Comparing Model Selection Criteria for Belief Networks

Tim Van Allen, Russ Greiner
Department of Computing Science
University of Alberta

August 14, 2000

Abstract

Learning belief network structure involves a trade-off between network complexity and goodness of fit. More complex structures allow for a better fit to the data, but suffer from a decreased ability to generalize to unseen data. This bias-variance trade-off may be operationalized by a scoring function on structures, called a model selection criterion. We describe various criteria for model selection: a prequential, or Bayesian, criterion, 2-fold cross validation, a bootstrap criterion, Akaike's information criterion, and a minimum description length criterion. We carry out an empirical comparison of these criteria and identify differences in how each one handles the bias-variance trade-off. We discuss some of the theoretical problems and issues involved in model selection.

1 Introduction

Belief networks (also known as Bayesian networks, graphical models) are commonly used to model joint probability distributions in expert systems. A graph is used to represent all direct dependencies between the variables. This structure both organizes reasoning and reduces the dimensionality of the parameter space to the point where it is feasible to estimate parameters from sample data. When network parameters have been learned from a random sample, those parameters are themselves random variables, adding an additional level of uncertainty, which may be represented by a posterior distribution over parameter values (Bayesian perspective) or by the sampling distribution of the unknown true parameters (Frequentist perspective).

The uncertainty in parameters is an issue when learning network structure. One can typically obtain an improved fit to the data by increasing the complexity of the structure, and thus the dimensionality of the parameter space; however, this increased representational power comes at a cost: namely the increase in parameter variance. Finding the appropriate balance between complexity and goodness of fit is a matter of handling this *bias-variance trade-off.* The trade-off may be operationalized as a scoring function on network structures known as a *model selection criterion.* Various criteria have been proposed for model selection — among them: Akaike's information criterion (AIC), cross-validation, bootstrapping, the marginalized log-likelihood, and minimum description length (MDL).

The choice of a model selection criterion reflects not only a paradigm for inductive inference (Frequentist: AIC, cross-validation, bootstrap; Bayesian: marginalized log-likelihood; Information-theoretic: MDL), but also prior beliefs and pragmatic considerations as well. Intuitions about the domain of induction and its goals, however, are often tacit and consequently difficult to express in the mathematical formalism of an evaluation function. It is therefore of interest to empirically explore the consequences of applying a particular criterion to a particular domain, especially in small sample contexts where model selection is so critical. Toward this end, we compare several model selection criteria (those mentioned above) applied to the task of learning belief networks. Through this comparison, we hope to shed light on some of the issues involved in choosing a criterion, and to make clearer the nature of the trade-off embodied in each.

Overview

Section 2 reviews related work and presents our notation. Section 3 introduces belief networks and their properties — in particular, issues involving parameter estimation and the posterior distribution over the parameter space. Section 4 introduces the model selection problem and the notion of a model selection criterion, then presents the criteria under consideration and discusses each one from a theoretical standpoint. In Section 5 we describe the experimental apparatus used to compare criteria on various model selection problems. Section 6 describes our results, and Section 7 is a discussion of those results and their implications. Appendix A gives details of several results that are used in the paper.

2 Preliminaries

Related Work

Early work on induction in AI tended to focus on symbolic manipulation and logical (truth-functional) representations. Pearl (1988) was instrumental in directing attention to probabilistic methods, and, in particular, to belief networks. Cooper and Herskovits (1992) established the basic results about the posterior distribution of belief network parameters, upon which we rely. There is a considerable literature on learning belief networks, and in particular, on learning their structure; see Heckerman (1995) for a detailed overview of the subject. Note that many researchers, including Lam and Bacchus (1994) Suzuki (1996), and Friedman and Goldszmidt (1996) explicitly use the MDL criterion (or something close to it) to evaluate candidate networks. We explore the behaviour of this MDL criterion, among others. Friedman and Yakhini (1996) carry out an analysis of the sample requirements for various complexity penalty approaches to belief net learning. While that work also addresses suitability of various selection criteria, its analysis is theoretical and based on asymptotic behaviour, and it only considers complexity penalization; by contrast we are empirically investigating small sample behaviour over a different class of criteria, including Bayesian, bootstrap and cross-validation criteria.

Linhart and Zucchini (1986) provide an overview of the general problem of model selection, covering AIC and cross-validation, but not MDL. Rissanen (1989) gives a detailed development of the *Minimum Description Length Principle*, which is the information-theoretic view of induction that the MDL criterion is based on. Schwarz (1978) gives an alternative derivation of the MDL criterion (therein referred to as BIC, the *Bayesian Information Criterion*) as a large-sample approximation of the Bayesian posterior criterion. Bozdogan (1987) gives the derivation of AIC and a discussion of its use. Kearns et al (1997) describe experiments similar to our own. They make a similar comparison between an MDL criterion and a cross-validation criterion, and report similar behaviour in a different context: learning a function $f : [a, b] \to \{0, 1\}$ from noisy examples, under zero-one loss.

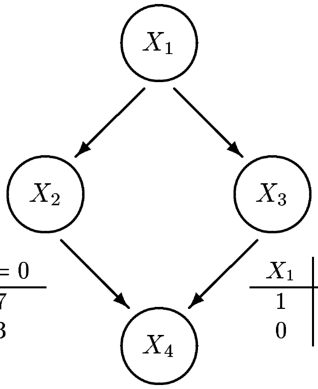
Notation

The notation used to express statements about probability distributions and random variables is frequently a source of confusion. Here we will introduce a notation that deviates slightly from existing conventions for reasons of clarity.

We need to represent random variables, probability distributions over random variables and the parameters of those distributions. Our formulae may be complicated by the fact that variables are often vector valued (discrete or continuous), and that a parameter may in certain cases be a random variable as well. Therefore, we will drop the distinction between parameters, variables, and random variables, and simply denote all variables with names that begin with upper case Roman or Greek letters. The names of symbolic constants will be in lower case Roman letters or Greek letters. A variable name in bold font will denote the set of values it may take on. So X is a random variable that takes a value $x \in \mathbf{X}$.

We will denote functions as their "signatures" — by an application to variables. Where variables are replaced by constants, this denotes a new function of lesser arity; obviously a function of zero arity is a constant. Function parameters may be divided into two groups by the "conditioning bar", which means that if all the parameters to the right of the bar are fixed while those to the right are left to vary, then the function is a distribution. Obviously, where variables are discrete, distributions are probability mass functions, otherwise they are probability density functions. To illustrate this with an example, suppose that we have a function $P(X, Y \mid A, B)$. $P(X, Y \mid a, b)$ denotes a distribution, and $P(x, y \mid a, b)$ denotes a probability (or probability density).

$X_1 = 1$	$X_1 = 0$
0.43	0.57



X_1	$X_2 = 1$	$X_2 = 0$
1	0.3	0.7
0	0.7	0.3

X_1	$X_3 = 1$	$X_3 = 0$
1	0.12	0.88
0	0.91	0.09

X_2	X_3	$X_4 = 1$	$X_4 = 0$
1	1	1.00	0.00
1	0	0.81	0.19
0	1	0.50	0.50
0	0	0.22	0.78

Figure 1: An example of a simple belief network.

For a discrete random variable X, we will use the notation $N(X)$ to denote the number of values X may take on. $\Pr[\, g(X) = x;\ P(X)\,]$ denotes the probability that $g(X) = x$ under the distribution $P(X)$; in general, any statement is allowed on the left. We will denote the expectation and variance of $g(X)$ with respect to a distribution $P(X)$ as $\mathrm{E}[\, g(X);\ P(X)\,]$ and $\mathrm{Var}[\, g(X);\ P(X)\,]$ respectively, with $\mathrm{SD}[g(X);\ P(X)]$ being the standard deviation. The frequency (number of occurrences) that X is assigned x in a random sample d will be denoted $\mathrm{Fr}[\, X = x;\ d\,]$. The conditioning bar will be used to denote conditional sentences appearing on the left in any of these expressions.

3 Belief Nets

Representation

A belief network is a representation for a joint distribution over a set of random variables $\{X_1 \ldots X_n\}$. It consists of a dependency graph and a set of conditional probability functions. The dependency graph is a directed acyclic graph, whose vertices are the random variables. This *network structure* represents the assumption that the joint distribution can be written as:

$$P(X_1 \ldots X_n) = \prod_{i=1}^{n} P_i(X_i \mid W_i)$$

where W_i denotes the parents of X_i in the graph. Formally, the network structure implies that X_i is independent of all non-descendents, given its parents. The resulting set of conditional independencies allows for a compact representation when the network is sparse.

In the most general case, the network variables may be continuous or discrete, and the conditional probability functions may be represented in a variety of ways. In this paper we are restricting our attention to discrete, finite-valued variables, and extensionally represented conditional probability functions, whose parameters are an exhaustive set of conditional probabilities that may be directly estimated from sample frequencies. In addition, we will assume that our sample data is composed of complete tuples (all network variables assigned a value).[1]

The simplest representation for the conditional probability functions is as *CP-tables* (see Figure 1). Here each $P_i(X_i \mid W_i)$ is represented by a number of *rows*, $P_i(X_i \mid w)$, one for each possible assignment (w) to W_i. In turn, each row is composed of a number of entries $P_i(x \mid w)$, one for each assignment (x) to X_i. Each row thus represents a multi-variate Bernoulli distribution. In some cases, greater economy of representation can be acheived by marginalizing over rows of a CP-table (Friedman & Goldszmidt, 1996), representing the conditional distribution as a "decision" tree (actually, a *distribution tree*, as the leaves are not decision nodes, but distributions). All of our results hold for any parameterization where each row $P_i(X_i \mid w)$ corresponds to a complete *or partial* assignment (w) to W_i. We then require that conditioning events for all rows be exhaustive and mutually exclusive. Note that one parameter per row is a *pseudo-parameter*, as it is a function of the other parameters in that row. However, it simplifies the presentation if we consider them all to be parameters of the distribution, and treat the dependency issue seperately.

