Constraint-based Bayesian Network Learning with Permutation Tests

Marco Scutari
marco.scutari@stat.unipd.it
Adriana Brogini
brogini@stat.unipd.it

Department of Statistical Sciences
University of Padova

June 15, 2010

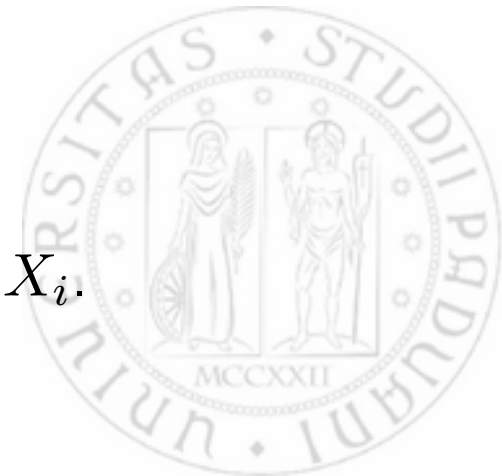
Bayesian networks: definitions

A Bayesian network $B = (\mathcal{G}, \mathrm{P})$ is a graphical model composed by:

- a directed acyclic graph $\mathcal{G} = (\mathbf{U}, A)$. Each node represents a random variable $X \in \mathbf{U}$ and the arcs in A specify the conditional dependence structure of $\mathbf{U}$.
- a global probability distribution $\mathrm{P}(\mathbf{U})$ defined over the variable set $\mathbf{U}$. It can be factorized into a set of local probability distributions of the form

$$\mathrm{P}(\mathbf{U}) = \prod_{X_i \in \mathbf{U}} \mathrm{P}(X_i \,|\, \Pi_{X_i}),$$

where Π_{X_i} is the set of the parents of the node X_i.



Learning Bayesian networks

Model selection (usually called learning) of a Bayesian network is also performed in two steps:

1. structure learning: finding a graph structure that encodes the conditional independence (CI) relationships in the data.
2. parameter learning: fitting the parameters of each local distribution given the graph structure selected in the previous step.

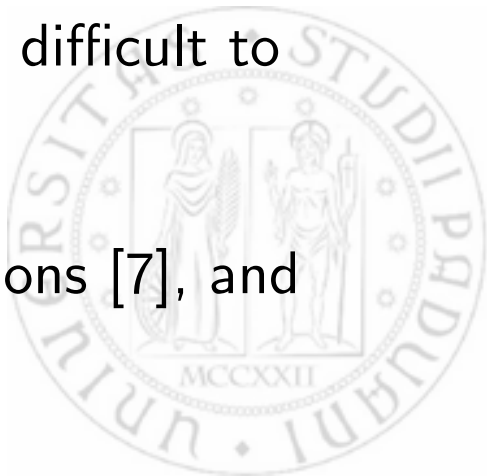
Most modern structure learning algorithms use conditional independence tests to find out CI constraints from data (constraint-based algorithms), sometimes together with goodness-of-fit scores (hybrid algorithms).

Parametric vs Permutation tests for structure learning

Proofs of correctness of structure learning algorithms assume that conditional independence tests do not incur in type I or type II errors [6, 8, 10]. This makes the use of parametric tests problematic because:

- most of them are asymptotic or approximate; but they are often applied in situations where convergence is problematic (high-dimensional data, "small n, large p" settings).
- they require distributional assumptions which are difficult to justify and rarely satisfied by real-world data.

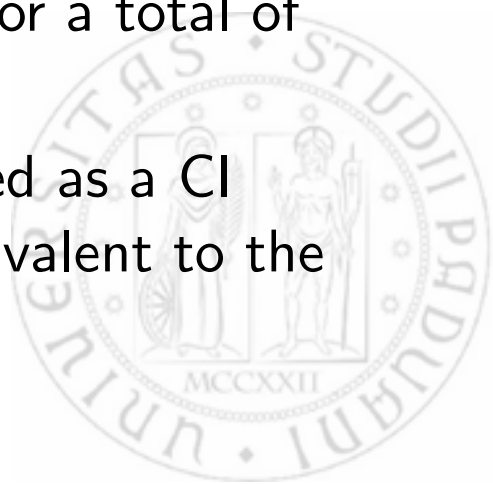
Permutation tests do not present any of these limitations [7], and therefore result in a more effective model selection.



