

A Comparison of Scientific and Engineering Criteria for
Bayesian Model Selection

David Heckerman
heckerma@microsoft.com

David Maxwell Chickering
dmax@microsoft.com

Technical Report
MSR-TR-96-12

Microsoft Research
Advanced Technology Division
Microsoft Corporation
One Microsoft Way
Redmond, WA 98052

Topics: graphical models, model uncertainty, model selection

1 Introduction

Suppose that the joint probability distribution over a set of random variables $\mathbf{X} = \{X_1, \dots, X_n\}$ is given by $p(\mathbf{X}|\boldsymbol{\theta}_m, \mathbf{m})$, where \mathbf{m} is a model with parameters $\boldsymbol{\theta}_m$. In addition, suppose that the true model and its parameters are unknown, but we nevertheless want to estimate the true distribution somehow given a random sample $D = \{\mathbf{x}_1, \dots, \mathbf{x}_N\}$ from the true distribution.

In the Bayesian approach to this problem, we define a discrete random variable \mathbf{M} whose states correspond to the possible true models, and encode our uncertainty about \mathbf{M} with the probabilities $p(\mathbf{M} = \mathbf{m})$. In this paper, we assume that there are a finite number of possible true models. For each possible model \mathbf{m} , we define the random (vector) variable Θ_m whose values correspond to the possible values of the parameters for \mathbf{m} . We encode our uncertainty about Θ_m using the probability distribution $p(\Theta_m|\mathbf{m})$. In this paper, we assume that $p(\Theta_m|\mathbf{m})$ is a probability density function.

Given random sample D , we compute the posterior distributions for \mathbf{M} and each Θ_m using Bayes' rule:

$$p(\mathbf{m}|D) = \frac{p(\mathbf{m})p(D|\mathbf{m})}{\sum_{m'} p(\mathbf{m}')p(D|\mathbf{m}')} \\ p(\boldsymbol{\theta}_m|D, \mathbf{m}) = \frac{p(\boldsymbol{\theta}_m|\mathbf{m})p(D|\boldsymbol{\theta}_m, \mathbf{m})}{p(D|\mathbf{m})}$$

where

$$p(D|\mathbf{m}) = \int p(D|\boldsymbol{\theta}_m, \mathbf{m}) p(\boldsymbol{\theta}_m|\mathbf{m}) d\boldsymbol{\theta}_m$$

We estimate the joint distribution for \mathbf{X} , by averaging over all possible models and their parameters:

$$p(\mathbf{x}|D) = \sum_m p(\mathbf{m}|D) \int p(\mathbf{x}|\boldsymbol{\theta}_m, \mathbf{m}) p(\boldsymbol{\theta}_m|D, \mathbf{m}) d\boldsymbol{\theta}_m \quad (1)$$

This approach is known as *Bayesian model averaging*.

In many real-world problems, the sum over possible models is intractable. A common approximation in these circumstances is to select a single “good” model \mathbf{m} , and to estimate the joint distribution for \mathbf{X} using

$$p(\mathbf{x}|D, \mathbf{m}) = \int p(\mathbf{x}|\boldsymbol{\theta}_m, \mathbf{m}) p(\boldsymbol{\theta}_m|D, \mathbf{m}) d\boldsymbol{\theta}_m$$

This approach is known as *Bayesian model selection*.

Model scores that define “good” models are commonly known as *criteria*. A criterion commonly used in Bayesian model selection is the logarithm of the relative posterior probability of the model $\log p(\mathbf{m}, D) = \log p(\mathbf{m}) + \log p(D|\mathbf{m})$. Under the assumption that the prior distribution for \mathbf{M} is uniform, an equivalent criterion is $\log p(D|\mathbf{m})$, the *log marginal likelihood* of the data given the model. In the remainder of this paper, we assume that $p(\mathbf{M})$ is uniform to simplify our presentation, although the generalization to non-uniform model priors is straightforward.

