Statistics and Computing (2000) **10,** 55–62

A comparison of scientific and engineering criteria for Bayesian model selection

DAVID MAXWELL CHICKERING and DAVID HECKERMAN

Microsoft Research, Redmond WA 98052-6399, USA
(dmax@microsoft.com), (heckerma@microsoft.com)

Submitted March 1998 and accepted June 1999

Given a set of possible models for variables **X** and a set of possible parameters for each model, the Bayesian "estimate" of the probability distribution for **X** given observed data is obtained by averaging over the possible models and their parameters. An often-used approximation for this estimate is obtained by selecting a single model and averaging over its parameters. The approximation is useful because it is computationally efficient, and because it provides a model that facilitates understanding of the domain. A common criterion for model selection is the posterior probability of the model. Another criterion for model selection, proposed by San Martini and Spezzafari (1984), is the predictive performance of a model for the next observation to be seen. From the standpoint of domain understanding, both criteria are useful, because one identifies the model that is most likely, whereas the other identifies the model that is the best predictor of the next observation. To highlight the difference, we refer to the posterior-probability and alternative criteria as the *scientific criterion* (SC) and *engineering criterion* (EC), respectively. When we are interested in predicting the next observation, the model-averaged estimate is at least as good as that produced by EC, which itself is at least as good as the estimate produced by SC. We show experimentally that, for Bayesian-network models containing discrete variables only, the predictive performance of the model average can be significantly better than those of single models selected by either criterion, and that differences between models selected by the two criterion can be substantial.

Keywords: model selection, model averaging, Bayesian selection criteria

1. Introduction

Suppose that the joint probability distribution over a set of variables $\mathbf{X} = \{X_1, \ldots, X_n\}$ is given by $p(\mathbf{x} \mid \boldsymbol{\theta}_m, \mathbf{m})$, where $\mathbf{m}$ is a model with parameters $\boldsymbol{\theta}_m$.[1] 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 $D = \{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$ from the true distribution.

In the Bayesian approach to this problem, we define a discrete variable $\mathbf{M}$ whose states correspond to the possible true models, and encode our uncertainty about $\mathbf{M}$ with the probabilities $p(\mathbf{m})$. In this paper, we assume that there are a finite number of possible true models.[2] For each possible model $\mathbf{m}$, we define the (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 $p(\boldsymbol{\theta}_m \mid \mathbf{m})$. We assume that $p(\boldsymbol{\theta}_m \mid \mathbf{m})$ is a probability density function. Given random sample D, we compute the posterior distributions for $\mathbf{M}$ and each Θ_m using Bayes' rule:

$$p(\mathbf{m} \mid D) = \frac{p(\mathbf{m})p(D \mid \mathbf{m})}{\Sigma_{m'} p(\mathbf{m}')p(D \mid \mathbf{m}')}$$

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

where

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

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

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

The approach is sometimes called *Bayesian model averaging.*

In many real-world problems, the sum over possible models is intractable. Or, even when the sum can be performed, the averaged model is difficult to interpret. In either of these circumstances, a common approach is to select a single "good" model $\mathbf{m}$, and to estimate the jont distribution for $\mathbf{X}$ using

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

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 $\log p(\mathbf{m}, D) = \log p(\mathbf{m}) + \log p(D \mid \mathbf{m})$. Under the assumption that the prior distribution for $\mathbf{M}$ is uniform, an equivalent criterion is $\log p(D \mid \mathbf{m})$, 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. The generalization of the mathematics 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

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

The term $p(\mathbf{x}_l \mid \mathbf{x}_1, \ldots, \mathbf{x}_{l-1}, \mathbf{m})$ 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 $\log p(\mathbf{x})$.[3] 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

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

Because we have not yet seen $\mathbf{x}_{N+1}$, we average this utility over all possible observations, obtaining the following criterion for model $\mathbf{m}$ given data D:

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

where $p(\mathbf{x}_{N+1} \mid D)$ is given by equation 1. We call this criterion the *engineering criterion* for reasons that we make clear in a moment. 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 $p(\mathbf{x}_{N+1} \mid D)$ and the posterior distribution determined by model $\mathbf{m}$.

In terms of model understanding, both criteria are useful. Using the log-marginal-likelihood criterion, we identify a model that is most likely to be true. Using the alternative criterion given by equation 2, we identify a model that is the best predictor of the next observation. To emphasize the difference between the two criteria, we refer to the log-marginal-likelihood criterion and equation 2 as the *scientific criterion* (SC) and *engineering criterion* (EC), respectively. In any given analysis, one or both models may provide insights about the domain.