Model validation: experimental setting

The impact of permutation tests on Bayesian network learning has been evaluated for the Max-Min Hill Climbing (MMHC) hybrid algorithm [9], which is one of the best performers up to date and has been extensively tested over a wide variety of data sets.
In particular:

- data sets have been generated from the ALARM network [2], which is often used a benchmark for testing structure learning algorithms. ALARM contains 37 dicrete nodes, for a total of 509 parameters.
- the G^2 log-likelihood ratio test [1] have been used as a CI test, with an $\alpha = 0.05$ threshold. G^2 is also equivalent to the mutual information CI test up to a constant [5].

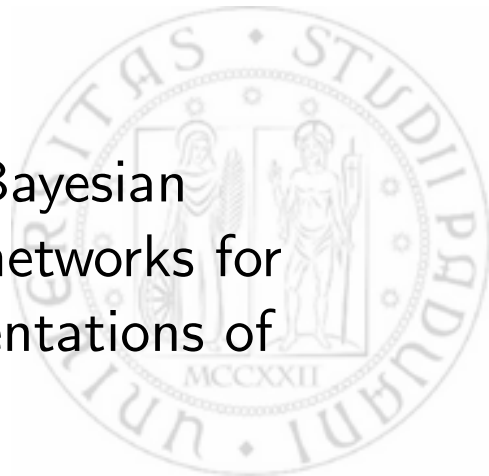


Model validation: goodness of fit

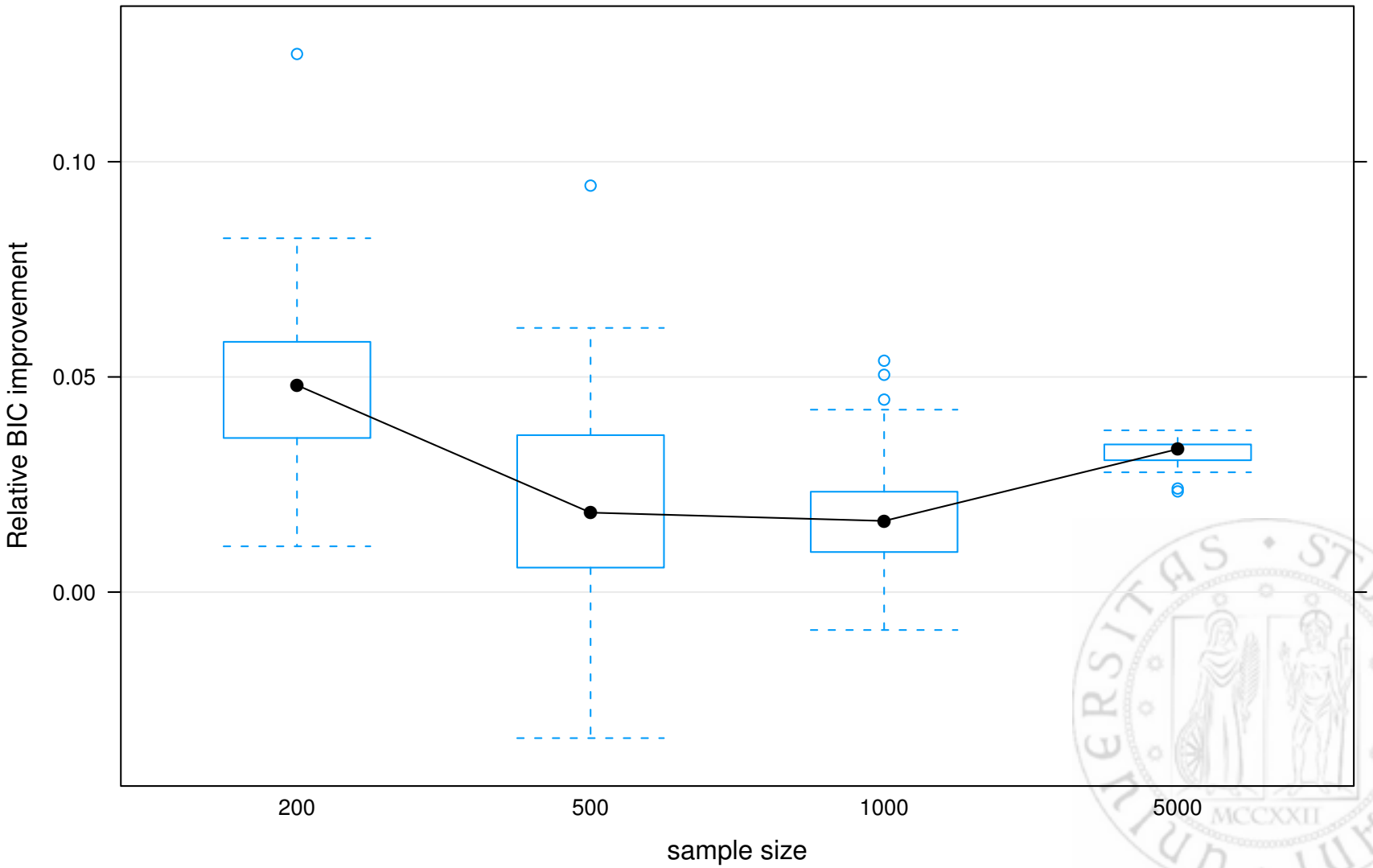
Goodness of fit has been measured with the following scores:

