Discriminative Model Selection for Belief Net Structures

Yuhong Guo and **Russ Greiner**
Department of Computing Science
University of Alberta
Edmonton, Alberta T6G 2E8
{ yuhong, greiner }@cs.ualberta.ca

Abstract

Bayesian belief nets (BNs) are often used for classification tasks, typically to return the most likely class label for a specified instance. Many BN-learners, however, attempt to find the BN that maximizes a different objective function — viz., likelihood, rather than classification accuracy — typically by first using some *model selection criterion* to identify an appropriate graphical structure, then finding good parameters for that structure. This paper considers a number of possible criteria for selecting the best structure, both generative (*i.e.*, based on likelihood; *BIC*, *BDe*) and discriminative (*i.e.*, Conditional *BIC* (*CBIC*), resubstitution Classification Error (*CE*) and Bias^2+Variance (*BV*)). We empirically compare these criteria against a variety of different "correct BN structures", both real-world and synthetic, over a range of complexities. We also explore different ways to set the parameters, dealing with two issues: (1) Should we seek the parameters that maximize likelihood versus the ones that maximize *conditional* likelihood? (2) Should we use (i) the *entire* training sample first to learn the best parameters and then to evaluate the models, versus (ii) only a partition for parameter estimation and another partition for evaluation (cross-validation)? Our results show that the discriminative *BV* model selection criterion is one of the best measures for identifying the optimal structure, while the discriminative *CBIC* performs poorly; that we should use the parameters that maximize likelihood; and that it is typically better to use cross-validation here.

Keywords: machine learning, Bayesian networks

1 Introduction

While belief networks (BNs, a.k.a. Bayesian networks, graphical models) are generative models, capable of modeling a joint probability distribution over a set of variables, they are typically used *discriminatively* for some classification task — *e.g.*, to predict the probability of some disease, given some specific evidence about the patient.[1] This has motivated the growing body of work on learning an effective *BN-classifier* from a datasample [ILLP05].

In general, learning an effective BN-classifier requires first finding a good BN structure (a.k.a. model, which represents the direct dependencies among the variables) then determining appropriate parameters for this model. The first step requires searching through a space of models, seeking the element that optimizes some *model selection criterion*. This paper investigates a number of criteria, towards determining which one works best in practice — *i.e.*, which will best identify the structure whose instantiation will minimize classification error on unseen data.

This is not a trivial challenge. While one can typically improve classification performance *on the training data* by increasing the complexity of the model, this usually increases the number of parameters that must be estimated. This typically increases parameter variance, which leads to inferior generalization performance — *i.e.*, worse performance on unseen data. A *model selection criterion* attempts to operationalize this balance between complexity and goodness of fit to training data, by providing a single number for each network structure. A good model selection criterion is especially important when we have limited training data, which is the standard case.

Van Allen and Greiner [VG00] evaluated several standard *generative* criteria, where the goal is a structure that produces the best fit to the underlying distribution (using likelihood, Eqn 3). We consider two of these: *BIC* [Sch78] and *BDe* [CH92].

As noted above, our overall goal is different, as we are seeking a structure that leads to good *discriminative* performance, *i.e.*, which has the best classification performance on unseen testing data. We therefore consider several *discriminative* criteria: Conditional *BIC* (*CBIC*), resubstitution Classification Error (*CE*), and Bias^2+Variance (*BV*).

When deciding on an appropriate structure, we need to consider how to instantiate its parameters (CPtables). This leads to two issues. (1) Should we use the parameters that optimize the simple likelihood of the data (*i.e.*, the standard *generative* approach), versus the *discriminative* parameters that optimize *conditional* likelihood [NJ01]. (2) The learner has access to a corpus of training data, both to find the best parameters for each structure, and also to evaluate the quality of this instantiated model. Should the learner use the *same* data for both tasks, or should it instead partition the training sample into two subsamples, and use the first for parameter instantiation, and the second for model selection, perhaps in a cross-validation fashion?

The rest of this section discusses related work. Section 2

[1] Personal conversation with B Boerlage, Norsys.

provides the framework for this paper, describing belief networks, our model selection criteria and parameter estimation. Section 3 presents our experimental setup and results. As our preliminary experimental results, on data from a real-world distribution, suggest the performance of each criterion may be related to complexity of the Markov blanket around the class variable, we therefore systematically explore the effectiveness of various model selection criteria across generative models with a wide range of Markov blanket complexities. The webpage [Gre05] contains additional information about the experiments reported here, as well as other related results.

Three preliminary comments: (1) There are many reasons to select some specific criteria, some of which relate more to prior assumptions and constraints, than to performance. In this paper, however, we are *only* concerned with eventual classification performance, as measured by Eqn 1. (2) While most of these criteria are known to be asymptotically correct, our interest is with the practical use of these criteria. We therefore focus on small sample sizes. (3) Our goal is to better understand model selection criteria, divorced with the search issues associated with learning itself. We therefore follow the standard framework for evaluating criteria [VG00; KMNR97]: consider only a small set of models, small enough that *each* can be evaluated.

