3.2 Finite Database Properties of Quality Measures

In the previous section, the Bayesian measure and the MDL measure has been compared as to their behavior for databases of infinite size. Since infinite size databases never occur, we are interested in non-asymptotic properties of these measures. The following theorem gives some insight on the behavior of the MDL measure.

Theorem 3.3 *Let U be a set of variables and let the prior distribution over all network structures over U be uniform. Let D be a database with N cases over U, $N \geq 10$. Let B_S be a network structure over U, with at least one parent set containing $\log N$ variables or more. Then, a network structure $B_{S'}$ exists such that $L(B_{S'}, D) > L(B_S, D)$.*

Proof: Consider the network stucture B_S. Let x_i be a variable in U that has more than $\log N$ variables in its parent set π_i in B_S. Now, let the network stucture $B_{S'}$ be obtained from B_S by deleting all incoming arcs for this variable x_i, that is, $\pi_i' = \emptyset$. Let K, r_i, and q_i as defined as before for B_S with the MDL measure and let K', r_i', and q_i' be defined for $B_{S'}$. We

prove the theorem by contradiction. Suppose that $B_{S'}$ is not preferred over B_S. We consider the difference $L(B_{S'}, D) - L(B_S, D)$, which equals

$$\log \frac{P(B_{S'})}{P(B_S)} - N \cdot (H(B_{S'}, D) - H(B_S, D)) - \frac{1}{2}(K' - K) \cdot \log N. \tag{5}$$

Since $B_{S'}$ is not preferred over B_S, this difference is not positive.

Because the prior distribution over all network structures is uniform, we have that the term $\log(P(B_{S'})/P(B_S)) = 0$, so this term may be deleted from (5).

Now consider the entropy terms. A property of entropy is that it is maximal $\log r_i$ and minimal zero. So, the minimum value of the difference of the entropy terms $-N \cdot (H(B_{S'}, D) - H(B_S, D))$ is $-N \cdot \log r_i$. Since (5) is not positive, the total difference in cost $-\frac{1}{2}(K' - K) \cdot \log N$ is at most $N \cdot \log r_i$. Therefore, the following inequality holds,

$$\frac{1}{2}(r_i - 1)q_i \cdot \log N - \frac{1}{2}(r_i - 1) \cdot \log N \leq N \cdot \log r_i,$$

where $q_i' = 1$ since $\pi_i' = \emptyset$. Division of this expression by $\frac{1}{2}(r_i - 1) \cdot \log N$ gives,

$$q_i - 1 \leq \frac{2N}{\log N} \cdot \frac{\log r_i}{r_i - 1}.$$

Using the inequality $\log x \leq (x - 1) \log e$, we find,

$$q_i - 1 \leq \frac{2N}{\log N} \log e.$$

Adding one gives,

$$q_i < \frac{2N}{\log N} \log e + 1.$$

Now, observe that the function $f(x) = \frac{2x}{\log x} \log e + 1 - x = \frac{2x - (x-1)\ln x}{\ln x}$ equals 0 for $x \approx 9.4$. Since $f'(x) = -\frac{\ln^2 x - 2\ln x + 2}{\ln^2 x} < 0$, we have that f is a descending function. So,

$$\frac{2N}{\log N} \log e + 1 < N,$$

and thus $q_i < N$. This contradicts the number of parents of x_i in B_S being larger than $\log N$. From this contradiction, we conclude that (5) is positive that $B_{S'}$ is preferred over B_S. □

Theorem 3.3 implies that good search algorithms that use the MDL measure will not select network structures that contain parent sets with more than $\log N$ parents, when N is the number of cases in the database used for learning. A similar result would be expected for the Bayesian measure. However, this is not true.