- the Bayesian Information Criterion (BIC) [4], which is a penalized likelihood score.
- the Bayesian Dirichlet equivalent (BDe) score [3], which is posterior Dirichlet distribution based on a uniform prior.
- the Structural Hamming Distance (SHD) score [9], which is an extension of Hamming's distance measure for undirected graphs.

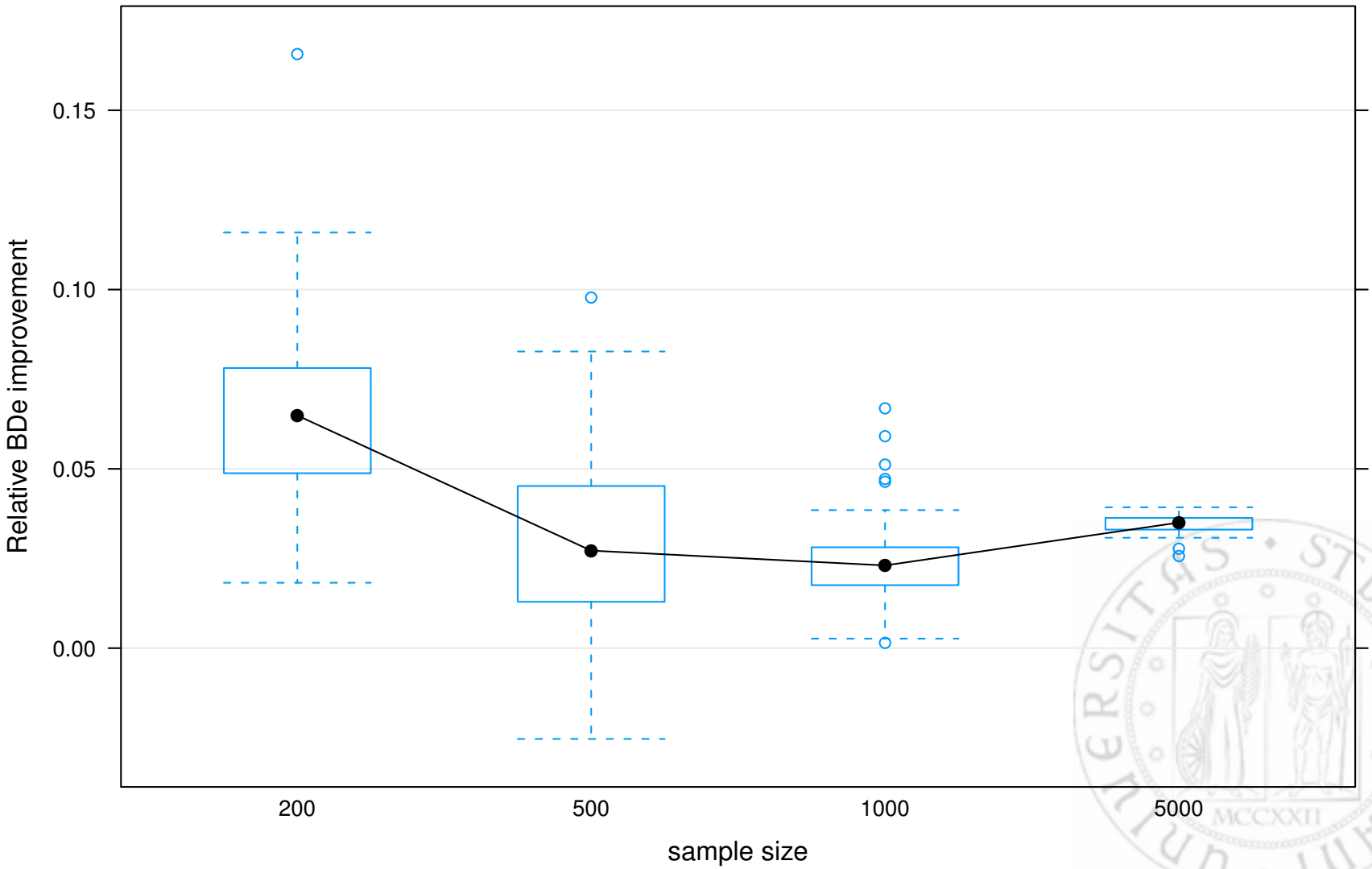
Each score has been computed on 4 sets of pairs of Bayesian networks learned from samples of different sizes (50 networks for each size) using parametric and permutation implementations of the G^2 CI test.