Parameter Estimation

Learning a belief network from sample data may involve learning both the dependency graph and the conditional probability functions. There may be additional structure within the conditional probability functions, as well as numeric parameters. The learning problem may therefore be quite complex. In this paper we are concerned only with the uncertainty caused by parameter estimation when the network structure is fixed.

Suppose we are given a random sample d of m i.i.d. instances $\langle d_1 \ldots d_m \rangle$ from the "true" distribution $P(X_1 \ldots X_n \mid \theta)$, where $\theta = \theta_1 \ldots \theta_k$ are the true parameters. The likelihood function is $L(d \mid m, \Theta) = \prod_{i=1}^{m} P(d_i \mid \Theta)$. Based on our assumptions about the parameterization, we can write this as: $\prod_{i=1}^{k} {\Theta_i}^{a_i}$, where each Θ_i corresponds to some $P_j(x \mid w)$, and $a_i = \mathrm{Fr}[\, X_j = x, W_j = w;\ d\,]$. Thus, $a = \langle a_1 \ldots a_k \rangle$ is a sufficient statistic for Θ.

An obvious estimate for θ is the *maximum likelihood estimate* $\hat{\theta} = \langle \hat{\theta}_1 \ldots \hat{\theta}_k \rangle$, where $\hat{\theta}_i = P_j(x \mid w) = a_i/(a_i + b_i)$ for $b_i = \mathrm{Fr}[\, X_j \neq x, W_j = w;\ d\,]$. However, this is undefined when $(a_i + b_i) = 0$; this corresponds to the case where a particular parental assignment is never observed in the evidence, and thus the parameters corresponding to the distribution conditional on that assignment must be assigned values in the absence of evidence (and consequently, without affecting the likelihood function). A natural choice is the maximum entropy distribution: $1/r$ for each of the r parameters of the distribution.

Maximum likelihood estimation is the classic Frequentist approach to estimation. The corresponding Bayesian approach is to use the *maximum a posteriori* (MAP) estimate. This is the value for Θ that maximizes $L(d \mid m, \Theta) f(\Theta \mid \alpha)$, where $f(\Theta \mid \alpha)$ is a prior distribution over $\boldsymbol{\Theta}$. For a uniform prior the MAP approach is equivalent to maximum likelihood estimation.

From a Bayesian perspective, however, the MAP model is not necessarily the ideal choice. An alternative is to use a weighted average over all models, called the *predictive distribution.* For a fixed network structure this corresponds to an average over all possible parameter values for that structure. In certain cases, a model can be constructed such that inferences based on that model are equivalent to inferences based on the predictive distribution. In fact, for a certain class of priors, belief nets have the nice property that such a model can easily be constructed. When the prior distribution is a product of Dirichlet distributions (see Wilks 1962), one for each CP-table row (recall that each row is a multivariate Bernoulli distribution), then the posterior distribution also has this form, and is given by the simple equation:

$$f(\Theta \mid a + \alpha) = \frac{L(d \mid m, \Theta) f(\Theta \mid \alpha)}{\int L(d \mid m, \Theta) f(\Theta \mid \alpha)\, \mathrm{d}\Theta}$$

Here $\alpha = \langle \alpha_1 \ldots \alpha_k \rangle$ is the parameter of the prior distribution — the concatenation of the individual Dirichlet parameters for each row. Simple updating with the statistic a ($a + \alpha$ signifies vector addition) is sufficient

[1] Obviously, these are very restrictive assumptions; however, they simplify the analysis and its presentation. Most of our results can be extended to more general domains.

to specify the posterior distribution, which, furthermore, differs from the prior only in this value, not in the form of f. This family of distributions is thus *conjugate* for belief nets, and for some applications we may prefer to view a belief net not as a model of a data-generating process, but as a representation of the posterior distribution (by storing Dirichlet distributions in the CP-tables, instead of probability vectors). Furthermore, if we use the parameter estimates given by:

$$\hat{\theta}_i = (a_i + \alpha_i)/(a_i + b_i + \alpha_i + \beta_i)$$

where β_i has the analogous interpretation to b_i, then inferences based on this model are the same as those based on the predictive distribution. Consequently, this is the natural estimate from a Bayesian perspective. Proofs of the above claims, as well as references, are given in Appendix A.

The Posterior Distribution of Θ

The posterior distribution of Θ is given by a product of Dirichlets, one for each row of each CP-table:

$$\begin{aligned} f(\Theta \mid a+\alpha) = \quad & \mathrm{Dir}(\Theta_1, \ldots \Theta_{r-1} \mid a_1 + \alpha_1, \ldots, a_r + \alpha_r) \\ & \mathrm{Dir}(\Theta_{r+1}, \ldots \Theta_{s-1} \mid a_{r+1} + \alpha_{r+1}, \ldots, a_s + \alpha_s) \\ & \vdots \\ & \mathrm{Dir}(\Theta_{t+1}, \ldots, \Theta_{k-1} \mid a_{t+1} + \alpha_{t+1}, \ldots, a_k + \alpha_k) \end{aligned}$$

where the first row of the first CP-table represents an $(r-1)$-variate Bernoulli distribution, the second row represents an $(s-r-1)$-variate distribution, and so on. Notice that the last "parameter" of each row has been excluded, as it is a function of the others. The posterior distribution contains all the information in the priors and the sample except the order of the data (which is non-informative, given the i.i.d. assumption).

4 Model Selection Criteria

Most learning problems can be cast in the following form. First, a *type* of model (such as Markov model, belief network, neural network, decision tree, etc.) is chosen, based on (often tacit) domain knowledge. Next, within this model type a *structure* is chosen. This structure defines a parametric class of models. Last, the parameters are estimated, or tuned/fit, based on a sample. The problem of choosing the structure, or parametric class of models, is called the *model selection* problem. Model selection can be viewed as an optimization problem where there are two separate issues: (1) how to search the space of structures, and (2) what function (model selection criterion) to optimize for. Here we explore the second issue: the choice of a model selection criterion.

A standard approach to parameter estimation is to maximize the likelihood of the data. Applying this principle to model selection, however, tends to result in the phenomenon of *overfitting.* A more complex model allows for a tighter fit to the data, thus increasing (up to some absolute maximum that depends on the data) the mode of the likelihood function over the parameter space, but also increasing the variance in the parameter estimates. There is a direct trade-off between representational power and the power to generalize to unseen data. This is called the *bias-variance trade-off.* More complex models can represent a wider range of data-generating processes, but require more data to learn. Thus, given a limited sample it is imperative to choose the appropriate model structure, to minimize the error of the model.[2]

Error Functions

A standard measure of training error is the negative log-likelihood:

$$\mathrm{DL}(d, \hat{\theta}, h) = -\log L(d \mid m, \hat{\theta}, h)$$

[2] Overfitting avoidance is generally not justifiable unless one has a prior expectation of simplicity. For a large class of problems there are no distinctions between algorithms that avoid overfitting and those that don't, independent of prior beliefs. In practice, however, avoiding overfitting is usually desirable, perhaps because the choice of model type reflects a prior belief that a parsimonious representation exists for this type. See Wolpert (1996a, 1996b) for more on this topic.

where $d = \langle d_1 \ldots d_m \rangle$ is a sample of size m, h is a model structure, $\hat{\theta}$ is the parameter vector for that model, and $L(d \mid m, \hat{\theta}, h)$ is the likelihood function.[3] We denote it as DL because it is the description length of the data given an optimal code based on $P(X \mid \hat{\theta}, h)$. When d is an i.i.d. sequence of values $d_1 \ldots d_m$ then:

$$\mathrm{DL}(d, \hat{\theta}, h) = -\sum_{i=1}^{m} \log P(d_i \mid \hat{\theta}, h)$$

KL-divergence (due to Kullback & Leibler, 1951; see also Cover & Thomas, 1991) is a standard measure of true error for distribution learning. If $P(X \mid \theta, g)$ is the "true" model, and $P(X \mid \hat{\theta}, h)$ is a hypothesized model, then the KL-divergence of $P(X \mid \hat{\theta}, h)$ from $P(X \mid \theta, g)$ is given by:

$$\mathrm{Err}[P(X \mid \theta, g);\ P(X \mid \hat{\theta}, h)] = \sum_{X} P(X \mid \theta, g) \log \frac{P(X \mid \theta, g)}{P(X \mid \hat{\theta}, h)}$$

This is the expected cost of encoding instances from $P(X \mid \theta, g)$ using a code based on $P(X \mid \hat{\theta}, h)$. Note that it can also be written as:

$$\mathrm{Err}[P(X \mid \theta, g);\ P(X \mid \hat{\theta}, h)] = \mathrm{E}[\mathrm{DL}(X, \hat{\theta}, h)] - \mathrm{E}[\mathrm{DL}(X, \theta, g)]$$

where the expectations are taken under $P(X \mid \theta, g)$. The second term is the *entropy* of $P(X \mid \theta, g)$. As it does not depend on $P(X \mid \hat{\theta}, h)$, the first term alone is sufficient to compare models. We can form an estimate of this first term:

$$\mathrm{Fit}(d, \hat{\theta}, h) = \frac{1}{m}\mathrm{DL}(d, \hat{\theta}, h)$$

The problem, however, is that if we use d to estimate $\hat{\theta}$ then $\hat{\theta}$ depends on d and so $\mathrm{Fit}(d, \hat{\theta}, h)$ is a biased estimator, underestimating $\mathrm{E}[\mathrm{DL}(X, \hat{\theta}, h)]$. As h grows more complex, the dimensionality of $\hat{\theta}$ increases and the more it can be tuned to d, increasing the bias in $\mathrm{Fit}(d, \hat{\theta}, h)$. Note that there are two different kinds of bias involved here: as models become more complex, they become less representationally biased (they can represent a larger class of distributions), but their parameters have higher variance under sampling, and so $\mathrm{Fit}(d, \hat{\theta}, h)$ becomes *more* biased as an estimator of $\mathrm{E}[\mathrm{DL}(X, \hat{\theta}, h)]$. Model selection criteria attempt to correct for this bias or avoid it altogether.