Consider a database D_n defined recursively by

$$D_1 = \begin{array}{cc} x_1 & y \\ 0 & 0 \\ 1 & 1 \end{array}$$

x_7	x_6	x_5	x_4	x_3	x_2	x_1	y	
0	0	0	0	0	0	0	0	
0	0	0	0	0	0	0	0	D_7
1	0	0	0	0	0	0	1	
1	0	0	0	0	0	0	1	D_6
1	1	0	0	0	0	0	0	
1	1	0	0	0	0	0	0	D_5
1	1	1	0	0	0	0	1	
1	1	1	0	0	0	0	1	D_4
1	1	1	1	0	0	0	0	
1	1	1	1	0	0	0	0	D_3
1	1	1	1	1	0	0	1	
1	1	1	1	1	0	0	1	D_2
1	1	1	1	1	1	0	0	
1	1	1	1	1	1	1	1	D_1

Figure 2: Example of databases D_1 up to D_7.

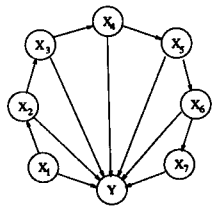

Figure 3: Most probable network for D_7.

where the rows represent the individual cases and the columns the values of the variables; database D_n is constructed from D_{n-1} by adding a column for an extra variable x_n filled with 1s. Further, two identical cases are added where $x_i = 0$ $i = 1, \ldots, n$ and $y = (n+1) \bmod 2$. For example, Figure 2 shows how the database D_7 is build from D_1 up to D_6.

Figure 3 shows the network structure that scores highest with the Bayesian measure for D_7; each node x_i $i = 1, \ldots, 7$ has a node x_{i-1} in their parent set except for node x_1 which has an empty parent set; node y has all nodes $x_1, \ldots, x_7$ in its parent set. This network was found by brute force; the Bayesian measure was calculated for all 1.138.779.265 possible networks. Obviously, the parent set of node y contains more than $\log(14) \approx 3.81$ parents.

To explain why the network stucture shown in Figure 3 is preferred over any other network stucture for the database D_7, we examine the behavior of the Bayesian measure on the database D_n. In general, The Bayesian measure will favor a parent set π_i for a variable x_i if knowledge of the values of variables in π_i gives information as to the distribution of the values x_i as long as the number of instantiations of the parent set in the database q_i is not too large. Now, let us consider the parent sets of the best structure B_S for D_n according to the Bayesian measure. In D_n each node x_i, $i = 2, \ldots, n-1$, may have x_{i-1} in its parent set because when x_{i-1} has the value 1, we find in the database that x_i is 1. In addition, x_{i+1} may be in π_i because when x_{i+1} is 0, we find in the database that x_i is 0. The only variable that would provide more information about the value of x_i is y; note that x_i's value is a function of the values of x_{i-1}, x_{i+1} and y for the cases in the database. However, adding y to the parent set of x_i would increase q_i considerably while the information obtained about the value of x_i would only increase for four cases in the database. Therefore, y will not be in x_i's parent set. For x_1 and x_n a similar argument holds, except that x_0 and x_{n+1} respectively are non-existent and need not be considered.

We now turn to node y. By an inductive argument, it can be shown that any parent set of size k for y scores worse than the parent set $\{x_{n-k+1}, \ldots, x_n\}$ (see appendix Lemma A.1 and A.2). Further, it can be shown that if $\pi_y = \{x_k, \ldots, x_n\}$, for some $1 < k \leq n$, then B_S cannot be the best structure for D_n since taking $\{x_{k-1}, \ldots, x_n\}$ for π_y results in a better quality according to the Bayesian measure (see appendix Lemma A.3). As a consequence, taking $\{x_1, \ldots, x_n\}$ for π_y results in the highest value for the Bayesian measure (see appendix Lemma A.4). Therefore, the best scoring parent set for y is the set $\{x_1, \ldots, x_n\}$. We conclude that for a database with N cases, a network with a parent set with $N/2$ variables can be assigned the highest quality by the Bayesian measure.

So, while the asymptotic behavior of the Bayesian and MDL measure on databases of infinite size is the same, this is not the true for practical cases where only a finite database is available.

4 Conclusions

In this paper, we have investigated the influence of quality measures on learning Bayesian belief networks for both infinite size databases and finite size databases.

