Discriminative learning of Bayesian networks via factorized conditional log-likelihood

Carvalho, Alexandra M.

2011

Carvalho , A M , Roos , T T , Oliveira , A L & Myllymäki , P 2011 , ' Discriminative learning of Bayesian networks via factorized conditional log-likelihood ' , Journal of Machine Learning Research , vol. 12 , pp. 2181-2210 . < http://jmlr.csail.mit.edu/papers/v12/carvalho11a.html >

http://hdl.handle.net/10138/28459

publishedVersion

Journal of Machine Learning Research 12 (2011) 2181-2210 Submitted 4/11; Published 7/11

Discriminative Learning of Bayesian Networks via Factorized Conditional Log-Likelihood

Alexandra M. Carvalho ASMC@INESC-ID.PT
Department of Electrical Engineering
Instituto Superior Técnico, Technical University of Lisbon
INESC-ID, R. Alves Redol 9
1000-029 Lisboa, Portugal

Teemu Roos TEEMU.ROOS@CS.HELSINKI.FI
Department of Computer Science
Helsinki Institute for Information Technology
P.O. Box 68
FI-00014 University of Helsinki, Finland

Arlindo L. Oliveira AML@INESC-ID.PT
Department of Computer Science and Engineering
Instituto Superior Técnico, Technical University of Lisbon
INESC-ID, R. Alves Redol 9
1000-029 Lisboa, Portugal

Petri Myllymäki PETRI.MYLLYMAKI@CS.HELSINKI.FI
Department of Computer Science
Helsinki Institute for Information Technology
P.O. Box 68
FI-00014 University of Helsinki, Finland

Editor: Russell Greiner

Abstract

We propose an efficient and parameter-free scoring criterion, the factorized conditional log-likelihood ($\hat{\text{f}}$CLL), for learning Bayesian network classifiers. The proposed score is an approximation of the conditional log-likelihood criterion. The approximation is devised in order to guarantee decomposability over the network structure, as well as efficient estimation of the optimal parameters, achieving the same time and space complexity as the traditional log-likelihood scoring criterion. The resulting criterion has an information-theoretic interpretation based on interaction information, which exhibits its discriminative nature. To evaluate the performance of the proposed criterion, we present an empirical comparison with state-of-the-art classifiers. Results on a large suite of benchmark data sets from the UCI repository show that $\hat{\text{f}}$CLL-trained classifiers achieve at least as good accuracy as the best compared classifiers, using significantly less computational resources.

Keywords: Bayesian networks, discriminative learning, conditional log-likelihood, scoring criterion, classification, approximation

1. Introduction

Bayesian networks have been widely used for classification, see Friedman et al. (1997), Grossman and Domingos (2004), Su and Zhang (2006) and references therein. However, they are often outperformed by much simpler methods (Domingos and Pazzani, 1997; Friedman et al., 1997). One of the likely causes for this appears to be the use of so called *generative learning* methods in choosing the Bayesian network structure as well as its parameters. In contrast to generative learning, where the goal is to be able to describe (or generate) the entire data, *discriminative learning* focuses on the capacity of a model to discriminate between different classes of instances. Unfortunately, discriminative learning of Bayesian network classifiers has turned out to be computationally much more challenging than generative learning. This led Friedman et al. (1997) to bring up the question: are there heuristic approaches that allow efficient discriminative learning of Bayesian network classifiers?

During the past years different discriminative approaches have been proposed, which tend to decompose the problem into two tasks: (i) discriminative structure learning, and (ii) discriminative parameter learning. Greiner and Zhou (2002) were among the first to work along these lines. They introduced a discriminative parameter learning algorithm, called the *Extended Logistic Regression* (ELR) algorithm, that uses gradient descent to maximize the *conditional log-likelihood* (CLL) of the class variable given the other variables. Their algorithm can be applied to an arbitrary Bayesian network structure. However, they only considered *generative* structure learning methods. Greiner and Zhou (2002) demonstrated that their parameter learning method, although computationally more expensive than the usual generative approach that only involves counting relative frequencies, leads to improved parameter estimates. More recently, Su et al. (2008) have managed to significantly reduce the computational cost by proposing an alternative discriminative parameter learning method, called the *Discriminative Frequency Estimate* (DFE) algorithm, that exhibits nearly the same accuracy as the ELR algorithm but is considerably more efficient.

Full structure and parameter learning based on the ELR algorithm is a burdensome task. Employing the procedure of Greiner and Zhou (2002) would require a new gradient descent for each candidate network at each search step, turning the method computationally infeasible. Moreover, even in parameter learning, ELR is not guaranteed to find globally optimal CLL parameters. Roos et al. (2005) have showed that globally optimal solutions can be guaranteed *only* for network structures that satisfy a certain graph-theoretic property, including for example, the naive Bayes and tree-augmented naive Bayes (TAN) structures (see Friedman et al., 1997) as special cases. The work by Greiner and Zhou (2002) supports this result empirically by demonstrating that their ELR algorithm is successful when combined with (generatively learned) TAN classifiers.

For discriminative structure learning, Kontkanen et al. (1998) and Grossman and Domingos (2004) propose to choose network structures by maximizing CLL while choosing parameters by maximizing the parameter posterior or the (joint) *log-likelihood* (LL). The *BNC algorithm* of Grossman and Domingos (2004) is actually very similar to the hill-climbing algorithm of Heckerman et al. (1995), except that it uses CLL as the primary objective function. Grossman and Domingos (2004) also experiment with full structure and parameter optimization for CLL. However, they found that full optimization does not produce better results than those obtained by the much simpler approach where parameters are chosen by maximizing LL.

The contribution of this paper is to present two criteria similar to CLL, but with much better computational properties. The criteria can be used for efficient learning of augmented naive Bayes

classifiers. We mostly focus on structure learning. Compared to the work of Grossman and Domingos (2004), our structure learning criteria have the advantage of being *decomposable*, a property that enables the use of simple and very efficient search heuristics. For the sake of simplicity, we assume a binary valued class variable when deriving our results. However, the methods are directly applicable to multi-class classification, as demonstrated in the experimental part (Section 5).

Our first criterion is the *approximated conditional log-likelihood* (aCLL). The proposed score is the minimum variance unbiased (MVU) approximation to CLL under a class of uniform priors on certain parameters of the joint distribution of the class variable and the other attributes. We show that for most parameter values, the approximation error is very small. However, the aCLL criterion still has two unfavorable properties. First, the parameters that maximize aCLL are hard to obtain, which poses problems at the parameter learning phase, similar to those posed by using CLL directly. Second, the criterion is not well-behaved in the sense that it sometimes diverges when the parameters approach the usual relative frequency estimates (maximizing LL).

In order to solve these two shortcomings, we devise a second approximation, the *factorized conditional log-likelihood* ($\hat{\text{f}}$CLL). The $\hat{\text{f}}$CLL approximation is uniformly bounded, and moreover, it is maximized by the easily obtainable relative frequency parameter estimates. The $\hat{\text{f}}$CLL criterion allows a neat interpretation as a sum of LL and another term involving the *interaction information* between a node, its parents, and the class variable; see Pearl (1988), Cover and Thomas (2006), Bilmes (2000) and Pernkopf and Bilmes (2005).

To gauge the performance of the proposed criteria in classification tasks, we compare them with several popular classifiers, namely, *tree augmented naive Bayes* (TAN), *greedy hill-climbing* (GHC), C4.5, *k-nearest neighbor* (*k*-NN), *support vector machine* (SVM) and *logistic regression* (LogR). On a large suite of benchmark data sets from the UCI repository, $\hat{\text{f}}$CLL-trained classifiers outperform, with a statistically significant margin, their generatively-trained counterparts, as well as C4.5, *k*-NN and LogR classifiers. Moreover, $\hat{\text{f}}$CLL-optimal classifiers are comparable with ELR induced ones, as well as SVMs (with linear, polynomial, and radial basis function kernels). The advantage of $\hat{\text{f}}$CLL with respect to these latter classifiers is that it is computationally as efficient as the LL scoring criterion, and considerably more efficient than ELR and SVMs.

The paper is organized as follows. In Section 2 we review some basic concepts of Bayesian networks and introduce our notation. In Section 3 we discuss generative and discriminative learning of Bayesian network classifiers. In Section 4 we present our scoring criteria, followed by experimental results in Section 5. Finally, we draw some conclusions and discuss future work in Section 6. The proofs of the results stated throughout this paper are given in the Appendix.

2. Bayesian Networks

In this section we introduce some notation, while recalling relevant concepts and results concerning discrete Bayesian networks.

Let X be a *discrete random variable* taking values in a countable set $\mathcal{X} \subset \mathbb{R}$. In all what follows, the domain $\mathcal{X}$ is finite. We denote an n-dimensional *random vector* by $\mathbf{X} = (X_1, \ldots, X_n)$ where each component X_i is a random variable over $\mathcal{X}_i$. For each variable X_i, we denote the elements of $\mathcal{X}_i$ by $x_{i1}, \ldots, x_{ir_i}$ where r_i is the number of values X_i can take. The probability that $\mathbf{X}$ takes value $\mathbf{x}$ is denoted by $P(\mathbf{x})$, conditional probabilities $P(\mathbf{x} \mid \mathbf{z})$ being defined correspondingly.

A *Bayesian network* (BN) is defined by a pair $B = (G, \Theta)$, where G refers to the graph structure, and Θ are the parameters. The structure $G = (V, E)$ is a *directed acyclic graph* (DAG) with vertices

(nodes) V, each corresponding to one of the random variables X_i, and edges E representing direct dependencies between the variables. The (possibly empty) set of nodes from which there is an edge to node X_i is called the set of the *parents* of X_i, and denoted by Π_{X_i}. For each node X_i, we denote the number of possible *parent configurations* (vectors of the parents' values) by q_i, the actual parent configurations being ordered (arbitrarily) and denoted by $w_{i1}, \ldots, w_{iq_i}$. The *parameters*, $\Theta = \{\theta_{ijk}\}_{i\in\{1,\ldots,n\},\, j\in\{1,\ldots,q_i\},k\in\{1,\ldots,r_i\}}$, determine the *local distributions* in the network via

$$P_B(X_i = x_{ik} \mid \Pi_{X_i} = w_{ij}) = \theta_{ijk}.$$

The local distributions define a unique joint probability distribution over $\mathbf{X}$ given by

$$P_B(X_1, \ldots, X_n) = \prod_{i=1}^{n} P_B(X_i \mid \Pi_{X_i}).$$

The conditional independence properties pertaining to the joint distribution are essentially determined by the network structure. Specifically, X_i is conditionally independent of its non-descendants given its parents Π_{X_i} in G (Pearl, 1988).

Learning unrestricted Bayesian networks from data under typical scoring criteria is NP-hard (Chickering et al., 2004). This result has led the Bayesian network community to search for the largest subclass of network structures for which there is an efficient learning algorithm. First attempts confined the network to tree structures and used Edmonds (1967) and Chow and Liu (1968) optimal branching algorithms. More general classes of Bayesian networks have eluded efforts to develop efficient learning algorithms. Indeed, Chickering (1996) showed that learning the structure of a Bayesian network is NP-hard even for networks constrained to have in-degree at most two. Later, Dasgupta (1999) showed that even learning an optimal *polytree* (a DAG in which there are not two different paths from one node to another) with maximum in-degree two is NP-hard. Moreover, Meek (2001) showed that identifying the best *path structure*, that is, a total order over the nodes, is hard. Due to these hardness results exact polynomial-time algorithms for learning Bayesian networks have been restricted to tree structures.

Consequently, the standard methodology for addressing the problem of learning Bayesian networks has become heuristic score-based learning where a *scoring criterion* ϕ is considered in order to quantify the capability of a Bayesian network to explain the observed data. Given data $D = \{\mathbf{y}_1, \ldots, \mathbf{y}_N\}$ and a scoring criterion ϕ, the task is to find a Bayesian network B that maximizes the score $\phi(B, D)$. Many search algorithms have been proposed, varying both in terms of the formulation of the search space (network structures, equivalence classes of network structures and orderings over the network variables), and in the algorithm to search the space (greedy hill-climbing, simulated annealing, genetic algorithms, tabu search, etc). The most common scoring criteria are reviewed in Carvalho (2009) and Yang and Chang (2002). We refer the interested reader to newly developed scoring criteria to the works of de Campos (2006) and Silander et al. (2010).