When we substitute $p(\mathbf{x}_{N+1} \mid D)$ for $p(\mathbf{x}_{N+1} \mid D, \mathbf{m})$ in equation 2, the engineering criterion obtains its maximum value. That is, the criterion is maximized when we make predictions using the model-averaged estimate. Also, as N approaches infinity, the model that includes the true distribution will have a posterior probability that approaches one, and we obtain $p(\mathbf{x}_{N+1} \mid D) = p(\mathbf{x}_{N+1} \mid D, \mathbf{m})$. Consequently, in this limit, the estimates produced by model averaging and by model selection using the two criterion coincide.

But what happens is the non-asymptotic regime? If we are interested in predicting the next observation, how much do we lose by using a single model instead of the model average? How much more do we lose by using a model selected by SC rather than one selected by EC? Alternatively, if we are interested in model understanding, how different are the models produced by the SC and EC criteria? In this paper, we investigate these questions in the context of Bayesian-network models for discrete variables.

2. Bayesian networks

A Bayesian network for a set of variables $\mathbf{X} = \{X_1, \ldots, X_n\}$ is the pair (S, P), where S is a directed acyclic 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

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

The local probability distributions P are the distributions corresponding to the terms in the product of equation 3.[4]

We can use Bayesian networks as models in the sense of Section 1 as follows. First, we suppose that the true joint distribution for $\mathbf{X}$ exhibits precisely the conditional independencies encoded by some structure S, but we are uncertain about the identity of S. We write $\mathbf{M} = \mathbf{m}_s$ when precisely the independence assertions implied by S hold in the true joint distribution. Second, for each model $\mathbf{m}_s$, 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

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

where $\boldsymbol{\theta}_i$ are the parameters for the local distribution associated with $\mathbf{X}_i$, and $\boldsymbol{\theta}_s = (\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_n)$ are the parameters for model $\mathbf{m}_s$ 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 kth possible state of X_i and the jth 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 $p(x_i \mid \mathbf{pa}_i, \boldsymbol{\theta}_i, \mathbf{m}_s)$ is a collection of multinomial distributions, one distribution for each state of $\mathbf{Pa}_i$:

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

where $\theta_{ijk} > 0$ 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

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

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

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

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

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

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

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

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 addition, it can be shown that

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

Table 1. *Sensitivity to generative model and sample size N. Models* $\mathbf{m}_{\text{sc}}$ *and* $\mathbf{m}_{\text{ec}}$ *and corresponding* $\Delta_r\text{EC}$ *values from randomly selected trials* ($\alpha = 8$)

