Appendix to Discriminative Model Selection for Belief Net Structures

March 27, 2005

The short submission "Discriminative Model Selection for Belief Net Structures" describes the challenge of learning a belief net structure that will produce a good classifier. Unfortunately, due to space limitations, it could only include some of our empirical data. This appendix provides a more comprehensive set of results.

In general, a model selection criterion $c_\chi(G,\ D)$ provides a score for each proposed belief net structure G given a database of complete instances D.

We compared the performance of 7 such model selection criteria $c_\chi(G,\ D)$

- 3 generative: AIC, MDL, BDe
- 4 discriminative: CAIC, CMDL, CE, BV

using several different procedures for estimating the parameters

- generatively (OFE) vs discriminatively (ELR)
- undivided sample (1S) vs divided sample (5CV)

across a number of different situations

- different "generative Markov Blanket" complexity
- different training sizes,

on both

- real world BNs, and
- synthetic BNs.

(Each of these terms are defined in the submission mentioned above.)

Appendix A provides a short discription of how we evaluate the quality of each criteria. Appendix B then provides preliminary explorations on a few "real world" belief nets. This motivated the type of experiments we performed on synthesized data. Appendix C explicitly compared the quality of the answer produced by each criteria in 20 different contexts — each formed by fixing a specific size of belief net structure and a particular way of instantiating the parameters. Finally, Appendix D compares the relative performances of the different ways to instantiate the parameters, for a fixed selection criteria. Here, we sequentially consider 3 different high-quality criteria.

A Evaluation Measure

For any instantiated structure G, let $err(\,G\,)$ be its "classification" (0/1) error on an unseen sample. (Here, most measures actually consider some specific instantiation of G's parameters; see below.) Over a set of structures $\mathcal{G}$, let

$$G^* \quad = \quad \operatorname*{argmin}_{G \in \mathcal{G}} err(\,G\,)$$

be the (instantiated) structure with the minimum error. Further, let

$$G^\chi \quad = \quad \operatorname*{argmin}_{G \in \mathcal{G}} c_\chi(G, D)$$

be the structure that c_χ thinks is best, wrt the datasample D.

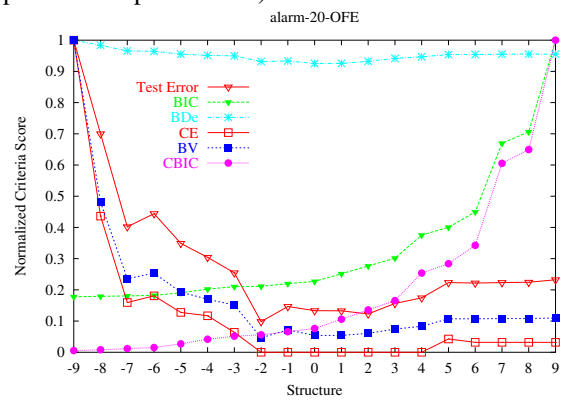
To measure the performance of the model selection criteria χ, we use *Relative Model Selection Error (RMSE)*, which is defined as the ratio of the test error of the selected structure over the minimum test error among the set of structures — *i.e.*,

$$RMSE(\chi) \quad = \quad \frac{err(\,G^\chi\,)}{err(\,G^*\,)}$$

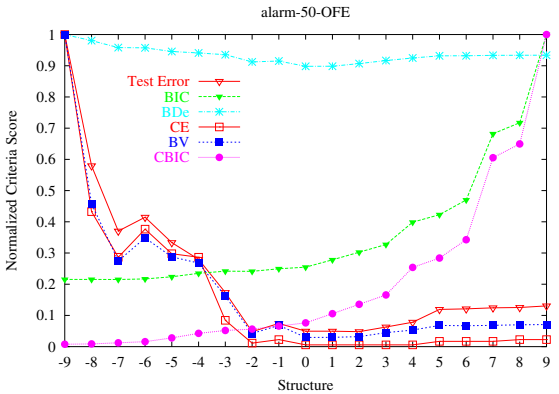
Each line on each graph below is the $RMSE(\chi)$ score for some criterion, as a function of the structure G considered, normalized to lie in the interval $[0, 1]$. We also include th "Test Error" score.

(This is based on a dataset of size $|S|$ drawn from the Alarm network, using the 1sample and OFE to produce the parameters.)