1.1 Related Work

There is a large literature on the general model selection problem, proposing a variety of schemes including *BIC* [Sch78] (related to Minimum Description Length [Ris89]) and *AIC* [Boz87]; our analysis (here and in [Gre05]) includes each of these schemes in the context of learning BN-structures.

There are also many papers on learning the structure of belief nets, but most focus on *generative* learning [Hec98]. [VG00] provides a comprehensive comparison of selection criteria when learning belief network structures *generatively*. While we borrow some of the techniques from these projects (and from [KMNR97]), recall our goal is learning the structure that is best for a *discriminative* classification task.

As noted earlier, belief nets are often used for this classification task. This dates back (at least) to NaïveBayes classifiers [DH73], and has continued with various approaches that include feature selection [LS94], and alternative structures [FGG97; CG99]. Kontkanen *et al.* [KMST99] compared several model selection criteria (unsupervised/supervised marginal likelihood, supervised prequential likelihood, cross validation) on a restricted subset of belief nets structures. Grossman and Domingos [GD04] presented an algorithm for discriminatively learning belief networks that used the conditional likelihood of the class variable given the evidence (Eqn 2) as the model selection criterion. Our work differs by proposing several new discriminative model selection criteria (including a variant of a generative criteria (*CBIC*), and another (*BV*) motivated by the classification task in general [Rip96]), and by providing a comprehensive comparison between classical generative model selection criteria and various discriminative criteria on the task of learning good structures for a BN-classifier.

2 Framework

2.1 Belief Network Classifiers

We assume there is a stationary underlying distribution $P(\,\cdot\,)$ over n (discrete) random variables $\mathcal{V} = \{V_1, \ldots, V_n\}$, which we encode as a "(Bayesian) belief net" (BN) — a directed acyclic graph $B = \langle \mathcal{V}, A, \Theta \rangle$, whose nodes $\mathcal{V}$ represent variables, and whose arcs A represent dependencies. Each node $D_i \in \mathcal{V}$ also includes a conditional-probability-table (CPtable) $\theta_i \in \Theta$ that specifies how D_i's values depend (stochastically) on the values of its immediate parents. In particular, given a node $D \in \mathcal{V}$ with immediate parents $\mathbf{F} \subset \mathcal{V}$, the parameter $\theta_{d|\mathbf{f}}$ represents the network's term for $P(\,D = d \,|\, \mathbf{F} = \mathbf{f}\,)$ [Pea88].

The user interacts with the belief net by asking *queries*, each of the form "What is $P(\,C = c \,|\, \mathbf{E} = \mathbf{e}\,)$?" — *e.g.*,

"What is $P\left(\texttt{Cancer = true} \,\middle|\, \begin{array}{l} \texttt{Gender=male} \\ \texttt{Smoke=true} \end{array} \right)$?"

— where $C \in \mathcal{V}$ is a single "class variable", $\mathbf{E} \subset \mathcal{V}$ is the subset of "evidence variables", and c (resp., $\mathbf{e}$) is a legal assignment to C (resp., $\mathbf{E}$).[2]

Given any unlabeled instance $\mathbf{E} = \mathbf{e}$, the belief net B will produce a distribution P_B over the values of the class variable; perhaps $P_B(\,\texttt{Cancer = true} \,|\, \mathbf{E} = \mathbf{e}\,) = 0.3$ and $P_B(\,\texttt{Cancer = false} \,|\, \mathbf{E} = \mathbf{e}\,) = 0.7$. In general, the associated H_B classifier system will then return the value

$$H_B(\mathbf{e}) = \underset{c}{\operatorname{argmax}} \{ P_B(\,C = c \,|\, \mathbf{E} = \mathbf{e}\,) \}$$

with the largest posterior probability — here return $H_B(\,\mathbf{e}\,) =$ false as $P_B(\,\texttt{Cancer = false} \,|\, \mathbf{E} = \mathbf{e}\,) > P_B(\,\texttt{Cancer = true} \,|\, \mathbf{E} = \mathbf{e}\,)$.

A good belief net classifier is one that produces the appropriate answers to these unlabeled queries. We will use "classification error" (aka "0/1" loss) to evaluate the resulting B-based classifier H_B

$$\text{err}(\,B\,) \quad = \sum_{\langle \mathbf{e}, c \rangle : H_B(\mathbf{e}) \neq c} P(\,\mathbf{e}\,,\,c\,) \qquad (1)$$

Our goal is a belief net B^* that minimizes this score, with respect to the true distribution $P(\,\cdot\,)$. While we do not know this distribution *a priori*, we can use a sample drawn from this distribution to help determine which belief net is optimal. We will use a training set S of $m = |S|$ complete instances, where the ith instance is represented as $\langle c^i,\, e^i_1, \ldots, e^i_n \rangle$. This paper focuses on the task of learning

[2] This paper focuses on the standard "machine learning" case, where all queries involve the same variable (*e.g.*, all queries ask about Cancer), and we assume the distribution of conditioning events matches the underlying distribution, which means there is a single distribution from which we can draw instances, which correspond to a set of labeled instances (aka "labeled queries"). See [GGS97] for an alternative position, and the challenges this requires solving.

the BN-structure $G = \langle \mathcal{V}, A \rangle$ that allows optimal classification performance (Eqn 1) on unseen examples.