N	$\mathbf{m}_{\text{sc}}$	$\Delta_r\text{EC}_{\text{sc}}(D)$	$\mathbf{m}_{\text{ec}}$	$\Delta_r\text{EC}_{\text{ec}}(D)$
Generative model: empty (no arcs)				
50	Empty	0.01	$X_1 \rightarrow X_4$	1.09
200	$X_1 \rightarrow X_4$	0.01	$X_3 \rightarrow X_1 \rightarrow X_4$	0.68
800	$X_3 \rightarrow X_1 \leftarrow X_4$	0.74	$X_3 \rightarrow X_1 \rightarrow X_4$	0.29
3200	Empty	0.00	Empty	0.10
Generative model: $X_1 \rightarrow X_2 \rightarrow X_3 \rightarrow X_4$				
50	$X_1 \rightarrow X_2 \rightarrow X_3 \rightarrow X_4$	0.11	$X_1 \rightarrow X_2 \rightarrow X_3 \rightarrow X_4$, $X_1 \rightarrow X_3$	0.07
200	$X_1 \rightarrow X_2 \rightarrow X_3 \rightarrow X_4$	0.05	$X_1 \rightarrow X_2 \rightarrow X_3 \rightarrow X_4$, $X_2 \rightarrow X_4$	0.00
800	$X_1 \rightarrow X_2 \rightarrow X_3 \rightarrow X_4$	0.00	$X_1 \rightarrow X_2 \rightarrow X_3 \rightarrow X_4$	0.00
3200	$X_1 \rightarrow X_2 \rightarrow X_3 \rightarrow X_4$	0.00	$X_1 \rightarrow X_2 \rightarrow X_3 \rightarrow X_4$	0.00
Generative model: $\{X_1, X_2, X_3\} \rightarrow X_4$				
50	$\{X_1, X_3, X_4\} \rightarrow X_2$, $X_4 \leftarrow X_1 \rightarrow X_3$	0.21	Complete	0.07
200	$\{X_1, X_2\} \rightarrow X_4$	0.00	$\{X_1, X_4\} \rightarrow X_2$, $X_3 \rightarrow X_4 \rightarrow X_1$	0.02
800	$\{X_1, X_2, X_3\} \rightarrow X_4$	0.00	$\{X_1, X_2, X_3\} \rightarrow X_4$	0.01
3200	$\{X_1, X_2, X_3\} \rightarrow X_4$	0.00	$\{X_1, X_2, X_3\} \rightarrow X_4$	0.00
Generative model: complete (no missing arcs)				
50	$X_1 \rightarrow X_2 \rightarrow X_3 \rightarrow X_4$	0.05	$X_1 \rightarrow X_2 \rightarrow X_3 \rightarrow X_4$, $X_1 \rightarrow X_3$	0.12
200	$X_1 \rightarrow X_2 \rightarrow X_3 \leftarrow X_4$	0.06	$X_1 \rightarrow X_2 \rightarrow X_3 \leftarrow X_4$, $X_2 \rightarrow X_4$	0.02
800	$\{X_1, X_2, X_3\} \rightarrow X_4$, $X_1 \rightarrow X_3 \rightarrow X_2$	0.07	$X_1 \rightarrow X_3 \rightarrow X_4 \rightarrow X_2$, $X_3 \rightarrow X_2$	0.01
3200	Complete	0.00	Complete	0.00

3. Experiments

Our goals are (1) to compare the accuracy of predictions based on the model average and models selected using the EC and SC criteria, and (2) to compare the models (structures) selected by the two criteria. To do so, we performed experiments based on both synthetic and real data.

In our experiments with synthetic data, we created several Bayesian networks (all with binary variables), and from them generated random data sets of various sizes. In all our experiments, we selected models using the two criteria, and compared the EC for both models with the maximum value for EC obtained by using the correct Bayesian prediction:

$$\mathrm{EC}_{\mathrm{opt}}(D) = \sum_{\mathbf{x}_{N+1}} p(\mathbf{x}_{N+1} \mid D) \log p(\mathbf{x}_{N+1} \mid D)$$

In particular, we computed

$$\Delta\mathrm{EC}_{\mathrm{ec}}(D) = \mathrm{EC}_{\mathrm{opt}}(D) - \mathrm{EC}(\mathbf{m}_{\mathrm{ec}}, D)$$

$$\Delta\mathrm{EC}_{\mathrm{sc}}(D) = \mathrm{EC}(\mathbf{m}_{\mathrm{ec}}, D) - \mathrm{EC}(\mathbf{m}_{\mathrm{sc}}, D)$$

where $\mathbf{m}_{\mathrm{sc}}$ and $\mathbf{m}_{\mathrm{ec}}$ were the models selected by SC and EC, respectively. Note that both differences are non-negative for any D. Because it was difficult to compare ΔEC values for different generative models, different priors, and different sample sizes,