Alarm
$|S| = 20$
1S, OFE



Alarm
$|S| = 50$
1S, OFE



B Real World Belief Nets

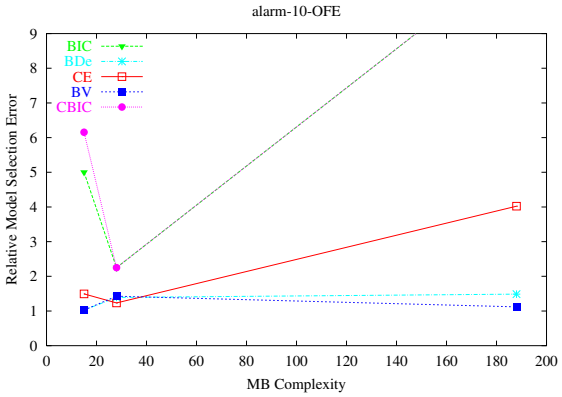
We first explored the performance of these measures on several real-world BNs — Alarm, Dog-Problem, . . . Those results motivated our current experimental designs.

For the Alarm BN, we picked 3 different variables as the class variable which implied 3 different Markov Blanket sizes

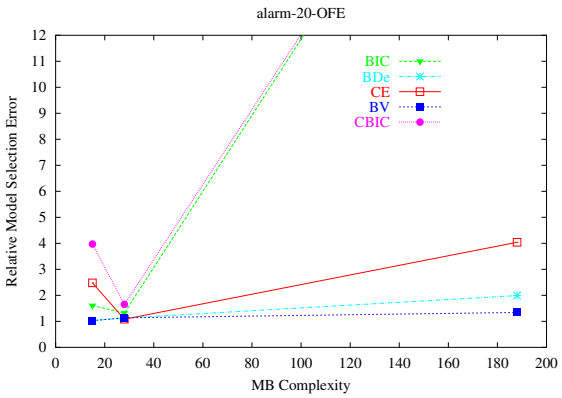
HREKG	15
Disconnect	28
VentLung	188

For each of these Class variables, we generate a structure sequence by adding or removing edges in the MarkovBlanket (MB) of the Class variable, and then evaluate the performance of the model selection criteria using this structure sequence. The results are. . .