Complexity Penalty Criteria

One way to (attempt to) correct for the bias in $\mathrm{Fit}(d, \hat{\theta}, h)$ is to add a complexity penalty to the function. The major difficulty here is determining what an appropriate penalty is. One cannot determine a priori how much bias there is in $\mathrm{Fit}(d, \hat{\theta}, h)$, as this depends on $P(X \mid \theta, g)$ as well as $P(X \mid \hat{\theta}, h)$. For each structure, there exists a continuum of possible biases in $\mathrm{Fit}(d, \hat{\theta}, h)$, from zero, when the true model has zero entropy and zero variance, on up to some maximum value, when the true model has maximum entropy. Therefore, one cannot justify a complexity penalty simply on the grounds that it is a bias correction; one must appeal to other considerations.

There are several well-known penalty functions, each motivated by different theoretical considerations. The Minimum Description Length (MDL) criterion is based on an information-theoretic view of induction as data compression; see Rissanen (1989) for a detailed development. It is equivalent to the *Bayesian Information Criterion* (BIC), which was introduced originally by Schwarz (1978) and given a Bayesian interpretation. The information-theoretic interpretation of the MDL criterion is as the length of an encoding of the sample as a two part code. The model defines a code for the sample, and one encodes the sample by first encoding the model, and then encoding the data using the optimal code given by the model.[4] If the model captures significant features of the data, this encoding will be considerably smaller than the original encoding of the sample. On the other hand, if the model represents too much about the sample, the encoding size will increase (Linhart & Zucchini, 1986). This trade-off is similar to the bias-variance trade-off.

[3] We leave the estimation of $\hat{\theta}$ as a black box for the time being — we return to this at the end of the section.

[4] Every probability distribution has an associated optimal code (Cover & Thomas 1991).

The MDL criterion we will use is given by:

$$\mathrm{MDL}(d,h) = \mathrm{Fit}(d,\hat{\theta},h) + \frac{k \log m}{2m}$$

where k is the dimension of $\hat{\theta}$: the number of *free* parameters of h. Recall that m is the sample size. This version differs from the standard form in that we have normalized everything by $1/m$ so we can compare it across sample sizes and with other criteria. Some low order terms (all positive) have been dropped as well (from the MDL criterion given by Rissanen; the BIC is exactly what we have above). It will be seen that this omission has no negative impact on the criterion.

Akaike's Information Criterion (AIC) comes from a different theoretical perspective. It is an explicit attempt to correct the overfitting bias. See Bozdogan (1987) for the derivation of the criterion. The complexity penalty is considerably smaller than MDL's. Our version of AIC is given by:

$$\mathrm{AIC}(d,h) = \mathrm{Fit}(d,\hat{\theta},h) + \frac{k \log e}{m}$$

where e is the base of the natural logarithm, and $\log e$ converts from nats to bits.

Validation Criteria

An alternative approach is to use part of the data to estimate the parameters and the rest of the data to estimate the error of the resulting model. There are drawbacks to this approach, however. There is a bias toward simplicity in the model, because the criterion estimates the error for a model with parameters based on a smaller sample than will ultimately be used.[5] Also, there will be increased variance in the criterion, as opposed to the naive $\mathrm{Fit}(d,\hat{\theta},h)$, because of the smaller sample size used to estimate the error. So this *train-test* validation trades the bias of the naive criterion for another bias and increased variance.

Cross-validation (Stone 1974) is a more sophisticated approach which seeks to alleviate these concerns. In cross-validation every datum in the sample is used both for training and testing. The sample is partitioned into r subsamples, each of which is used once for testing and $(r-1)$ times for training. Each subsample is tested on the model that results from training on the remaining $(r-1)$ subsamples. This is called *r-fold* cross validation. The most extreme case is m-fold (*leave-one-out*) cross-validation. We will use the simplest version, 2-fold cross-validation, which is given by:

$$\mathrm{XV}(d,h) = \frac{1}{m}\left[\mathrm{DL}(d',\hat{\theta}(d''),h) + \mathrm{DL}(d'',\hat{\theta}(d'),h)\right]$$

where d' and d'' partition d into two equal-size subsamples, and $\hat{\theta}(\cdot)$ denotes the parameter estimated from the data given in the parentheses.

Cross-validation has its problems as well. First of all, the meaning of the criterion is not explicit — it is an average of many individual estimates which are not independent if $r > 2$. Leave-one-out cross-validation is not *asymptotically minimal*: in the limit of sample size it may not prefer the simplest unbiased hypothesis (Stone 1977). Worst of all, it may be very expensive to compute, and time is at a premium when searching a large space of structures.

Bootstrap Criteria

Bootstrap methods are similar to validation methods, but instead of withholding data from the original sample, a training or test sample is generated by *resampling* (Efron 1982). The idea is that, by adding noise through resampling, the variance of the original sampling process is simulated. One problem with the non-parametric bootstrap is that the empirical variance of the bootstrapped sample tends to be lower than that of the original sample — if the bootstrap is recursively applied over and over, with asymptotic probability 1 it will result in a homogenous sample. Another problem is that the empirical distribution in the sample

[5] When a structure has been chosen, all available data is used to estimate the parameters of the final model.

will differ from the true distribution, and thus will have a different variance under sampling.[6] The formula below shows the bootstrap criterion we will be using, where the training sample has been bootstrapped and the original sample is used for testing.

$$\mathrm{Boot}(d,h) = \mathrm{Fit}(d, \hat{\theta}(d'), h)$$

Here d' is a sample of size m, generated from d by resampling with replacement.

A Bayesian Approach?

Overfitting results from the naive maximum-likelihood approach because the likelihood of the data is based on a single (maximum-likelihood) model in the model space for each structure. As this space increases, the mode of the likelihood function cannot decrease, and tends to increase. The Bayesian approach to prediction, by contrast, involves averaging over all models according to the posterior distribution over models. For a given structure, one can compute the marginal likelihood of the data by integrating over the parameter space. The integral under the likelihood function (coupled to the prior distribution over models), rather than its mode, may thus be used as a way to evaluate structures.[7] This mitigates the effects of increasing complexity: as complexity increases the likelihood function becomes more peaked, increasing its mode but potentially decreasing the integral underneath after a point. The bias-variance trade-off is handled implicitly — it is reflected in the shape of the likelihood function.

This, however, is a Frequentist justification for a putatively Bayesian procedure. The notion of overfitting is not really defined in the Bayesian paradigm: upon seeing the data, you simply compute the posterior distribution over models and use this to carry out inference. A possible Bayesian approach is to parameterize the model space by first defining a hyper-parameter that ranges over structures; then the marginalized likelihood is the data-dependent part of the posterior distribution over structures. Ideally, prediction would involve marginalizing over all structures according to this posterior, but for pragmatic reasons one could approximate this by choosing the structure with the maximum posterior probability, given the data. (See Heckerman 1996.)

The use of a hyper-parameter ranging over structures, however, must be viewed simply as a device to define the prior over models. The model class for the most complex structure contains every possible model, and so "structure" is a superfluous syntactic notion. A uniform prior over structures, coupled with a uniform prior over the parameters for each structure, results in a very *non-uniform* prior over the space of models, favouring those models which appear in more structures (exactly those representable in simpler structures). Perhaps this is desirable, being a reflection of the intuition behind choosing a particular model type and representational framework. But it is open to the criticism of being ad hoc, as well as the classic objection to Bayesian methods as being subjective — because many incompatible uniform priors may be proposed as representing identical states of ignorance (see Howson 1997).