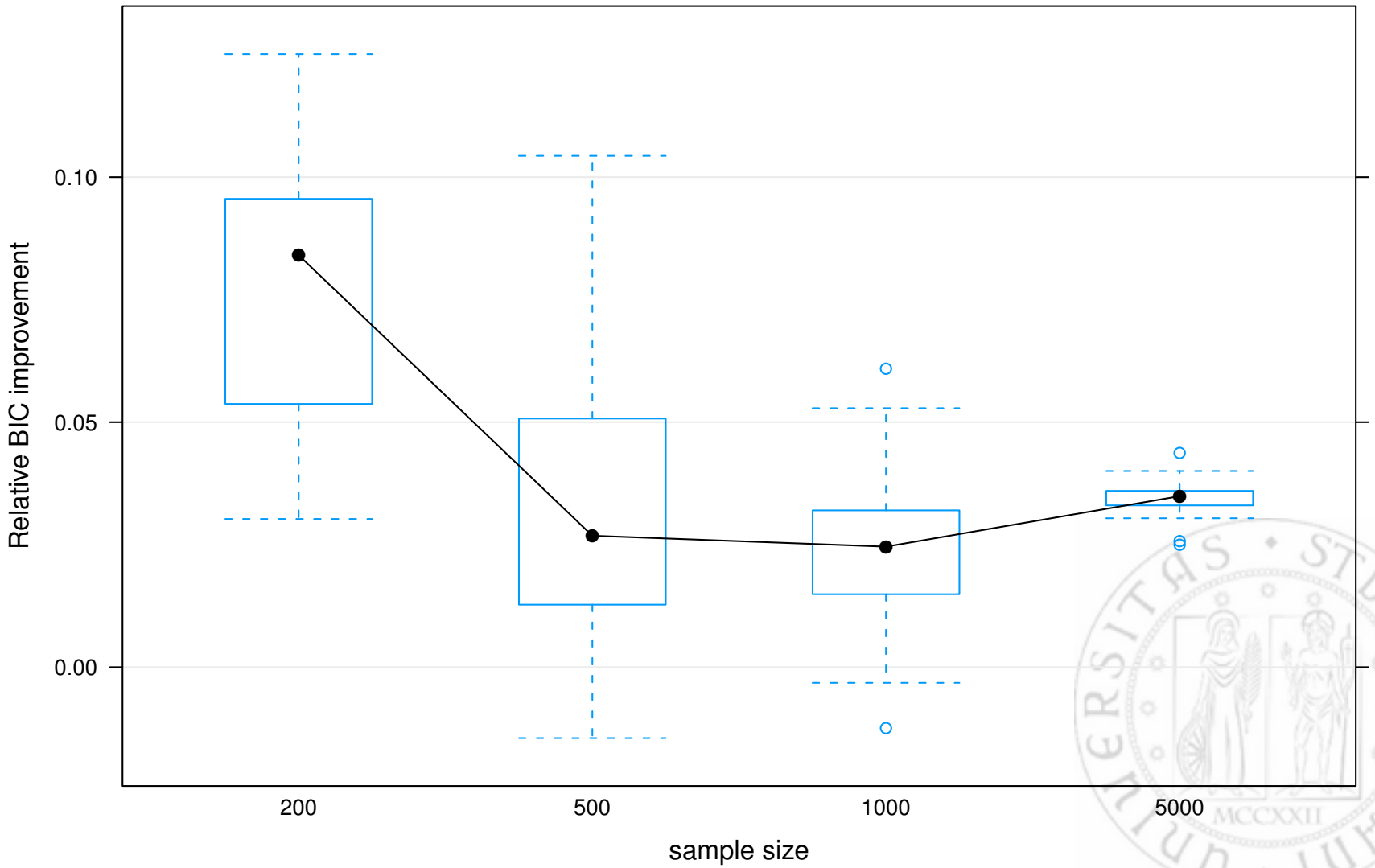
Effect on the BIC score of fitted networks