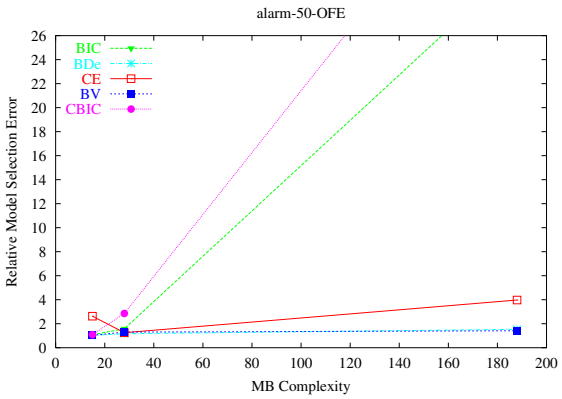
Alarm
$|S| = 10$
1S, OFE



Alarm
$|S| = 20$
1S, OFE

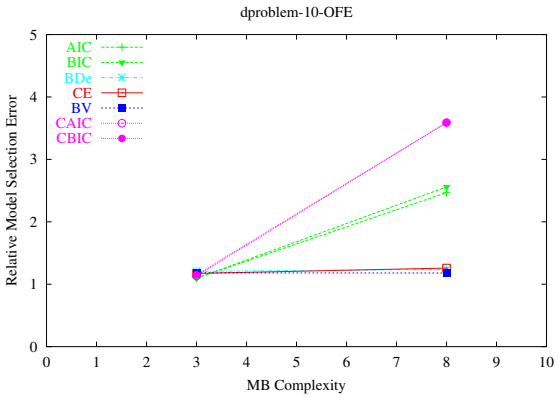


Alarm
$|S| = 50$
1S, OFE

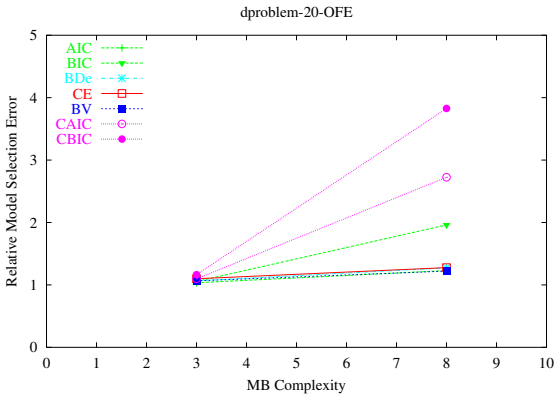


We performed similar experiments on the Dog-Problem BN, producing

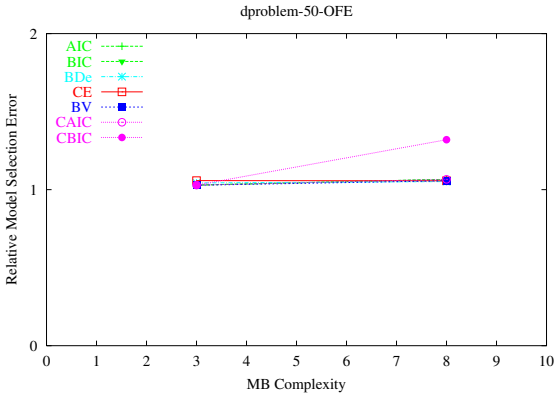
Dog-Problem
$|S| = 10$
1S, OFE



Dog-Problem
$|S| = 20$
1S, OFE



Dog-Problem
$|S| = 50$
1S, OFE



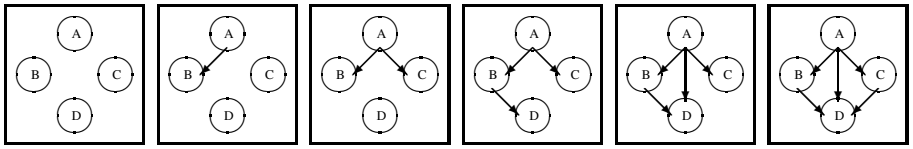


Figure 1: Sequence of structures; (d) is the original structure

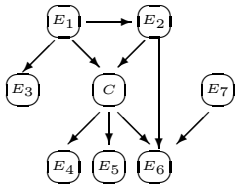


Figure 2: Example of a Belief Net Structure

C Experiments on Synthesized Data

These results suggest the relative performance of the criteria may change in different situations, so we further conducted some experiments on synthetic data by controlling the different situation factors.

Procedure:

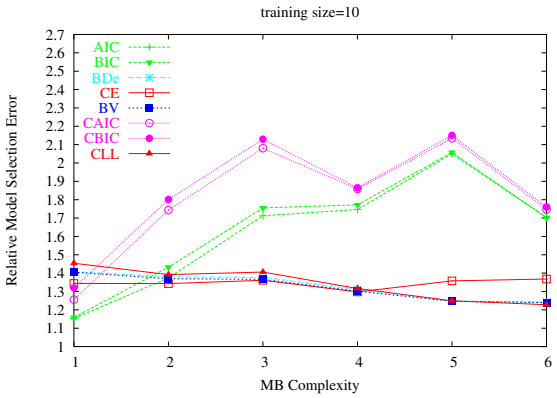
- Randomly Generate 6 groups of structures with increasing complexity (#edges or free parameters in the class Markov Blanket).
  [Here, we consider 7 and 15 variables.]
  Each group contains 30 structures that have the same complexity level.
- For each structure G, generate a sequence of structures to imitate model selection sequence by adding or deleting edges in the Markov Blanket.

  See Figure 1. (Although our actual graphs will be more complicated — see Figure 2.)
- For each sequence, for each sample size m, run the experiment 20 times:
  - Produce a "target" instantiation of G's parameters, Θ
  - Generate data sample of size m from $\langle G, \Theta \rangle$.
  - Ask each criteria to evualate each structure, based on this sample.

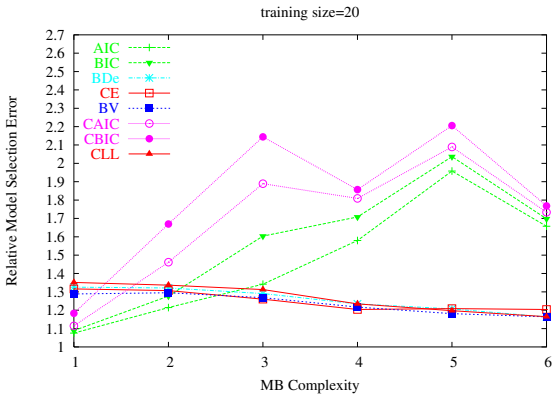
Each graph plots the 7 criteria shown above, and also plots the "Conditional Log Likelihood", CLL, of the data.