One could carry out a Bayesian analysis of model selection as a decision problem, where the objective is to minimize expected loss. Where KL-divergence is the loss function, and $f(\Theta, G \mid d)$ is the posterior over models, the problem reduces to computing:

$$\begin{aligned}
&\mathrm{argmin}_{\hat{\theta},h} \int_{\Theta,G} \left[\sum_X P(X \mid \Theta,G) \log P(X \mid \Theta,G) - P(X \mid \Theta,G) \log P(X \mid \hat{\theta},h)\right] \cdot f(\Theta,G \mid d) \\
= \quad &\mathrm{argmin}_{\hat{\theta},h} \int_{\Theta,G} \left[\sum_X P(X \mid \Theta,G) \log P(X \mid \Theta,G) - \sum_X P(X \mid \Theta,G) \log P(X \mid \hat{\theta},h)\right] \cdot f(\Theta,G \mid d) \\
= \quad &\mathrm{argmin}_{\hat{\theta},h} \int_{\Theta,G} \left[-\sum_X P(X \mid \Theta,G) \log P(X \mid \hat{\theta},h)\right] \cdot f(\Theta,G \mid d) \\
= \quad &\mathrm{argmin}_{\hat{\theta},h} \; -\sum_X \left[\int_{\Theta,G} P(X \mid \Theta,G) f(\Theta,G \mid d)\right] \cdot \log P(X \mid \hat{\theta},h)
\end{aligned}$$

Based on a fundamental result of information theory due to Shannon (see Cover & Thomas 1991), a minimum is attained at the entropy of the predictive distribution when:

$$P(X \mid \hat{\theta}, h) = \int_{\Theta,G} P(X \mid \Theta, G) f(\Theta, G \mid d)$$

[6] It might be possible to correct for some of these problems by measuring and correcting for the difference in variance between the two samples.

[7] Prior distributions are implicit where not mentioned.

Thus, choosing the predictive distribution is the Bayes-optimal strategy for minimizing KL-divergence.

This is a sort of "no free lunch" theorem (Wolpert, 1996a) for model selection, as it follows that any attempt to avoid overfitting reflects a prior bias toward simplicity. For example, if one's prior belief is reflected by a uniform prior over all possible models, one ought to use the *most complex* structure, with a uniform prior over its parameter space, and then take the average model from the posterior that results from Bayesian updating, which is the predictive distribution for this prior. Of course, one might be interested in other properties of a learning algorithm besides expected loss (like minimax properties, for example).

It is interesting to note that the marginalized likelihood has an interpretation as a validation criterion. For belief networks the marginalized likelihood is given by (see Appendix A):

$$\mathrm{ML}(d \mid m, h) = \prod_{i=1}^{m} \int_{\Theta} P(d_i \mid \Theta, h) f(\Theta \mid h, d_1 \ldots d_{i-1})$$

This *prequential* criterion (Dawid & Vovk, 1999) involves iteratively computing the probability of each datum given those already seen by marginalizing over the parameters using the current distribution over parameter space, then updating that distribution with the current datum. Each datum is used once for testing and once for training, but the testing precedes the training, so each datum's error score is based on parameters it has not been used to estimate, as in validation criteria. For belief networks, because using the average model under the posterior distribution is equivalent to integration over the parameter space (again, see Appendix A for details), it may be efficiently computed. We use the following form as a model selection criterion:

$$\mathrm{Preq}(d, h) = \frac{1}{m} \sum_{i=1}^{m} -\log P(d_i \mid \hat{\theta}(d_1 \ldots d_{i-1}), h)$$

where $\hat{\theta}(d_1 \ldots d_{i-1})$ is the mean parameter vector after updating with $d_1 \ldots d_{i-1}$.

KL-Divergence and $\hat{\theta}$

KL-divergence does not lend itself to maximum-likelihood methods. As we have seen, it is minimized by the predictive distribution, not the MAP hypothesis. More importantly, using maximum-likelihood parameters virtually guarantees infinite KL-divergence for all but the simplest models. It is therefore unrealistic to use maximum-likelihood estimates when comparing criteria under the KL-divergence loss function. Therefore, in our experiments we will use the the average model (predictive distribution) for each structure, under uniform priors. In other words: $\hat{\theta}_i = (a_i + 1)/(b_i + r)$ for all i, when $\hat{\theta}_i = \Pr[X_j = x \mid W_j = w]$ and X_j ranges over r values. One consequence of using this parameter estimation method is that the asymptotic arguments used to justify the AIC and MDL criteria no longer strictly hold; the AIC is designed to correct the bias in maximum-likelihood estimates, while the MDL criterion is based on optimizing the precision in the parameters, assuming maximum-likelihood estimation. Another consequence is that $\mathrm{Fit}(d, \hat{\theta}, h)$ is not necessarily monotonically decreasing as model complexity increases — using the average model instead of the ML-model has a smoothing effect which avoids some overfitting automatically. In certain cases, as we will see, this is sufficient to remove the bias in $\mathrm{Fit}(d, \hat{\theta}, h)$.

5 Empirical Study

We carried out a series of experiments where we chose a true model, drew a sample from it, and evaluated the criteria across a set of hypothesis network structures. We also computed the true error of each structure with the parameters estimated from the sample. This allowed us to compare the behaviour of the criteria across a range of situations, by comparing their evaluations to the true error function (in particular, comparing the minima).

Criteria

We compared the following criteria described in the previous section:

- The "Bayesian" prequential criterion (Preq).
- 2-fold cross-validation (XV).
- A bootstrap criterion (Boot).
- Akaike's information criterion (AIC).
- The Minimum Description Length criterion (MDL).

Each criterion takes a hypothesis structure and a sample of m data as input, and returns a real number being the estimated error per datum. Each uses m training iterations and m testing iterations, where a training iteration is Bayesian updating with a single datum, and a testing iteration is computing the negative log-likelihood of a single datum. Thus, all required the same amount of computational resources, and each assumed that the average model (predictive distribution) would be used for each structure.

True Models

We considered the following domains:

- Random networks: We generated networks on 10 binary variables with 10, 20, and 30 links. In each case, we generated the parameters from a uniform distribution over the parameter space.
- The Alarm network (Bienlich et al 1989): This is a benchmark network commonly used in empirical studies on belief networks, constructed by medical experts for monitoring patients in intensive care units. It has 37 variables each with 2-4 possible values, 46 links (very sparse), and 509 parameters.
- The Insurance network (Binder et al 1997): This is another widely used benchmark network. It represents car insurance risks. It has 27 variables, 52 links and over 1400 parameters.[8]

Hypotheses

For each true model, we generated sets of hypotheses as follows. Each set of hypotheses was a sequence of network structures that included the true structure. Starting from the simplest network, with no links, each hypothesis in the sequence was constructed from the previous by adding a link. Eventually the true structure appeared, followed by progressively more complex networks with "redundant" links. (We constructed these sequences from the true structure by randomly deleting and adding links.) By construction, the model class defined by each hypothesis properly contained the model classes of all preceding hypotheses. Thus, the bias with respect to the true model was decreasing up to the point where the true structure appeared, and after that all error could be attributed to variance. We did this so that we could observe the behaviour of the criteria as bias decreased and variance increased.

Experimental Design

Each experiment consisted of the following steps:

1. Fix a true model.
2. Generate a sequence of candidate network structures.
3. Generate a sample from the true distribution.
4. Evaluate the criteria on that sample, across all structures.
5. Compute the sample error, $\mathrm{Fit}(d, \hat{\theta}, h)$, and the true error $\mathrm{E}[\mathrm{DL}(X, h, \hat{\theta})]$ for each structure.

[8] The Alarm and Insurance networks are available at the UCI machine learning repository.

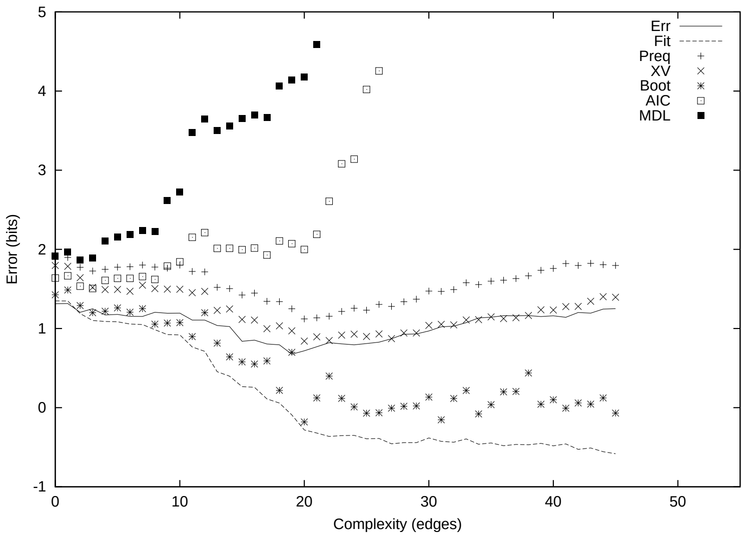


Figure 2: Random Network Case Study: $t = 20$, $m = 50$

We looked at individual experiments as well as gathering summary statistics from large numbers of experiments — the case studies allowed us to identify patterns of behaviour; the comprehensive studies allowed us to quantify some aspects of those patterns. Since we were only interested in the issue of the criteria used for model selection, not the search strategy, we based our comprehensive results on selecting the structure that minimized each criterion, then comparing the true error of those preferred structures.

6 Results

Figure 2 shows an experiment on a randomly generated true model, for a sample of size 50. Here the true distribution was represented by a 20-edge dependency graph on 10 binary variables, where each conditional distribution was randomly generated from a $\text{Beta}(1, 1)$ (uniform Beta) distribution. The x-axis shows the complexity of the hypothesis networks in terms of number of links, while the y-axis shows the error in terms of bits. The entropy of the true distribution has been subtracted from all y-values to scale them down and to allow for comparisons between different true distributions.[9] In each graph all criteria are plotted, as are the true error (Err) and the sample error (Fit).