Score-based learning algorithms can be extremely efficient if the employed scoring criterion is decomposable. A scoring criterion ϕ is said to be *decomposable* if the score can be expressed as a sum of local scores that depends only on each node and its parents, that is, in the form

$$\phi(B, D) = \sum_{i=1}^{n} \phi_i(\Pi_{X_i}, D).$$

One of the most common criteria is the *log-likelihood* (LL), see Heckerman et al. (1995):

$$\mathrm{LL}(B \mid D) \quad = \quad \sum_{t=1}^{N} \log P_B(y_t^1, \ldots, y_t^n) = \sum_{i=1}^{n} \sum_{j=1}^{q_i} \sum_{k=1}^{r_i} N_{ijk} \log \theta_{ijk},$$

which is clearly decomposable.

The *maximum likelihood* (ML) parameters that maximize LL are easily obtained as the *observed frequency estimates* (OFE) given by

$$\hat{\theta}_{ijk} = \frac{N_{ijk}}{N_{ij}}, \tag{1}$$

where N_{ijk} denotes the number of instances in D where $X_i = x_{ik}$ and $\Pi_{X_i} = w_{ij}$, and $N_{ij} = \sum_{k=1}^{r_i} N_{ijk}$. Plugging these estimates back into the LL criterion yields

$$\widehat{\mathrm{LL}}(G \mid D) \quad = \quad \sum_{i=1}^{n} \sum_{j=1}^{q_i} \sum_{k=1}^{r_i} N_{ijk} \log \left(\frac{N_{ijk}}{N_{ij}} \right).$$

The notation with G as the argument instead of $B = (G, \Theta)$ emphasizes that once the use of the OFE parameters is decided upon, the criterion is a function of the network structure, G, only.

The $\widehat{\mathrm{LL}}$ scoring criterion tends to favor complex network structures with many edges since adding an edge never decreases the likelihood. This phenomenon leads to *overfitting* which is usually avoided by adding a complexity penalty to the log-likelihood or by restricting the network structure.

3. Bayesian Network Classifiers

A *Bayesian network classifier* is a Bayesian network over $\mathbf{X} = (X_1, \ldots, X_n, C)$, where C is a class variable, and the goal is to classify instances $(X_1, \ldots, X_n)$ to different classes. The variables $X_1, \ldots, X_n$ are called *attributes*, or *features*. For the sake of computational efficiency, it is common to restrict the network structure. We focus on *augmented naive Bayes classifiers*, that is, Bayesian network classifiers where the class variable has no parents, $\Pi_C = \emptyset$, and all attributes have at least the class variable as a parent, $C \in \Pi_{X_i}$ for all X_i.

For convenience, we introduce a few additional notations that apply to augmented naive Bayes models. Let the class variable C range over s distinct values, and denote them by $z_1, \ldots, z_s$. Recall that the parents of X_i are denoted by Π_{X_i}. The parents of X_i without the class variable are denoted by $\Pi^*_{X_i} = \Pi_{X_i} \setminus \{C\}$. We denote the number of possible configurations of the parent set $\Pi^*_{X_i}$ by q_i^*; hence, $q_i^* = \prod_{X_j \in \Pi^*_{X_i}} r_j$. The j'th configuration of $\Pi^*_{X_i}$ is represented by w^*_{ij}, with $1 \leq j \leq q_i^*$. Similarly to the general case, local distributions are determined by the corresponding parameters

$$\begin{aligned} &P(C = z_c) = \theta_c, \\ &P(X_i = x_{ik} \mid \Pi^*_{X_i} = w^*_{ij}, C = z_c) = \theta_{ijck}. \end{aligned}$$

We denote by N_{ijck} the number of instances in the data D where $X_i = x_{ik}$, $\Pi^*_{X_i} = w^*_{ij}$ and $C = z_c$. Moreover, the following short-hand notations will become useful:

$$\begin{aligned} N_{ij*k} &= \sum_{c=1}^{s} N_{ijck}, \qquad & N_{ij*} &= \sum_{k=1}^{r_i} \sum_{c=1}^{s} N_{ijck}, \\ N_{ijc} &= \sum_{k=1}^{r_i} N_{ijck}, \qquad & N_c &= \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{q_i^*} \sum_{k=1}^{r_i} N_{ijck}. \end{aligned}$$

Finally, we recall that the total number of instances in the data D is N.

The ML estimates (1) become now

$$\hat{\theta}_c = \frac{N_c}{N}, \qquad \text{and} \qquad \hat{\theta}_{ijck} = \frac{N_{ijck}}{N_{ijc}}, \tag{2}$$

which can again be plugged into the LL criterion:

$$\begin{aligned} \widehat{\mathrm{LL}}(G \mid D) &= \sum_{t=1}^{N} \log P_B(y_t^1, \ldots, y_t^n, c_t) \\ &= \sum_{c=1}^{s} \left(N_c \log \left(\frac{N_c}{N} \right) + \sum_{i=1}^{n} \sum_{j=1}^{q_i} \sum_{k=1}^{r_i} N_{ijck} \log \left(\frac{N_{ijck}}{N_{ijc}} \right) \right). \end{aligned} \tag{3}$$

As mentioned in the introduction, if the goal is to discriminate between instances belonging to different classes, it is more natural to consider the *conditional log-likelihood* (CLL), that is, the probability of the class variable given the attributes, as a score:

$$\mathrm{CLL}(B \mid D) = \sum_{t=1}^{N} \log P_B(c_t \mid y_t^1, \ldots, y_t^n).$$

Friedman et al. (1997) noticed that the log-likelihood can be rewritten as

$$\mathrm{LL}(B \mid D) = \mathrm{CLL}(B \mid D) + \sum_{t=1}^{N} \log P_B(y_t^1, \ldots, y_t^n). \tag{4}$$

Interestingly, the objective of generative learning is precisely to maximize the whole sum, whereas the goal of discriminative learning consists on maximizing only the first term in (4). Friedman et al. (1997) attributed the underperformance of learning methods based on LL to the term $\mathrm{CLL}(B \mid D)$ being potentially much smaller than the second term in Equation (4). Unfortunately, CLL does not decompose over the network structure, which seriously hinders structure learning, see Bilmes (2000); Grossman and Domingos (2004). Furthermore, there is no closed-form formula for optimal parameter estimates maximizing CLL, and computationally more expensive techniques such as ELR are required (Greiner and Zhou, 2002; Su et al., 2008).

4. Factorized Conditional Log-Likelihood Scoring Criterion

The above shortcomings of earlier discriminative approaches to learning Bayesian network classifiers, and the CLL criterion in particular, make it natural to explore good approximations to the CLL that are more amenable to efficient optimization. More specifically, we now set out to construct approximations that are *decomposable*, as discussed in Section 2.

4.1 Developing a New Scoring Criterion

For simplicity, assume that the class variable is binary, $C = \{0, 1\}$. For the binary case the conditional probability of the class variable can then be written as

$$P_B(c_t \mid y_t^1, \ldots, y_t^n) = \frac{P_B(y_t^1, \ldots, y_t^n, c_t)}{P_B(y_t^1, \ldots, y_t^n, c_t) + P_B(y_t^1, \ldots, y_t^n, 1 - c_t)}. \tag{5}$$

For convenience, we denote the two terms in the denominator as

$$\begin{aligned} U_t &= P_B(y_t^1,\ldots,y_t^n,c_t) \qquad \text{and} \\ V_t &= P_B(y_t^1,\ldots,y_t^n,1-c_t), \end{aligned} \tag{6}$$

so that Equation (5) becomes simply

$$P_B(c_t \mid y_t^1,\ldots,y_t^n) = \frac{U_t}{U_t+V_t}.$$

We stress that both U_t and V_t depend on B, but for the sake of readability we omit B in the notation. Observe that while $(y_t^1,\ldots,y_t^n,c_t)$ is the t'th sample in the data set D, the vector $(y_t^1,\ldots,y_t^n,1-c_t)$, which we call the *dual sample* of $(y_t^1,\ldots,y_t^n,c_t)$, may or may not occur in D.

The log-likelihood (LL), and the conditional log-likelihood (CLL) now take the form

$$\begin{aligned} \mathrm{LL}(B \mid D) &= \sum_{t=1}^{N} \log U_t, \qquad \text{and} \\ \mathrm{CLL}(B \mid D) &= \sum_{t=1}^{N} \log U_t - \log(U_t+V_t). \end{aligned}$$

Recall that our goal is to derive a decomposable scoring criterion. Unfortunately, even though $\log U_t$ decomposes, $\log(U_t+V_t)$ does not.

Now, let us consider approximating the log-ratio

$$f(U_t,V_t) = \log\left(\frac{U_t}{U_t+V_t}\right),$$

by functions of the form

$$\hat{f}(U_t,V_t) = \alpha \log U_t + \beta \log V_t + \gamma,$$

where α, β, and γ are real numbers to be chosen so as to minimize the approximation error. Since the accuracy of the approximation obviously depends on the values of U_t and V_t as well as the constants α, β, and γ, we need to make some assumptions about U_t and V_t in order to determine suitable values of α, β and γ. We explicate one possible set of assumptions, which will be seen to lead to a good approximation for a very wide range of U_t and V_t. We emphasize that the role of the assumptions is to aid in arriving at good choices of the constants α, β, and γ, after which we can dispense with the assumptions—they need not, in particular, hold true exactly.

Start by noticing that $R_t = 1-(U_t+V_t)$ is the probability of observing neither the t'th sample nor its dual, and hence, the triplet (U_t,V_t,R_t) are the parameters of a trinomial distribution. We assume, for the time being, that no knowledge about the values of the parameters (U_t,V_t,R_t) is available. Therefore, it is natural to assume that (U_t,V_t,R_t) follows the uniform Dirichlet distribution, Dirichlet$(1,1,1)$, which implies that

$$(U_t,V_t) \sim \mathrm{Uniform}(\Delta^2), \tag{7}$$

where $\Delta^2 = \{(x,y) : x+y \leq 1 \text{ and } x,y \geq 0\}$ is the 2-simplex set. However, with a brief reflection on the matter, we can see that such an assumption is actually rather unrealistic. Firstly, by inspecting the total number of possible observed samples, we expect, R_t to be relatively large (close to 1).

In fact, U_t and V_t are expected to become exponentially small as the number of attributes grows. Therefore, it is reasonable to assume that

$$U_t, V_t \leq p < \frac{1}{2} < R_t$$

for some $0 < p < \frac{1}{2}$. Combining this constraint with the uniformity assumption, Equation (7), yields the following assumption, which we will use as a basis for our further analysis.

Assumption 1 There exists a small positive $p < \frac{1}{2}$ such that

$$(U_t, V_t) \sim \text{Uniform}(\Delta^2)|_{U_t, V_t \leq p} = \text{Uniform}([0,p] \times [0,p]).$$

Assumption 1 implies that U_t and V_t are uniform i.i.d. random variables over $[0,p]$ for some (possibly unknown) positive real number $p < \frac{1}{2}$. (See Appendix B for an alternative justification for Assumption 1.) As we show below, we do not need to know the actual value of p. Consequently, the envisaged approximation will be robust to the choice of p.

We obtain the best fitting approximation $\hat{f}$ by the least squares method.

Theorem 1 Under Assumption 1, the values of α, β and γ that minimize the *mean square error* (MSE) of $\hat{f}$ w.r.t. f are given by

$$\alpha = \frac{\pi^2 + 6}{24}, \tag{8}$$

$$\beta = \frac{\pi^2 - 18}{24}, \text{ and} \tag{9}$$

$$\gamma = \frac{\pi^2}{12 \ln 2} - \left(2 + \frac{(\pi^2 - 6)\log p}{12}\right), \tag{10}$$

where log is the binary logarithm and ln is the natural logarithm.

We show that the mean difference between $\hat{f}$ and f is zero for all values of p, that is, $\hat{f}$ is *unbiased*.[1] Moreover, we show that $\hat{f}$ is the approximation with the lowest variance among unbiased ones.

Theorem 2 The approximation $\hat{f}$ with α, β, γ defined as in Theorem 1 is the *minimum variance unbiased* (MVU) approximation of f.

Next, we derive the standard error of the approximation $\hat{f}$ in the square $[0,p] \times [0,p]$, which, curiously, does not depend on p. To this end, consider

$$\mu = E[f(U_t, V_t)] = \frac{1}{2\ln(2)} - 2$$

which is a negative value; as it should since f ranges over $(-\infty, 0]$.

1. Herein we apply the nomenclature of estimation theory in the context of approximation. Thus, an approximation is *unbiased* if $E[\hat{f}(U_t,V_t) - f(U_t,V_t)] = 0$ for all p. Moreover, an approximation is *(uniformly) minimum variance unbiased* if the value $E[(\hat{f}(U_t,V_t) - f(U_t,V_t))^2]$ is the lowest for all unbiased approximations and values of p.