C.1 Results#A1: 7-Variable BNs, 1S, OFE

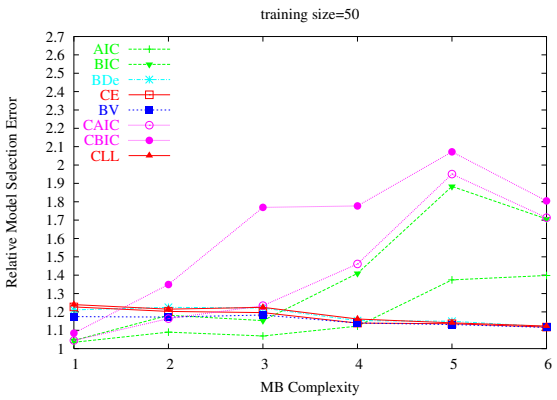
$|S| = 10$



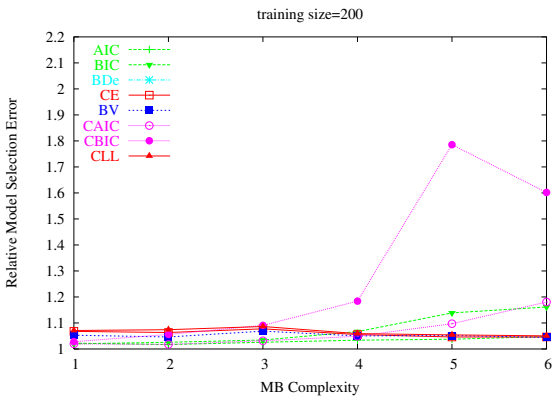
$|S| = 20$



$|S| = 50$

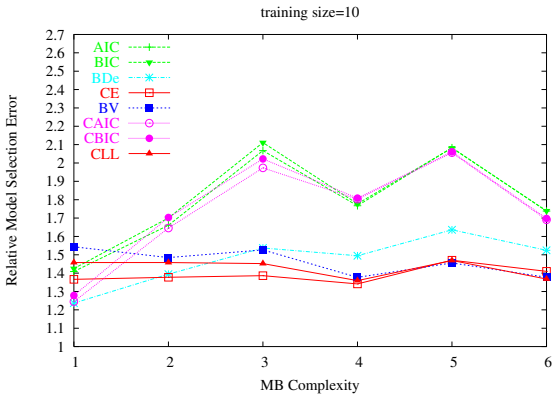


$|S| = 200$

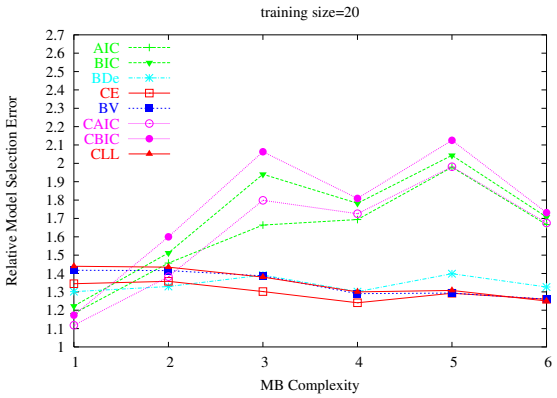


C.2 Results#A2: 7-Variable BNs, 1S, ELR

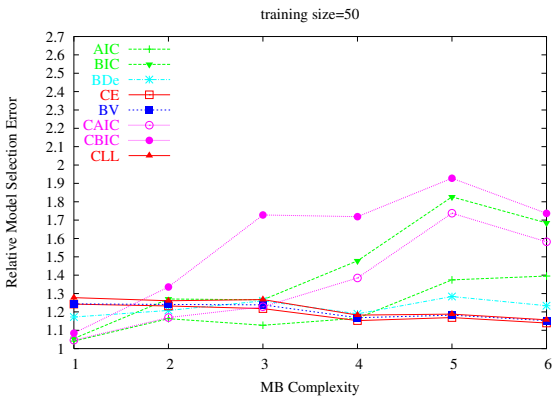
$|S| = 10$



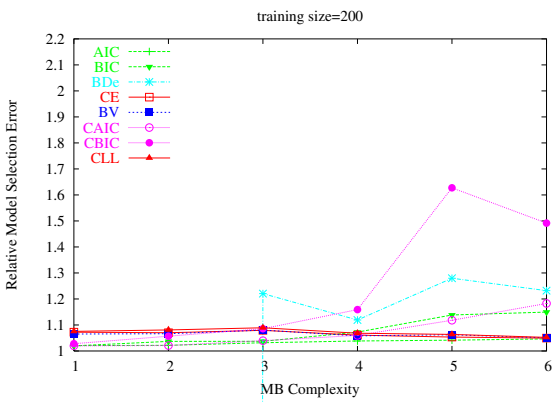
$|S| = 20$



$|S| = 50$

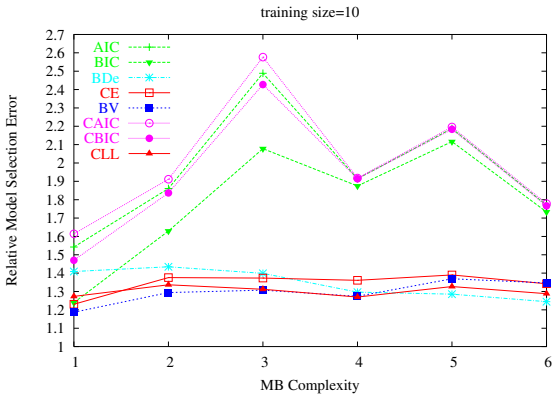