From left to right across the graph, the hypothesis complexity is increasing, and the true error decreases until it reaches a minimum near the true structure ($t = 20$), where it begins to climb back up. We can observe the bias-variance trade-off in action: bias decreases until the 20-link network, while variance is increasing continuously. The sample error (Fit), here shown as the dashed line, continues to fall past the true structure, showing that a naive maximum-likelihood approach to choosing a structure would lead to overfitting here. By observing the criteria across the range of hypotheses, for different sample sizes (m = 50, 100, 150), we can get some idea of how they are handling the bias-variance trade-off. Figures 3 and 4 show the same experiment when the sample size is increased to 100 and 150 data, respectively.

Ideally, we would like a criterion to have the same shape, or at least the same minimum, as the true error function. By observing how the criteria differ from this unrealizable ideal, in different learning contexts, we get an idea of how they handle the bias-variance trade-off.

Both the MDL and AIC criteria tend to underfit at first, and then converge on the true network as more data are provided. MDL, in particular, has a strong bias for simplicity that requires a lot of data to counteract. In addition, a small number of additional data can radically change the evaluation of a structure

[9] Note that this makes some of the scores negative, as the empirical entropy tends to underestimate true entropy, but this does not affect our comparisons.

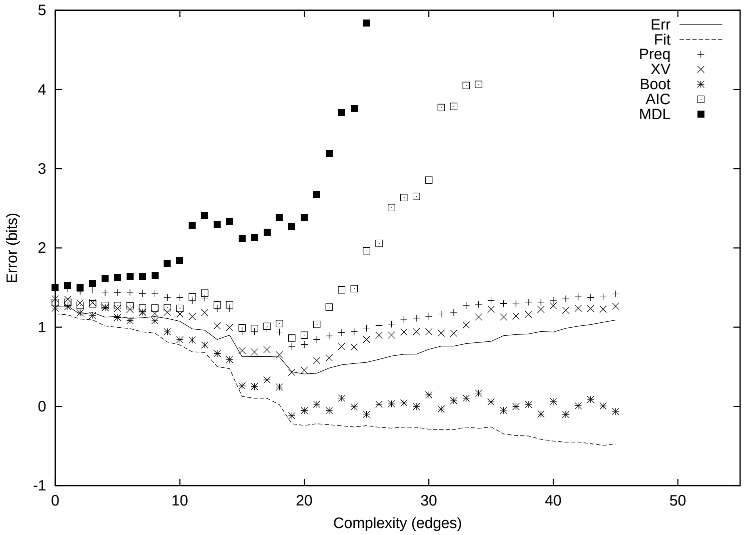


Figure 3: Random Network Case Study: $t = 20$, $m = 100$

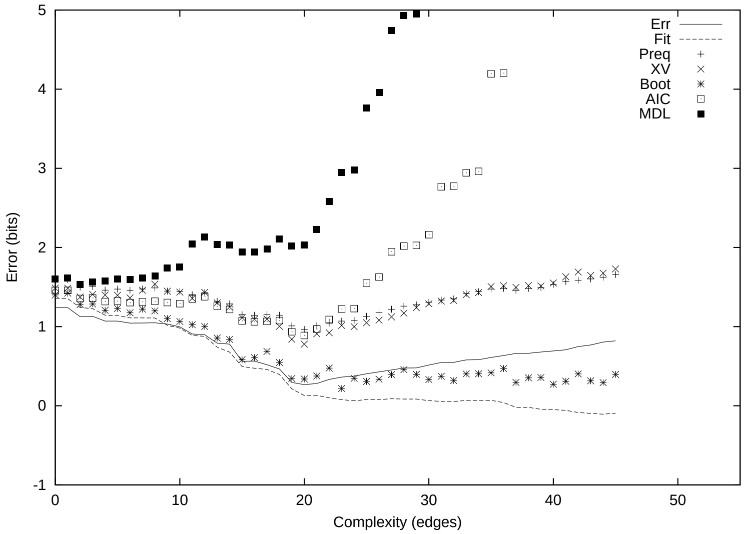


Figure 4: Random Network Case Study: $t = 20$, $m = 150$

— notice the difference in the AIC scores from Figure 2 to Figure 3.

The bootstrap criterion, by contrast, tends to overfit, a tendency which does not diminish as sample size increases. This is because it appears to be determined more by the sample error than by the true error. Here we should mention that the bootstrap criterion plotted here is less "smooth" than the other data-dependent criteria; this is due to the fact that we generated a new bootstrap sample each time we evaluated a hypothesis, and thus the difference in the score between two points depends not only on the structure, but also on the training sample generated. The general pattern of behaviour is clear, however: the bootstrap criterion is more reflective of the sample error than the true error.

The prequential criterion and 2-fold cross-validation give very close evaluations, across the range of hypotheses and sample sizes. They also match best with the shape of the true error function, having minima and maxima at identical or nearby structures. The prequential criterion appears to be slightly less sensitive to underlying fluctuations in the true error.

In addition to randomly generated distributions, we considered "natural" distributions defined by benchmark networks. Here we give results on two widely used benchmarks: the Alarm and Insurance networks. The former has 37 variables and 46 links; the latter 27 variables and 52 links. Figures 5 and 6 show the pattern observed on the Alarm network, while Figures 7 and 8 show the same results for the Insurance network. First of all, note that the sample error (Fit) is not monotonically decreasing as hypothesis complexity increases. If $\hat{\theta}$ was a maximum-likelihood estimate, then Fit would necessarily be monotonic, because by construction each successive hypothesis structure defines a class of models which contains the model class of all predecessors. The effect observed here is due to the smoothing effect of choosing the *average* a posteriori model within each class, instead of the *maximum* a posteriori (or maximum-likelihood) model. This is interesting because many papers have been published describing results of structure-learning experiments on the Alarm and Insurance networks, where various methods were used to avoid overfitting, (see Liu et al, 1998, for example) when in fact it is virtually impossible to overfit on these networks.[10] We believe that the difference observed here between the random networks and these benchmark networks is due to the fact that the benchmark networks are locally low entropy — the conditional probability distributions for each node tend to be highly skewed. On these networks all the data-dependent criteria worked fairly well, including the bootstrap, because its determination by the sample error was not a deficit. The complexity-penalty criteria underfit; again, the MDL criterion was much slower to converge than AIC.

In our comprehensive studies we attempted to measure the difference in error that would result from using each criterion to pick a structure. We carried out experiments of the form described above, but instead of plotting the criteria across all hypotheses, we used each criterion to select a hypothesis by taking its argmin across the hypotheses under consideration. We then compared the true error of this preferred hypothesis (for each criterion) with the minimum true error obtained by any hypothesis (the ideal criterion, of course, being the true error). The additional true error of the preferred hypothesis was our dependent variable. On random networks we carried out experiments for several combinations of truth complexity (t) and sample size (m). For each assignment to t and m we carried out 30 experiments, generating a new true model and hypothesis sequence each time. We summarize these experiments by giving the average additional true error for each criterion. Table 1 displays our results for random networks; Tables 2 and 3 present our results on the Alarm and Insurance networks; where we carried out identical experiments to those described on random networks, except of course that the true model was fixed. Again, each cell represents 30 experiments.

7 Discussion

Model selection involves dealing with fundamental aspects of inductive inference. The no free lunch theorems show that there is no justification for avoiding overfitting that is independent of prior beliefs and/or pragmatic considerations. In other words, we avoid overfitting either because we have prior beliefs that the truth tends to be simple, or because we prefer simpler models for pragmatic reasons — they are easier to remember, explain and reason with. Ideally, we could pose model selection as a Bayesian decision problem — in the first case our bias for simplicity would be reflected in the prior; in the second case our bias for simplicity would be reflected in the loss function. However, it is often the case that we have no clear idea of how to

[10] This also demonstrates the perils of over-reliance on benchmark problems.