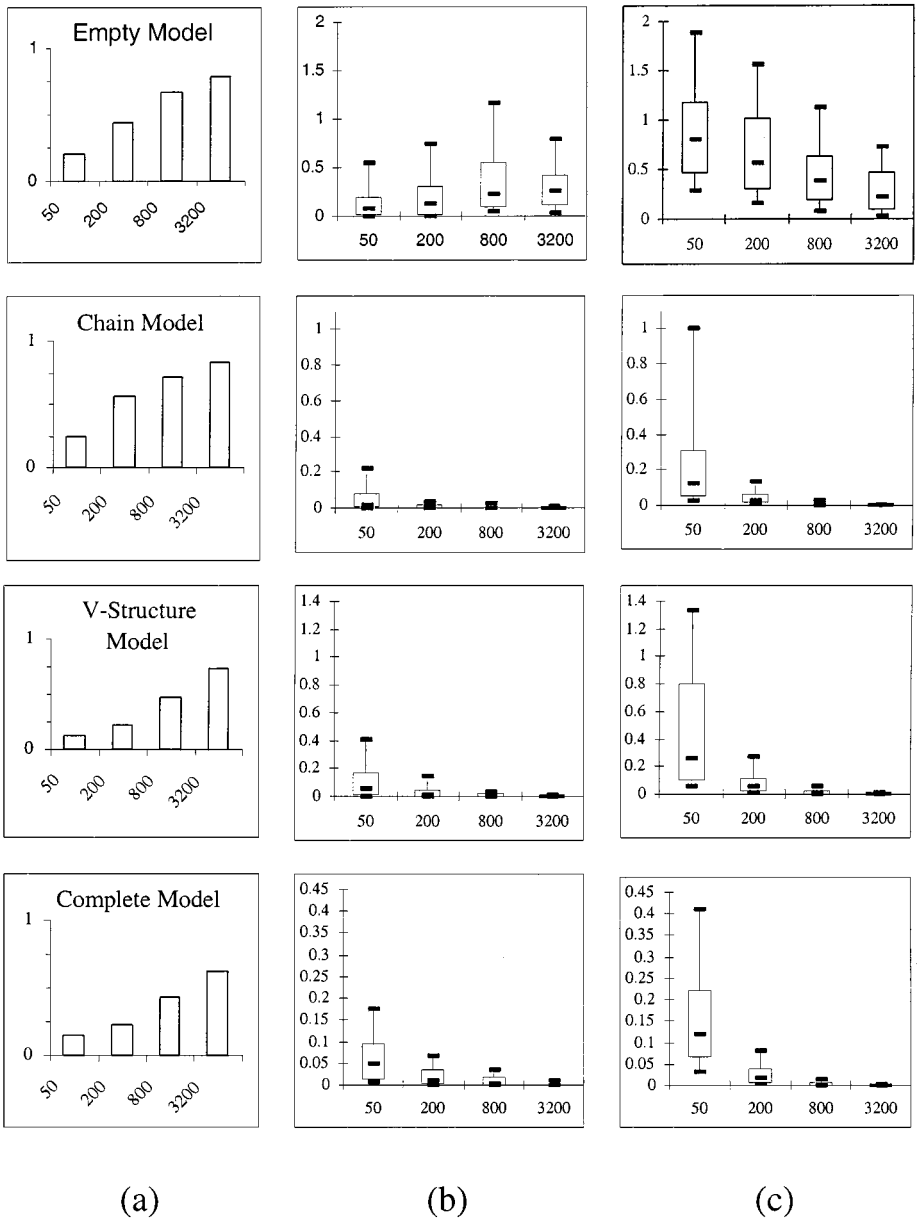

Fig. 1. *Sensitivity to generative model and sample size. (a) Fraction of trials in which* $\mathbf{m}_{\mathrm{sc}}$ *and* $\mathbf{m}_{\mathrm{ec}}$ *are equivalent. (b) Box plots of* $\Delta_r \mathrm{EC}_{\mathrm{sc}}(D)$ *for the non-equivalent models. (c) Box plots of* $\Delta_r \mathrm{EC}_{\mathrm{ec}}(D)$ *for all models*

we compared only the relative differences

$$\Delta_r \mathrm{EC}_{\mathrm{ec}}(D) = \frac{\Delta \mathrm{EC}_{\mathrm{ec}}(D)}{\mathrm{sd(EC, D)}}$$

$$\Delta_r \mathrm{EC}_{\mathrm{sc}}(D) = \frac{\Delta \mathrm{EC}_{\mathrm{sc}}(D)}{\mathrm{sd(EC, D)}}$$

where sd(EC,D) is the (equal-weight) standard deviation of EC($\mathbf{m}$, D) over all models. Also, because the number of possible Bayesian-network structures for n variables is more than exponential in n, we performed our experiments only for small n ($n = 3, \ldots, 6$).

We performed three experiments with synthetic data. In our first experiment, we examined the effect of sample size and generative network structure on predictive performance, while fixing priors and the number of variables ($n = 4$). We selected several generative Bayesian-network models of varying complexity: (1) the empty model, containing no arcs, (2) the Markov-chain model $X_1 \rightarrow X_2 \rightarrow X_3 \rightarrow X_4$, (3) a multiple *v-structure* model $\{X_1, X_2, X_3\} \rightarrow X_4$, and (4) the complete model for the ordering (X_1, X_2, X_3, X_4). For each generative model and for each of four sample sizes ranging from $N = 50$ to 3200, we ran a series of 100 trials. In each trial, we first sampled the parameters for the generative model from a uniform distribution, then generated a random data set of the appropriate size, and next identified the best model under both criteria. To compute the criteria for a given model and data set, we used uniform priors for network structure and Dirichlet parameter priors with $\alpha_{ijk} = 8/r_i q_i$ for all i, j, and k. Note that, given our model and parameter priors, the scientific (and engineering) criteria for two Markov equivalent structures are equal (e.g., Heckerman *et al.* (1995)). Thus, each criterion selects an equivalence class of models.