$|S| = 200$

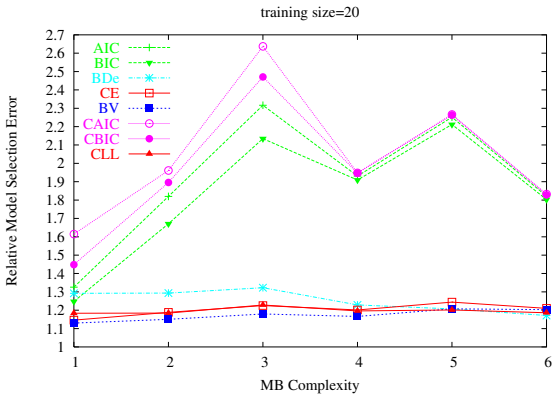


C.3 Results#A3: 7-Variable BNs, 5CV, OFE

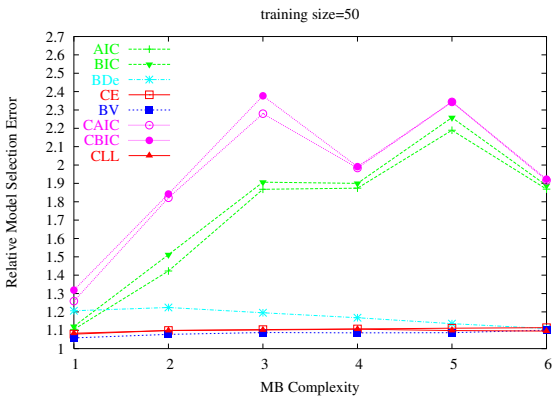
$|S| = 10$



$|S| = 20$

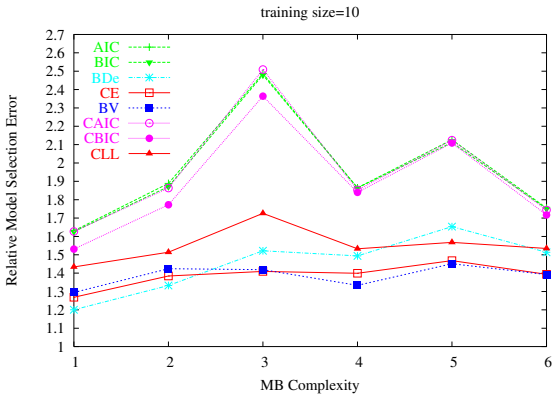


$|S| = 50$

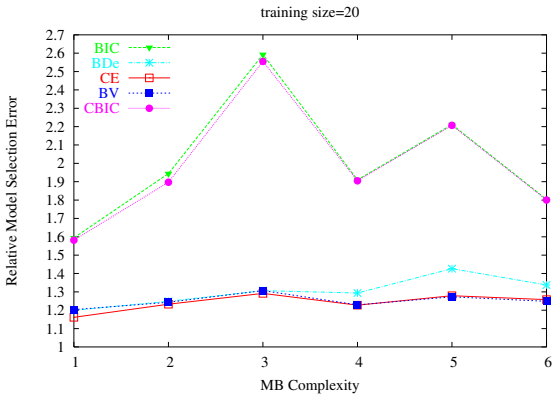


C.4 Results#A4: 7-Variable BNs, 5CV, ELR

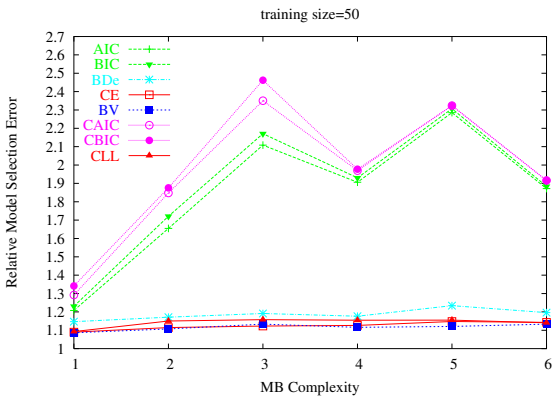
$|S| = 10$



$|S| = 20$

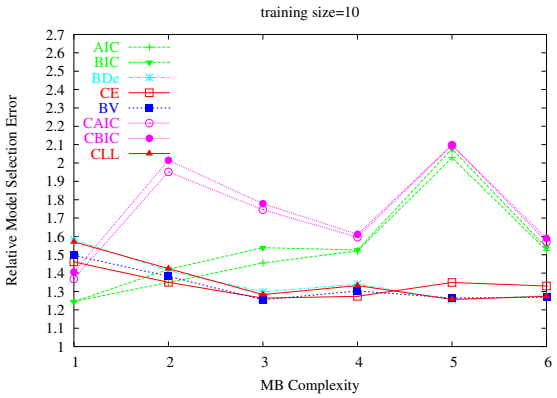


$|S| = 50$