The log-marginal-likelihood criterion has the following interesting interpretation described by Dawid (1984). From the chain rule of probability, we have

$$\log p(D|\mathbf{m}) = \sum_{l=1}^N \log p(\mathbf{x}_l|\mathbf{x}_1, \dots, \mathbf{x}_{l-1}, \mathbf{m})$$

The term $p(\mathbf{x}_l|\mathbf{x}_1, \dots, \mathbf{x}_{l-1}, \mathbf{m})$ is the prediction for \mathbf{x}_l made by model \mathbf{m} after averaging over its parameters. The log of this term can be thought of as the score or utility for this prediction under the scoring rule or utility function $\log p(\mathbf{x})$.¹ Thus, a model with the highest log marginal likelihood is also a model that is the best sequential predictor of the data D under the log scoring rule.

This observation suggests an alternative criterion for choosing \mathbf{m} . Rather than select a model that is the best sequential predictor of the data we have seen, we can select a model that is the best predictor of the *next* observation we will see, given the data we have seen. Using again the log scoring rule, the utility to maximize is

$$\log p(\mathbf{x}_{N+1}|D, \mathbf{m})$$

Because we have not yet seen the next observation, we average this utility over all possible ones, obtaining the following criterion for model \mathbf{m} given data D :

$$EC(\mathbf{m}, D) = \sum_{\mathbf{x}_{N+1}} p(\mathbf{x}_{N+1}|D) \log p(\mathbf{x}_{N+1}|D, \mathbf{m}) \quad (2)$$

where $p(\mathbf{x}_{N+1}|D)$ is given by Equation 1. This criterion, first suggested by Chow (1981) and made more precise by San Martini and Spezzaferri (1984), is the negative cross entropy between the correct posterior distribution $p(\mathbf{x}_{N+1}|D)$ and the posterior distribution determined by model \mathbf{m} .

Whereas the log-marginal-likelihood criterion selects a model that is most likely to be true, this alternative criterion selects a model that is the best predictor of the next observation. Thus, we sometimes refer to these two scoring functions as a *scientific criterion* (SC) and *engineering criterion* (EC), respectively.

¹An axiomatic characterization of this proper scoring rule is given by Bernardo (1979).

When we are interested in prediction, EC is a better criterion than SC. Unfortunately, the use of EC is impractical for model selection, because its computation involves the sum over models in Equation 1. In particular, if we could perform this sum, we would be better off making predictions using Equation 1. Thus, in practice, we use predictions based on SC as an approximation for predictions based on EC, which in turn is an approximation for predictions based on model averaging.

When N is large, these approximations are good. In particular, as N increases, the probability of the model \mathbf{m} that is closest to truth (in the KL sense) will approach one, and we obtain $p(\mathbf{x}_{N+1}|D) = p(\mathbf{x}_{N+1}|D, \mathbf{m})$. In this situation, both SC and EC will select this model. The latter observation follows from Equation 2 and the fact that cross entropy between two probability distributions is minimized when the distributions are equal. We know of no theoretical characterizations for these approximations, however, when N is small. Draper (1993) and Madigan et al. (1996) provide experimental comparisons of predictions based on SC and model averaging. In this paper, we perform additional experimental comparisons of the type, and also compare predictions based on EC with those based on SC and model averaging. We perform these comparisons in the context of Bayesian-network models for discrete variables.

2 Bayesian Networks

A Bayesian network for a set of random variables $\mathbf{X} = \{X_1, \dots, X_n\}$ is the pair (S, P) , where S is an acyclic directed graph, which we call the *structure* of the Bayesian network, and P is a set of *local probability distributions*. The nodes in S are in one-to-one correspondence with the variables \mathbf{X} . We use X_i to denote both the variable and its corresponding node, and \mathbf{Pa}_i to denote the parents of node X_i in S as well as the variables corresponding to those parents. The lack of possible arcs in S reflect conditional independence assertions. In particular, given structure S , the joint probability distribution for \mathbf{X} is given by

$$p(\mathbf{x}) = \prod_{i=1}^n p(x_i | \mathbf{pa}_i) \quad (3)$$

The local probability distributions P are the distributions corresponding to the terms in the product of Equation 3.²

We can use Bayesian networks as models in the sense of Section 1 as follows. First, we suppose that the true joint distribution for \mathbf{X} factors according to some structure S , but we are uncertain about the identity of S . We write $\mathbf{M} = \mathbf{m}_s$ when the true distribution