Table 1 shows, for each generative model and for each sample size, the two models selected and the corresponding relative scores for a single (randomly selected) trial. In Fig. 1, we use histograms and box plots to summarize the results from the 100 trials. Figure 1(a) shows the fraction of trials in which the models selected using the two criteria are equivalent. Figure 1(b) shows a box plot of $\Delta_r \mathrm{EC}_{\mathrm{sc}}$ for all trials in which the two criteria do *not* yield equivalent models. Figure 1(c) shows a box plot of $\Delta_r \mathrm{EC}_{\mathrm{ec}}$ for all trials. In all box plots, the top and bottom whiskers extend to the 90 and 10 percentiles, respectively.

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

Table 2. *Sensitivity to generative model and equivalent sample size α of the parameter priors. Models $\mathbf{m}_{\mathrm{sc}}$ and $\mathbf{m}_{\mathrm{ec}}$ and corresponding Δ_rEC values from randomly selected trials* ($\alpha = 8$)

α	$\mathbf{m}_{\mathrm{sc}}$	$\Delta_r \mathrm{EC}_{\mathrm{sc}}(D)$	$\mathbf{m}_{\mathrm{ec}}$	$\Delta_r \mathrm{EC}_{\mathrm{ec}}(D)$
Generative model: empty, $N = 50$, $n = 4$				
2	Empty	0.00	$X_3 \rightarrow X_4$	0.30
8	Empty	0.09	$X_3 \rightarrow X_1 \leftarrow X_2$	1.51
32	$X_1 \rightarrow X_2 \rightarrow X_3$	0.29	$X_1 \rightarrow X_2 \rightarrow X_3 \rightarrow X_4$	3.38
128	$X_1 \rightarrow X_2 \rightarrow X_3 \rightarrow X_4$	0.00	$X_1 \rightarrow X_2 \rightarrow X_3 \rightarrow X_4$	1.89
Generative model: $X_1 \rightarrow X_2 \rightarrow X_3 \rightarrow X_4$, $N = 50$, $n = 4$				
2	$X_3 \rightarrow X_4$	0.00	$X_3 \rightarrow X_4$	0.56
8	$X_3 \rightarrow X_4$	0.09	$X_1 \rightarrow X_2 \rightarrow X_4 \rightarrow X_3$	0.54
32	$X_3 \rightarrow X_4$	0.08	$X_1 \rightarrow X_2 \rightarrow X_4 \rightarrow X_3$	0.52
128	$X_3 \rightarrow X_4$	0.26	$X_1 \rightarrow X_2 \rightarrow X_3 \rightarrow X_4$, $X_1 \rightarrow X_3$	0.92
Generative model: $\{X_1, X_2, X_3\} \rightarrow X_4$, $N = 50$, $n = 4$				
2	Empty	0.00	Empty	0.22
8	Empty	0.07	$X_1 \rightarrow X_2 \leftarrow X_3$	0.75
32	$X_1 \rightarrow X_2 \rightarrow X_3$	0.14	$X_1 \rightarrow X_2 \rightarrow X_3 \rightarrow X_4$, $X_1 \rightarrow X_3, X_2 \rightarrow X_4$	1.39
128	$X_1 \rightarrow X_2 \rightarrow X_3$	0.25	$X_1 \rightarrow X_2 \rightarrow X_3 \rightarrow X_4$, $X_2 \rightarrow X_4$	1.54
Generative model: complete $N = 50$, $n = 4$				
2	$X_1 \rightarrow X_3 \rightarrow X_4$	0.00	$X_1 \rightarrow X_3 \rightarrow X_4$	0.35
8	$X_1 \rightarrow X_3 \rightarrow X_4$	0.08	$X_2 \rightarrow X_3 \rightarrow X_4$ $\{X_3, X_4\} \rightarrow X_1$	0.41
32	$X_1 \rightarrow X_3 \rightarrow X_4 \leftarrow X_2$ $X_2 \rightarrow X_3$	0.12	$X_2 \rightarrow X_3 \rightarrow X_2 \leftarrow X_4$ $\{X_2, X_3\} \rightarrow X_4$	0.55
128	$X_1 \rightarrow X_3 \leftarrow X_2 \rightarrow X_4$ $X_3 \rightarrow X_4$	0.18	$X_1 \leftarrow X_3 \leftarrow X_2 \rightarrow X_4$ $X_3 \rightarrow X_4 \rightarrow X_1$	1.08