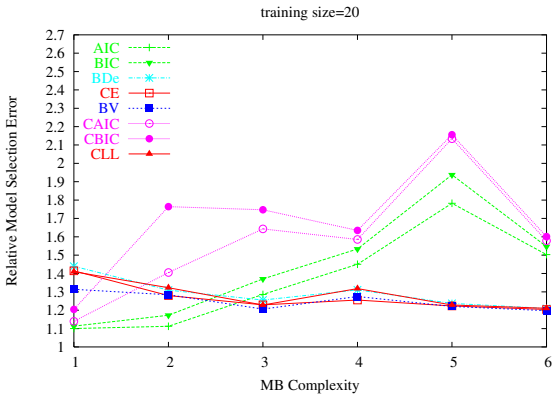


C.5 Results#A5: 15-Variable BNs, 1S, OFE

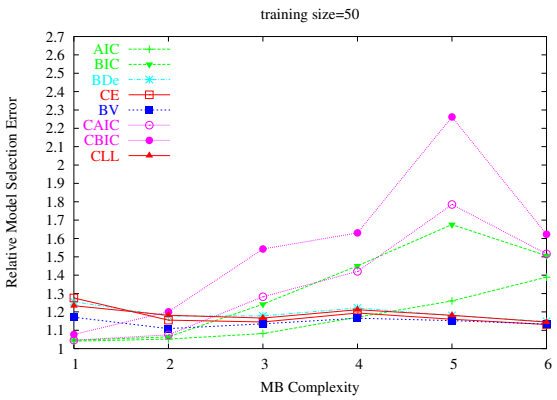
$|S| = 10$



$|S| = 20$

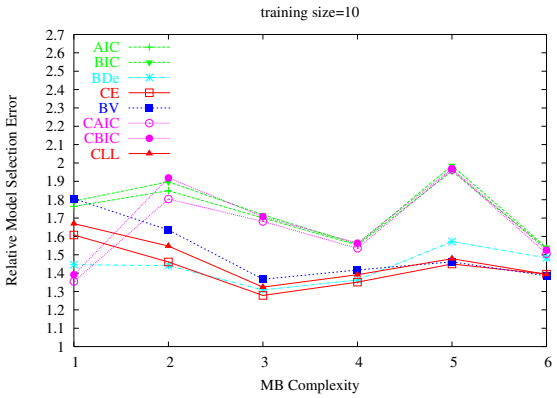


$|S| = 50$

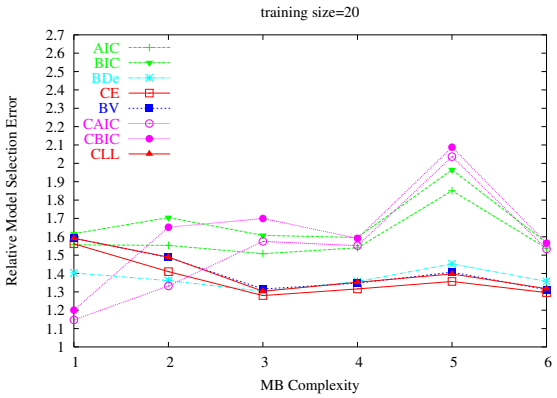


C.6 Results#A6: 15-Variable BNs, 1S, ELR

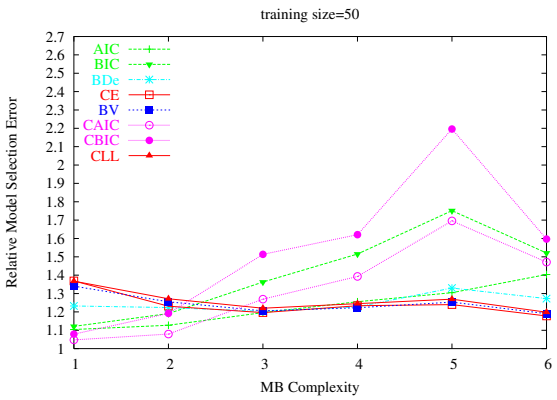
$|S| = 10$



$|S| = 20$



$|S| = 50$



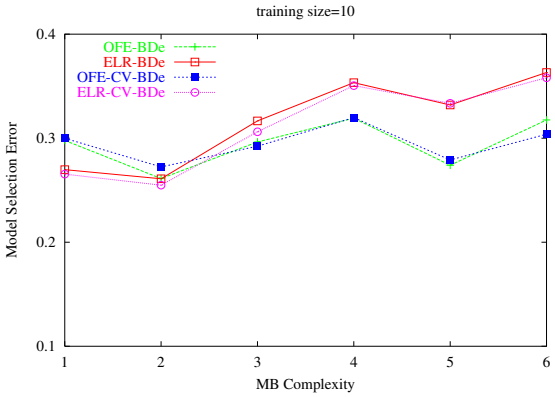
D Comparing Different Ways to Estimate Parameters

The results so far show the relative scores of the different selection criteria, for a single parameter-estimation process — one of "1S" vs "5CV", and one of "OFE" vs "ELR". To compare the different parameter-estimation processes, we focused on a single model selection criterion and sample size, and explicitly compared the quality of the answer produced.

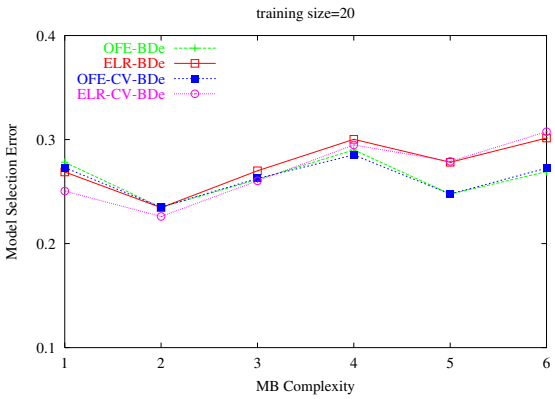
In each case, considered 7-Variable BN.