We have shown that the Bayesian measure and the minimum description length (MDL) measure have some properties for infinite databases in common. When a given topological ordering on the set of variables is assumed, both measures prefer minimal I-maps obeying this ordering. Furthermore, both measures prefer minimum I-maps, that is, I-maps for which a minimal number of probabilities needs to be specified in order to construct the joint prob-

ability distribution of the belief network. Thus, as a P-map exists, then a P-map will be overwhelmingly preferred over any non P-map. This property justifies the use of quality measures for recovering causality.

However, the behavior of the two measures differ on finite databases. For the MDL measure the sizes of the parent sets in a preferred network structure are bounded by the logarithm of the database size; for the Bayesian measure, this property does not hold, in fact, a parent set may be as large as half the database size. This is due to the MDL measure assigning a cost to each probability that has to be estimated in a Bayesian belief network with the network stucture at hand. The Bayesian measure assigns costs only to probabilities for instantiations of parent sets that appear in the database.

In Bayesian belief network learning applications where a database is relatively small compared to the number of variables, it is very likely that the database contains parts that make the Bayesian measure behave as described, resulting in network stuctures with an enormous number of arcs.

It would be interesting to investigate the behavior of Bayesian measures with other priors than Cooper and Herskovits use; instead of a uniform prior over all probability tables, a uniform prior over all network stuctures may be an interesting alternative.

Acknowledgements

I would like to thank Linda van der Gaag for her helpful comments that improved the presentation of this paper considerably.

Appendix

Let for the four lemmas that follow D_n be a database as defined in Section 3.2 with nodes $\{x_1, \ldots, x_n, y\}$. Let $m_y(\pi_y)$ be the contribution to the Bayesian measure when y has parent set π_y, defined by

$$m_y(\pi_y) = \prod_{j=1}^{q} \frac{N_{j0}! N_{j1}!}{(N_j + 1)!}.$$

where q is the number of unique instantiations of π_y that can be found in the database, N_{jk} the number of cases for which y takes value k and π_y is instantiated as w_j, and $N_j = N_{j0} + N_{j1}$.

Further, we consider two parent sets π_y and π'_y. Let q', N'_{j0}, N'_{j1}, and N'_j be the values of q, N_{j0}, N_{j1}, and N_j applied to π'_y.

Lemma A.1 *Let* $\pi'_y = \{x_{j-k}, x_{j-k+1}, \ldots, x_j\}$ $(k \geq 0,\ k < j < n)$ *and let* $\pi_y = \{x_{n-k}, x_{n-k+1}, \ldots, x_n\}$ *then* $m(\pi'_y) < m(\pi_y)$.

Proof: Let b be the number of counts of 0's for the last instantiation of π_y at the bottom of the database as depicted in Figure 2 and let a be the number of 0's for the first instantiation for which $N_j \neq 2$. We can write $m(\pi'_y)$ as

$$\frac{a!(a(\pm\mathbf{2}))!}{(2a + 1(\pm\mathbf{2}))!} \left(\frac{2!}{3!}\right)^{k-1} \frac{b!(b(+\mathbf{2}))!}{(2b + 1(+\mathbf{2}))!}, \tag{6}$$

where $(\pm\mathbf{2})$ are optional terms representing the case that the number of cases with $y = 1$ more, less, or equal to $y = 0$ for the first instantiation of π'_y. The term $(+\mathbf{2})$ is an optional

term representing that there may be two cases more with $y = 1$ than with $y = 0$ for the last instantiation. So, there are six ways to interpret Formula 6. Likewise, we can write for $m(\pi_y)$,

$$\left(\frac{2!}{3!}\right)^k \frac{(a+b+d-2)!(a+b-d(\pm\mathbf{2})(+\mathbf{2}))!}{(2a+2b+1-2(\pm\mathbf{2})(+\mathbf{2}))!}, \tag{7}$$

where d is 0 or 2 to represent that it is possible that in the first instantiation, the number of cases with $y = 0$ may vary from $y = 1$. We show that (7) is an upper bound of (6) and thus that $m(\pi'_y) < m(\pi_y)$. By writing $x^{\underline{y}}$ for $\frac{x!}{(x-y)!}$ and grouping common terms, we can write (7) as,