Conditional Likelihood: Given a sample S, the empirical "log conditional likelihood" of a belief net B is

$$\mathit{LCL}^{(S)}(\,B\,) \;=\; \frac{1}{|S|} \sum_{\langle \mathbf{e},c\rangle \in S} \log(\,P_B(\,c\,|\,\mathbf{e}\,)\,) \qquad (2)$$

where $P_B(\,c\,|\,\mathbf{e}\,)$ represents the conditional probability produced by the belief network B. Previous researchers [MN89; FGG97] have noted that maximizing this score will typically produce a classifier that comes close to minimizing the classification error (Eqn 1).

Note that this $\mathit{LCL}^{(S)}(\,B\,)$ formula is significantly different from the (empirical) "log likelihood" function

$$\mathit{LL}^{(S)}(\,B\,) \;=\; \frac{1}{|S|} \sum_{\langle \mathbf{e},c\rangle \in S} \log(\,P_B(\,c, \mathbf{e}\,)\,) \qquad (3)$$

used as part of many *generative* BN-learning algorithms [FGG97].

We will measure the complexity of the BN B as the number of free parameters in the network

$$k(B) \;=\; \sum_{i=1}^{n} (|V_i| - 1) \prod_{F \in \mathrm{Pa}(\,V_i\,)} |F| \qquad (4)$$

where $|V|$ is the number of values of any variable V, and Pa(V) is the set of immediate parents of the node V.

For a belief network structure, given a completely instantiated tuple, a variable C is only dependent on the variables in its immediate Markov Blanket [Pea88], which is defined as the union of C's direct parents, C's direct children and all direct parents of C's direct children. We define $k_C(\,B\,)$ as the number of parameters in C's Markov blanket, within B, using an obvious analogue to Eqn 4.

2.2 Generative Model Selection Criteria

Most of the *generative* criteria begin with the average empirical log likelihood of the data, Eqn 3, as $\mathit{LL}^{(S')}(\,B\,)$ on *unseen* data S' is useful as an unbiased estimate of the average generative quality of the distribution B. *BIC* uses this measure on *training* data, but attempts to avoid overfitting by adding a "regularizing" term that penalizes complex structures:

$$\mathit{BIC}^{(S)}(\,B\,) \;=\; -\mathit{LL}^{(S)}(\,B\,) \;+\; \frac{k(B)\,\log|S|}{2\,|S|}$$

Another generative model selection criterion is the marginal likelihood — averaged over all possible CPtable values (in the Bayesian framework):

$$\mathit{BDe}^{(S)}(\,B\,) = \prod_{i=1}^{n} \prod_{j=1}^{q_i} \frac{\Gamma(\alpha_{ij})}{\Gamma(\alpha_{ij} + a_{ij})} \prod_{k=1}^{|V_i|} \frac{\Gamma(\alpha_{ijk} + a_{ijk})}{\Gamma(\alpha_{ijk})},$$

where $q_i = \prod_{F \in \mathrm{Pa}(\,V_i\,)} |F|$ is the number of states of the parents of variable V_i, α_{ijk} are the Dirichlet prior parameters (here set to 1), $\alpha_{ij} = \sum_{k=1}^{|V_i|} \alpha_{ijk}$, and a_{ijk} are the empirical counts — *i.e.*, the number of instances in the datasample S where the ith variable V_i takes its kth value and V_i's parents take their jth value.

2.3 Discriminative Model Selection Criteria

The *CBIC* (conditional *BIC*) criterion is a discriminative analogue of the generative *BIC* criterion, which differs by using log *conditional* likelihood to measure "training error" and by using $k_C(\,B\,)$ rather than $k(B)$ as the number of parameters.

$$\mathit{CBIC}^{(S)}(\,B\,) \;=\; -\mathit{LCL}^{(S)}(\,B\,) + \frac{k_C(\,B\,)\log|S|}{2\,|S|}$$

As we use classification error on testing data to measure a BN-classifier's performance, we include its classification error (*CE*) on *training data* as a discriminative model selection criterion.

$$\mathit{CE}^{(S)}(\,B\,) \;=\; \frac{|\{\langle \mathbf{e}, c\rangle \in S \mid H_B(\mathbf{e}) \neq c\}|}{|S|} \qquad (5)$$