Theorem 3 The approximation $\hat{f}$ with α, β, and γ defined as in Theorem 1 has *standard error* given by

$$\sigma = \sqrt{\frac{36+36\pi^2-\pi^4}{288\ln^2(2)} - 2} \approx 0.352$$

and *relative standard error* $\sigma/|\mu| \approx 0.275$.

Figure 1 illustrates the function f as well as its approximation $\hat{f}$ for $(U_t, V_t) \in [0,p] \times [0,p]$ with $p = 0.05$. The approximation error, $f - \hat{f}$ is shown in Figure 2. While the properties established in Theorems 1–3 are useful, we find it even more important that, as seen in Figure 2, the error is close to zero except when either U_t or V_t approaches zero. Moreover, we point out that the choice of $p = 0.05$ used in the figure is not important: having chosen another value would have produced identical graphs except in the scale of the U_t and V_t. In particular, the scale and numerical values on the vertical axis (i.e., in Figure 2, the error) would have been precisely the same.

Using the approximation in Theorem 1 to approximate CLL yields

$$\begin{aligned}
\mathrm{CLL}(B \mid D) &\approx \sum_{t=1}^{N} \alpha \log U_t + \beta \log V_t + \gamma \\
&= \sum_{t=1}^{N} (\alpha+\beta) \log U_t - \beta \log \left(\frac{U_t}{V_t} \right) + \gamma \\
&= (\alpha+\beta)\mathrm{LL}(B \mid D) - \beta \sum_{t=1}^{N} \log \left(\frac{U_t}{V_t} \right) + N\gamma, \qquad (11)
\end{aligned}$$

where constants α, β and γ are given by Equations (8), (9) and (10), respectively. Since we want to maximize $\mathrm{CLL}(B \mid D)$, we can drop the constant $N\gamma$ in the approximation, as maxima are invariant under monotone transformations, and so we can just maximize the following formula, which we call the *approximate conditional log-likelihood* (aCLL):

$$\begin{aligned}
\mathrm{aCLL}(B \mid D) &= (\alpha+\beta)\mathrm{LL}(B \mid D) - \beta \sum_{t=1}^{N} \log \left(\frac{U_t}{V_t} \right) \\
&= (\alpha+\beta)\,\mathrm{LL}(B \mid D) - \beta \sum_{i=1}^{n} \sum_{j=1}^{q_i^*} \sum_{k=1}^{r_i} \sum_{c=0}^{1} N_{ijck} \log \left(\frac{\theta_{ijck}}{\theta_{ij(1-c)k}} \right) \\
&\qquad -\beta \sum_{c=0}^{1} N_c \log \left(\frac{\theta_c}{\theta_{(1-c)}} \right). \qquad (12)
\end{aligned}$$

The fact that $N\gamma$ can be removed from the maximization in (11) is most fortunate, as we eliminate the dependency on p. An immediate consequence of this fact is that we do not need to know the actual value of p in order to employ the criterion.

We are now in the position of having constructed a *decomposable* approximation of the conditional log-likelihood score that was shown to be very accurate for a wide range of parameters U_t and V_t. Due to the dependency of these parameters on Θ, that is, the parameters of the Bayesian network B (recall Equation (6)), the score still requires that a suitable set of parameters is chosen. Finding the parameters maximizing the approximation is, however, difficult; apparently as difficult as finding parameters maximizing the CLL directly. Therefore, whatever computational advantage

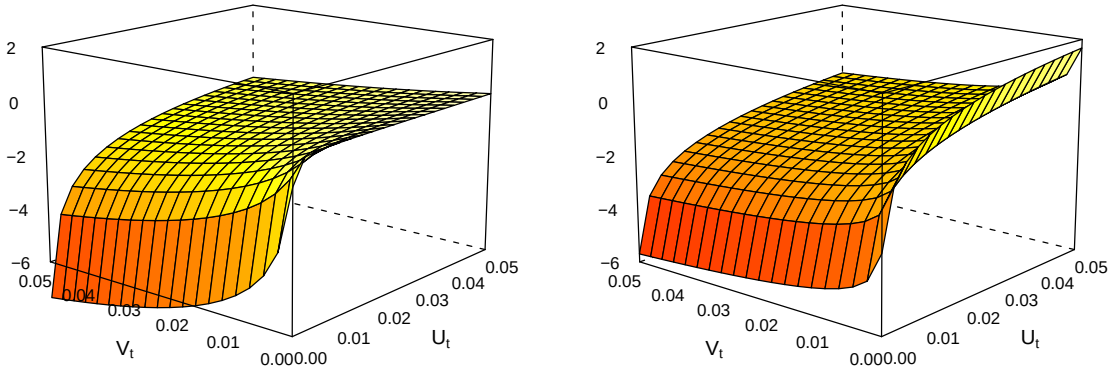

Figure 1: Comparison between $f(U_t, V_t) = \log\left(\frac{U_t}{U_t+V_t}\right)$ (left), and $\hat{f}(U_t, V_t) = \alpha \log U_t + \beta \log V_t + \gamma$ (right). Both functions diverge (to $-\infty$) as $U_t \to 0$. The latter diverges (to $+\infty$) also when $V_t \to 0$. For the interpretation of different colors, see Figure 2 below.

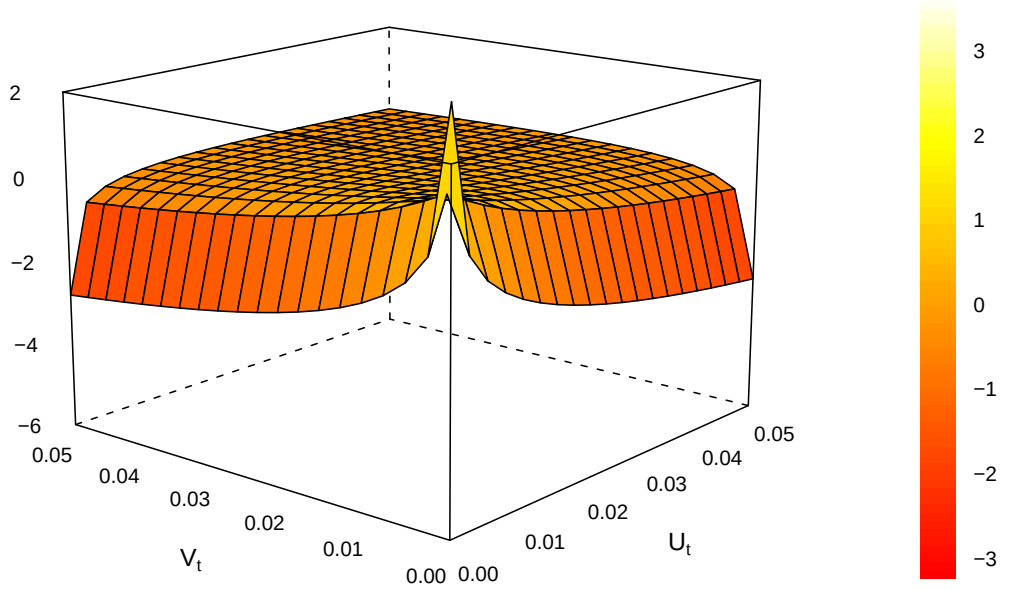

Figure 2: Approximation error: the difference between the exact value and the approximation given in Theorem 1. Notice that the error is symmetric in the two arguments, and diverges as $U_t \to 0$ or $V_t \to 0$. For points where neither argument is close to zero, the error is small (close to zero).

is gained by decomposability, it would seem to be dwarfed by the expensive parameter optimization phase.

Furthermore, trying to use the OFE parameters in aCLL may lead to problems since the approximation is undefined at points where either U_t or V_t is zero. To better see why this is the case, substitute the OFE parameters, Equation (2), into the aCLL criterion, Equation (12), to obtain

$$\hat{\mathrm{a}}\mathrm{CLL}(G \mid D) = (\alpha+\beta)\,\widehat{\mathrm{LL}}(G \mid D) - \beta \sum_{i=1}^{n} \sum_{j=1}^{q_i^*} \sum_{k=1}^{r_i} \sum_{c=0}^{1} N_{ijck} \log \left(\frac{N_{ijck} N_{ij(1-c)}}{N_{ijc} N_{ij(1-c)k}} \right) - \beta \sum_{c=0}^{1} N_c \log \left(\frac{N_c}{N_{1-c}} \right). \tag{13}$$

The problems are associated with the denominator in the second term. In LL and CLL criteria, similar expressions where the denominator may be zero are always eliminated by the OFE parameters since they are always multiplied by zero, see, for example, Equation (3), where $N_{ijc} = 0$ implies $N_{ijck} = 0$. However, there is no guarantee that $N_{ij(1-c)k}$ is non-zero even if the factors in the numerator are non-zero, and hence the division by zero may lead to actual indeterminacies.

Next, we set out to resolve these issues by presenting a well-behaved approximation that enables easy optimization of both structure (via decomposability), as well as parameters.

4.2 Achieving a Well-Behaved Approximation

In this section, we address the singularities of aCLL under OFE by constructing an approximation that is well-behaved.

Recall aCLL in Equation (12). Given a fixed network structure, the parameters that maximize the first term, $(\alpha+\beta)\mathrm{LL}(B \mid D)$, are given by OFE. However, as observed above, the second term may actually be unbounded due to the appearance of $\theta_{ij(1-c)k}$ in the denominator. In order to obtain a well-behaved score, we must therefore make a further modification to the second term. Our strategy is to ensure that the resulting score is *uniformly bounded* and *maximized by OFE parameters*. The intuition behind this is that we can thus guarantee not only that the score is well-behaved, but also that parameter learning is achieved in a simple and efficient way by using the OFE parameters—solving both of the aforementioned issues with the aCLL score. As it turns out, we can satisfy our goal while still retaining the discriminative nature of the score.

The following result is of importance in what follows.

Theorem 4 Consider a Bayesian network B whose structure is given by a fixed directed acyclic graph, G. Let $f(B \mid D)$ be a score defined by

$$f(B \mid D) = \sum_{i=1}^{n} \sum_{j=1}^{q_i^*} \sum_{k=1}^{r_i} \sum_{c=0}^{1} N_{ijck} \left(\lambda \log \left(\frac{\theta_{ijck}}{\frac{N_{ijc}}{N_{ij*}} \theta_{ijck} + \frac{N_{ij(1-c)}}{N_{ij*}} \theta_{ij(1-c)k}} \right) \right), \tag{14}$$

where λ is an arbitrary positive real value. Then, the parameters Θ that maximize $f(B \mid D)$ are given by the observed frequency estimates (OFE) obtained from G.

The theorem implies that by replacing the second term in (12) by (a non-negative multiple of) $f(B \mid D)$ in Equation (14), we get a criterion where both the first and the second term are maximized

by the OFE parameters. We will now proceed to determine a suitable value for the parameter λ appearing in Equation (14).

To clarify the analysis, we introduce the following short-hand notations:

$$\begin{aligned} A_1 &= N_{ijc}\theta_{ijck}, & A_2 &= N_{ijc}, \\ B_1 &= N_{ij(1-c)}\theta_{ij(1-c)k}, & B_2 &= N_{ij(1-c)}. \end{aligned} \tag{15}$$

With simple algebra, we can rewrite the logarithm in the second term of Equation (12) using the above notations as

$$\log\left(\frac{N_{ijc}\theta_{ijck}}{N_{ij(1-c)}\theta_{ij(1-c)k}}\right) - \log\left(\frac{N_{ijc}}{N_{ij(1-c)}}\right) = \log\left(\frac{A_1}{B_1}\right) - \log\left(\frac{A_2}{B_2}\right). \tag{16}$$

Similarly, the logarithm in (14) becomes

$$\begin{aligned} &\lambda\log\left(\frac{N_{ijc}\theta_{ijck}}{N_{ijc}\theta_{ijck}+N_{ij(1-c)}\theta_{ij(1-c)k}}\right)+\rho-\lambda\log\left(\frac{N_{ijc}}{N_{ijc}+N_{ij(1-c)}}\right)-\rho \\ &=\lambda\log\left(\frac{A_1}{A_1+B_1}\right)+\rho-\lambda\log\left(\frac{A_2}{A_2+B_2}\right)-\rho, \end{aligned} \tag{17}$$

where we used $N_{ij*} = N_{ijc} + N_{ij(1-c)}$; we have introduced the constant ρ that was added and subtracted without changing the value of the expression for a reason that will become clear shortly. By comparing Equations (16) and (17), it can be seen that the latter is obtained from the former by replacing expressions of the form $\log(\frac{A}{B})$ by expressions of the form $\lambda\log(\frac{A}{A+B})+\rho$.