$$\frac{a!(a(\pm\mathbf{2}))!}{(2a+1(\pm\mathbf{2}))!}\left(\frac{2!}{3!}\right)^{k-1}\left(\frac{2!}{3!}\frac{(a+b+d-2)^{\underline{b+d-2}}(a+b-d(\pm\mathbf{2})(+\mathbf{2}))^{\underline{b-d(+2)}}}{(2a+2b-1(\pm\mathbf{2})(+\mathbf{2}))^{\underline{2b-2(+2)}}}\right). \tag{8}$$

The first two terms are exactly the same as in (6). So, let us concentrate on the third term,

$$\frac{(a+b+d-2)^{\underline{b+d-2}}(a+b-d(\pm\mathbf{2})(+\mathbf{2}))^{\underline{b-d(+2)}}}{(2a+2b-1(\pm\mathbf{2})(+\mathbf{2}))^{\underline{2b-2(+2)}}}\frac{2!}{3!}. \tag{9}$$

We proceed by comparing 'peeling' terms of the form $\frac{x \cdot y}{w \cdot z}$ from (9) starting with as high as possible values of $x \cdot y$ and $w \cdot z$. We can write (9) as,

$$\frac{(a+b+d-2)(a+b-d(\pm\mathbf{2})(+\mathbf{2}))}{(2a+2b-1(\pm\mathbf{2})(+\mathbf{2}))(2a+2b-2(\pm\mathbf{2})(+\mathbf{2}))} \cdot \frac{(a+b+d-3)^{\underline{b+d-3}}(a+b-d-1(\pm\mathbf{2})(+\mathbf{2}))^{\underline{b-d-1(+2)}}}{(2a+2b-3(\pm\mathbf{2})(+\mathbf{2}))^{\underline{2b-4(+2)}}}\frac{2!}{3!}. \tag{10}$$

By inspection, we find that,

$$\frac{(a+b+d-2)(a+b-d(\pm\mathbf{2})(+\mathbf{2}))}{(2a+2b-1(\pm\mathbf{2})(+\mathbf{2}))(2a+2b-2(\pm\mathbf{2})(+\mathbf{2}))} > \frac{b(b(+\mathbf{2}))}{(2b+1(+\mathbf{2}))(2b(+\mathbf{2}))},$$

and therefore, (10) is larger than,

$$\frac{b(b(+\mathbf{2}))}{(2b+1(+\mathbf{2}))(2b(+\mathbf{2}))}\frac{(a+b+d-3)^{\underline{b+d-3}}(a+b-d-1(\pm\mathbf{2})(+\mathbf{2}))^{\underline{b-d-1(+2)}}}{(2a+2b-3(\pm\mathbf{2})(+\mathbf{2}))^{\underline{2b-4(+2)}}}\frac{2!}{3!}.$$

Note that the last part is of the same form as (9) but now with $b-1$ instead of b, thus we can repeat this step. After $b-2$ times applying this step, we get as lower bound of (10),

$$\frac{b^{\underline{b-2}}(b(+\mathbf{2}))^{\underline{b-2}}}{(2b+1(+\mathbf{2}))^{\underline{2b-4}}}\frac{(a+d-2)^{\underline{d}}(a-d(\pm\mathbf{2})(+\mathbf{2}))^{\underline{-d(+2)}}}{(2a+3(\pm\mathbf{2})(+\mathbf{2}))^{\underline{2(+2)}}}\frac{2!}{3!}. \tag{11}$$

By inspection, we find that,

$$\frac{(a+d-2)^{\underline{d}}(a-d(\pm\mathbf{2})(+\mathbf{2}))^{\underline{-d(+2)}}}{(2a+3(\pm\mathbf{2})(+\mathbf{2}))^{\underline{2(+2)}}}\frac{2!}{3!} > \frac{2!(2(+\mathbf{2}))!}{5(+\mathbf{2})!}.$$