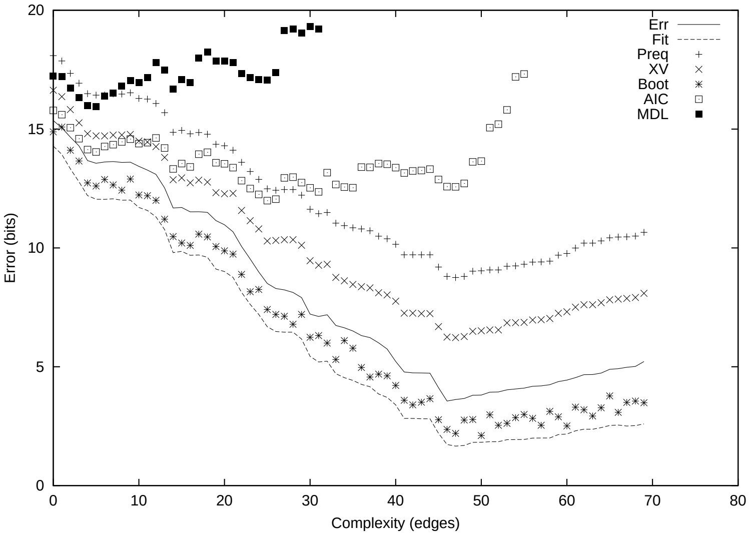


Figure 5: Alarm Network Case Study: $m = 50$

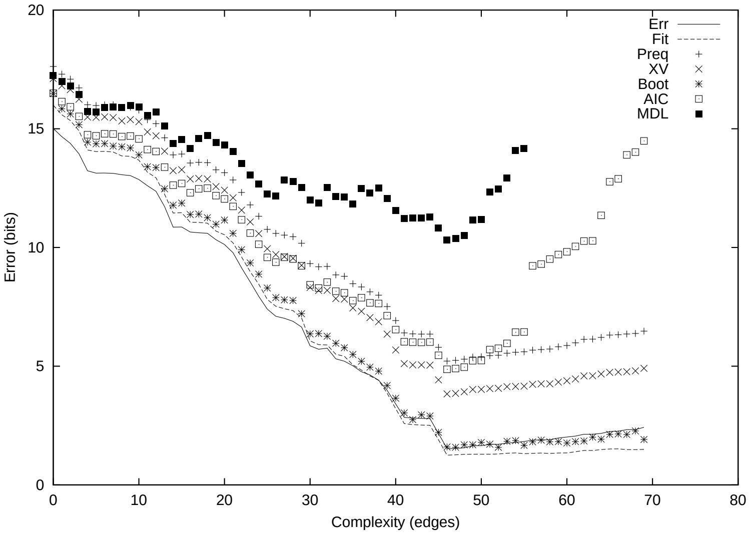


Figure 6: Alarm Network Case Study: $m = 150$

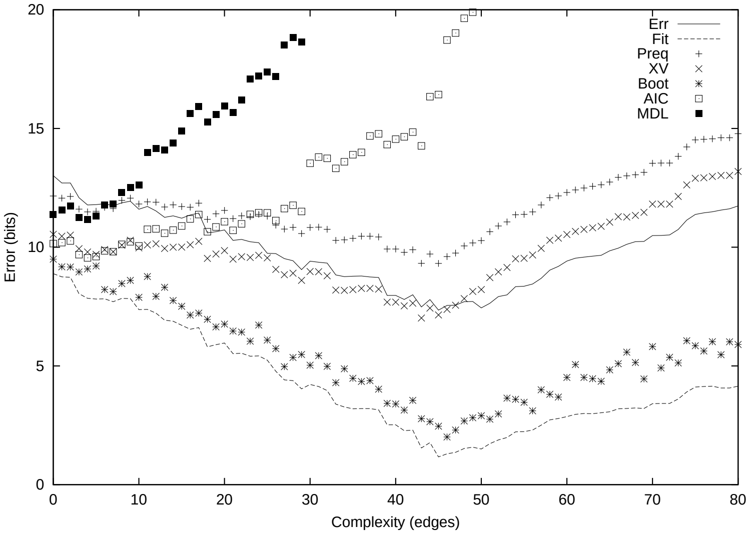


Figure 7: Insurance Network Case Study: $m = 50$

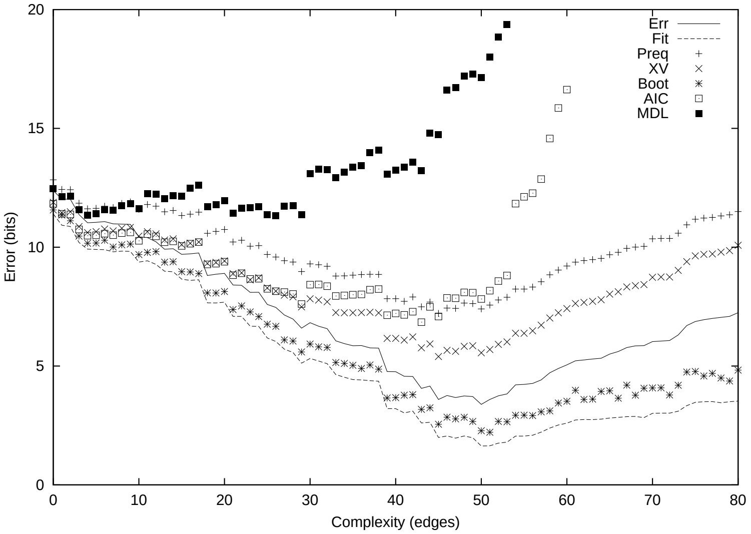


Figure 8: Insurance Network Case Study: $m = 150$

Table 1: Results for Random Networks

t	m	Fit	Preq	XV	Boot	AIC	MDL
10	50	0.673424	0.045774	0.085914	0.526975	0.087324	0.253013
10	100	0.547661	0.005703	0.025173	0.454687	0.017614	0.130488
10	150	0.487156	0.007373	0.016239	0.363291	0.008755	0.047132
10	200	0.453187	0.003962	0.007292	0.311022	0.003405	0.020010
20	50	0.524959	0.058641	0.104019	0.390290	0.479635	0.765795
20	100	0.451487	0.020641	0.037211	0.352972	0.157028	0.643342
20	150	0.432322	0.015944	0.020506	0.304437	0.093406	0.486259
20	200	0.370868	0.003800	0.017478	0.305357	0.061886	0.317950
30	50	0.255416	0.047017	0.101118	0.177463	0.765639	0.981904
30	100	0.259497	0.016463	0.021493	0.173163	0.598727	0.969809
30	150	0.244873	0.008740	0.027909	0.158132	0.455095	0.889597
30	200	0.223711	0.008547	0.016779	0.163472	0.357849	0.866408

Table 2: Results for Alarm Network

m	Fit	Preq	XV	Boot	AIC	MDL
50	0.002055	0.002055	0.002055	0.130730	3.386938	9.996832
100	0.007444	0.000000	0.000473	0.094789	0.000000	4.657260
150	0.008488	0.000608	0.002463	0.045147	0.000608	0.000000
200	0.003116	0.000000	0.000582	0.033719	0.000000	0.000000

Table 3: Results for Insurance Network

m	Fit	Preq	XV	Boot	AIC	MDL
50	0.190568	0.171794	0.126868	0.273680	4.692191	4.768936
100	0.010237	0.025124	0.024360	0.086121	1.243281	6.565789
150	0.000000	0.167360	0.129376	0.057607	0.669900	3.597372
200	0.010371	0.243592	0.147450	0.071585	0.512019	3.464004

represent our intuitions in the prior or loss function. For this reason, it is of interest to consider properties of various proposed model selection criteria, to see how well or how poorly they capture our intuitions.

From our observations, we conclude that the MDL and AIC criteria display a strong bias for simplicity that may require considerable data to outweigh. This is particularly true for MDL. The amount of data required is much more than the amount of data required by other criteria to get a good estimate of the true error of the structure by other criteria. Also, a small amount of additional data can radically affect the evaluation of a structure, and it is impossible to predict in advance how much data is required to evaluate a given structure (as this depends on the true distribution).

In defense of the MDL criterion, however, it was not derived specifically to avoid overfitting (although that is a benefit frequently claimed for it), but simply to implement the MDL principle, which takes an information-theoretic view of induction orthogonal to the Bayesian and Frequentist paradigms.[11] Furthermore, it was designed with asymptotic consistency and minimality in mind, not small sample behaviour. Nevertheless, one should be aware when applying the MDL principle in this context that one may end up seriously underfitting the data.

The prequential criterion and 2-fold cross-validation displayed similar behaviour to one another. Both converged quite rapidly in shape to the form of the true error function. This makes them good functions to optimize for if one's goal is minimizing expected error and one's prior expectations of simplicity are modest. The bootstrap criterion, by contrast, proved to be a better estimator of sample error than true error. Perhaps our formulation was naive and a better one could be obtained (for minimal additional computation?), however, we would not recommend the use of a bootstrap criterion based on these results.

Aside from the behaviour of particular criteria, we observed some interesting phenomena relating to the general problem of model selection in belief networks. Very little data seems to be required to get a good evaluation of a network, even if the true distribution is reasonably complex; for example, a sample size of less than 50 was sufficient to get a good evaluation of structures for the Alarm network, which has 37 variables and 46 dependencies between them. This is surprising, given that recent work in the area of learning belief network structure (see Liu et al, 1998, for example) uses as many as 10,000 data to learn this network. As search spaces for model selection are very large, one could probably do much better by evaluating more networks on fewer data. Also, we noted that the practice of taking the average model (which is Bayes-optimal within a fixed model class) handles overfitting to some extent. In fact, although the Alarm network has been used in several studies of structure-learning, using the sample error of the average model to choose a structure results in *no overfitting* for this learning problem. This is explained by the smoothing effect of the prior, which increases as the model complexity increases.

We cannot conclude from this research that any one criterion is better than another, unless we specify exactly what loss function and prior beliefs we have. We can suggest caution in applying complexity-penalty criteria such as MDL and AIC (particularly MDL), as they may lead to underfitting the data — they appear to be risky criteria. In some situations they are dominated by the complexity penalty, and thus almost independent of the data. This is probably not desirable in most learning contexts. The bootstrap criterion, on the other hand, seems to lead to overfitting, as it is governed more by the sample error than by the true error. The prequential criterion and 2-fold cross-validation seem to be relatively risk-free and assumption-light criteria suitable for a wide range of problems.

It might be possible to rehabilitate the complexity penalty criteria or the bootstrap criterion by a more careful formulation. Recall that we were not using maximum-likelihood parameter estimates because this almost certainly guarantees an infinite KL-divergence for more complex hypotheses. We used the predicitive distribution instead, which is the Bayes-optimal model within the model class for each structure. The MDL and AIC criteria were designed for maximum-likelihood estimates (although this is a subtlety that is frequently ignored). On the other hand, these criteria cannot be viewed as correcting bias anyway, for it is impossible to predict, a priori, the bias in sample error as an estimate of true error. (This can only be established asymptotically, which, as we have seen, is a rather meaningless assertion, especially when we can derive two different criteria this way.) Our bootstrap criterion also might be improved, by correcting somehow for the expected reduction in the empirical variance of the bootstrapped sample. These would merely be attempts to make these criteria more like the cross-validation and prequential criteria, however, so the value of this exercise is not clear.