We can simplify the two-variable approximation to a single variable one by taking $W = \frac{A}{A+B}$. In this case we have that $\frac{A}{B} = \frac{W}{1-W}$, and so we propose to apply once again the least squares method to approximate the function

$$g(W) = \log\left(\frac{W}{1-W}\right)$$

by

$$\hat{g}(W) = \lambda\log W + \rho.$$

The role of the constant ρ is simply to translate the approximate function to better match the target $g(W)$.

As in the previous approximation, here too it is necessary to make assumptions about the values of A and B (and thus W), in order to find suitable values for the parameters λ and ρ. Again, we stress that the sole purpose of the assumption is to guide in the choice of the parameters.

As both A_1, A_2, B_1, and B_2 in Equation (15) are all non-negative, the ratio $W_i = \frac{A_i}{A_i+B_i}$ is always between zero and one, for both $i \in \{1,2\}$, and hence it is natural to make the straightforward assumption that W_1 and W_2 are uniformly distributed along the unit interval. This gives us the following assumption.

Assumption 2 We assume that

$$\begin{aligned} \frac{N_{ijc}\theta_{ijck}}{N_{ijc}\theta_{ijck}+N_{ij(1-c)}\theta_{ij(1-c)k}} &\sim \text{Uniform}(0,1), \quad \text{and} \\ \frac{N_{ijc}}{N_{ijc}+N_{ij(1-c)}} &\sim \text{Uniform}(0,1). \end{aligned}$$

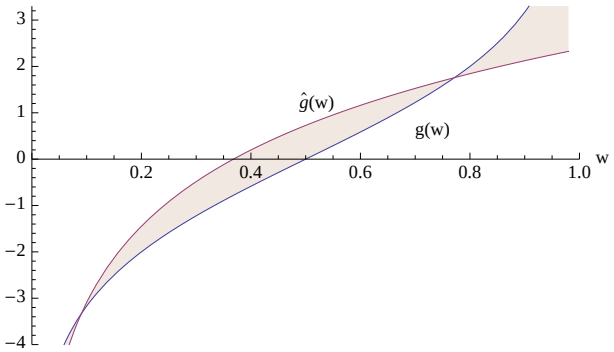

Figure 3: Plot of $g(w) = \log\left(\frac{w}{1-w}\right)$ and $\hat{g}(w) = \lambda \log w + \rho$.

Herein, it is worthwhile noticing that although the previous assumption was meant to hold for general parameters, in practice, we know in this case that OFE will be used. Hence, Assumption 2 reduces to

$$\frac{N_{ijck}}{N_{ij*k}} \sim \text{Uniform}(0,1), \text{ and } \frac{N_{ijc}}{N_{ij*}} \sim \text{Uniform}(0,1).$$

Under this assumption, the mean squared error of the approximation can be minimized analytically, yielding the following solution.

Theorem 5 Under Assumption 2, the values of λ and ρ that minimize the mean square error (MSE) of $\hat{g}$ w.r.t. g are given by

$$\lambda = \frac{\pi^2}{6}, \text{ and} \tag{18}$$

$$\rho = \frac{\pi^2}{6\ln 2}. \tag{19}$$

Theorem 6 The approximation $\hat{g}$ with λ and ρ defined as in Theorem 5 is the minimum variance unbiased (MVU) approximation of g.

In order to get an idea of the accuracy of the approximation $\hat{g}$, consider the graph of $\log\left(\frac{w}{1-w}\right)$ and $\lambda \log w + \rho$ in Figure 3. It may appear problematic that the approximation gets worse as w tends to one. However this is actually unavoidable since that is precisely where âCLL diverges, and our goal is to obtain a criterion that is uniformly bounded.

To wrap up, we first rewrite the logarithm of the second term in Equation (12) using formula (16), and then apply the above approximation to both terms to get

$$\log\left(\frac{\theta_{ijck}}{\theta_{ij(1-c)k}}\right) \approx \lambda \log\left(\frac{N_{ijc}\theta_{ijck}}{N_{ijc}\theta_{ijck} + N_{ij(1-c)}\theta_{ij(1-c)k}}\right) + \rho - \lambda \log\left(\frac{N_{ijc}}{N_{ij*}}\right) - \rho, \tag{20}$$

where ρ cancels out. A similar analysis can be applied to rewrite the logarithm of the third term in Equation (12) leading to

$$\log\left(\frac{\theta_c}{\theta_{(1-c)}}\right) = \log\left(\frac{\theta_c}{1-\theta_c}\right) \approx \lambda \log \theta_c + \rho. \tag{21}$$

Plugging the approximations of Equations (20) and (21) into Equation (12) gives us finally the *factorized conditional log-likelihood* (fCLL) score:

$$\begin{aligned}
&\mathrm{fCLL}(B \mid D) = (\alpha+\beta)\,\mathrm{LL}(B \mid D) \\
&-\beta\lambda \sum_{i=1}^{n}\sum_{j=1}^{q_i^*}\sum_{k=1}^{r_i}\sum_{c=0}^{1} N_{ijck}\left(\log\left(\frac{N_{ijc}\theta_{ijck}}{N_{ijc}\theta_{ijck}-N_{ij(1-c)}\theta_{ij(1-c)k}}\right)-\log\left(\frac{N_{ijc}}{N_{ij*}}\right)\right) \\
&-\beta\lambda\sum_{c=0}^{1} N_c\log\theta_c-\beta N\rho.
\end{aligned} \tag{22}$$

Observe that the third term of Equation (22) is such that

$$-\beta\lambda\sum_{c=0}^{1} N_c\log\theta_c = -\beta\lambda N\sum_{c=0}^{1}\frac{N_c}{N}\log\theta_c, \tag{23}$$

and, since $\beta<0$, by Gibbs inequality (see Lemma 8 in the Appendix at page 2204) the parameters that maximize Equation (23) are given by the OFE, that is, $\hat{\theta}_c=\frac{N_c}{N}$. Therefore, by Theorem 4, given a fixed structure, the maximizing parameters of fCLL are easily obtained as OFE. Moreover, the fCLL score is clearly decomposable.

As a final step, we plug in the OFE parameters, Equation (2), into the fCLL criterion, Equation (22), to obtain

$$\begin{aligned}
\hat{\mathrm{f}}\mathrm{CLL}(G \mid D) \;=\; &(\alpha+\beta)\,\widehat{\mathrm{LL}}(B \mid D)-\beta\lambda\sum_{i=1}^{n}\sum_{j=1}^{q_i^*}\sum_{k=1}^{r_i}\sum_{c=0}^{1} N_{ijck}\left(\log\left(\frac{N_{ijck}}{N_{ij*k}}\right)-\log\left(\frac{N_{ijc}}{N_{ij*}}\right)\right) \\
&-\beta\lambda\sum_{c=0}^{1} N_c\log\left(\frac{N_c}{N}\right)-\beta N\rho,
\end{aligned} \tag{24}$$

where we also use the OFE parameters in the log-likelihood $\widehat{\mathrm{LL}}$. Observe that we can drop the last two terms in Equation (24) as they become constants for a given data set.

4.3 Information-Theoretical Interpretation

Before we present empirical results illustrating the behavior of the proposed scoring criteria, we point out that the $\hat{\mathrm{f}}$CLL criterion has an interesting information-theoretic interpretation based on *interaction information*. We will first rewrite LL in terms of conditional mutual information, and then, similarly, rewrite the second term of $\hat{\mathrm{f}}$CLL in Equation (24) in terms of interaction information.

As Friedman et al. (1997) point out, the local contribution of the i'th variable to $\mathrm{LL}(B \mid D)$ (recall Equation (3)) is given by

$$\begin{aligned}
N\sum_{j=1}^{q_i^*}\sum_{c=0}^{1}\sum_{k=1}^{r_i}\frac{N_{ijck}}{N}\log\left(\frac{N_{ijck}}{N_{ijc}}\right) \;&=\; -NH_{\hat{P}_D}(X_i \mid \Pi^*_{X_i}, C) \\
&=\; -NH_{\hat{P}_D}(X_i \mid C)+NI_{\hat{P}_D}(X_i;\Pi^*_{X_i} \mid C),
\end{aligned} \tag{25}$$

where $H_{\hat{P}_D}(X_i \mid \ldots)$ denotes the *conditional entropy*, and $I_{\hat{P}_D}(X_i;\Pi^*_{X_i} \mid C)$ denotes the *conditional mutual information*, see Cover and Thomas (2006). The subscript $\hat{P}_D$ indicates that the information

theoretic quantities are evaluated under the joint distribution $\hat{P}_D$ of $(\vec{X}, C)$ induced by the OFE parameters.

Since the first term on the right-hand side of (25) does not depend on $\Pi^*_{X_i}$, finding the parents of X_i that maximize $\mathrm{LL}(B \mid D)$ is equivalent to choosing the parents that maximize the second term, $NI_{\hat{P}_D}(X_i; \Pi^*_{X_i} \mid C)$, which measures the information that $\Pi^*_{X_i}$ provides about X_i when the value of C is known.

Let us now turn to the second term of the $\hat{\mathrm{f}}$CLL score in Equation (24). The contribution of the i'th variable in it can also be expressed in information theoretic terms as follows:

$$\begin{aligned} -\beta\lambda N \left(H_{\hat{P}_D}(C \mid X_i, \Pi^*_{X_i}) - H_{\hat{P}_D}(C \mid \Pi^*_{X_i}) \right) &= \beta\lambda N I_{\hat{P}_D}(C; X_i \mid \Pi^*_{X_i}) \\ &= \beta\lambda N \left(I_{\hat{P}_D}(C; X_i; \Pi^*_{X_i}) + I_{\hat{P}_D}(C; X_i) \right), \end{aligned} \tag{26}$$

where $I_{\hat{P}_D}(C; X_I; \Pi^*_{X_i})$ denotes the *interaction information* (McGill, 1954), or the *"co-information"* (Bell, 2003); for a review on the history and use of interaction information in machine learning and statistics, see Jakulin (2005).

Since $I_{\hat{P}_D}(X_i; C)$ on the last line of Equation (26) does not depend on $\Pi^*_{X_i}$, finding the parents of X_i that maximize the sum amounts to maximizing the interaction information. This is intuitive, since the interaction information measures the increase—or the decrease, as it can also be negative—of the mutual information between X_i and C when the parent set $\Pi^*_{X_i}$ is included in the model.

All said, the $\hat{\mathrm{f}}$CLL criterion can be written as

$$\hat{\mathrm{f}}\mathrm{CLL}(G \mid D) = \sum_{i=1}^{n} \left[(\alpha + \beta) N I_{\hat{P}_D}(X_i; \Pi^*_{X_i} \mid C) - \beta\lambda N I_{\hat{P}_D}(C; X_i; \Pi^*_{X_i}) \right] + const, \tag{27}$$

where *const* is a constant independent of the network structure and can thus be omitted. To get a concrete idea of the trade-off between the first two terms, the numerical values of the constants can be evaluated to obtain

$$\hat{\mathrm{f}}\mathrm{CLL}(G \mid D) \approx \sum_{i=1}^{n} \left[0.322\, N I_{\hat{P}_D}(X_i; \Pi^*_{X_i} \mid C) + 0.557\, N I_{\hat{P}_D}(C; X_i; \Pi^*_{X_i}) \right] + const. \tag{28}$$

Normalizing the weights shows that the first term that corresponds to the behavior of the LL criterion, Equation (25), has proportional weight of approximately 36.7 percent, while the second term that gives $\hat{\mathrm{f}}$CLL criterion its discriminative nature has the weight 63.3 percent.[2]

In addition to the insight provided by the information-theoretic interpretation of $\hat{\mathrm{f}}$CLL, it also provides a practically most useful corollary: the $\hat{\mathrm{f}}$CLL criterion is score equivalent. A scoring criterion is said to be *score equivalent* if it assigns the same score to all network structures encoding the same independence assumptions, see Verma and Pearl (1990), Chickering (2002), Yang and Chang (2002) and de Campos (2006).

Theorem 7 The $\hat{\mathrm{f}}$CLL criterion is score equivalent for augmented naive Bayes classifiers.

The practical utility of the above result is due to the fact that it enables the use of powerful algorithms, such as the tree-learning method by Chow and Liu (1968), in learning TAN classifiers.