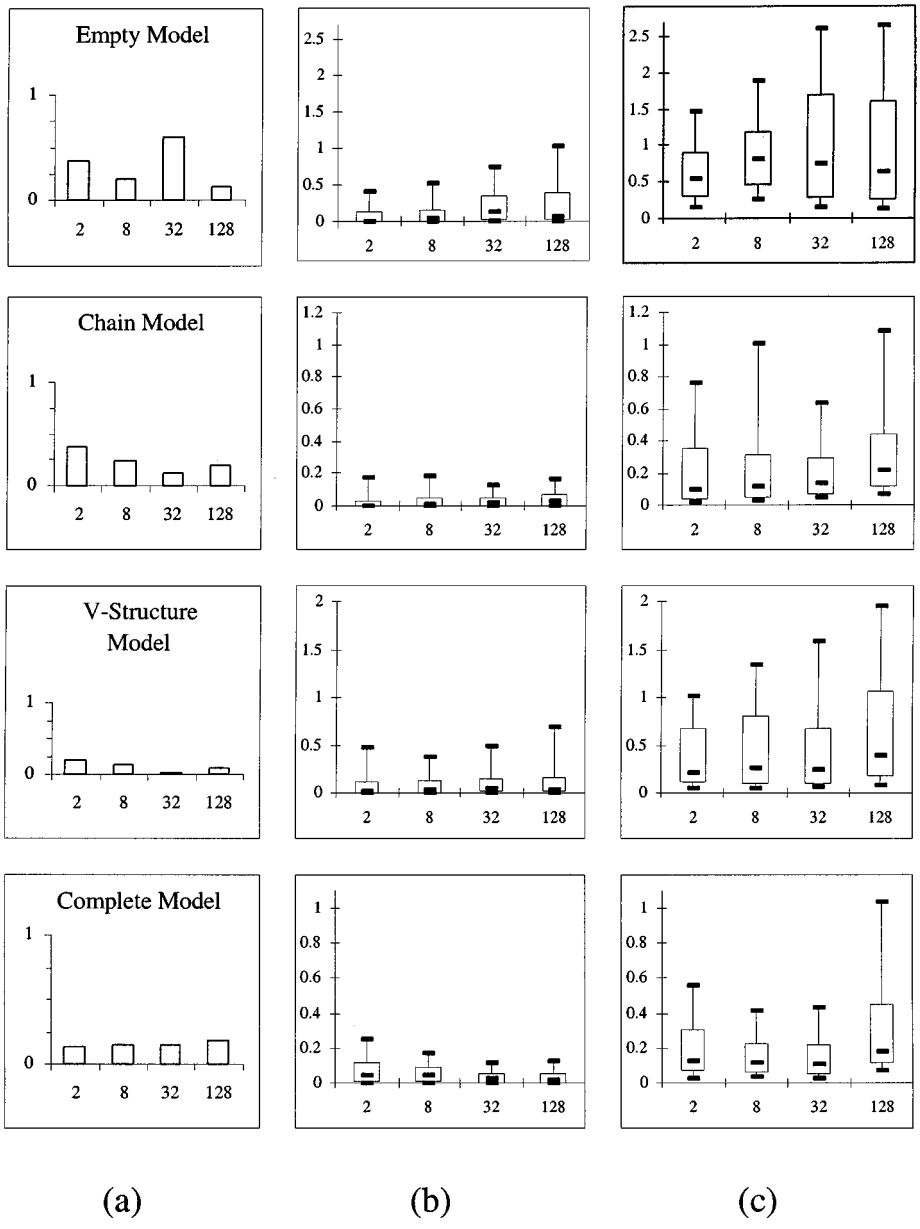

Fig. 2. *Sensitivity to generative model and α. (a) Fraction of trials in which* $\mathbf{m}_{sc}$ *and* $\mathbf{m}_{ec}$ *are equivalent. (b) Box plots of* $\Delta_r \mathrm{EC}_{sc}(D)$ *for the non-equivalent models. (c) Box plots of* $\Delta_r \mathrm{EC}_{ec}(D)$ *for all models*