²Sometimes, an additional causal interpretation is given to the arcs in S . Namely, an arc from X_i to X_j reflects the assertion that X_i is a direct cause of X_j (Spirtes et al., 1993; Pearl, 1995).

factors according to S .³ Second, we parameterize the local probability distributions with a finite number of parameters. Explicitly conditioning on the model and its parameters, we rewrite Equation 3 as

$$p(\mathbf{x}|\boldsymbol{\theta}_s, \mathbf{m}_s) = \prod_{i=1}^n p(x_i|\mathbf{pa}_i, \boldsymbol{\theta}_i, \mathbf{m}_s)$$

where $\boldsymbol{\theta}_i$ are the parameters for the local distribution associated with \mathbf{X}_i , and $\boldsymbol{\theta}_s = (\boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_n)$ are the parameters for the structure as a whole.

In this paper, we concentrate on the case where every variable in \mathbf{X} is discrete. Let x_i^k and \mathbf{pa}_i^j denote the k th possible state of X_i and the j th possible state of \mathbf{Pa}_i , respectively. Also, let r_i and q_i denote the number of possible states of X_i and \mathbf{Pa}_i , respectively. We further specialize to the case where $p(x_i|\mathbf{pa}_i, \boldsymbol{\theta}_i, \mathbf{m}_s)$ for each state of \mathbf{Pa}_i is a multinomial distribution:

$$p(x_i^k|\mathbf{pa}_i^j, \boldsymbol{\theta}_i, \mathbf{m}_s) = \theta_{ijk}$$

such that $\theta_{ijk} > 0$ for all i, j , and k , and $\sum_{k=1}^{r_i} \theta_{ijk} = 1$ for all i and j . Given these parameters, we define the vector combinations

$$\boldsymbol{\theta}_{ij} = (\theta_{ijk})_{k=1}^{r_i} \quad \boldsymbol{\theta}_i = (\boldsymbol{\theta}_{ij})_{j=1}^{q_i}$$

The scientific and engineering criteria can be computed efficiently and in closed form assuming (1) the parameters $\boldsymbol{\theta}_{ij}$ are mutually independent:

$$p(\boldsymbol{\theta}_s|\mathbf{m}_s) = \prod_{i=1}^n \prod_{j=1}^{q_i} p(\boldsymbol{\theta}_{ij}|\mathbf{m}_s)$$

(2) each parameter set $\boldsymbol{\theta}_{ij}$ has a Dirichlet distribution:

$$p(\boldsymbol{\theta}_{ij}|\mathbf{m}_s) = c \cdot \prod_{k=1}^{r_i} \theta_{ijk}^{\alpha_{ijk}-1}$$

where $\alpha_{ijk} > 0$ for every i, j , and k , and c is a normalization constant, and (3) data is complete—that is, there are no missing observations. Under these assumptions, several researchers (e.g., Cooper and Herskovits, 1992) have shown that

$$p(\mathbf{x}_{N+1}|D, \mathbf{m}_s) = \prod_{i=1}^n \frac{\alpha_{ijk} + N_{ijk}}{\alpha_{ij} + N_{ij}}$$

where $X_i = x_i^k$ and $\mathbf{Pa}_i = \mathbf{pa}_i^j$ in \mathbf{x}_{N+1} (k and j depend on i), N_{ijk} is the number observations in D in which $X_i = x_i^k$ and $\mathbf{Pa}_i = \mathbf{pa}_i^j$, $N_{ij} = \sum_{k=1}^{r_i} N_{ijk}$, and $\alpha_{ij} = \sum_{k=1}^{r_i} \alpha_{ijk}$. In

³We use the causal interpretation of Bayesian-network structure so that different structures correspond to mutually exclusive events. We treat this issue more carefully in a later version.

addition, it can be shown that

$$p(D|\mathbf{m}_s) = \prod_{i=1}^n \prod_{j=1}^{q_i} \frac{\Gamma(\alpha_{ij})}{\Gamma(\alpha_{ij} + N_{ij})} \cdot \prod_{k=1}^{r_i} \frac{\Gamma(\alpha_{ijk} + N_{ijk})}{\Gamma(\alpha_{ijk})}$$