2. The particular linear combination of the two terms in Equation (28) brings out the question what would happen in only one of the terms was retained, or equivalently, if one of the weights was set to zero. As mentioned above, the first term corresponds to the LL criterion, and hence, setting the weight of the second term to zero would reduce the criterion to LL. We also experimented with a criterion where only the second term is retained but this was observed to yield poor results; for details, see the additional material at `http://kdbio.inesc-id.pt/~asmc/software/fCLL.html`.

4.4 Beyond Binary Classification and TAN

Although âCLL and f̂CLL scoring criteria were devised having in mind binary classification tasks, their application in multi-class problems is straightforward.[3] For the case of f̂CLL, the expression (24) does not involve the dual samples. Hence, it can be trivially adapted for non-binary classification tasks. On the other hand, the score âCLL in Equation (13) does depend on the dual samples. To adapt it for multi-class problems, we considered $N_{ij(1-c)k} = N_{ij*k} - N_{ijck}$ and $N_{ij(1-c)} = N_{ij} - N_{ijc}$.

Finally, we point out that despite being derived under the augmented naive Bayes model, the f̂CLL score can be readily applied to models where the class variable is *not* a parent of some of the attributes. Hence, we can use it as a criterion for learning more general structures. The empirical results below demonstrate that this indeed leads to good classifiers.

5. Experimental Results

We implemented the f̂CLL scoring criterion on top of the Weka open-source software (Hall et al., 2009). Unfortunately, the Weka implementation of the TAN classifier assumes that the learning criterion is score equivalent, which rules out the use of our âCLL criterion. Non-score-equivalent criteria require the Edmonds' maximum branchings algorithm that builds a maximal *directed* spanning tree (see Edmonds 1967, or Lawler 1976) instead of an undirected one obtained by the Chow-Liu algorithm (Chow and Liu, 1968). Edmonds' algorithm had already been implemented by some of the authors (see Carvalho et al., 2007) using Mathematica 7.0 and the Combinatorica package (Pemmaraju and Skiena, 2003). Hence, the âCLL criterion was implemented in this environment. All source code and the data sets used in the experiments, can be found at fCLL web page.[4]

We evaluated the performance of âCLL and f̂CLL scoring criteria in classification tasks comparing them with state-of-the-art classifiers. We performed our evaluation on the same 25 benchmark data sets used by Friedman et al. (1997). These include 23 data sets from the UCI repository of Newman et al. (1998) and two artificial data sets, *corral* and *mofn*, designed by Kohavi and John (1997) to evaluate methods for feature subset selection. A description of the data sets is presented in Table 1. All continuous-valued attributes were discretized using the supervised entropy-based method by Fayyad and Irani (1993). For this task we used the Weka package.[5] Instances with missing values were removed from the data sets.

The classifiers used in the experiments were:
GHC2: Greedy hill climber classifier with up to 2 parents.
TAN: Tree augmented naive Bayes classifier.
C4.5: C4.5 classifier.
k-NN: *k*-nearest neighbor classifier, with $k = 1, 3, 5$.
SVM: Support vector machine with linear kernel.
SVM2: Support vector machine with polynomial kernel of degree 2.

3. As suggested by an anonymous referee, the techniques used in Section 4.1 for deriving the âCLL criterion can be generalized to the multi-class case as well as to other distributions in addition to the uniform one in a straightforward manner by applying results from regression theory. We plan to explore such generalizations of both the âCLL and f̂CLL criteria in future work.
4. fCLL web page is at `http://kdbio.inesc-id.pt/~asmc/software/fCLL.html`.
5. Discretization was done using `weka.filters.supervised.attribute.Discretize`, with default parameters. This discretization improved the accuracy of all classifiers used in the experiments, including those that do not necessarily require discretization, that is, C4.5 *k*-NN, SVM, and LogR.

	Data Set	Features	Classes	Train	Test
1	australian	15	2	690	CV-5
2	breast	10	2	683	CV-5
3	chess	37	2	2130	1066
4	cleve	14	2	296	CV-5
5	corral	7	2	128	CV-5
6	crx	16	2	653	CV-5
7	diabetes	9	2	768	CV-5
8	flare	11	2	1066	CV-5
9	german	21	2	1000	CV-5
10	glass	10	7	214	CV-5
11	glass2	10	2	163	CV-5
12	heart	14	2	270	CV-5
13	hepatitis	20	2	80	CV-5
14	iris	5	3	150	CV-5
15	letter	17	26	15000	5000
16	lymphography	19	4	148	CV-5
17	mofn-3-7-10	11	2	300	1024
18	pima	9	2	768	CV-5
19	satimage	37	6	4435	2000
20	segment	20	7	1540	770
21	shuttle-small	10	7	3866	1934
22	soybean-large	36	19	562	CV-5
23	vehicle	19	4	846	CV-5
24	vote	17	2	435	CV-5
25	waveform-21	22	3	300	4700

Table 1: Description of data sets used in the experiments.

SVMG: Support vector machine with Gaussian (RBF) kernel.
LogR: Logistic regression.

Bayesian network-based classifiers (GHC2 and TAN) were included in different flavors, differing in the scoring criterion used for structure learning (LL, âCLL, $\hat{\text{f}}$CLL) and the parameter estimator (OFE, ELR). Each variant along with the implementation used in the experiments is described in Table 2. Default parameters were used in all cases unless explicitly stated. Excluding TAN classifiers obtained with the ELR method, we improved the performance of Bayesian network classifiers by smoothing parameter estimates according to a Dirichlet prior (see Heckerman et al., 1995). The smoothing parameter was set to 0.5, the default in Weka. The same strategy was used for TAN classifiers implemented in Mathematica. For discriminative parameter learning with ELR, parameters were initialized to the OFE values. The gradient descent parameter optimization was terminated using *cross tuning* as suggested in Greiner et al. (2005).

Three different kernels were applied in SVM classifiers: (i) a linear kernel of the form $K(\mathbf{x}_i, \mathbf{x}_j) = \mathbf{x}_i^T \mathbf{x}_j$; (ii) a polynomial kernel of the form $K(\mathbf{x}_i, \mathbf{x}_j) = (\mathbf{x}_i^T \mathbf{x}_j)^2$; and (iii) a Gaussian (radial basis

Classifier	Struct.	Param.	Implementation
GHC2	LL	OFE	`HillClimber` (P=2) implementation from Weka
GHC2	$\hat{\text{f}}$CLL	OFE	`HillClimber` (P=2) implementation from Weka
TAN	LL	OFE	`TAN` implementation from Weka
TAN	LL	ELR	`TAN` implementation from Greiner and Zhou (2002)
TAN	âCLL	OFE	`TAN` implementation from Carvalho et al. (2007)
TAN	$\hat{\text{f}}$CLL	OFE	`TAN` implementation from Weka
C4.5			`J48` implementation from Weka
1-NN			`IBk` (K=1) implementation from Weka
3-NN			`IBk` (K=3) implementation from Weka
5-NN			`IBk` (K=5) implementation from Weka
SVM			`SMO` implementation from Weka
SVM2			`SMO` with `PolyKernel` (E=2) implementation from Weka
SVMG			`SMO` with `RBFKernel` implementation from Weka
LogR			`Logistic` implementation from Weka

Table 2: Classifiers used in the experiments.

function) kernel of the form $K(\mathbf{x}_i,\mathbf{x}_j) = \exp(-\gamma||\mathbf{x}_i - \mathbf{x}_j||^2)$. Following established practice (see Hsu et al., 2003), we used a grid-search on the penalty parameter C and the RBF kernel parameter γ, using cross-validation. For linear and polynomial kernels we selected C from $[10^{-1}, 1, 10, 10^2]$ by using 5-fold cross-validation on the training set. For the RBF kernel we selected C and γ from $[10^{-1}, 1, 10, 10^2]$ and $[10^{-3}, 10^{-2}, 10^{-1}, 1, 10]$, respectively, by using 5-fold cross-validation on the training set.

The accuracy of each classifier is defined as the percentage of successful predictions on the test sets in each data set. As suggested by Friedman et al. (1997), accuracy was measured via the holdout method for larger training sets, and via stratified five-fold cross-validation for smaller ones, using the methods described by Kohavi (1995). Throughout the experiments, we used the same cross-validation folds for every classifier. Scatter plots of the accuracies of the proposed methods against the others are depicted in Figure 4 and Figure 5. Points above the diagonal line represent cases where the method shown in the vertical axis performs better than the one on the horizontal axis. Crosses over the points depict the standard deviation. The standard deviation is computed according to the binomial formula $\sqrt{acc \times (1-acc)/m}$, where acc is the classifier accuracy and, for the cross-validation tests, m is the size of the data set. For the case of holdout tests, m is the size of the test set. Tables with the accuracies and standard deviations can be found at the fCLL webpage.

We compare the performance of the classifiers using Wilcoxon signed-rank tests, using the same procedure as Grossman and Domingos (2004). This test is applicable when paired classification accuracy differences, along the data sets, are independent and non-normally distributed. Alternatively, a paired t-test could be used, but as the Wilcoxon signed-rank test is more conservative than the paired t-test, we apply the former. Results are depicted in Table 3 and Table 4. Each entry of Table 3 and Table 4 gives the Z-score and p-value of the significance test for the corresponding pairs

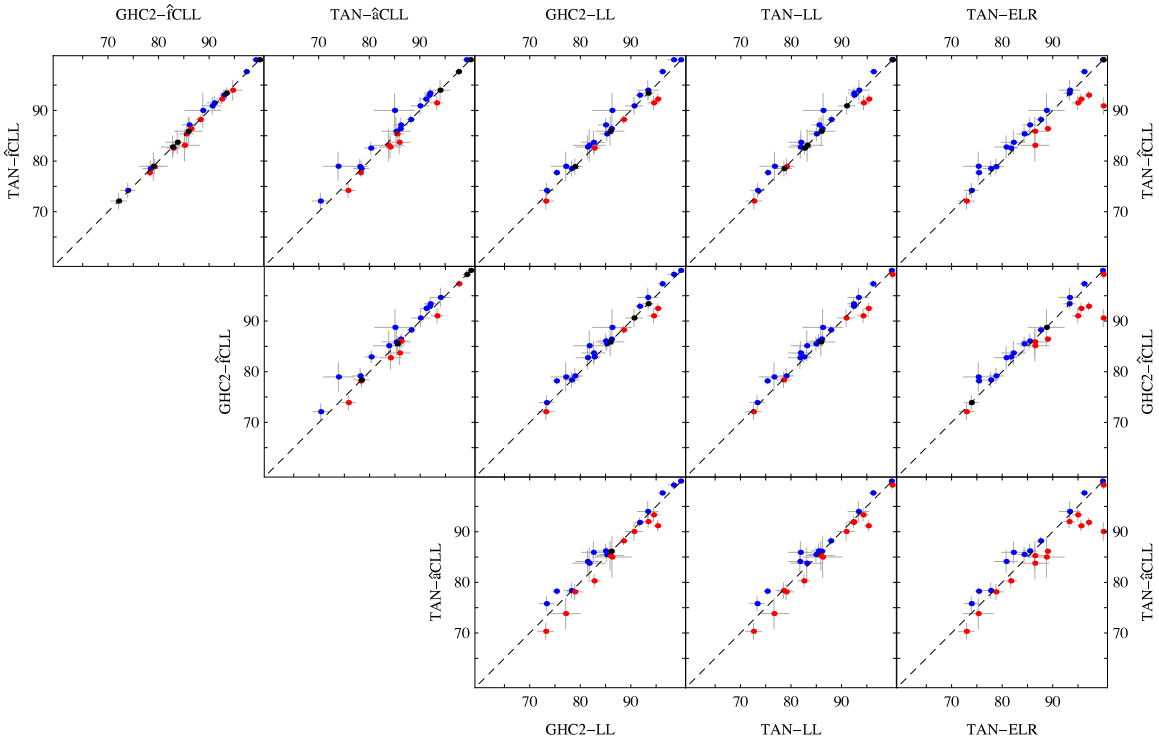

Figure 4: Scatter plots of the accuracy of Bayesian network-based classifiers.

of classifiers. The arrow points towards the learning algorithm that yields superior classification performance. A double arrow is used if the difference is significant with p-value smaller than 0.05.