[Rip96] proves that the expected mean-square-error of a classifier corresponds to "Bias^2+Variance",

$$\mathit{BV}^{(S)}(\,B\,) \;=\;$$

$$\frac{1}{|S|} \sum_{\langle c,\mathbf{e}\rangle \in S} [\,t(\,c\,|\,\mathbf{e}\,) \;-\; P_B(\,c\,|\,\mathbf{e}\,)]^2 \;+\; \hat{\sigma}^2[P_B(\,c\,|\,\mathbf{e}\,)]$$

where the "true" response $t(\,c\,|\,\mathbf{e}\,)$ corresponds to the empirical frequency within the training data:

$$t(\,c\,|\,\mathbf{e}\,) \;=\; \frac{\#_S(C\!=\!c, \mathbf{E}\!=\!\mathbf{e})}{\#_S(\mathbf{E}\!=\!\mathbf{e})}$$

where $\#_S(\mathbf{E} = \mathbf{e})$ is the number of instances in training set S that match this (partial) assignment, and we use the (Bayesian) variance estimate provided in [VGH01]:

$$\hat{\sigma}^2[P_B(\,c\,|\,\mathbf{e}\,)] \;=\;$$

$$\sum_{\theta_{D|\mathbf{f}} \in \Theta} \frac{1}{n_{D|\mathbf{f}}} \left[\begin{array}{l} \displaystyle\sum_{d \in D} \frac{1}{\theta_{d|\mathbf{f}}} [P_B(\,d, \mathbf{f}, c\,|\,\mathbf{e}\,) - P_B(\,c\,|\,\mathbf{e}\,)\,P_B(\,d, \mathbf{f}\,|\,\mathbf{e}\,)]^2 \\ -\; (P_B(\,\mathbf{f}, c\,|\,\mathbf{e}\,) \;-\; P_B(\,c\,|\,\mathbf{e}\,)\,P_B(\,\mathbf{f}\,|\,\mathbf{e}\,)\,)^2 \end{array} \right]$$

which requires summing over the CPtable rows $\theta_{D=d|\mathbf{F}=\mathbf{f}}$, and uses $n_{D|\mathbf{F}=\mathbf{f}} = 1 + |D| + \#_S(\mathbf{F}\!=\!\mathbf{f})$ as the "effective sample size" of the conditioning event for this row.[3]

[3] Note this is done from a Bayesian perspective, where we first identify each CPtable row with a Dirichlet-distributed random variable, then compute its posterior based on the training sample, and finally use these posterior distributions of the CPtable rows to compute the distribution over the response to the $B_\Theta(\,c\,|\,\mathbf{e}\,)$, which is after-all just a function of those random variables, $\Theta = \{\theta_{d|\mathbf{f}}\}$.

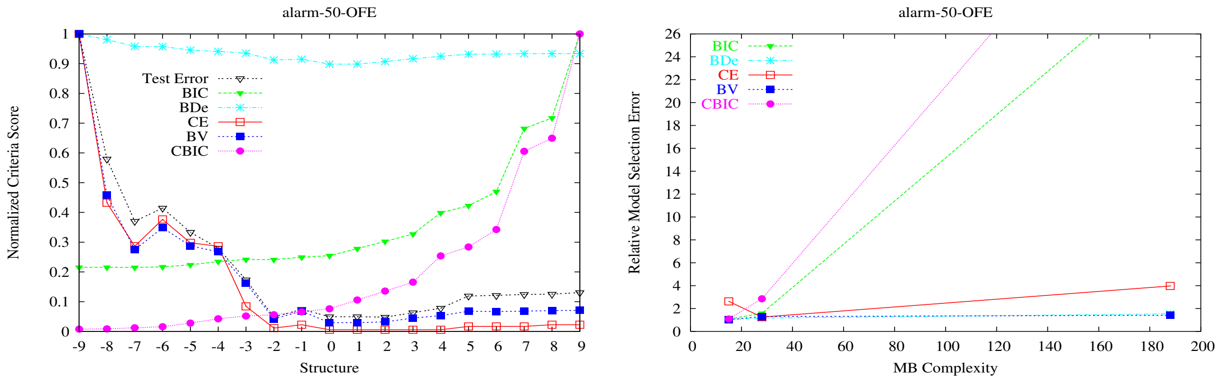

Figure 1: (a) Criteria Score, as function of Structure; (b) Relative Score of Various Criteria, over 3 queries in ALARM Network.

2.4 How to Instantiate the Parameters

As mentioned above, a belief net includes both a structure *and a set of parameters* for that structure. Given complete training data, the standard parameter learning algorithm, OFE, sets the value of each parameter to its empirical frequency in the datasample, with a Laplacian correction:

$$\theta_{D=d|\mathbf{F}=\mathbf{f}} \quad = \quad \frac{\#_S(D=d,\mathbf{F}=\mathbf{f})+1}{\#_S(\mathbf{F}=\mathbf{f})\ +\ |D|}$$

Cooper and Herskovits [CH92] prove these generative values correspond to the mean posterior of a distribution whose prior was a uniform Dirichlet; moreover, they optimize the *likelihood* of the data, Eqn 3, for the given structure.