Therefore, (11) is larger than,

$$\frac{b^{\underline{b-2}}(b(+\mathbf{2}))^{\underline{b-2}}}{(2b+1(+\mathbf{2}))^{\underline{2b-4}}}\frac{2!(2(+\mathbf{2}))!}{(5(+\mathbf{2})!)},$$

which is equal to $\frac{b!(b(+\mathbf{2}))!}{(2b+1(+\mathbf{2}))!}$. So, we have shown that (9) is larger than $\frac{b!(b(+\mathbf{2}))!}{(2b+1(+\mathbf{2}))!}$. Therefore, (8) is larger than (6) which completes the proof. □

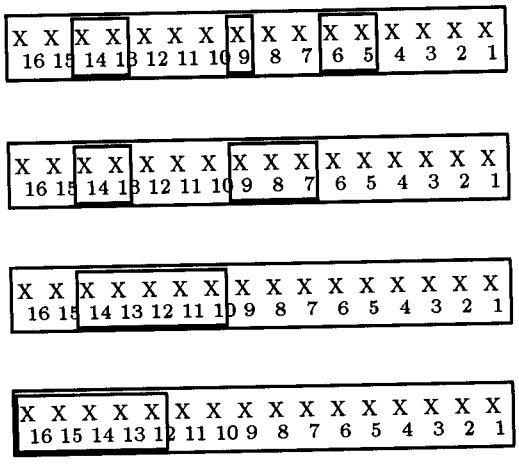

Figure 4: Example of parent set of y

Lemma A.2 *Let* $\pi_y = \{x_{n-k}, x_{n-k+1}, \ldots, x_n\}$ *and* $\pi'_y \subset \{x_1, \ldots, x_n\}$, $|\pi'_y| = k+1 > 0$, $\pi'_y \neq \pi_y$ *then* $m(\pi'_y) < m(\pi_y)$.

Proof: Regard parent set π'_y as groups of consecutive nodes. In Figure 4, an example with four groups is shown for D_{16}. Consider the score of the parent set on top, and the one right below. In comparing their contribution to the Bayesian measure, all instantiations where $x_8 = 0$ need not be considered. And in fact, one could act as if D_8 was used.

By Lemma A.1, we find that the parent set on top scores lower than the one right below. By the same argument, this parent set scores less than the one below it, and the one on the bottom scores highest of them all.

So, by shifting groups of nodes in the parent set we find parent sets that score better and better after each shift, where the parent set π_y has highest score.

□

Lemma A.3 *Let* $\pi'_y = \{x_{n-k}, x_{n-k+1}, \ldots, x_n\}$ $(k < n-1)$ *and let* $\pi_y = \pi'_y \cup \{x_{n-k-1}\}$ *then* $m(\pi'_y) < m(\pi_y)$.

Proof: By definition, π_y is equal to π'_y united with x_{n-k-1}. Note that $\{w_1, \ldots, w_q\}$ differs from $\{w'_1, \ldots, w'_{q'}\}$ only for the case that all $x_{n-k}, \ldots, x_n$ are 1 and x_{n-k-1} is 0. So,

$$m(\pi'_y) - m(\pi_y) = \prod_{j'=1}^{q'} \frac{N'_{j0}! N'_{j1}!}{(N'_j+1)!} - \prod_{j=1}^{q} \frac{N_{j0}! N_{j1}!}{(N_j+1)!}.$$

Let b be the number of counts of 0's for the last instantiation of π_y at the bottom of the database as depicted in Figure 2. Since all instantiation are the same for $x_{n-k} = 0$, this can be written as

$$C\left(\frac{b!(b(+\mathbf{2}))!}{(2b+1(+\mathbf{2}))!} - \frac{2!}{3!} \cdot \frac{b!(b-2(+\mathbf{2}))!}{(2b-2(+\mathbf{2}))!}\right),$$

where C is a positive constant. By bringing common terms out of the brackets, we get for $k < n-2$,