Over all, TAN-f̂CLL-OFE and GHC-f̂CLL-OFE performed the best (Tables 3–4). They outperformed C4.5, the nearest neighbor classifiers, and logistic regression, as well as the generatively-trained Bayesian network classifiers, TAN-LL-OFE and GHC-LL-OFE, all differences being statistically significant at the $p < 0.05$ level. On the other hand, TAN-âCLL-OFE did not stand out compared to most of the other methods. Moreover, TAN-f̂CLL-OFE and GHC-f̂CLL-OFE classifiers fared sightly better than TAN-LL-ELR and the SVM classifiers, although the difference was not statistically significant. In these cases, the only practically relevant factor is computational efficiency.

To roughly characterize the computational complexity of learning the various classifiers, we measured the total time required by each classifier to process all the 25 data sets.[6] Most of the methods only took a few seconds ($\sim 1-3$ seconds), except for TAN-âCLL-OFE which took a few minutes ($\sim 2-3$ minutes), SVM with linear kernel which took some minutes ($\sim 17-18$ minutes), TAN-LL-ELR and SVM with polynomial kernel which took a few hours ($\sim 1-2$ hours) and, finally, logistic regression and SVM with RBF kernel which took several hours ($\sim 18-32$ hours). In the case of TAN-âCLL-OFE, the slightly increased computation time was likely caused by the Mathematica package, which is not intended for numerical computation. In theory, the computational complexity of TAN-âCLL-OFE is of the same order as TAN-LL-OFE or TAN-f̂CLL-

6. Reporting the total time instead of the individual times for each data set will emphasize the significance of the larger data sets. However, the individual times were in accordance with the general conclusion drawn from the total time.

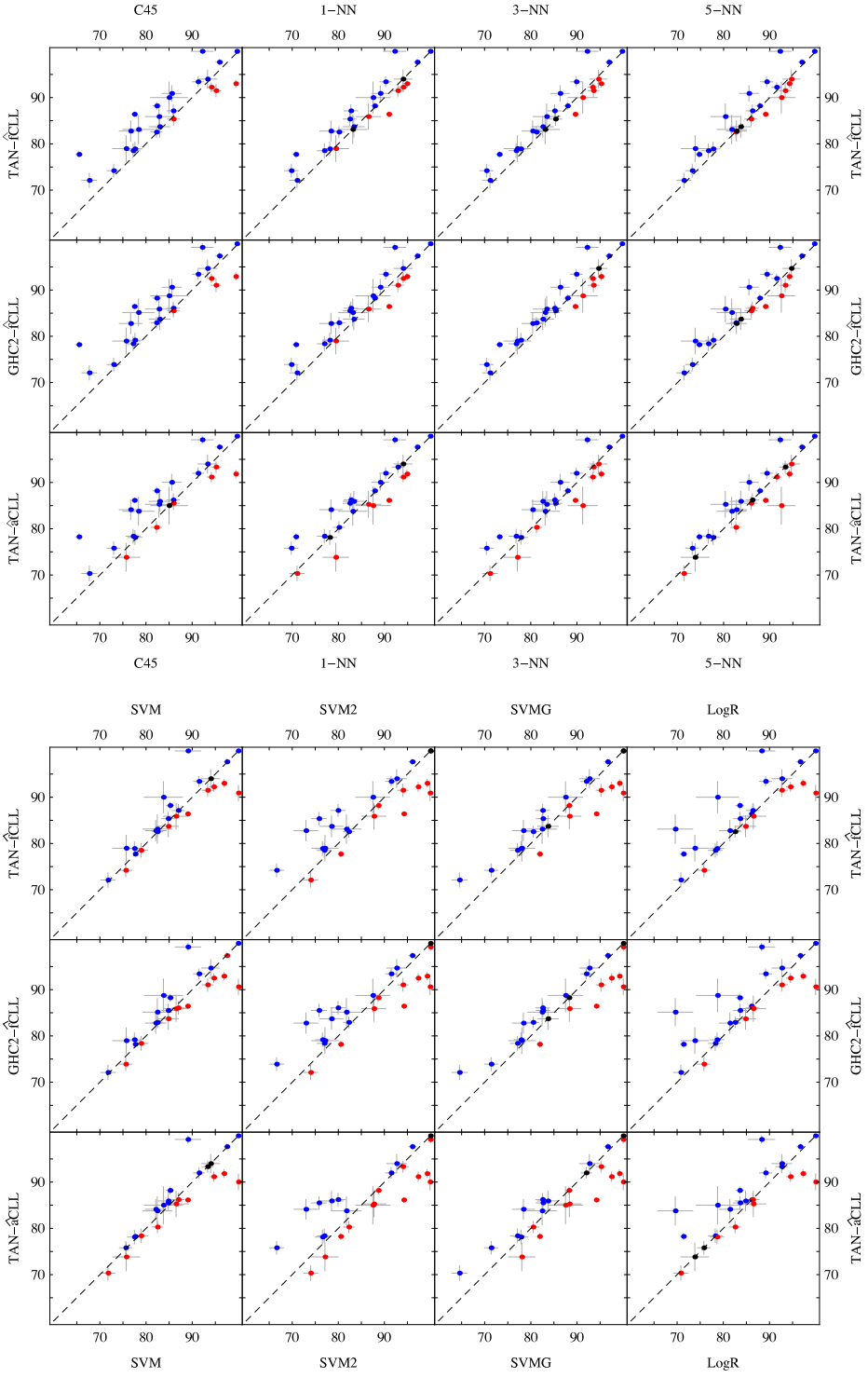

Figure 5: The accuracy of the proposed methods vs. state-of-the-art classifiers.

Classifier	GHC2	TAN	GHC2	TAN	TAN
Struct.	$\hat{\text{f}}$CLL	âCLL	LL	LL	LL
Param.	OFE	OFE	OFE	OFE	ELR
TAN	0.37	1.44	2.13	2.13	0.31
$\hat{\text{f}}$CLL	0.36	0.07	0.02	0.02	0.38
OFE	$\leftarrow$	$\leftarrow$	$\Leftarrow$	$\Leftarrow$	$\leftarrow$
GHC2		1.49	2.26	2.21	0.06
$\hat{\text{f}}$CLL		0.07	0.01	0.01	0.48
OFE		$\leftarrow$	$\Leftarrow$	$\Leftarrow$	$\leftarrow$
TAN			0.04	-0.34	-1.31
âCLL			0.48	0.37	0.10
OFE			$\leftarrow$	$\uparrow$	$\uparrow$

Table 3: Comparison of the Bayesian network classifiers against each other, using the Wilcoxon signed-rank test. Each cell of the array gives the Z-score (top) and the corresponding p-value (middle). Arrow points towards the better method, double arrow indicates statistical significance at level $p < 0.05$.

Classifier	C4.5	1-NN	3-NN	5-NN	SVM	SVM2	SVMG	LogR
TAN	3.00	2.25	2.16	2.07	0.43	0.61	0.21	1.80
$\hat{\text{f}}$CLL	$<$0.01	0.01	0.02	0.02	0.33	0.27	0.42	0.04
OFE	$\Leftarrow$	$\Leftarrow$	$\Leftarrow$	$\Leftarrow$	$\leftarrow$	$\leftarrow$	$\leftarrow$	$\Leftarrow$
GHC2	3.00	2.35	2.20	2.19	0.39	0.74	0.11	1.65
$\hat{\text{f}}$CLL	$<$0.01	$<$0.01	0.01	0.01	0.35	0.23	0.45	0.05
OFE	$\Leftarrow$	$\Leftarrow$	$\Leftarrow$	$\Leftarrow$	$\leftarrow$	$\leftarrow$	$\leftarrow$	$\Leftarrow$
TAN	2.26	1.34	1.17	1.31	-0.40	-0.29	-0.55	1.37
âCLL	0.01	0.09	0.12	0.09	0.35	0.38	0.29	0.09
OFE	$\Leftarrow$	$\leftarrow$	$\leftarrow$	$\leftarrow$	$\uparrow$	$\uparrow$	$\uparrow$	$\leftarrow$

Table 4: Comparison of the Bayesian network classifiers against other classifiers. Conventions identical to those in Table 3.

OFE: $O(n^2 \log n)$ in the number of features and linear in the number of instances, see Friedman et al. (1997).

Concerning TAN-LL-ELR, the difference is caused by the discriminative parameter learning method (ELR), which is computationally expensive. In our experiments, TAN-LL-ELR was 3 order of magnitude slower than TAN-$\hat{\text{f}}$CLL-OFE. Su and Zhang (2006) report a difference of 6 orders of magnitude, but different data sets were used in their experiments. Likewise, the high computational cost of SVMs was expected. Selection of the regularization parameter using cross-tuning further

increases the cost. In our experiments, SVMs were clearly slower than $\hat{\text{f}}$CLL-based classifiers. Furthermore, in terms of memory, SVMs with polynomial and RBF kernels, as well as logistic regression, required that the available memory was increased to 1 GB of memory, whereas all other classifiers coped with the default 128 MB.

6. Conclusions and Future Work

We proposed a new decomposable scoring criterion for classification tasks. The new score, called factorized conditional log-likelihood, $\hat{\text{f}}$CLL, is based on an approximation of conditional log-likelihood. The new criterion is decomposable, score-equivalent, and allows efficient estimation of both structure and parameters. The computational complexity of the proposed method is of the same order as the traditional log-likelihood criterion. Moreover, the criterion is specifically designed for discriminative learning.

The merits of the new scoring criterion were evaluated and compared to those of common state-of-the-art classifiers, on a large suite of benchmark data sets from the UCI repository. Optimal $\hat{\text{f}}$CLL-scored tree-augmented naive Bayes (TAN) classifiers, as well as somewhat more general structures (referred to above as GHC2), performed better than generatively-trained Bayesian network classifiers, as well as C4.5, nearest neighbor, and logistic regression classifiers, with statistical significance. Moreover, $\hat{\text{f}}$CLL-optimized classifiers performed better, although the difference is not statistically significant, than those where the Bayesian network parameters were optimized using an earlier discriminative criterion (ELR), as well as support vector machines (with linear, polynomial and RBF kernels). In comparison to the latter methods, our method is considerably more efficient in terms of computational cost, taking 2 to 3 orders of magnitude less time for the data sets in our experiments.

Directions for future work include: studying in detail the asymptotic behavior of $\hat{\text{f}}$CLL for TAN and more general models; combining our intermediate approximation, aCLL, with discriminative parameter estimation (ELR); extending aCLL and $\hat{\text{f}}$CLL to mixture models; and applications in data clustering.

Acknowledgments

The authors are grateful to the invaluable comments by the anonymous referees. The authors thank Vtor Rocha Vieira, from the Physics Department at IST/TULisbon, for his enthusiasm in cross-checking the analytical integration of the first approximation, and Mrio Figueiredo, from the Electrical Engineering at IST/TULisbon, for his availability in helping with concerns that appeared with respect to this work.

The work of AMC and ALO was partially supported by FCT (INESC-ID multiannual funding) through the PIDDAC Program funds. The work of AMC was also supported by FCT and EU FEDER via project PneumoSyS (PTDC/SAU-MII/100964/2008). The work of TR and PM was supported in part by the Academy of Finland (Projects MODEST and PRIME) and the European Commission Network of Excellence PASCAL.

Availability: Supplementary material including program code and the data sets used in the experiments can be found at `http://kdbio.inesc-id.pt/~asmc/software/fCLL.html`.