The ELR algorithm [GZ02; GSSZ05], by contrast, attempts to find the parameters that optimize the discriminative *conditional likelihood* score. (This algorithm extends logistic regression as it applies to arbitrary network structures, while standard logistic regression corresponds to naive bayes.)

In either case, the learner has access to a training sample S, to use as it wishes when producing the optimal structure. A simple model selection process will use the "undivided sample" approach: Use *all* of S when finding the appropriate instantiation of the parameters, then compute a score for this instantiated structure, based again on S. Note this "1Sample" approach was the motivation for many of the scoring criteria. We compare this to the obvious "cross-validation" approach (5CV): first partition the data into 5 subsets, and for each subset S_i, use the other 4 subsets to fit parameters then compute the score of the result on S_i. We repeat this 5 times, and average the scores. (Note "score" refers to the actual criterion, which is not necessarily Eqn 5.)

3 Empirical Studies

This section reports on our empirical studies that compare the 5 model selection criteria mentioned above, to help determine when (if ever) to use each, and to investigate the parameter-instantiation issues listed in Section 2.4. We therefore asked each of the criteria to identify the appropriate structure, across a range of situations. Section 3.1 first explains how we will run each experiment, and how we will evaluate the results. Section 3.2 presents our first study, on a real-world distribution. Here, we use OFE (rather than ELR) to instantiate the parameters, and use "1Sample" (*i.e.*, only a single undivided training sample, for both instantiating the parameters and for evaluating the resulting instantiated network). This data suggests that the complexity of the generative model may strongly affect which criterion works best. The remaining subsections explore this. Section 3.3 (resp., 3.4) considers the performance of the selection criteria on a set of synthetic models with a range of complexities, using 1Sample (resp., 5CV sample). Section 3.5 then considers model selection when ELR is used to find parameters; and Section 3.6 explicitly compares the different approaches to instantiating the parameters.

3.1 Experimental Setup

In each experiment, we have a specific "true" distribution $P(\cdot)$ — *i.e.*, correct BN-structure and parameters — that we either download, or synthesize. We produce a set of possible candidate models by modifying the true structure; see below. We also generate a number of complete datasamples from the true $P(\cdot)$, of various sizes. For each training sample we then run each of the selection criteria in the appropriate context: 1Sample vs 5CV, and OFE vs ELR. Each criteria produces a single number for each candidate structure. Figure 1(a) shows this, in the context of the ALARM [BSCC89] network (Section 3.2).[4] Each criteria then identifies the structure it considers best — *i.e.*, with the lowest score. Here, for example, *CBIC* would select the structure labeled "-9", *BIC* would pick "-7", *BV* would select "+1" and *BDe*, "0". (These numbers correspond to the number of arcs added, or deleted, to the initial structure. Hence, the original structure is labeled "0".) For each criteria χ, let B^χ be this selected structure, instantiated appropriately. We then compute the error of each B^χ, based on a hold-out sample S' of size $|S'| = 1000$, generated from $P(\cdot)$ — *i.e.*, $err^{(S')}(\,B^\chi\,)$.

We also let $B^* = \text{argmin}_B\{err^{(S')}(\,B\,)\}$ denote the best structure. (See the "Test Error" line in Figure 1(a); notice

[4] Each measure is normalized to fit between 0 (best) and 1.

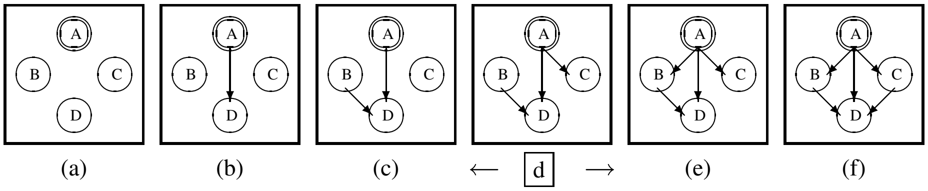

Figure 2: Sequence of structures; (d) is the original structure

this picks "+2". That is, the structure that is best for a particular sample need not be the original structure!) We measure the performance of each criterion χ by its "relative model selection error", $err^{(S')}(\, B^{\chi}\,)/err^{(S')}(\, B^{*}\,)$. For a given sample size, we compute the average over 20 repeated trials, each time using a different training set. This ratio will be 1 for a perfect criteria; in general, it will be higher. [5]

Proper model selection is most challenging, and hence more relevant, when given limited training data; this paper therefore focuses on a very small training sample — *e.g.*, of 20 instances. (The results for other sizes were similar; see [Gre05].)

