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
A Bayesian network $B = (G, P)$ is a graphical model composed by:

- a directed acyclic graph $G = (U, A)$. Each node represents a random variable $X \in U$ and the arcs in $A$ specify the conditional dependence structure of $U$.

- a global probability distribution $P(U)$ defined over the variable set $U$. It can be factorized into a set of local probability distributions of the form

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

where $\Pi_{X_i}$ is the set of the parents of the node $X_i$. 

Marco Scutari and Adriana Brogini

University of Padova
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).
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.
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 discrete 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].
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

Sample size: 200, 500, 1000, 5000

Relative BIC improvement: 0.00, 0.05, 0.10

Marco Scutari and Adriana Brogini
Effect on the BDe score of fitted networks

![Graph showing the effect of sample size on BDe score improvement.](image)
Effect on the BIC score, predictive goodness-of-fit

Marco Scutari and Adriana Brogini

Universita degli Studi di Padova
Effect on the BDe score, predictive goodness-of-fit

Marco Scutari and Adriana Brogini University of Padova
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

A. Agresti.
Categorical Data Analysis.

The ALARM Monitoring System: A Case Study with Two Probabilistic Inference Techniques for Belief Networks.

D. Heckerman, D. Geiger, and D. M. Chickering.
Learning Bayesian Networks: The Combination of Knowledge and Statistical Data.

D. Koller and N. Friedman.

S. Kullback.
Information Theory and Statistics.

D. Margaritis.
Learning Bayesian Network Model Structure from Data.
F. Pesarin.

*Multivariate Permutation Tests with Applications in Biostatistics.*

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

Algorithms for Large Scale Markov Blanket Discovery.

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

The Max-Min Hill-Climbing Bayesian Network Structure Learning Algorithm.

T. S. Verma and J. Pearl.

Equivalence and Synthesis of Causal Models.