3 Experiments

As mentioned, our goal is to compare the accuracy of predictions based on SC, EC, and model averaging. To do so, we create several Bayesian networks, and from them generate random data sets of various sizes. We then select models using the two criteria, and compare the EC for both models with the maximum value for EC obtained by using the correct prediction:

$$EC_{\text{opt}} = \sum_{\mathbf{x}_{N+1}} p(\mathbf{x}_{N+1}|D) \log p(\mathbf{x}_{N+1}|D)$$

In computing the criteria for a given model, we assume (1) uniform priors on models, and (2) Dirichlet parameter priors with $\alpha_{ijk} = 8/r_i q_i$ for all i, j , and k . Because the number of possible Bayesian-network structures for n variables is more than exponential in n , we perform our experiments for small n ($n = 4$) only.

Results are shown in Table 1. The number in parentheses under each EC value corresponds to $(EC(\mathbf{m}, D) - \text{ave}(EC))/\text{sd}(EC)$, where $\text{ave}(EC)$ and $\text{sd}(EC)$ are the simple average and standard deviations of $EC(\mathbf{m}, D)$ over all models, respectively. Given our structure and parameter priors, the scientific (and engineering) criteria for two Markov equivalent structures are equal. Thus, each criterion selects an equivalence class of structures. In the table, we report a representative acyclic directed graph from each selected class. Note that all *complete structures*—that is, structures containing no missing arcs—are Markov equivalent. Because space is limited, we do not give the local probability distributions for the generative models.

4 Discussion

Our results confirm the conclusions of Draper (1993) and Madigan (1996) that model averaging sometimes produces substantially better predictions than does model selection. We also see that, when using model selection to choose a predictive model, SC is a good approximation for EC.

The results also confirm our argument that the two criteria select the same models when the sample size becomes sufficiently large. More interesting, for small sample sizes, we find that the engineering criterion tends to select models that are more complex than those selected by the scientific criterion. A simple explanation for this difference is that, when

Table 1: Structures selected by SC and EC given data of sample size N generated by various four-variable Bayesian networks. The structures selected by SC and EC are denoted \mathbf{m}_{sc} and \mathbf{m}_{ec} , respectively.

generative structure: empty (no arcs)					
N	\mathbf{m}_{sc}	$-\text{EC}(\mathbf{m}_{sc}, D)$	\mathbf{m}_{ec}	$-\text{EC}(\mathbf{m}_{ec}, D)$	$-\text{EC}_{opt}$
50	empty	2.44176 (1.20)	$X_1 \rightarrow X_4$	2.44170 (1.21)	2.43526 (2.30)
200	$X_1 \rightarrow X_4$	2.16721 (1.45)	$X_1 \rightarrow X_3, X_1 \rightarrow X_4$	2.16720 (1.46)	2.16559 (2.15)
800	$X_3 \rightarrow X_1 \leftarrow X_4$	2.09369 (0.54)	$X_3 \rightarrow X_1 \rightarrow X_4$	2.09275 (1.27)	2.09239 (1.56)
3200	empty	2.08585 (1.38)	empty	2.08585 (1.38)	2.08580 (1.48)

generative structure: $X_1 \rightarrow X_2 \rightarrow X_3 \rightarrow X_4$					
N	\mathbf{m}_{sc}	$-\text{EC}(\mathbf{m}_{sc}, D)$	\mathbf{m}_{ec}	$-\text{EC}(\mathbf{m}_{ec}, D)$	$-\text{EC}_{opt}$
50	$X_1 \rightarrow X_2 \rightarrow X_3 \rightarrow X_4$	2.02613 (1.39)	$X_1 \rightarrow X_2 \rightarrow X_3 \rightarrow X_4,$ $X_1 \rightarrow X_3, X_3 \rightarrow X_4$	2.01905 (1.50)	2.01433 (1.58)
200	$X_1 \rightarrow X_2 \rightarrow X_3 \rightarrow X_4$	1.43916 (1.32)	$X_1 \rightarrow X_2 \rightarrow X_3 \rightarrow X_4,$ $X_2 \rightarrow X_4$	1.43558 (1.37)	1.43486 (1.38)
800	$X_1 \rightarrow X_2 \rightarrow X_3 \rightarrow X_4$	1.43057 (1.32)	$X_1 \rightarrow X_2 \rightarrow X_3 \rightarrow X_4$	1.43057 (1.32)	1.43034 (1.32)
3200	$X_1 \rightarrow X_2 \rightarrow X_3 \rightarrow X_4$	1.34462 (1.26)	$X_1 \rightarrow X_2 \rightarrow X_3 \rightarrow X_4$	1.34462 (1.26)	1.34461 (1.26)