Generating Sequence of Structures: Given a true BN-structure G^*, we generate a sequence of BN-structure candidates with decreasing/increasing complexities, as follows:

1. Starting from the original structure, sequentially remove one randomly-selected arc from the Markov blanket (MB) of the class variable, to generate a series of structures whose class variable has decreasing MB size.
2. Starting from the original structure, sequentially add one randomly-selected arc to the Markov blanket of the class variable, to generate a series of structures whose class variable has increasing MB size.[6]

See Figure 2, where a is the class variable. Here (d) is the starting point, and we produce (c), (b) and (a) by deleting existing arcs, and produce (e) and (f) by adding new arcs.

Each experiment will begin with its own "true structure" G^*. When G^* includes few arcs, the generated set of candidate structures will include very small structures — including one with 0 arcs, another with 1 arc, and so forth. Notice these correspond to the structures that a standard learning algorithm would consider, as it "grows" a structure from nothing.

3.2 Exp I: Real-World Distr'n, 1Sample, OFE

Our preliminary investigations examined several real-world belief nets; here we focus on ALARM [BSCC89]. We considered three different variables to serve as the class variable C_i, which produced three different query forms, whose Markov blankets $k_{C_i}(\,\text{ALARM}\,)$ had a wide range in size: 15, 28, 188.

As outlined above, we computed the relative model selection error for each criterion, $err^{(S')}(\, B^{\chi}\,)/err^{(S')}(\, B^{*}\,)$. Figure 1(b) is the result when we used a sample of size $m = 50$. We found that *BV* performed well throughout, with *BDe* being very close; but the other measures were generally inferior. ([Gre05] shows similar performances on other sample sizes, and for various queries on different networks.)

3.3 Exp II: Synthetic Distribution, 1Sample, OFE

We observed different behavior of the various selection criteria as we varied the complexity of the Markov blanket around the class variable. To further explore this, we generated a set of synthetic networks, whose class variables could have arbitrary Markov blanket complexity (aka "MB complexity"). We will use the networks here and below.

We first randomly generated six groups of belief network structures with varying Markov blanket complexity, where each group includes 30 structures. We sequentially made each of these the gold standard, used to generate datasamples.

We used the experimental apparatus described in Section 3.1 to test the behavior of each criterion, across a spectrum of complexities and a range of sample sizes. The graph in Figure 3(a) show the results for belief networks with seven variables, over a sample of size $m = 20$, using an undivided training sample (1Sample), and the generative for estimating parameters (OFE). The complexity (on the X axis, from 1 to 6) represents the six group of structures, with increasing MB complexity.

This plot shows that the *BV*, *CE* and *BDe* criteria perform comparably across the MB complexities, and each is typically (far) superior to *BIC* and *CBIC*. ([Gre05] shows this holds for other training sizes as well.)

The *BIC* and *CBIC* criteria perform well only when the MB complexity is very small, otherwise, they perform very poorly. Our experiments reveal why: These criteria have too strong a preference for simple structures, as they almost always pick the simplest structure in the sequence, irrespective of the data. (Notice this data *was* sufficient to tell the other measures to prefer other larger structures.) This is consistent with the [VG00] observation that the *BIC* criterion seriously underfits — indeed, for small samples, it almost invariably produced no arcs. This suggests the complexity penalty term for *BIC*/*CBIC* may be too big, and not appropriate for belief network on most cases. Given that *CBIC*'s

[5] We considered an alternative evaluation method: simply measure how often each method "correctly" selected the original structure. We prefer our evaluation measure for two reasons: First, the original structure might not be the optimal structure, given this datasample. Second, we wanted to quantify how bad the loss was.

[6] Note that adding one more arc might increase the MB size by more than one, since adding one arc into the Markov blanket may cause some other arcs of the original network to now become part of the Markov blanket of the class variable. Similarly removing a single arc may reduce the MB size by more than 1.