[11] We *can* criticize it as the BIC, however.

Future Work

There are many avenues for future research. It would be interesting to consider other loss functions, other representations for belief networks, other model types such as Markov models or decision trees, and additional criteria. The experimental framework we have developed could be used to explore these and other issues. As well, we would like to carry out a more thorough theoretical analysis of the model selection problem in general. There are other interesting properties of learning algorithms besides expected error. We would like to compare these criteria in terms of asymptotic consistency, minimality, and convergence rates. Minimax properties are another interesting basis for a comparison: if an adversary could choose the true model, knowing your model selection strategy, how could you minimize the expected error? (Assuming, of course, your adversary chooses the model that maximizes your expected error under the sampling distribution of that model.) It would be worthwhile to complement the perspective taken here by these alternative theoretical perspectives.

8 Conclusion

The paradox of model selection remains. Why should we choose simpler models? For apparently benign prior assumptions the Bayes-optimal model is the predictive distribution for the most complex belief network. Yet, this is somewhat like throwing salt into the sea. Are "natural" distributions low-entropy? Do they have simple representations in our favourite encoding schemes? What do we believe? And what do we want our models to do, anyway?

From our empirical study we draw the following conclusions. First, we can arrange the criteria under consideration in order of their bias toward simplicity: MDL, AIC, XV, Preq, Boot, and (Fit). We noted that the MDL and AIC criteria displayed a bias for simplicity that was of an entirely different form from that displayed by the data-dependent criteria: it grew exponentially and thus was largely determined by the complexity penalty when the structure was large and the sample was small. We saw that it was difficult to predict the behaviour of these criteria on a given problem, as they were quite sensitive to the amount of data provided. In addition, we noted that the bootstrap criterion that we used seemed to be determined more by the sample error than the true error, and consequently did little to avoid overfitting. Another interesting aspect of our empirical study was the way that the Bayesian practice of using the average model for each structure, rather than the maximum-likelihood model, made the sample error non-decreasing in hypothesis complexity, and for the benchmark networks we considered, the sample error was a reasonably good criterion for model selection. We also noted that very modest-sized samples were sufficient to get good evaluations of these networks, as well as our random networks. This has implications for model selection search strategies, where there is a trade-off between getting a good evaluation of candidate structures and visiting more structures. It suggests that evaluating on a subset of the data might be very effective.

It is certainly desirable to base one's choice of criterion on an understanding of the issues and trade-offs in model selection, and on how various criteria differ in their approach to handling these trade-offs. Through an empirical comparison of model selection criteria, we have shed light on the behaviour of a broad class of criteria, spanning a range of different paradigms of inductive inference.

Acknowledgements

We would like to thank the National Science and Engineering Research Council of Canada for their financial support.

References

Beinlich, I., Suermont, G., Chavez, R. & Cooper, G. (1989). The ALARM monitoring system: A case study with two probabilistic inference techniques for belief networks. *Proceedings of the 2nd European Conference on AI and Medicine*, Springer Verlag.

Binder, J., Koller, D., Russell, S. & Kanazawa, K. (1997). Adaptive probabilistic networks with hidden variables. *Machine Learning 29.*

Bozdogan, H. (1987). Model selection and Akaike's information criterion (AIC): The general theory and its analytical extensions. *Psychometrika, 52, 3*, 345–370.

Cooper, G., & Herskovits, E. (1992). A Bayesian method for the induction of probabilistic networks from data. *Machine Learning, 9,* 309–347.

Cover, T. & Thomas, J. (1991). *Elements of Information Theory.* John Wiley & Sons.

Dawid, P. & Lauritzen, S. (1993). Hyper markov laws in the statistical analysis of decomposable graphical models, *Annals of Statistics*, 1993.

Dawid, P. & Vovk, V. (1999). Prequential probability: Principles and properties. *Bernoulli 5*, 125-162.

Efron, B. (1982). *The Jackknife, the Bootstrap and other Resampling Plans.* Society for Industrial and Applied Mathematics, Philadelphia.

Friedman, N., & Goldszmidt, M. (1996a). Learning Bayesian networks with local structure. *Proceedings of the 12th Conference on Uncertainty in Artificial Intelligence*, Morgan Kaufmann.

Friedman, N., & Yakhini, Z. (1996b). On the sample complexity of learning Bayesian networks. *Proceedings of the 12th Conference on Uncertainty in Artificial Intelligence*, Morgan Kaufmann.

Heckerman, D. (1995). *A tutorial on learning with Bayesian networks* Technical Report MSR-TR-95-06, Microsoft Research.

Heckerman, D. (1996). A comparison of scientific and engineering criteria for Bayesian model selection. Technical Report MSR-TR-96-12, Microsoft Research.

Howson, C. (1997). A logic of induction. *Philosophy of Science*, 64, (pp. 268-290).

Kearns, M., Mansour, Y., Ng, A. Y., & Ron, D. (1997). An experimental and theoretical comparison of model selection methods. *Machine Learning*, 14.

Kullback, S. & Leibler, R. (1951). On information and sufficiency. *Annals of Mathematical Statistics 22*, 79–86.

Lam, W., & Bacchus, F. (1994). Learning Bayesian belief networks: An approach based on the MDL principle. *Computational Intelligence*, *10*, 269–293.

Linhart, H., & Zucchini, W. (1986). *Model selection.* New York: John Wiley & Sons.

Liu, J., Chang, K. & Zhou, J. (1998). A hybrid convergent method for learning probabilistic networks. *Proceedings of the 12th Biennial Conference of the Canadian Society for Computational Studies of Intelligence*, 393–410.

Newman, T. & Odell, P. (1971). *The Generation of Random Variates*, Vol. 21 of Griffen's Statistical Monographs and Courses, Griffen, London, 1971.

Pearl, J. (1988). *Probabilistic reasoning in intelligent systems: Networks of plausible inference.* Morgan Kaufmann.

Rissanen, J. (1989). *Stochastic complexity in statistical inquiry.* World Scientific.

Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics*, *6*, 461–464.

Stone, M. (1974). Cross-validatory choice and assessment of statistical predictions. *J. Royal Statistical Society, Series B, 36*, 44-47.

Stone, M. (1977). Asymptotics for and against cross-validation. *Biometrika, 64, 1*, 29-35.

Suzuki, J. (1996). Learning bayesian belief networks based on the minimum description length principle: An efficient algorithm using the branch and bound technique. *Machine Learning 13.*

Wilks, S. (1962). *Mathematical Statistics*, John Wiley & Sons, New York.

Wolpert, D. (1996a). The lack of a priori distinctions between learning algorithms. Available at `ftp://ftp.santafe.edu/pub/dhw_ftp/`.

Wolpert, D. (1996b). The existence of a priori distinctions between learning algorithms. Available at `ftp://ftp.santafe.edu/pub/dhw_ftp/`.

A Properties of $f(\Theta \mid a+\alpha)$

Here we demonstrate some important properties of the posterior distribution $f(\Theta \mid a+\alpha)$, which follows Bayesian updating of the prior distribution. In particular, we claim the following properties:

- Each row of a CP-table is independently Dirichlet-distributed in the posterior distribution.
- The model in which each $\hat{\theta}_i$ is the mean of the posterior distribution is the average model: inferences based on this model are equivalent to those based on the predictive distribution.

We prove these properties here for the special case where all network variables are binary, and CP-tables are complete; the generalization to n-ary variables and marginalized CP-tables is conceptually straightforward, but the notation becomes cumbersome. See also Cooper and Herskovits (1992) who make these claims without proof (their proofs are available in a technical report). Dawid and Lauritzen (1993) present more general results. We present these results here to make the paper self-contained.

First we derive the likelihood function $L(d \mid m, \Theta)$. Because d is a sequence $d_1 \ldots d_m$ of i.i.d. instances of $X_1 \ldots X_n$, we can write:

$$L(d \mid m, \Theta) = \prod_{i=1}^{m} P(d_i \mid \Theta)$$

And, because we interpret our model as a conditional factorization of a joint probability distribution, we can write:

$$P(d_i \mid \Theta) = \prod_{j=1}^{n} P_j(x_{ij} \mid w_{ij})$$

where x_{ij} is the assignment to X_j in d_i, and w_{ij} is the joint assignment to W_j in d_i.

Since $P_j(x_{ij} \mid w_{ij})$ is a network parameter, or pseudo-parameter, we can rearrange the factors as:

$$L(d \mid m, \Theta) = \prod_{i=1}^{k} {\Theta_i}^{a_i}$$

where a_i (as previously defined) is the observed frequency in the sufficient statistic for Θ.