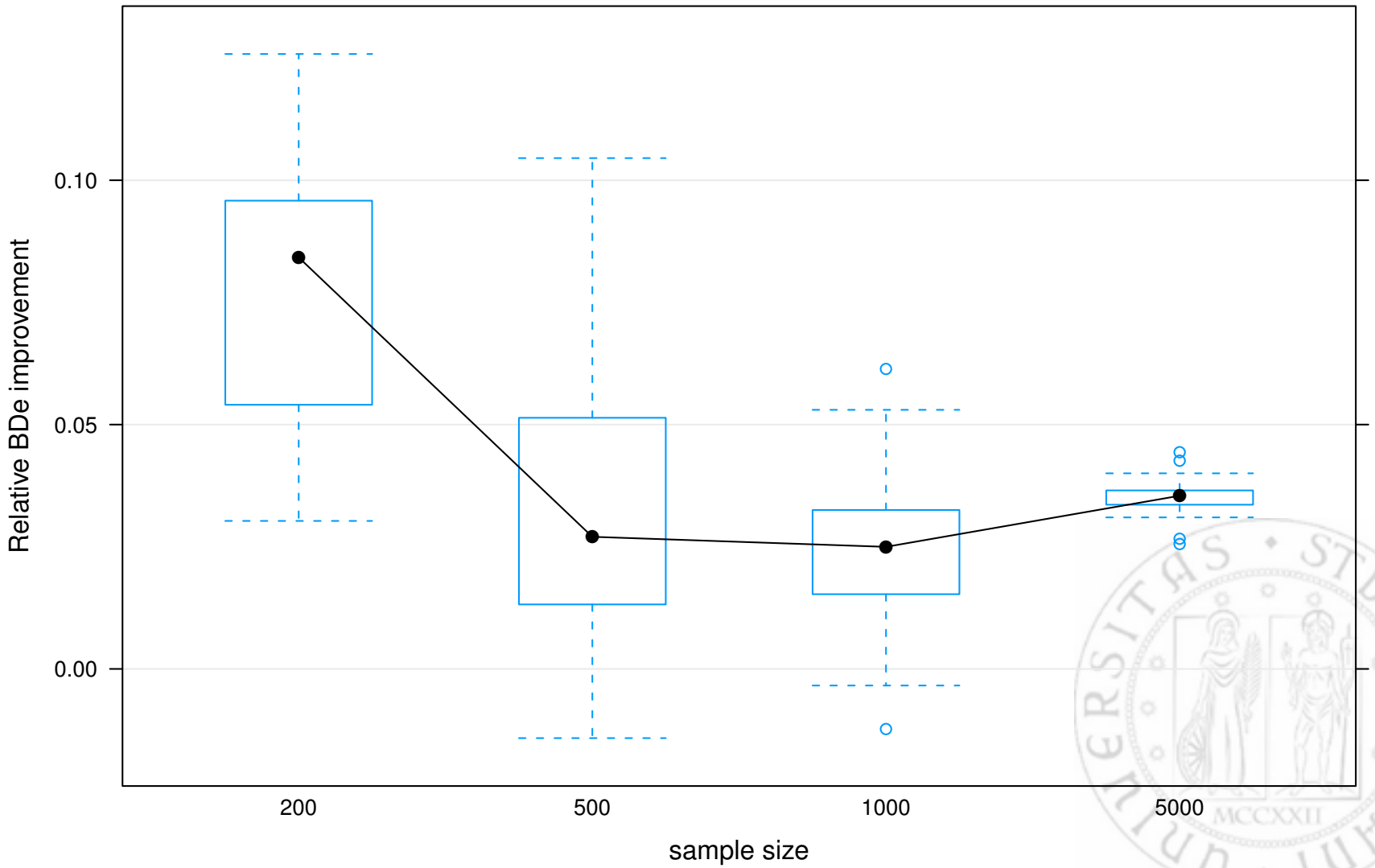
Effect on the BDe score of fitted networks



