A Comparison of Scientific and Engineering Criteria for Bayesian Model Selection

David Heckerman and David Maxwell Chickering
Microsoft Research
Redmond WA 98052-6399
heckerma@microsoft.com, dmax@microsoft.com

Abstract

Given a set of possible model structures for variables $\mathbf{X}$ and a set of possible parameters for each structure, the Bayesian "estimate" of the probability distribution for $\mathbf{X}$ given observed data is obtained by averaging over the possible model structures and their parameters. An often-used approximation for this estimate is obtained by selecting a single model structure 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, differences in predictive performance between the model-averaged estimate and EC and between EC and SC can be substantial.

1 Introduction

Suppose that the joint probability distribution over a set of random variables $\mathbf{X} = \{X_1, \ldots, X_n\}$ 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 $D = \{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$ 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 $p(\Theta_m|\mathbf{m})$. We assume that $p(\Theta_m|\mathbf{m})$ is a probability density function. Given random sample D, we compute the posterior distributions for each $\mathbf{M}$ and 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

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

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

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

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 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

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

The term $p(\mathbf{x}_l|\mathbf{x}_1, \ldots, \mathbf{x}_{l-1}, \mathbf{m})$ 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 $\log p(\mathbf{x})$.[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 $\mathbf{x}_{N+1}$, we average this utility over all possible observations, obtaining the following criterion for model $\mathbf{m}$ given data D:

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

where $p(\mathbf{x}_{N+1}|D)$ 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 $p(\mathbf{x}_{N+1}|D)$ and the posterior distribution determined by model $\mathbf{m}$.

When we substitute $p(\mathbf{x}_{N+1}|D)$ for $p(\mathbf{x}_{N+1}|D, \mathbf{m})$ 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 probability of the model $\mathbf{m}$ that is closest to truth (in the KL sense) will approach one,[2] and we obtain $p(\mathbf{x}_{N+1}|D) = p(\mathbf{x}_{N+1}|D, \mathbf{m})$. Consequently, in this limit, the estimates produced by model averaging and by model selection using the two criterion coincide.

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 to 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.

In contrast, if we are interested in predicting the next observation, then by definition, the model-averaged estimate is at least as good as that produced by EC, which in turn is at least as good as the estimate produced by SC. Thus, an important question arises: How much do we loose by using EC instead of model averaging, or by using SC instead of EC? When N is large, we loose nothing, as we have discussed. But what happens for small N? In this paper, we investigate this question in the context of Bayesian-network models for discrete variables.

We note that model selection using EC is more expensive than model averaging, because the former computation requires that we first determine the model-averaged estimate $p(\mathbf{x}_{N+1}|D)$. Therefore, at first glance, there appears to be no reason to investigate the predictive performance of the EC estimate. Nonetheless, if we find that EC significantly outperforms SC, then we have a reason to look for a more efficient criterion that approximates EC.

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

[2] When there is more than one such model, our conclusions still hold, although the argument is more detailed.

2 Bayesian Networks

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

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

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

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 factors according to S.[4] 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 $\boldsymbol{\theta}_s = (\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_n)$ 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 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 $p(x_i|\mathbf{pa}_i, \boldsymbol{\theta}_i, \mathbf{m}_s)$ 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

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

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

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

(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 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 model averaging, EC, and SC. To do so, we created several Bayesian networks, and from them generated random data sets of various sizes. We then 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:

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

In particular, we computed

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

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

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

[4] We use the causal interpretation of Bayesian-network structure so that different structures correspond to mutually exclusive events. Heckerman et al. (1994) describe an acausal interpretation that partitions models into mutually exclusive equivalence classes.

where $\mathbf{m}_{\text{sc}}$ and $\mathbf{m}_{\text{ec}}$ were the structures selected by SC and EC, respectively. Note that both differences are non-negative for any D. Because it was difficult to compare values for ΔEC across generative models, we also computed the relative differences

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

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

where sd(EC,D) is the (equal-weight) standard deviation of $\text{EC}(\mathbf{m}, D)$ 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 $(n = 2, \ldots, 6)$.

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 network structures of varying complexity: (1) the empty graph, containing no arcs, (2) the Markov chain $X_1 \rightarrow X_2 \rightarrow X_3 \rightarrow X_4$, and (3) the complete graph for the ordering (X_1, X_2, X_3, X_4). We then sampled the parameters of each graph from a uniform distribution. Next, from each of these generative models, we sampled data at random with sample sizes ranging from $N = 50$ to 3200. To compute the criteria for a given model, 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. Results are shown in Table 1. Note that, given our structure 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 structures. In the table, we report a representative directed acyclic graph from each selected class.

The results 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 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 only the empty generative structure. Results are shown in Table 2. We see that the models selected and predictive scores are sensitive only to large variations in α.