Next we derive the posterior density function for Θ. Baye's theorem gives us:

$$f(\Theta \mid a+\alpha) = \frac{L(d \mid m, \Theta) f(\Theta \mid \alpha)}{\int L(d \mid m, \Theta) f(\Theta \mid \alpha)\ \mathrm{d}\Theta}$$

We can take the parameters to be jointly Beta-distributed in pairs, ($\Theta_{i+1} = 1 - \Theta_i$ when i is odd) and given our previous equality for the likelihood function, write the posterior as:

$$f(\Theta \mid a+\alpha) = \frac{{\Theta_1}^{a_1}{\Theta_2}^{a_2}\mathrm{Be}(\Theta_1 \mid \alpha_1, \alpha_2)\cdots{\Theta_{k-1}}^{a_{k-1}}{\Theta_k}^{a_k}\mathrm{Be}(\Theta_{k-1} \mid \alpha_{k-1}, \alpha_k)}{\int\cdots\int {\Theta_1}^{a_1}{\Theta_2}^{a_2}\mathrm{Be}(\Theta_1 \mid \alpha_1, \alpha_2)\cdots{\Theta_{k-1}}^{a_{k-1}}{\Theta_k}^{a_k}\mathrm{Be}(\Theta_{k-1} \mid \alpha_{k-1}, \alpha_k)\ \mathrm{d}\theta_1\cdots\ \mathrm{d}\theta_{k-1}}$$

Using the definition of the Beta distribution, and our assumptions about the form of Θ, we can write each ${\Theta_i}^{a_i}{\Theta_{i+1}}^{a_{i+1}}\mathrm{Be}(\Theta_i \mid \alpha_i, \alpha_{i+1})$ as ${\Theta_i}^{a_i}(1-\Theta_i)^{a_{i+1}}{\Theta_i}^{\alpha_i-1}(1-\Theta_i)^{\alpha_{i+1}-1}D(\alpha_i, \alpha_{i+1})$, and thus as: ${\Theta_i}^{a_i+\alpha_i-1}(1-\Theta_i)^{a_{i+1}+\alpha_{i+1}-1}D(\alpha_i, \alpha_{i+1})$. We can then cancel the Dirichlet integrals in the top with those on the bottom, and factor out the integration to get:

$$f(\Theta \mid a+\alpha) = \frac{\Theta_1^{a_1+\alpha_1-1}(1-\Theta_1)^{a_2+\alpha_2-1}\cdots\Theta_{k-1}^{a_{k-1}+\alpha_{k-1}-1}(1-\Theta_{k-1})^{a_k+\alpha_k-1}}{\int_0^1 {\Theta_1}^{a_1+\alpha_1-1}(1-\Theta_1)^{a_2+\alpha_2-1}\ \mathrm{d}\Theta_1\cdots\int_0^1 {\Theta_{k-1}}^{a_{k-1}+\alpha_{k-1}-1}(1-\Theta_{k-1})^{a_k+\alpha_k-1}\ \mathrm{d}\Theta_{k-1}}$$

and since the integrals in the denominator are equal to Beta functions (Dirichlet integrals with two parameters) it follows that:

$$f(\Theta \mid a+\alpha) = \mathrm{Be}(\Theta_1 \mid a_1+\alpha_1, a_2+\alpha_2)\cdots\mathrm{Be}(\Theta_{k-1} \mid a_{k-1}+\alpha_{k-1}, a_k+\alpha_k)$$

a product of independent Beta distributions, one for each row of a CP-table $\langle\Theta_i, \Theta_{i+1}\rangle$ (for i odd), as claimed. Furthermore, the posterior parameters are updated by just adding the statistic a to the prior parameters, α. This proves the first claim.

The second claim follows easily from the first. The predictive distribution is given by:

$$Q(X_1, \ldots, X_n \mid \dot{\alpha}) = \int P(X_1 \ldots X_n \mid \Theta) f(\Theta \mid \dot{\alpha})\, \mathrm{d}\Theta$$

where we use the notation $\dot{\alpha}$ to denote the sufficient statistic $a + \alpha$. Our second claim is that:

$$Q(X_1, \ldots, X_n \mid \dot{\alpha}) = P(X_1 \ldots X_n \mid \hat{\theta})$$

where $\hat{\theta}$ is the mean of the posterior distribution, given by $\hat{\theta}_i = \dot{\alpha}_i/(\dot{\alpha}_i + \dot{\beta}_i)$ Note that, for a particular full instantiation $x_1 \ldots x_n$ $P(x_1 \ldots x_n \mid \Theta)$ depends on at most one parameter per CP-table row. So the predictive distribution is given by a product of n independent factors of the form:

$$\int \Theta_i \mathrm{Be}(\Theta_i \mid \dot{\alpha}_i, \dot{\beta}_i)\, \mathrm{d}\Theta_i = \int \frac{{\Theta_i}^{\dot{\alpha}_i}(1-\Theta_i)^{\dot{\beta}_i - 1}}{D(\dot{\alpha}_i, \dot{\beta}_i)}\, \mathrm{d}\Theta_i = \frac{D(\dot{\alpha}_i + 1, \dot{\beta}_i)}{D(\dot{\alpha}_i, \dot{\beta}_i)}$$

$$= \frac{\Gamma(\dot{\alpha}_i + 1)\Gamma(\dot{\beta}_i)\Gamma(\dot{\alpha}_i + \dot{\beta}_i)}{\Gamma(\dot{\alpha}_i + \dot{\beta}_i + 1)\Gamma(\dot{\alpha}_i)\Gamma(\dot{\beta}_i)} = \frac{(\dot{\alpha}_i)\Gamma(\dot{\alpha}_i)\Gamma(\dot{\alpha}_i + \dot{\beta}_i)}{(\dot{\alpha}_i + \dot{\beta}_i)\Gamma(\dot{\alpha}_i + \dot{\beta}_i)\Gamma(\dot{\alpha}_i)} = \frac{\dot{\alpha}_i}{\dot{\alpha}_i + \dot{\beta}_i}$$

as claimed.

We can also define a predictive likelihood function $PL(d_1 \ldots d_m \mid m, \dot{\alpha})$:

$$\begin{array}{rcl} PL(d \mid m, \dot{\alpha}) & = & \int \prod_{i=1}^{m} P(d_i \mid \Theta) f(\Theta \mid \dot{\alpha})\, \mathrm{d}\Theta \\ & = & \int \prod_{i=1}^{k} {\Theta_i}^{\dot{a}_i} f(\Theta \mid \dot{\alpha})\, \mathrm{d}\Theta \\ & = & \int {\Theta_1}^{\dot{a}_1}(1-\Theta_1)^{\dot{b}_1} \mathrm{Be}(\Theta_1 \mid \dot{\alpha}_1, \dot{\beta}_1)\, \mathrm{d}\Theta_1 \\ & & \vdots \\ & & \int {\Theta_{k-1}}^{\dot{a}_{k-1}}(1-\Theta_{k-1})^{\dot{b}_{k-1}} \mathrm{Be}(\Theta_{k-1} \mid \dot{\alpha}_{k-1}, \dot{\beta}_{k-1})\, \mathrm{d}\Theta_{k-1} \end{array}$$

where the dotted $\dot{a}$ and $\dot{b}$ are the observed frequencies in d, and where, as previously, we take each pair of parameters, $\langle\Theta_i, \Theta_{i+1}\rangle$ (i odd) to be jointly Beta-distributed. Then, for each i simple algebra gives us:

$$\int {\Theta_i}^{\dot{a}_i}(1-\Theta_i)^{\dot{b}_i} \mathrm{Be}(\Theta_i \mid \dot{\alpha}_i, \dot{\beta}_i)\, \mathrm{d}\Theta_i = \int \frac{{\Theta_i}^{\dot{a}_i + \dot{\alpha}_i - 1}(1-\Theta_i)^{\dot{b}_i + \dot{\beta}_i - 1}}{D(\dot{\alpha}_i, \dot{\beta}_i)}\, \mathrm{d}\Theta_i = \frac{D(\dot{a}_i + \dot{\alpha}_i, \dot{b}_i + \dot{\beta}_i)}{D(\dot{\alpha}_i, \dot{\beta}_i)}$$

$$= \frac{\Gamma(\dot{a}_i + \dot{\alpha}_i)\Gamma(\dot{b}_i + \dot{\beta}_i)\Gamma(\dot{\alpha}_i + \dot{\beta}_i)}{\Gamma(\dot{a}_i + \dot{\alpha}_i + \dot{b}_i + \dot{\beta}_i)\Gamma(\dot{\alpha}_i)\Gamma(\dot{\beta}_i)} = \frac{(\dot{a}_i + \dot{\alpha}_i - 1)\cdots\dot{\alpha}_i\Gamma(\dot{\alpha}_i)(\dot{b}_i + \dot{\beta}_i - 1)\cdots\dot{\beta}_i\Gamma(\dot{\beta}_i)\Gamma(\dot{\alpha}_i + \dot{\beta}_i)}{(\dot{a}_i + \dot{\alpha}_i + \dot{b}_i + \dot{\beta}_i - 1)\cdots(\dot{\alpha}_i + \dot{\beta}_i)\Gamma(\dot{\alpha}_i + \dot{\beta}_i)\Gamma(\dot{\alpha}_i)\Gamma(\dot{\beta}_i)}$$

$$= \frac{(\dot{a}_i + \dot{\alpha}_i - 1)\cdots\dot{\alpha}_i(\dot{b}_i + \dot{\beta}_i - 1)\cdots\dot{\beta}_i}{(\dot{a}_i + \dot{\alpha}_i + \dot{b}_i + \dot{\beta}_i - 1)\cdots(\dot{\alpha}_i + \dot{\beta}_i)}$$

And we can see from this that the predictive likelihood is given simply by iteratively calculating the probability of each instance of the sample, then updating the parameters, and repeating the process using the resulting model (as we can rearrange the factors on the top and bottom of the fraction above to be in the appropriate sequence).