generative structure: complete (no missing arcs)					
N	\mathbf{m}_{sc}	$-\text{EC}(\mathbf{m}_{sc}, D)$	\mathbf{m}_{ec}	$-\text{EC}(\mathbf{m}_{ec}, D)$	$-\text{EC}_{opt}$
50	$X_1 \rightarrow X_2 \rightarrow X_3 \leftarrow X_4$	2.28328 (0.91)	$X_1 \rightarrow X_2 \rightarrow X_3 \rightarrow X_4,$ $X_1 \rightarrow X_3$	2.28003 (0.96)	2.27277 (1.08)
200	$X_1 \rightarrow X_2 \rightarrow X_3 \leftarrow X_4$	2.23339 (1.15)	$X_1 \rightarrow X_2 \rightarrow X_3 \leftarrow X_4,$ $X_2 \rightarrow X_4$	2.23031 (1.20)	2.22906 (1.22)
800	$X_2 \rightarrow X_3 \rightarrow X_4,$ $X_1 \rightarrow X_4, X_2 \rightarrow X_4$	2.20401 (1.17)	$X_2 \rightarrow X_3 \rightarrow X_4,$ $X_1 \rightarrow X_4$	2.20014 (1.23)	2.19934 (1.24)
3200	complete	2.15849 (1.33)	complete	2.15849 (1.33)	2.15846 (1.33)

using EC, we reward a prediction based on all N observations. In contrast, when using SC, we reward predictions based on $0, 1, 2, \dots, N - 1$ observations—that is, less data. Thus, EC will tend to select more complex models, because it can afford to do so without overfitting the data. An alternative argument, due to Wray Buntine (personal communication), is as follows. When using EC, we choose the model that is closest (in the KL sense) to the correct posterior distribution for \mathbf{x} . This correct distribution is an average over models, some of which are more complicated than the most likely model (i.e., the model selected when using SC). Consequently, when using EC, we tend to select a model that is more complex than the most likely model.

Of course, for our conclusions to be trusted, we need to test more structures, priors, and data sets. In a later version of this paper, we shall do so. In addition, we shall discuss the connections between the SC and EC and the well-known BIC (Schwarz, 1978) and AIC (Akaike, 1973).

References

- Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In *Second International Symposium on Information Theory*, page 267. Akademiai Kiado, Budapest.
- Bernardo, J. (1979). Expected information as expected utility. *Annals of Statistics*, 7:686–690.
- Chow, G. (1981). A comparison of the information and posterior probability criteria for model selection. *Journal of Econometrics*, 16:21–33.
- Cooper, G. and Herskovits, E. (1992). A Bayesian method for the induction of probabilistic networks from data. *Machine Learning*, 9:309–347.
- Dawid, P. (1984). Present position and potential developments: some personal views. statistical theory. the prequential approach (with Discussion). *Journal of the Royal Statistical Society A*, 147:178–292.
- Draper, D. (1993). Assessment and propagation of model uncertainty. Technical Report 124, Department of Statistics, University of California, Los Angeles.
- Madigan, D., Raftery, A., Volinsky, C., and Hoeting, J. (1996). Bayesian model averaging. In *Proceedings of the AAAI Workshop on Integrating Multiple Learned Models*, Portland, OR.

- Pearl, J. (1995). Causal diagrams for empirical research. *Biometrika*, 82:669–710.
- SanMartini, A. and Spezzaferri, F. (1984). A predictive model selection criterion. *Journal of the Royal Statistical Society, B*, 46:296–303.
- Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics*, 6:461–464.
- Spirtes, P., Glymour, C., and Scheines, R. (1993). *Causation, Prediction, and Search*. Springer-Verlag, New York.