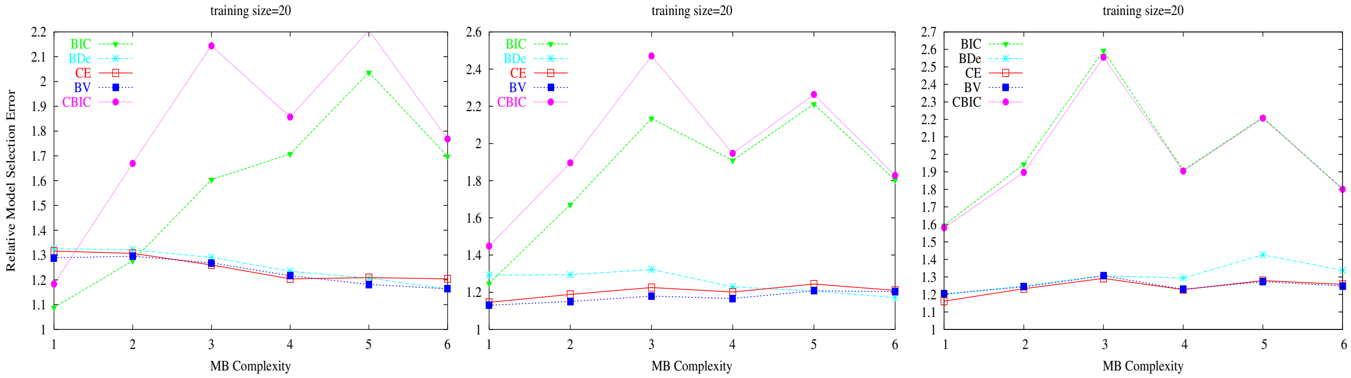

Figure 3: For 7-Variable BN, $m = 20$: (a) Exp II: 1Sample, OFE; (b) Exp III: 5CV, OFE; (c) Exp IV: 5CV, ELR

penalty term $k_c(B)$ is smaller than *BIC*'s $k(B)$, we were at first surprised that *CBIC* typically performed even *worse* than *BIC*. However, we attribute this to the observation that the $-LCL^{(S)}(B)$ term of *CBIC* is so much smaller than the $-LL^{(S)}(B)$ term of *BIC* that even the relatively small $k_c(B)$ is very influential, which increases *CBIC* tendency to underfit. ([Gre05] shows that increasing the training set size does improve the performance of those complexity penalized criteria.)

We attribute the good performance of the generative *BDe* measure to the observation that it is actually *averaging* over all parameter values, rather than being forced to pick a particular set of parameters, which could be problematic given a small training dataset.

3.4 Exp III: Synthetic Distribution, 5CV, OFE

Figure 3(b) shows the results of the 5CV variant (still using OFE), again on the 7-variable case with $m = 20$ training instances. In general, we can see that *BV* is often the best, closely followed by *CE*, then often *BDe*. Once again, we see that *BIC* and *CBIC* perform significantly worse; even with 5CV, they continue to select the simplest structure in almost all cases.

Here, the *CE* score corresponds to the standard 5-fold CV. Note that it does not always produce the best response; (5fold) *BV* is typically better!

3.5 Exp IV: Synthetic Distribution, 5CV, ELR

Here we ran the same experiments, but using ELR rather than OFE to instantiate the parameters. (Given that 5CV typically worked better than 1Sample, we only show the 5CV results here — see Section 3.6.) As shown in Figure 3(c), we again see that *BV* and *CE* appear to be the best, then *BDe*; *CBIC* and *BIC* are again not even close.

3.6 Comparing Parameter Estimation Regimes

The graphs shown above provide the *relative* model selection error for the different model selection criteria, for a fixed "context" (here, "way of instantiating the parameters"). While this is appropriate for our within-context comparisons, it does not allow us to compare these different contexts, to determine which leads to the smallest absolute error.

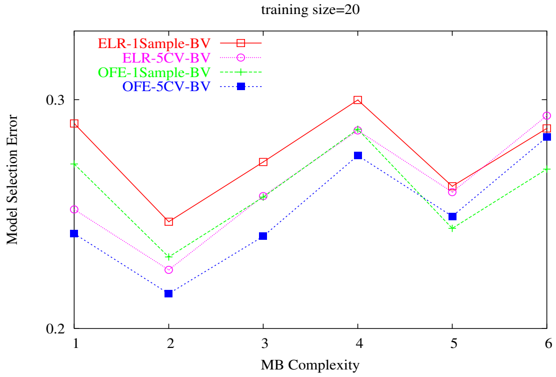

Figure 4: Different ways to instantiate parameters; Average Raw Performance of *BV* (7-Variable BN, $m = 20$)

We therefore computed the average 0/1-classification error (Eqn 1) obtained when using *BV* criterion (which was one of the top measures), for each of the 4 different ways to instantiate parameters: 1Sample vs 5CV and OFE vs ELR. Figure 4 shows the results, across a range of MB complexities. We note first that the optimal approach was always based on OFE rather than ELR; note this is consistent with [GD04]. In fact, it appears that 1Sample+ELR is tracking 1Sample+OFE, but at a slightly inferior error; and that 5CV+ELR is similarly tracking 5CV+OFE, For both ELR and OFE, we see that 5CV does better than the 1Sample variant for low MB complexity, but the situation reverses as the MB complexity increases.

We obtained similar results when we performed similar experiments with other model selection criteria, and for each, across several sample sizes $m \in \{10, 20, 50\}$, and also on a set of larger belief networks (*e.g.*, with 15 variables); see [Gre05].

3.7 Other Experiments

We also computed the (resubstitution) log likelihood criteria itself (Eqn 3) in each of these cases, as this measure is often used in the literature. But as it did not perform as well as the other criteria, we do not show those results here. We also omit both the *AIC* criterion (Akaike Information Criteria) and its discriminative analogue *CAIC*, as they behaved very similarly to *BIC* and *CBIC*. See [Gre05].