D.1 Absolute Differences: BDe criterion

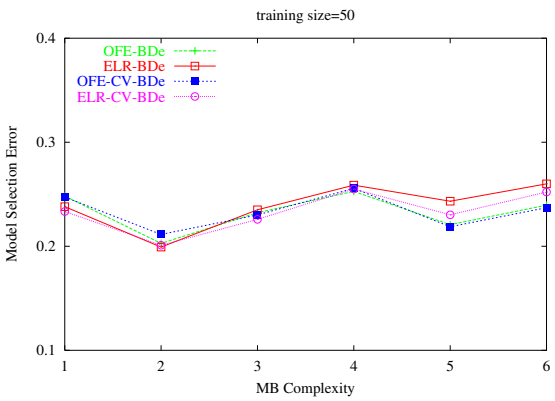
$|S| = 10$



$|S| = 20$

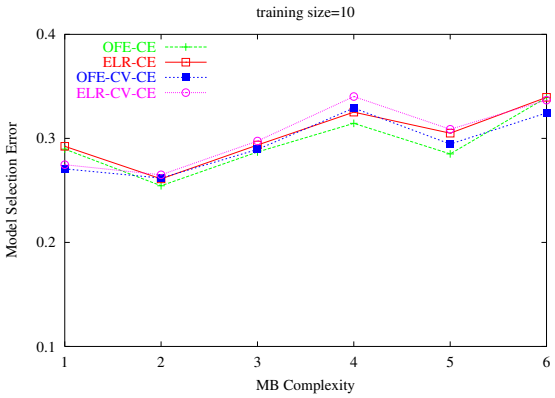


$|S| = 50$

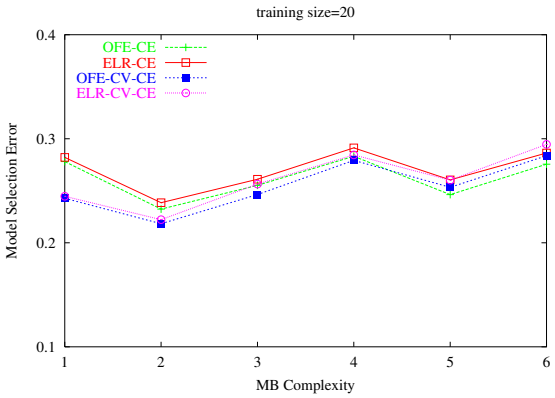


D.2 Absolute Differences: CE criterion

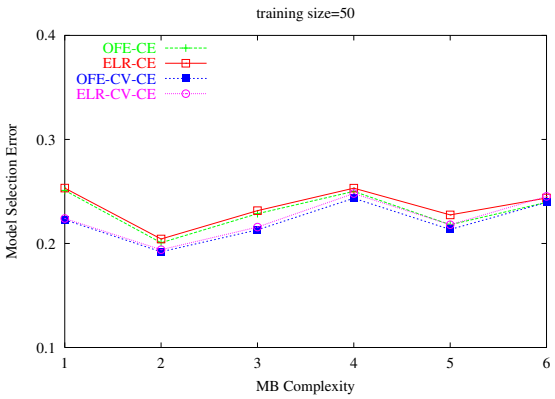
$|S| = 10$



$|S| = 20$

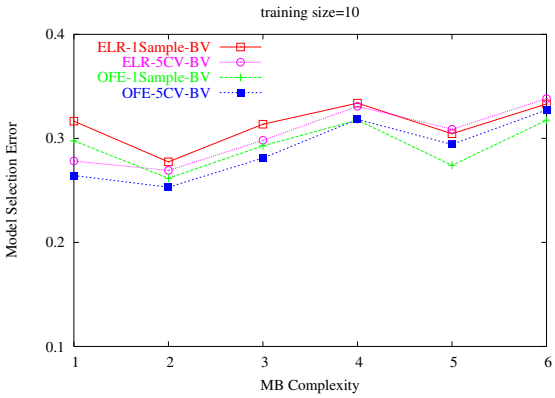


$|S| = 50$

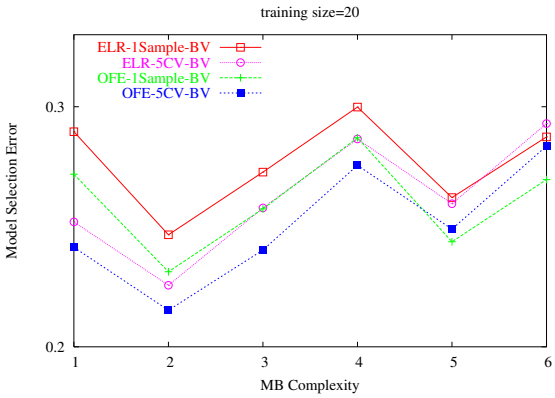
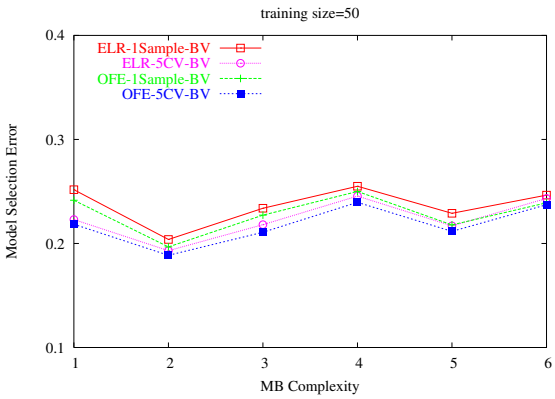


D.3 Absolute Differences: BV criterion

$|S| = 10$



$|S| = 20$



$|S| = 50$