$$C'\left(\frac{(b(+\mathbf{2}))(b-1(+\mathbf{2}))}{(2b+1(+\mathbf{2}))(2b(+\mathbf{2}))} - \frac{2!}{3!}\right),$$

which by inspection is less than 0. And, if $k = n-2$ we get,

$$C'\left(\frac{1!3!}{5!} - \frac{2!}{3!}\frac{1!1!}{3!}\right) = C'\left(\frac{1}{20} - \frac{1}{18}\right),$$

which also is less than 0. □

Lemma A.4 *Let $\pi'_y \subsetneq \{x_1, \ldots, x_n\}$ and let $\pi_y = \{x_1, \ldots, x_n\}$ then $m(\pi'_y) < m(\pi_y)$.*

Proof: Follows directly from the previous three lemmas: for any $\pi'_y \subsetneq \{x_1, \ldots, x_n\}$ with $|\pi_y| = k-1$ we have the parent set $\pi''_y = \{x_{n-k}, \ldots, x_n\}$ such that $m(\pi'_y) < m(\pi''_y)$ by Lemma A.2. By repetitively applying Lemma A.3, thus extending π''_y node by node, we find that $m(\pi''_y) < m(\pi_y)$. □

References

[1] B. Abramson and A. Finizza. Using belief networks to forecast oil prices. *International Journal of Forecasting*, 7:299–315, 1991.

[2] I. Beinlich, H. Suermondt, R. Chavez, and G. Cooper. The alarm monitoring system: a case study with two probabilistic inference techniques for belief networks. In *Proceedings Artificial Intelligence in Medical Care*, pages 247–256, 1989.

[3] R.R. Bouckaert. Belief network construction using the minimum description length principle. In *Proceedings ECSQARU*, pages 41–48, 1993.

[4] R.R. Bouckaert. Belief network construction using the minimum description length principle. Technical Report RUU-CS-94-27, Utrecht University, The Netherlands, 1993.

[5] G.F. Cooper and E. Herskovits. A Bayesian method for the induction of probabilistic networks from data. *Machine Learning*, pages 309–347, 1992.

[6] A.P. Dawid. Conditional independence in statistical theory. *Journal of the Royal Statistical Society B*, 41(1):1–31, 1979.

[7] M. Henrion. An introduction to algorithms for inference in belief nets. In *Proceedings Uncertainty in Artificial Intelligence 6*, pages 129–138, 1990.

[8] E. Herskovits. *Computer-based probabilistic-network construction.* PhD thesis, Section of Medical Informatics, University of Pittsburgh, 1991.

[9] W. Lam and F. Bacchus. Learning Bayesian belief networks, an approach based on the MDL principle. *Computational Intelligence*, 10(4), 1994.

[10] S.L. Lauritzen and D.J. Spiegelhalter. Local computations with probabilities on graphical structures and their applications to expert systems (with discussion). *Journal of the Royal Statistical Society B*, 50:157–224, 1988.

[11] J. Pearl. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference.* Morgan Kaufman, inc., San Mateo, CA, 1988.

[12] J. Pearl, D. Geiger, and T. Verma. The logic of influence diagrams. In R.M. Oliver and J.Q. Smith, editors, *Influence Diagrams, Belief Nets and Decision Analysis*, pages 67–87. John Wiley & Sons Ltd., 1990.

[13] M.A. Shwe, B. Middleton, D.E. Heckerman, M. Henrion, E. Horvitz, H. Lehmann, and G. Cooper. Probabilistic diagnosis using a reformulation of the INTERNIST-1/QMR knowledge base I: The probabilistic model and inference algorithms. *Methods of Information in Medicine*, 30:241–255, 1991.

[14] J. Suzuki. A construction of Bayesian networks from databases based on the MDL principle. In *Proceedings Uncertainy in Artificial Intelligence 9*, 1993.

[15] D. Wedelin. *Efficient Algorithms for Probabilistic Inference, Combinatorial Optimization and the Discovery of Causal Structure from Data.* PhD thesis, Department of Computer Sciences, Chalmers University of Technology Götenborg, Sweden, 1993.