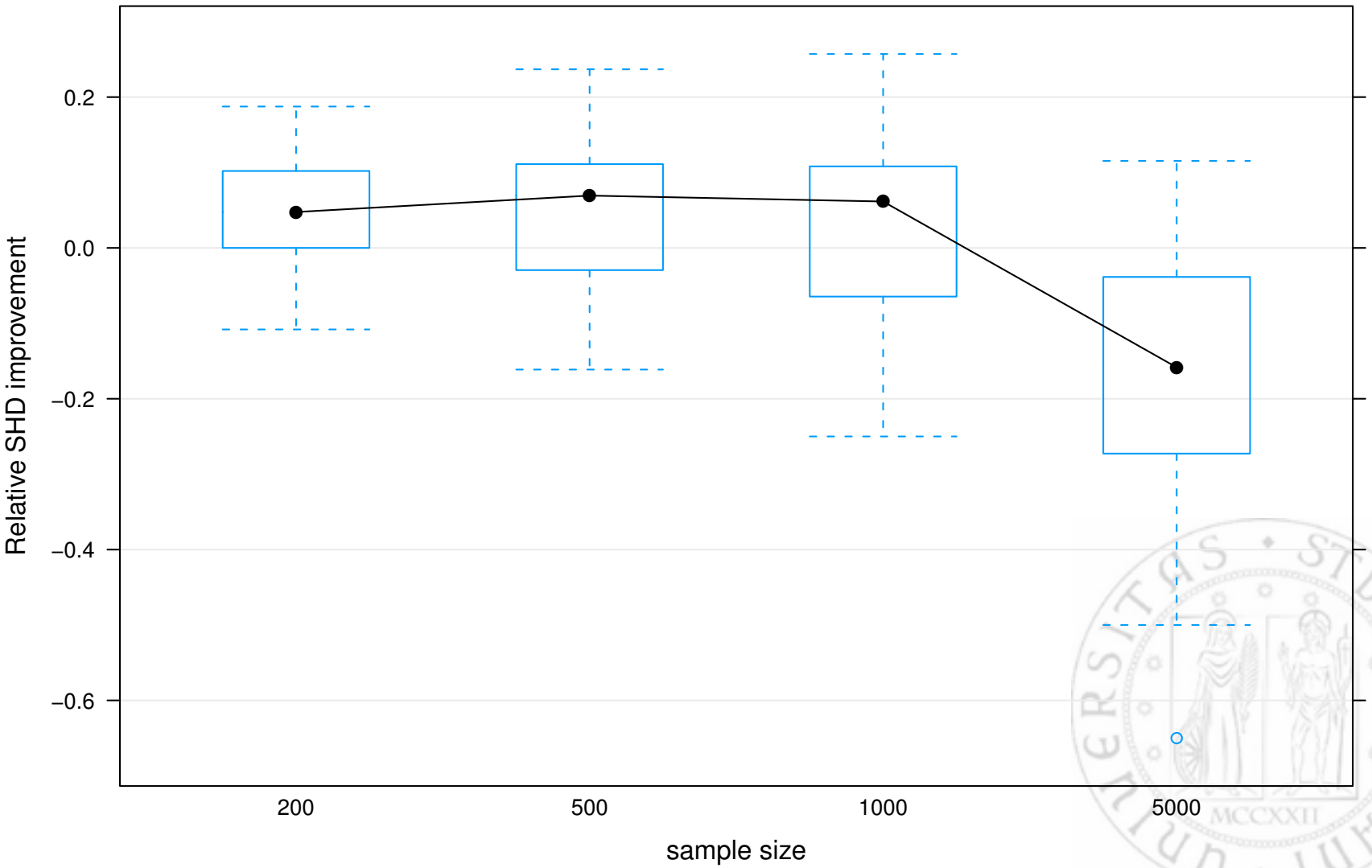
Effect on the BIC score, predictive goodness-of-fit