4 Conclusions

Belief nets are often used as classifiers. When learning the structure for such a BN-classifier, it is useful to have a criterion for evaluating the different candidate structures that corresponds to this objective. We proposed a number of novel *discriminative* model selective criteria, one (*CBIC*) being an analogue of a standard generative criterion (commonly used when learning generative models), and another (*BV*) motivated by the familiar discriminative approach of decomposing error into bias and variance components. We then evaluated these methods, along with the generative ones, across a number of different situations: over queries of different complexities and different ways to use the training sample — 1Sample vs 5CV, and OFE VS ELR.

As our underlying task is discriminative, we had anticipated that perhaps all of the discriminative methods would work well. This was only partly true: while one discriminate method *BV* is among the best criteria, another (*CBIC*) performed very poorly.[7] We also expected 5CV to be uniformly superior to the 1Sample approach. While our empirical evidence shows that this was not always true, we note that even when 5CV was inferior, it was never much worse. Finally, based on [GD04], we were not surprised that using the discriminative way to set the parameters (ELR) did not dominate the generative approach (OFE).

Our main contributions are defining the *BV* criterion, and providing empirical evidence that it performs effectively, especially for small samples. (While the *CBIC* criterion is also discriminative, our empirical evidence argues strongly against using this measure.) Based on our data, we also recommend (1) using the simpler OFE approach to estimating the parameters, and (2) using the standard 5CV approach.

Acknowledgments Both authors thank Dale Schuurmans for his advice on this project, Ajit Singh for the code for computing variances, and NSERC and the Alberta Ingenuity Centre for Machine Learning for funding.

References

[Boz87] H. Bozdogan. Model selection and Akaike's Information Criterion (AIC): the general theory and its analytical extensions. *Psychometrica*, 52, 1987.

[BSCC89] 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 *European Conference on Artificial Intelligence in Medicine*, 1989.

[CG99] J. Cheng and R. Greiner. Comparing Bayesian network classifiers. In *UAI*, 1999.

[CH92] G. Cooper and E. Herskovits. A Bayesian method for the induction of probabilistic networks from data. *Machine Learning Journal*, 9:309–347, 1992.

[DH73] R. Duda and P. Hart. *Pattern Classification and Scene Analysis*. Wiley, 1973.

[FGG97] N. Friedman, D. Geiger, and M. Goldszmidt. Bayesian network classifiers. *Machine Learning Journal*, 29, 1997.

[GD04] D. Grossman and P Domingos. Learning Bayesian network classifiers by maximizing conditional likelihood. In *ICML2004*, 2004.

[GGS97] R. Greiner, A. Grove, and D. Schuurmans. Learning Bayesian nets that perform well. In *UAI-97*, 1997.

[Gre05] 2005. http://www.cs.ualberta.ca/∼greiner-/DiscriminativeModelSelection.

[GSSZ05] 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), 2005.

[GZ02] R. Greiner and W. Zhou. Structural extension to logistic regression. In *AAAI*, 2002.

[Hec98] D. Heckerman. A tutorial on learning with Bayesian networks. In *Learning in Graphical Models*, 1998.

[ILLP05] I. Inza, P. Larranaga, J. Lozano, and J. Pena. *Special Issue of* Machine Learning Journal*: Probabilistic Graphical Models for Classification*, volume 59. May 2005.

[KMNR97] M. Kearns, Y. Mansour, A. Ng, and D. Ron. An experimental and theoretical comparison of model selection methods. *Machine Learning*, 27, 1997.

[KMST99] P. Kontkanen, P. Myllymäki, T. Silander, and H. Tirri. On supervised selection of Bayesian networks. In *UAI99*, 1999.

[LS94] P. Langley and S. Sage. Induction of selective bayesian classifiers. In *UAI-94*, 1994.

[MN89] P. McCullagh and J. Nelder. *Generalized Linear Models*. Chapman and Hall, 1989.

[NJ01] A. Ng and M. Jordan. On discriminative versus generative classifiers: A comparison of logistic regression and naive Bayes. In *NIPS*, 2001.

[Pea88] J. Pearl. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann, 1988.

[Rip96] B. Ripley. *Pattern Recognition and Neural Networks*. Cambridge University, 1996.

[Ris89] J. Rissanen. *Stochastic complexity in statistical inquiry*. World Scientific, 1989.

[Sch78] G. Schwartz. Estimating the dimension of a model. *Annals of Statistics*, 6:461–464, 1978.

[VG00] T. Van Allen and R. Greiner. Model selection criteria for learning belief nets. In *ICML*, 2000.

[VGH01] T. Van Allen, R. Greiner, and P. Hooper. Bayesian error-bars for belief net inference. In *UAI01*, 2001.

[7] This is not particularly surprising, given that its generative form *BIC* is known to seriously underfit [VG00].