Appendix A. Detailed Proofs

Proof (Theorem 1) We have that

$$\begin{aligned}
S_p(\alpha,\beta,\gamma) &= \int_0^p \int_0^p \frac{1}{p^2}\left(\log\left(\frac{x}{x+y}\right) - (\alpha\log(x)+\beta\log(y)+\gamma)\right)^2 \mathrm{d}y\,\mathrm{d}x \\
&= \frac{1}{12\ln(2)^2}(-\pi^2(-1+\alpha+\beta) \\
&\quad +6(2+4\alpha^2+4\beta^2-4\ln(2)-2\gamma\ln(2)+4\ln(2)^2+8\gamma\ln(2)^2+2\gamma^2\ln^2(2) \\
&\quad +\beta(5-4(2+\gamma)\ln(2))+\alpha(1+4\beta-4(2+\gamma)\ln(2))) \\
&\quad -12(\alpha+\beta)(1+2\alpha+2\beta-4\ln(2)-2\gamma\ln(2))\ln(p)+12(\alpha+\beta)^2\ln^2(p)).
\end{aligned}$$

Moreover, $\nabla.S_p = 0$ iff

$$\begin{aligned}
\alpha &= \frac{\pi^2+6}{24}, \\
\beta &= \frac{\pi^2-18}{24}, \\
\gamma &= \frac{\pi^2}{12\ln(2)} - \left(2+\frac{(\pi^2-6)\log(p)}{12}\right),
\end{aligned}$$

which coincides exactly with (8), (9) and (10), respectively. Now to show that (8), (9) and (10) define a global minimum, take $\delta = (\alpha\log(p)+\beta\log(p)+\gamma)$ and notice that

$$\begin{aligned}
S_p(\alpha,\beta,\gamma) &= \int_0^p \int_0^p \frac{1}{p^2}\left(\log\left(\frac{x}{x+y}\right) - (\alpha\log(x)+\beta\log(y)+\gamma)\right)^2 \mathrm{d}y\,\mathrm{d}x \\
&= \int_0^1 \int_0^1 \frac{1}{p^2}\left(\log\left(\frac{px}{px+py}\right) - (\alpha\log(px)+\beta\log(py)+\gamma)\right)^2 p^2\mathrm{d}y\,\mathrm{d}x \\
&= \int_0^1 \int_0^1 \left(\log\left(\frac{x}{x+y}\right) - (\alpha\log(x)+\beta\log(y)+(\alpha\log(p)+\beta\log(p)+\gamma))\right)^2 \mathrm{d}y\,\mathrm{d}x \\
&= \int_0^1 \int_0^1 \left(\log\left(\frac{x}{x+y}\right) - (\alpha\log(x)+\beta\log(y)+\delta)\right)^2 \mathrm{d}y\,\mathrm{d}x \\
&= S_1(\alpha,\beta,\delta).
\end{aligned}$$

So, S_p has a minimum at (8), (9) and (10) iff S_1 has a minimum at (8), (9) and

$$\delta = \frac{\pi^2}{12\ln(2)} - 2.$$

The Hessian of S_1 is

$$\begin{pmatrix}
\frac{4}{\ln^2(2)} & \frac{2}{\ln^2(2)} & -\frac{2}{\ln(2)} \\
\frac{2}{\ln^2(2)} & \frac{4}{\ln^2(2)} & -\frac{2}{\ln(2)} \\
-\frac{2}{\ln(2)} & -\frac{2}{\ln(2)} & 2
\end{pmatrix}$$

and its eigenvalues are

$$\begin{aligned} rcle_1 &= \frac{3+\ln^2(2)+\sqrt{9+2\ln^2(2)+\ln(2)^4}}{\ln^2(2)}, \\ e_2 &= \frac{2}{\ln^2(2)}, \\ e_3 &= \frac{3+\ln^2(2)-\sqrt{9+2\ln^2(2)+\ln(2)^4}}{\ln^2(2)}, \end{aligned}$$

which are all positive. Thus, S_1 has a local minimum in (α,β,δ) and, consequently, S_p has a local minimum in (α,β,γ). Since $\nabla.S_p$ has only one zero, (α,β,γ) is a global minimum of S_p. □

Proof (Theorem 2) We have that

$$\int_0^p \int_0^p \frac{1}{p^2}\left(\log\left(\frac{x}{x+y}\right)-(\alpha\log(x)+\beta\log(y)+\gamma)\right)\mathrm{d}y\,\mathrm{d}x = 0$$

for α,β and γ defined as in (8), (9) and (10). Since the MSE coincides with the variance for any unbiased estimator, the proposed approximation is the one with minimum variance. □

Proof (Theorem 3) We have that

$$\sqrt{\int_0^p \int_0^p \frac{1}{p^2}\left(\log\left(\frac{x}{x+y}\right)-(\alpha\log(x)+\beta\log(y)+\gamma)\right)^2\mathrm{d}y\,\mathrm{d}x} = \sqrt{\frac{36+36\pi^2-\pi^4}{288\ln^2(2)}-2}$$

for α,β and γ defined as in (8), (9) and (10), which concludes the proof. □

For the proof of Theorem 4, we recall Gibb's inequality.

Lemma 8 (Gibb's inequality) Let $P(x)$ and $Q(x)$ be two probability distributions over the same domain, then

$$\sum_x P(x)\log(Q(x)) \leq \sum_x P(x)\log(P(x)).$$

Proof (Theorem 4) We now take advantage of Gibb's inequality to show that the parameters that maximize the $f(B\,|\,D)$ are those given by the OFE. Observe that

$$\begin{aligned} f(B\,|\,D) &= \lambda\sum_{i=1}^{n}\sum_{j=1}^{q_i^*}\sum_{k=1}^{r_i}\sum_{c=0}^{1} N_{ijck}\log\left(\frac{N_{ijc}\theta_{ijck}}{N_{ijc}\theta_{ijck}+N_{ij(1-c)}\theta_{ij(1-c)k}}\right)-\log\left(\frac{N_{ijc}}{N_{ij*}}\right) \\ &= K+\lambda\sum_{i=1}^{n}\sum_{j=1}^{q_i^*}\sum_{k=1}^{r_i} N_{ij*k}\sum_{c=0}^{1}\frac{N_{ijck}}{N_{ij*k}}\log\left(\frac{N_{ijc}\theta_{ijck}}{N_{ijc}\theta_{ijck}+N_{ij(1-c)}\theta_{ij(1-c)k}}\right), \end{aligned} \tag{29}$$

where K is a constant that does not depend on the parameters θ_{ijck}, and therefore, can be ignored. Moreover, if we take the OFE for the parameters, we have

$$\hat{\theta}_{ijck} = \frac{N_{ijkc}}{N_{ijc}} \text{ and } \hat{\theta}_{ij(1-c)k} = \frac{N_{ijk(1-c)}}{N_{ij(1-c)}}.$$

By plugging the OFE estimates in (29) we obtain

$$\begin{aligned}\hat{f}(G \mid D) &= K + \lambda \sum_{i=1}^{n} \sum_{j=1}^{q_i^*} \sum_{k=1}^{r_i} N_{ij*c} \sum_{c=0}^{1} \frac{N_{ijck}}{N_{ij*k}} \log \left(\frac{N_{ijc} \frac{N_{ijck}}{N_{ijc}}}{N_{ijc} \frac{N_{ijck}}{N_{ijc}} + N_{ij(1-c)} \frac{N_{ij(1-c)k}}{N_{ij(1-c)}}} \right) \\ &= K + \lambda \sum_{i=1}^{n} \sum_{j=1}^{q_i} \sum_{k=1}^{r_i} N_{ij*k} \sum_{c=0}^{1} \frac{N_{ijck}}{N_{ij*k}} \log \left(\frac{N_{ijck}}{N_{ij*k}} \right).\end{aligned}$$

According to Gibb's inequality, this is the maximum value that $f(B \mid D)$ can attain, and therefore, the parameters that maximize $f(B \mid D)$ are those given by the OFE. □

Proof (Theorem 5) We have that

$$S(\lambda, \rho) = \int_0^1 \left(\log \left(\frac{x}{1-x} \right) - (\lambda \log(x) + \rho) \right)^2 \mathrm{d}x = \frac{6\lambda^2 + \pi^2 + 3\rho^2 \ln^2(2) - \lambda \left(\pi^2 + 6\rho \ln(2) \right)}{3 \ln^2(2)}.$$

Moreover $\nabla . S = 0$ iff

$$\begin{aligned}\lambda &= \tfrac{\pi^2}{6}, \\ \rho &= \tfrac{\pi^2}{6 \ln(2)},\end{aligned}$$

which coincides with (18) and (19), respectively. The Hessian of S is

$$\begin{pmatrix} \frac{4}{\ln^2(2)} & -\frac{2}{\ln(2)}, \\ -\frac{2}{\ln(2)} & 2 \end{pmatrix}$$

with eigenvalues

$$\frac{2 + \ln^2(2) \pm \sqrt{4 + \ln^4(2)}}{\ln^2(2)}$$

which are both positive. Hence, there is only one minimum, and (λ, ρ) is the global minimum. □

Proof (Theorem 6) We have that

$$\int_0^1 \left(\log \left(\frac{x}{1-x} \right) - (\lambda \log(x) + \rho) \right) \mathrm{d}x = 0$$

for λ and ρ defined as in Equations (18) and (19). Since the MSE coincides with the variance for any unbiased estimator, the proposed approximation is the one with minimum variance. □

Proof (Theorem 7) By Theorem 2 in Chickering (1995), it is enough to show that for graphs G_1 and G_2 differing only on reversing one covered edge, we have that $\hat{\text{f}}\text{CLL}(G_1 \mid D) = \hat{\text{f}}\text{CLL}(G_2 \mid D)$.

Assume that $X \to Y$ occurs in G_1 and $Y \to X$ occurs in G_2 and that $X \to Y$ is covered, that is, $\Pi_Y^{G_1} = \Pi_X^{G_1} \cup \{X\}$. Since we are only dealing with augment naive Bayes classifiers, X and Y are different from C and so we also have $\Pi_Y^{*G_1} = \Pi_X^{*G_1} \cup \{X\}$. Moreover, take G_0 to be the graph G_1 without the edge $X \to Y$ (which is the same as graph G_2 without the edge $Y \to X$). Then, we have that $\Pi_X^{*G_0} = \Pi_Y^{*G_0} = \Pi^{*G_0}$ and, moreover, the following equalities hold:

$$\begin{aligned} \Pi_X^{*G_1} &= \Pi^{*G_0}; & \Pi_Y^{*G_2} &= \Pi^{*G_0}; \\ \Pi_Y^{*G_1} &= \Pi^{*G_0} \cup \{X\}; & \Pi_X^{*G_2} &= \Pi^{*G_0} \cup \{Y\}. \end{aligned}$$

Since $\hat{\text{f}}\text{CLL}$ is a local scoring criterion, $\hat{\text{f}}\text{CLL}(G_1 \mid D)$ can be computed from $\hat{\text{f}}\text{CLL}(G_0 \mid D)$ taking only into account the difference in the contribution of node Y. In this case, by Equation (27), it follows that

$$\begin{aligned} \hat{\text{f}}\text{CLL}(G_1 \mid D) &= \hat{\text{f}}\text{CLL}(G_0 \mid D) - ((\alpha+\beta)NI_{\hat{P}_D}(Y;\Pi^{*G_0} \mid C) - \beta\lambda NI_{\hat{P}_D}(Y;\Pi^{*G_0};C)) \\ &\quad +((\alpha+\beta)NI_{\hat{P}_D}(Y;\Pi_Y^{*G_1} \mid C) - \beta\lambda NI_{\hat{P}_D}(Y;\Pi_Y^{*G_1};C)) \\ &= \hat{\text{f}}\text{CLL}(G_0 \mid D) + (\alpha+\beta)N(I_{\hat{P}_D}(Y;\Pi^{*G_0} \cup \{X\} \mid C) - I_{\hat{P}_D}(Y;\Pi^{*G_0} \mid C)) \\ &\quad -\beta\lambda N(I_{\hat{P}_D}(Y;\Pi^{*G_0} \cup \{X\};C) - I_{\hat{P}_D}(Y;\Pi^{*G_0};C)) \end{aligned}$$

and, similarly, that

$$\begin{aligned} \hat{\text{f}}\text{CLL}(G_2 \mid D) &= \hat{\text{f}}\text{CLL}(G_0 \mid D) + (\alpha+\beta)N(I_{\hat{P}_D}(X;\Pi^{*G_0} \cup \{Y\} \mid C) - I_{\hat{P}_D}(X;\Pi^{*G_0} \mid C)) + \\ &\quad -\beta\lambda N(I_{\hat{P}_D}(X;\Pi^{*G_0} \cup \{Y\};C) - I_{\hat{P}_D}(X;\Pi^{*G_0};C)). \end{aligned}$$

To show that $\hat{\text{f}}\text{CLL}(G_1 \mid D) = \hat{\text{f}}\text{CLL}(G_2 \mid D)$ it suffices to prove that

$$I_{\hat{P}_D}(Y;\Pi^{*G_0} \cup \{X\} \mid C) - I_{\hat{P}_D}(Y;\Pi^{*G_0} \mid C) = I_{\hat{P}_D}(X;\Pi^{*G_0} \cup \{Y\} \mid C) - I_{\hat{P}_D}(X;\Pi^{*G_0} \mid C) \tag{30}$$

and that

$$I_{\hat{P}_D}(Y;\Pi^{*G_0} \cup \{X\};C) - I_{\hat{P}_D}(Y;\Pi^{*G_0};C) = I_{\hat{P}_D}(X;\Pi^{*G_0} \cup \{Y\};C)) - I_{\hat{P}_D}(X;\Pi^{*G_0};C). \tag{31}$$

