Who Learns Better Bayesian Network Structures

Constraint-Based, Score-based or Hybrid Algorithms?

Marco Scutari[1]
Catharina Elisabeth Graafland[2]
José Manuel Gutiérrez[2]

[1]Department of Statistics
University of Oxford, UK
scutari@stats.ox.ac.uk

[2]Institute of Physics of Cantabria (CSIC-UC)
Santander, Spain

September 11, 2018

Outline

Bayesian network Structure learning is defined by the combination of a statistical criterion and an algorithm that determines how the criterion is applied to the data. After removing the confounding effect of different choices for the statistical criterion, we ask the following questions:

Q1 *Which of constraint-based and score-based algorithms provide the most accurate structural reconstruction?*

Q2 *Are hybrid algorithms more accurate than constraint-based or score-based algorithms?*

Q3 *Are score-based algorithms slower than constraint-based and hybrid algorithms?*

Classes of Structure Learning Algorithms

Structure learning consists in finding the DAG $\mathcal{G}$ that encodes the dependence structure of a data set $\mathcal{D}$ with n observations. Algorithms for this task fall into one three classes:

- Constraint-based algorithms identify conditional independence constraints with statistical tests, and link nodes that are not found to be independent.
- Score-based algorithms are applications of general optimisation techniques; each candidate DAG is assigned a network score maximise as the objective function.
- Hybrid algorithms have a *restrict* phase implementing a constraint-based strategy to reduce the space of candidate DAGs; and a *maximise* phase implementing a score-based strategy to find the optimal DAG in the restricted space.

Conditional Independence Tests and Network Scores

For discrete BNs, the most common test is the log-likelihood ratio test

$$\mathrm{G}^2(X, Y \mid \mathbf{Z}) = 2 \log \frac{\mathrm{P}(X \mid Y, \mathbf{Z})}{\mathrm{P}(X \mid \mathbf{Z})} = 2 \sum_{i=1}^{R} \sum_{j=1}^{C} \sum_{k=1}^{L} n_{ijk} \log \frac{n_{ijk} n_{++k}}{n_{i+k} n_{+jk}},$$

has an asymptotic $\chi^2_{(R-1)(C-1)L}$ distribution. For GBNs,

$$\mathrm{G}^2(X, Y \mid \mathbf{Z}) = n \log(1 - \rho^2_{XY \mid \mathbf{Z}}) \stackrel{.}{\sim} \chi^2_1.$$

As for network scores, the Bayesian Information criterion

$$\mathrm{BIC}(\mathcal{G}; \mathcal{D}) = \sum_{i=1}^{N} \left[\log \mathrm{P}(X_i \mid \Pi_{X_i}) - \frac{|\Theta_{X_i}|}{2} \log n \right],$$

is a common choice for both discrete BNs and GBNs, as it provides a simple approximation to $\log \mathrm{P}(\mathcal{G} \mid \mathcal{D})$. $\log \mathrm{P}(\mathcal{G} \mid \mathcal{D})$ itself is available in closed form as BDeu and BGeu [5, 4].

Score- and Constraint-Based Algorithms Can Be Equivalent

Cowell [3] famously showed that constraint-based and score-based algorithms can select identical discrete BNs.

1. He noticed that the G^2 test in has the same expression as a score-based network comparison based on the log-likelihoods $\log \mathrm{P}(X \mid Y, \mathbf{Z}) - \log \mathrm{P}(X \mid \mathbf{Z})$ if we take $\mathbf{Z} = \Pi_X$.
2. He then showed that these two classes of algorithms are equivalent if we assume a fixed, known topological ordering and we use log-likelihood and G^2 as matching statistical criteria.

We take the same view that the algorithms and the statistical criteria they use are separate and complementary in determining the overall behaviour of structure learning. We then want to remove the confounding effect of choices for the statistical criterion from our evaluation of the algorithms.

Constructing Matching Tests and Scores

Consider two DAGs $\mathcal{G}^+$ and $\mathcal{G}^-$ that differ by a single arc $X_j \to X_i$. In a score-based approach, we can compare them using BIC:

$$\mathrm{BIC}(\mathcal{G}^+; \mathcal{D}) > \mathrm{BIC}(\mathcal{G}^-; D) \Rightarrow$$
$$2 \log \frac{\mathrm{P}(X_i \mid \Pi_{X_i} \cup \{X_j\})}{\mathrm{P}(X_i \mid \Pi_{X_i})} > \left(|\Theta_{X_i}^{\mathcal{G}^+}| - |\Theta_{X_i}^{\mathcal{G}^-}|\right) \log n$$