In the third experiment, we examined the effects of domain size (n) on model selection. For each n, we created a generative model from the empty network structure with parameters sampled from the uniform distribution. For all trials, we used $\alpha = 8$ and $N = 50$. Whereas conclusions from the previous experiments were not sensitive to the randomness in the parameter values and the data, results in this experiment were sensitive. Thus, for each n, we computed ΔEC and Δ_rEC for eight parameter–data sets. Table 3 summarizes the results. We see that absolute differences in predictive performance grow with domain size for a fixed sample size, but that relative differences are fairly insensitive to sample size. Also, $\Delta\text{EC}_{\text{ec}}(D)$ is typically larger than $\Delta\text{EC}_{\text{sc}}(D)$, and $\Delta_r\text{EC}_{\text{ec}}(D)$ is typically larger than $\Delta_r\text{EC}_{\text{sc}}(D)$.

Overall, our results confirm the conclusions of Draper (1993) and Madigan et al. (1996) that model averaging sometimes produces substantially better predictions than does model selection using SC. In addition, we see that, when using model selection to choose a predictive model, EC can perform significantly better than SC. As we have discussed, EC is not practical for model selection. Nonetheless, our observations suggest that if (1) we want to predict the next observation, (2) the sample size is not in the asymptotic regime, and (3) model averaging is not practical, then we should look for a model-selection criterion other than SC as an approximation for EC.

Acknowledgments

We thank Wray Buntine and Chris Meek for useful discussions.

References

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.

Table 1: Network structures $\mathbf{m}_{\text{sc}}$ and $\mathbf{m}_{\text{ec}}$ and corresponding predictive performance for various generative network structures and sample sizes N.

generative structure: empty (no arcs), $n = 4$					
N	$\mathbf{m}_{\text{sc}}$	$\Delta\text{EC}_{\text{sc}}(D)$	$\mathbf{m}_{\text{ec}}$	$\Delta\text{EC}_{\text{ec}}(D)$	-EC_{opt}
50	empty	0.00006 (0.01)	$X_1 \to X_4$	0.00644 (1.09)	2.43526
200	$X_1 \to X_4$	0.00001 (0.01)	$X_3 \to X_1 \to X_4$	0.00161 (0.68)	2.16559
800	$X_3 \to X_1 \leftarrow X_4$	0.00094 (0.74)	$X_3 \to X_1 \to X_4$	0.00036 (0.29)	2.09239
3200	empty	0 (0)	empty	0.00000 (0.10)	2.08580

generative structure: $X_1 \to X_2 \to X_3 \to X_4$, $n = 4$					
N	$\mathbf{m}_{\text{sc}}$	$\Delta\text{EC}_{\text{sc}}(D)$	$\mathbf{m}_{\text{ec}}$	$\Delta\text{EC}_{\text{ec}}(D)$	-EC_{opt}
50	$X_1 \to X_2 \to X_3 \to X_4$	0.00708 (0.11)	$X_1 \to X_2 \to X_3 \to X_4$, $X_1 \to X_3, X_3 \to X_4$	0.00472 (0.07)	2.01433
200	$X_1 \to X_2 \to X_3 \to X_4$	0.00358 (0.05)	$X_1 \to X_2 \to X_3 \to X_4$, $X_2 \to X_4$	0.00072 (0.00)	1.43558
800	$X_1 \to X_2 \to X_3 \to X_4$	0 (0)	$X_1 \to X_2 \to X_3 \to X_4$	0.00023 (0.00)	1.43034
3200	$X_1 \to X_2 \to X_3 \to X_4$	0 (0)	$X_1 \to X_2 \to X_3 \to X_4$	0.00001 (0.00)	1.34461

generative structure: complete (no missing arcs), $n = 4$					
N	$\mathbf{m}_{\text{sc}}$	$\Delta\text{EC}_{\text{sc}}(D)$	$\mathbf{m}_{\text{ec}}$	$\Delta\text{EC}_{\text{ec}}(D)$	-EC_{opt}
50	$X_1 \to X_2 \to X_3 \to X_4$	0.00325 (0.05)	$X_1 \to X_2 \to X_3 \to X_4$, $X_1 \to X_3$	0.00726 (0.12)	2.28003
200	$X_1 \to X_2 \to X_3 \leftarrow X_4$	0.00308 (0.06)	$X_1 \to X_2 \to X_3 \leftarrow X_4$, $X_2 \to X_4$	0.00125 (0.02)	2.23031
800	$X_1 \to X_3 \to X_2$, $\{X_1, X_2, X_3\} \to X_4$	0.00387 (0.07)	$X_1 \to X_3 \to X_4 \to X_2$, $X_3 \to X_2$	0.00080 (0.01)	2.19934
3200	complete	0 (0)	complete	0.00003 (0.00)	2.15846