We start by showing (30). In this case, by definition of conditional mutual, we have that

$$\begin{aligned} &I_{\hat{P}_D}(Y;\Pi^{*G_0} \cup \{X\} \mid C) - I_{\hat{P}_D}(Y;\Pi^{*G_0} \mid C) = \\ &= H_{\hat{P}_D}(Y \mid C) + H_{\hat{P}_D}(\Pi^{*G_0} \cup \{X\} \mid C) - H_{\hat{P}_D}(\Pi^{*G_0} \cup \{X,Y\} \mid C) - H_{\hat{P}_D}(Y \mid C) + \\ &\qquad -H_{\hat{P}_D}(\Pi^{*G_0} \mid C) + H_{\hat{P}_D}(\Pi^{*G_0} \cup \{Y\} \mid C) \\ &= -H_{\hat{P}_D}(\Pi^{*G_0} \mid C) + H_{\hat{P}_D}(\Pi^{*G_0} \cup \{X\} \mid C) + H_{\hat{P}_D}(\Pi^{*G_0} \cup \{Y\} \mid C) - H_{\hat{P}_D}(\Pi^{*G_0} \cup \{X,Y\} \mid C) \\ &= I_{\hat{P}_D}(X;\Pi^{*G_0} \cup \{Y\} \mid C) - I_{\hat{P}_D}(X;\Pi^{*G_0} \mid C). \end{aligned}$$

Finally, each term in (31) is, by definition, given by

$$\begin{aligned} I_{\hat{P}_D}(Y;\Pi^{*G_0} \cup \{X\};C) &= I_{\hat{P}_D}(Y;\Pi^{*G_0} \cup \{X\} \mid C) - \underbrace{I_{\hat{P}_D}(Y;\Pi^{*G_0} \cup \{X\})}_{E_1} \\ I_{\hat{P}_D}(Y;\Pi^{*G_0};C) &= I_{\hat{P}_D}(Y;\Pi^{*G_0} \mid C) - \underbrace{I_{\hat{P}_D}(Y;\Pi^{*G_0})}_{E_2} \\ I_{\hat{P}_D}(X;\Pi^{*G_0} \cup \{Y\};C) &= I_{\hat{P}_D}(X;\Pi^{*G_0} \cup \{Y\} \mid C) - \underbrace{I_{\hat{P}_D}(X;\Pi^{*G_0} \cup \{Y\})}_{E_3} \\ I_{\hat{P}_D}(X;\Pi^{*G_0};C) &= I_{\hat{P}_D}(X;\Pi^{*G_0} \mid C) - \underbrace{I_{\hat{P}_D}(X;\Pi^{*G_0})}_{E_4}. \end{aligned}$$

Since by definition of mutual information we have that

$$\begin{aligned}
&I_{\hat{P}_D}(Y;\Pi^{*G_0}\cup\{X\}) - I_{\hat{P}_D}(Y;\Pi^{*G_0}) = \\
&= H_{\hat{P}_D}(Y) + H_{\hat{P}_D}(\Pi^{*G_0}\cup\{X\}) - H_{\hat{P}_D}(\Pi^{*G_0}\cup\{X,Y\}) - H_{\hat{P}_D}(Y) - H_{\hat{P}_D}(\Pi^{*G_0}) + \\
&\qquad + H_{\hat{P}_D}(\Pi^{*G_0}\cup\{Y\}) \\
&= -H_{\hat{P}_D}(\Pi^{*G_0}) + H_{\hat{P}_D}(\Pi^{*G_0}\cup\{X\}) + H_{\hat{P}_D}(\Pi^{*G_0}\cup\{Y\}) - xH_{\hat{P}_D}(\Pi^{*G_0}\cup\{X,Y\}) \\
&= I_{\hat{P}_D}(X;\Pi^{*G_0}\cup\{Y\}) - I_{\hat{P}_D}(X;\Pi^{*G_0}),
\end{aligned}$$

we know that $E_1 - E_2 = E_3 - E_4$. Thus, to prove the identity (31) it remains to show that

$$I_{\hat{P}_D}(Y;\Pi^{*G_0}\cup\{X\} \mid C) - I_{\hat{P}_D}(Y;\Pi^{*G_0} \mid C) = I_{\hat{P}_D}(X;\Pi^{*G_0}\cup\{Y\} \mid C) - I_{\hat{P}_D}(X;\Pi^{*G_0} \mid C),$$

which was already shown (in Equation (30)). This concludes the proof. □

Appendix B. Alternative Justification for Assumption 1

Observe that in the case at hand, we have some information about U_t and V_t, namely the number of times, say N_{U_t} and N_{V_t}, respectively, that U_t and V_t occur in the data set D. Moreover, we also have the number of times, say $N_{R_t} = N - (N_{U_t} + N_{V_t})$, that R_t is found in D. Given these observations, the posterior distribution of (U_t, V_t) under a uniform prior is

$$(U_t, V_t) \sim \text{Dirichlet}(N_{U_t} + 1, N_{V_t} + 1, N_{R_t} + 1). \tag{32}$$

Furthermore, we know that N_{U_t} and N_{V_t} are, in general, a couple (or more) orders of magnitude smaller than N_{R_t}. Due to this fact, most of all probability mass of (32) is found in the square $[0, p] \times [0, p]$ for some small p.

Take as an example the (typical) case where $N_{U_t} = 1$, $N_{V_t} = 0$, $N = 500$ and

$$p = E[U_t] + \sqrt{Var[U_t]} \approx E[V_t] + \sqrt{Var[V_t]},$$

and compare the cumulative distribution of Uniform$([0, p] \times [0, p])$ with the cumulative distribution of Dirichlet$(N_{U_t} + 1, N_{V_t} + 1, N_{R_t} + 1)$. (We provide more details in the supplementary material webpage.) Whenever N_{R_t} is much larger than N_{U_t} and N_{V_t}, the cumulative distribution Dirichlet$(N_{U_t} + 1, N_{V_t} + 1, N_{R_t} + 1)$ is close to that of the uniform distribution Uniform$([0, p] \times [0, p])$ for some small p, and hence, we obtain approximately Assumption 1.

Concerning independence, and by assuming that the distribution of (U_t, V_t) is given by Equation (32), it results from the neutrality property of the Dirichlet distribution that

$$V_t \perp\!\!\!\perp \frac{U_t}{1 - V_t}.$$

Since V_t is very small we have

$$V_t \perp\!\!\!\perp \frac{U_t}{1 - V_t} \approx U_t.$$

Therefore, it is reasonable to assume that U_t and V_t are (approximately) independent.

References

A. J. Bell. The co-information lattice. In *Proc. ICA'03*, pages 921–926, 2003.

J. Bilmes. Dynamic Bayesian multinets. In *Proc. UAI'00*, pages 38–45. Morgan Kaufmann, 2000.

A. M. Carvalho. Scoring function for learning Bayesian networks. Technical report, INESC-ID Tec. Rep. 54/2009, 2009.

A. M. Carvalho, A. L. Oliveira, and M.-F. Sagot. Efficient learning of Bayesian network classifiers: An extension to the TAN classifier. In M. A. Orgun and J. Thornton, editors, *Proc. IA'07*, volume 4830 of *LNCS*, pages 16–25. Springer, 2007.

D. M. Chickering. A transformational characterization of equivalent Bayesian network structures. In *Proc. UAI'95*, pages 87–98. Morgan Kaufmann, 1995.

D. M. Chickering. Learning Bayesian networks is NP-complete. In D. Fisher and H.-J. Lenz, editors, *Learning from Data: AI and Statistics V*, pages 121–130. Springer, 1996.

D. M. Chickering. Learning equivalence classes of Bayesian-network structures. *Journal of Machine Learning Research*, 2:445–498, 2002.

D. M. Chickering, D. Heckerman, and C. Meek. Large-sample learning of Bayesian networks is NP-hard. *Journal of Machine Learning Research*, 5:1287–1330, 2004.

C. K. Chow and C. N. Liu. Approximating discrete probability distributions with dependence trees. *IEEE Transactions on Information Theory*, 14(3):462–467, 1968.

T. Cover and J. Thomas. *Elements of Information Theory*. John Wiley & Sons, 2006.

S. Dasgupta. Learning polytrees. In *Proc. UAI'99*, pages 134–141. Morgan Kaufmann, 1999.

L. M. de Campos. A scoring function for learning Bayesian networks based on mutual information and conditional independence tests. *Journal of Machine Learning Research*, 7:2149–2187, 2006.

P. Domingos and M. J. Pazzani. On the optimality of the simple Bayesian classifier under zero-one loss. *Machine Learning*, 29(2–3):103–130, 1997.

J. Edmonds. Optimum branchings. *Journal of Research of the National Bureau of Standards*, 71B: 233–240, 1967.

U. M. Fayyad and K. B. Irani. Multi-interval discretization of continuous-valued attributes for classification learning. In *Proc. IJCAI'93*, pages 1022–1029. Morgan Kaufmann, 1993.

N. Friedman, D. Geiger, and M. Goldszmidt. Bayesian network classifiers. *Machine Learning*, 29 (2-3):131–163, 1997.

R. Greiner and W. Zhou. Structural extension to logistic regression: Discriminative parameter learning of belief net classifiers. In *Proc. AAAI/IAAI'02*, pages 167–173. AAAI Press, 2002.

R. Greiner, X. Su, B. Shen, and W. Zhou. Structural extension to logistic regression: Discriminative parameter learning of belief net classifiers. *Machine Learning*, 59(3):297–322, 2005.

D. Grossman and P. Domingos. Learning Bayesian network classifiers by maximizing conditional likelihood. In *Proc. ICML'04*, pages 46–53. ACM Press, 2004.

M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten. The WEKA data mining software: An update. *SIGKDD Explorations*, 11(1):10–18, 2009.

D. Heckerman, D. Geiger, and D. M. Chickering. Learning Bayesian networks: The combination of knowledge and statistical data. *Machine Learning*, 20(3):197–243, 1995.

C.-W. Hsu, C.-C. Chang, and C.-J. Lin. A practical guide to support vector classification. Technical report, Department of Computer Science, National Taiwan University, 2003.

A. Jakulin. *Machine Learning Based on Attribute Interactions*. PhD thesis, University of Ljubljana, 2005.

R. Kohavi. A study of cross-validation and bootstrap for accuracy estimation and model selection. In *Proc. IJCAI'95*, pages 1137–1145. Morgan Kaufmann, 1995.

R. Kohavi and G. H. John. Wrappers for feature subset selection. *Artificial Intelligence*, 97(1-2): 273–324, 1997.

P. Kontkanen, P. Myllymäki, T. Silander, and H. Tirri. BAYDA: Software for Bayesian classification and feature selection. In *Proc. KDD'98*, pages 254–258. AAAI Press, 1998.

E. Lawler. *Combinatorial Optimization: Networks and Matroids*. Dover, 1976.

W. J. McGill. Multivariate information transmission. *Psychometrika*, 19:97–116, 1954.

C. Meek. Finding a path is harder than finding a tree. *Journal of Artificial Intelligence Research*, 15:383–389, 2001.

D. J. Newman, S. Hettich, C. L. Blake, and C. J. Merz. UCI repository of machine learning databases, 1998. URL `http://www.ics.uci.edu/~mlearn/MLRepository.html`.

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

S. V. Pemmaraju and S. S. Skiena. *Computational Discrete Mathematics: Combinatorics and Graph Theory with Mathematica*. Cambridge University Press, 2003.

F. Pernkopf and J. A. Bilmes. Discriminative versus generative parameter and structure learning of Bayesian network classifiers. In *Proc. ICML'05*, pages 657–664. ACM Press, 2005.

T. Roos, H. Wettig, P. Grünwald, P. Myllymäki, and H. Tirri. On discriminative Bayesian network classifiers and logistic regression. *Machine Learning*, 59(3):267–296, 2005.

T. Silander, T. Roos, and P. Myllymäki. Learning locally minimax optimal Bayesian networks. *International Journal of Approximate Reasoning*, 51(5):544–557, 2010.

J. Su and H. Zhang. Full Bayesian network classifiers. In *Proc. ICML'06*, pages 897–904. ACM Press, 2006.

J. Su, H. Zhang, C. X. Ling, and S. Matwin. Discriminative parameter learning for Bayesian networks. In *Proc ICML'08*, pages 1016–1023. ACM Press, 2008.

T. Verma and J. Pearl. Equivalence and synthesis of causal models. In *Proc. UAI'90*, pages 255–270. Elsevier, 1990.

S. Yang and K.-C. Chang. Comparison of score metrics for Bayesian network learning. *IEEE Transactions on Systems, Man, and Cybernetics, Part A*, 32(3):419–428, 2002.