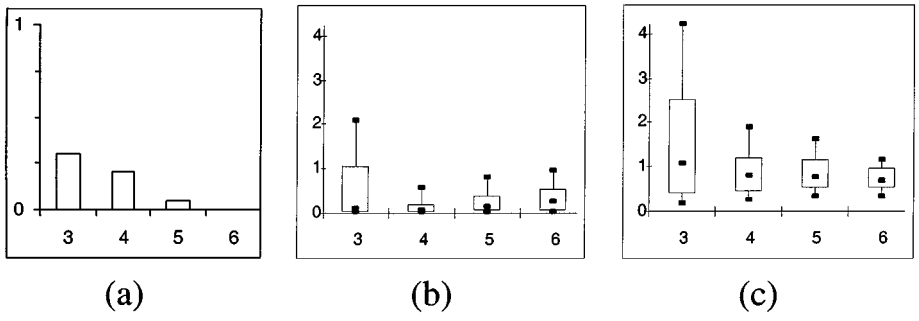

Fig. 3. *Sensitivity to number of variables n. (a) Fraction of trials in which* $\mathbf{m}_{sc}$ *and* $\mathbf{m}_{ec}$ *are equivalent. (b) Box plots of* $\Delta_r \mathrm{EC}_{sc}(D)$ *for the non-equivalent models. (c) Box plots of* $\Delta_r \mathrm{EC}_{ec}(D)$ *for all models*

explanation for this difference is that, when using EC, we reward a prediction based on all N observations. In contrast, when using SC, we reward predictions based on $0, 1, 2, \ldots, N-1$ 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.

In our second experiment, we investigated the sensitivity of model selection to parameter priors. We proceeded as in the first experiment, except that we used priors $\alpha_{ijk} = \alpha/r_i q_i$ for various values of the equivalent sample size α. Also, we used a single sample size $N = 50$. Example models and scores from a single trial are shown in Table 2. Summaries for the 100 trials are given in Fig. 2. We see that, for each generative model, the predictive scores for the models selected by each criterion are fairly insensitive to variations in α.

In the third experiment, we examined the effects of domain size (n) on model selection. For $n = 3$, 4, 5, and 6, we created a generative model from the empty network structure with parameters sampled from the uniform distribution. For each generative model, we ran 100 trials as before, using $\alpha = 8$ and $N = 50$ for all trials. Figure 3 summarizes the results. We see that the median relative performance of the model average with respect to the single model selected by the EC criterion is fairly insenstive to n. In contrast, the chances that the models selected by the two criteria are equivalent decrease with increasing n. This latter finding is reasonable – as n increases, the number of possible models increases dramatically.

In our experiment with real data, we examined the data set of Sewell and Shah (1968), who investigated factors that influence the intention of high school students to attend college. They measured the following variables for 10,318 Wisconsin high school seniors: *Sex* (SEX): male, female; *Socioeconomic Status* (SES): low, lower middle, upper middle, high; *Intelligence Quotient* (IQ): low, lower middle, upper middle, high; *Parental Encouragement* (PE): low, high; and *College Plans* (CP): yes, no. For this experiment, we restricted the possible models as follows: neither SEX nor SES were allowed to have any parents, and CP was not allowed to have children. We used a uniform model prior over all models that were deemed possible.

For this data set, the two criteria yielded the same model, which is shown in Fig. 4, and the relative score $\Delta_r \mathrm{EC}_{\mathrm{ec}}(D)$ for the model was small. We tested the sensitivity to sample size for this data set by selecting 100 random subsets of the data for each of four different sample sizes. Figure 5 summarizes the results. Note that, for this domain, the two criterion usually result in the same model, even with a sample size as small as 50.