Table 2: Sensitivity to the equivalent sample size α of the parameter priors.

generative structure: empty, $N = 50, n = 4$					
α	$\mathbf{m}_{\mathrm{sc}}$	$\Delta\mathrm{EC}_{\mathrm{sc}}(D)$	$\mathbf{m}_{\mathrm{ec}}$	$\Delta\mathrm{EC}_{\mathrm{ec}}(D)$	-$\mathrm{EC}_{\mathrm{opt}}$
2	empty	0 (0)	empty	0.00425 (0.31)	2.33847
8	empty	0.00006 (0.01)	$X_1 \rightarrow X_4$	0.00644 (1.09)	2.43526
32	empty	0.00186 (1.05)	$X_2 \rightarrow X_3 \rightarrow X_4 \rightarrow X_1$	0.00312 (1.75)	2.61296
128	$X_2 \rightarrow X_3 \leftarrow X_4$	0.00050 (1.31)	$X_1 \rightarrow X_2 \rightarrow X_3 \rightarrow X_4$	0.00073 (1.90)	2.73972

generative structure: empty, $N = 200, n = 4$					
α	$\mathbf{m}_{\mathrm{sc}}$	$\Delta\mathrm{EC}_{\mathrm{sc}}(D)$	$\mathbf{m}_{\mathrm{ec}}$	$\Delta\mathrm{EC}_{\mathrm{ec}}(D)$	-$\mathrm{EC}_{\mathrm{opt}}$
2	empty	0.00024 (0.06)	$X_1 \rightarrow X_4,$	0.00147 (0.35)	2.11330
8	$X_1 \rightarrow X_4$	0.00001 (0.01)	$X_3 \rightarrow X_1 \rightarrow X_4$	0.00161 (0.68)	2.16559
32	$X_1 \rightarrow X_4$	0.00029 (0.16)	$X_3 \rightarrow X_1 \rightarrow X_4 \rightarrow X_2$	0.00137 (0.75)	2.31139
128	$X_2 \rightarrow X_4 \rightarrow X_1$	0.00005 (0.03)	$X_3 \rightarrow X_1 \rightarrow X_4 \rightarrow X_2$	0.00062 (0.24)	2.55624

generative structure: empty, $N = 800, n = 4$					
α	$\mathbf{m}_{\mathrm{sc}}$	$\Delta\mathrm{EC}_{\mathrm{sc}}(D)$	$\mathbf{m}_{\mathrm{ec}}$	$\Delta\mathrm{EC}_{\mathrm{ec}}(D)$	-$\mathrm{EC}_{\mathrm{opt}}$
2	empty	0.00045 (0.16)	$X_1 \rightarrow X_3$	0.00044 (0.16)	2.07698
8	$X_3 \rightarrow X_1 \leftarrow X_4$	0.00094 (0.74)	$X_3 \rightarrow X_1 \rightarrow X_4$	0.00036 (0.29)	2.09239
32	$X_3 \rightarrow X_1 \leftarrow X_4$	0.00031 (0.43)	$X_3 \rightarrow X_1 \rightarrow X_4$	0.00024 (0.35)	2.14696
128	$X_3 \rightarrow X_1 \rightarrow X_4$	0 (0)	$X_3 \rightarrow X_1 \rightarrow X_4$	0.00019 (0.09)	2.29627

Table 3: Sensitivity to number of variables n. Numbers shown are mean ± s.d. over eight trials.

generative structure: empty, $N = 50$, $\alpha = 8$				
n	$\Delta\mathrm{EC}_{\mathrm{sc}}(D) \times 10^3$	$\Delta\mathrm{EC}_{\mathrm{ec}}(D) \times 10^3$	$\Delta_r\mathrm{EC}_{\mathrm{sc}}(D)$	$\Delta_r\mathrm{EC}_{\mathrm{ec}}(D)$
2	0.91 ± 1.45	0.88 ± 0.72	0.80 ± 1.10	1.50 ± 1.41
3	1.13 ± 1.22	2.65 ± 1.36	0.52 ± 0.78	0.91 ± 0.78
4	1.06 ± 1.04	6.30 ± 3.23	0.37 ± 0.71	1.06 ± 0.58
5	3.60 ± 6.56	13.1 ± 8.06	0.23 ± 0.36	0.88 ± 0.74
6	8.16 ± 8.76	17.0 ± 5.68	0.36 ± 0.41	0.71 ± 0.33

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.

Heckerman, D., Geiger, D., and Chickering, D. (1994). Learning Bayesian networks: The combination of knowledge and statistical data. In *Proceedings of Tenth Conference on Uncertainty in Artificial Intelligence,* Seattle, WA, pages 293–301. Morgan Kaufmann.

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.

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