Effect on the BDe score, predictive goodness-of-fit

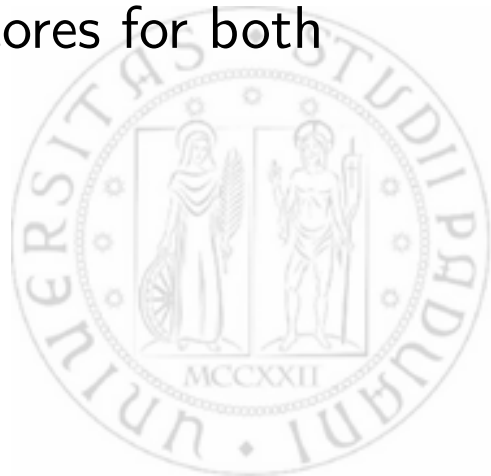


Effect on Structural Hamming Distance (SHD)



Conclusions

- The correctness of structure learning algorithms depends heavily on the performance of the underlying CI tests.
- Parametric tests are problematic in many real-world settings in which Bayesian networks are used ("small n, large p").
- Model selection based on permutation tests consistently produces networks with higher BIC and BDEu scores for both small and moderately large sample sizes.



References

References I

A. Agresti.

Categorical Data Analysis.
Wiley, 2002.

I.A Beinlich, H. J. Suermondt, R. M. Chavez, and G. F. Cooper.

The ALARM Monitoring System: A Case Study with Two Probabilistic Inference Techniques for Belief Networks.
In *Proceedings of the 2nd European Conference on Artificial Intelligence in Medicine*, pages 247–256, 1989.

D. Heckerman, D. Geiger, and D. M. Chickering.

Learning Bayesian Networks: The Combination of Knowledge and Statistical Data.
Machine Learning, 20(3):197–243, 1995.
Available as Technical Report MSR-TR-94-09.

D. Koller and N. Friedman.

Probabilistic Graphical Models: Principles and Techniques.
MIT Press, 2009.

S. Kullback.

Information Theory and Statistics.
Wiley, 1959.

D. Margaritis.

Learning Bayesian Network Model Structure from Data.
PhD thesis, School of Computer Science, Carnegie-Mellon University, May 2003.
Available as Technical Report CMU-CS-03-153.

References II

F. Pesarin.

Multivariate Permutation Tests with Applications in Biostatistics.
Wiley, 2001.

I. Tsamardinos, C. F. Aliferis, and A. Statnikov.

Algorithms for Large Scale Markov Blanket Discovery.
In *Proceedings of the 16th International Florida Artificial Intelligence Research Society Conference*, pages 376–381, 2003.

I. Tsamardinos, L. E. Brown, and C. F. Aliferis.

The Max-Min Hill-Climbing Bayesian Network Structure Learning Algorithm.
Machine Learning, 65(1):31–78, 2006.

T. S. Verma and J. Pearl.

Equivalence and Synthesis of Causal Models.
Uncertainty in Artificial Intelligence, 6:255–268, 1991.