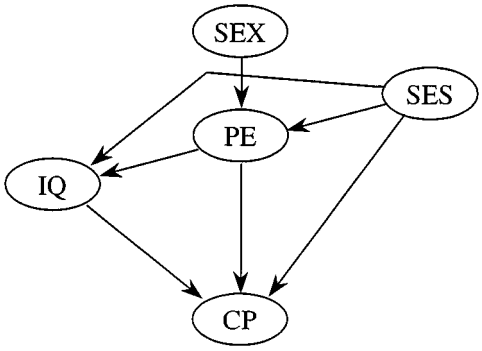

Fig. 4. *Model selected using both* SC *and* EC *criteria for the Sewell and Shah (1968) data*

4. Conclusions

Our results confirm the conclusions of Draper (1993) and Madigan *et al.* (1996) that model averaging produces substantially better predictions than does model selection using SC when sample sizes are small. Thus, when we are interested in prediction, we should use model averaging (or perhaps a Monte-Carlo approximation to model averaging) rather than model selection whenever feasible.

In situations where there is a good reason to select a single model – for example, a desire to understand the domain – we have found that the EC and SC criterion may produce substantially different models, again when sample sizes are small. Consequently, we should be careful to appreciate this difference and use the appropriate criterion.

Finally, the EC criterion is computationally infeasible in all but trivial situations. Monte-Carlo approximations for both the

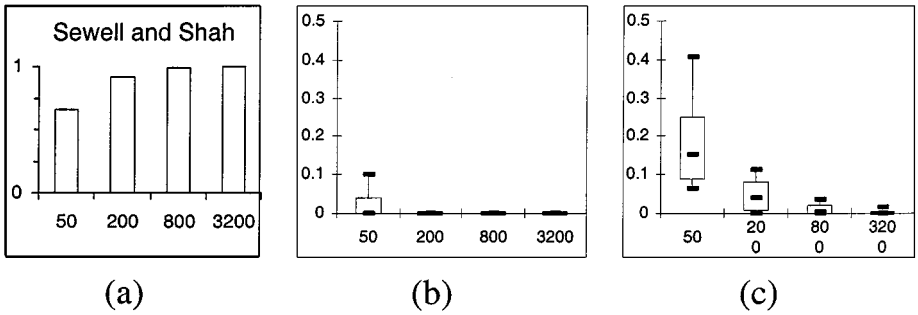

Fig. 5. *Sensitivity to sample size for Sewell and Shah (1968) data. (a) Fraction of trials in which* $\mathbf{m}_{\mathrm{sc}}$ *and* $\mathbf{m}_{\mathrm{ec}}$ *are equivalent. (b) Box plots of* $\Delta_r \mathrm{EC}_{\mathrm{sc}}(D)$ *for the non-equivalent models. (c) Box plots of* $\Delta_r \mathrm{EC}_{\mathrm{ec}}(D)$ *for all models*

average over all possible models in equation 1 and the sum over possible states of $\mathbf{X}_{N+1}$ in equation 2 offer promise, but work is needed to determine the quality of these approximations.

Acknowledgments

We thank Wray Buntine, Dan Geiger, and Chris Meek for useful discussions.

Notes

1. A technical point worth mentioning is our use of the term variable and its relationship to the standard definition of a random variable. A *discrete random variable* is a function $X : \Omega \rightarrow E$ where E is a discrete set such that $\{\omega \mid X(\omega) = x\} \in \mathcal{A}$ for every $x \in E$ where $\mathcal{A}$ is a σ-field and Ω is a sample space of a probability space $(\Omega, \mathcal{A}, P)$. We use the term discrete *variable*, as common to much of the literature on graphical models, to mean a function $X_i : \Omega \rightarrow E_i$, parallel to the usual definition of a random variable, but without fixing a specific probability measure P. A model $\mathbf{m}$ for a set of discrete variables $\mathbf{X}$ is simply a set of probability measures on the Cartesian product $\times_i E_i$. Once a particular probability measure from $\mathbf{m}$ is picked, a variable in our sense becomes a random variable in the usual sense. A similar comment applies to sets $\mathbf{X}$ that include continuous variables.
2. In the nomenclature of Bernardo and Smith (1994), we take the *M-closed* view.
3. An axiomatic characterization of this proper scoring rule is given by Bernardo (1979).
4. Sometimes, 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).

References

Bernardo J. 1979. Expected information as expected utility. Annals of Statistics 7: 686–690.

Bernardo J. and Smith A. 1994. Bayesian Theory. New York, John Wiley and Sons.

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. 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.

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

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.

Sewell W. and Shah V. 1968. Social class, parental encouragement, and educational aspirations. American Journal of Sociology 73: 559–572.

Spirtes P., Glymour C., and Scheines R. 1993. Causation, Prediction, and Search. New York, Springer-Verlag.