which is equivalent to testing the conditional independence of X_i and X_j given Π_{X_i} using the G^2 test, just with a different significance threshold. We will call this test $\mathrm{G}^2_{\mathrm{BIC}}$ and use it as the matching statistical criterion for BIC to compare different learning algorithms. For discrete BNs, starting from $\log \mathrm{P}(\mathcal{G} \mid \mathcal{D})$ we get

$$\log \mathrm{P}(\mathcal{G}^+ \mid \mathcal{D}) > \log \mathrm{P}(\mathcal{G}^- \mid \mathcal{D}) \Rightarrow \log \mathrm{BF} = \log \frac{\mathrm{P}(\mathcal{G}^+ \mid \mathcal{D})}{\mathrm{P}(\mathcal{G}^- \mid \mathcal{D})} > 0,$$

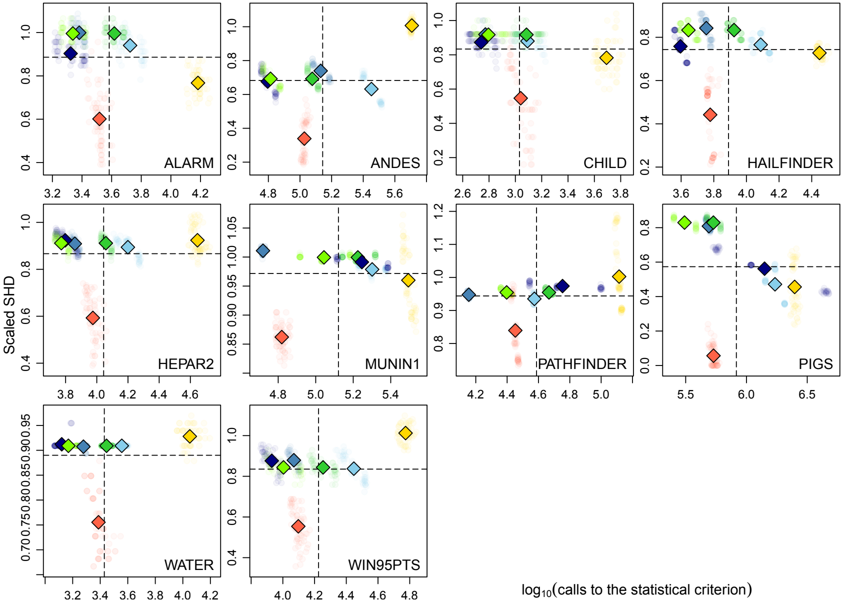
which uses Bayes factors as matching tests for BDeu.

A Simulation Study

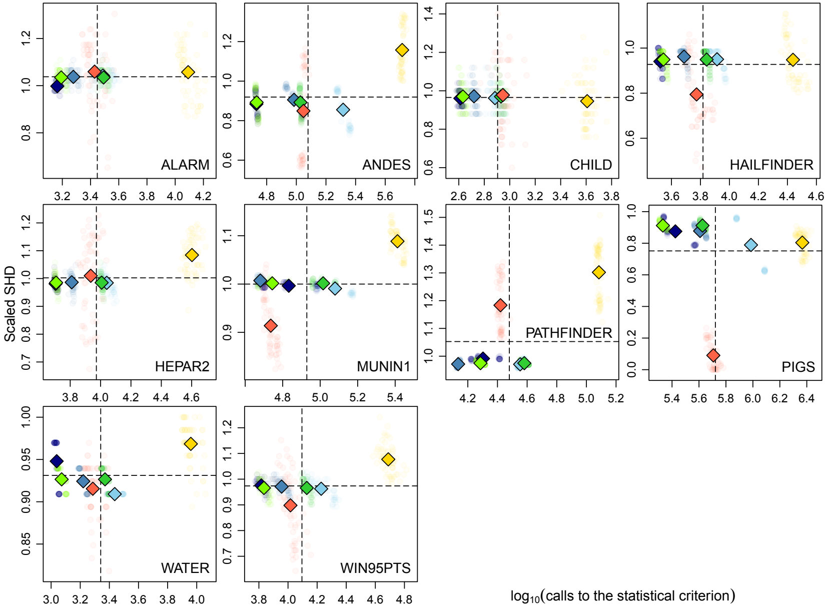
We assess three constraint-based algorithms (PC [2], GS [6], Inter-IAMB [13]), two score-based algorithms (tabu search, simulated annealing [7] for BIC, GES [1] for $\log \mathrm{BDeu}$) and two hybrid algorithms (MMHC [10], RSMAX2 [9]) on 14 reference networks [8]. For each BN:

1. We generate 20 samples of size $n/|\Theta| = 0.1$, 0.2, 0.5 (small samples), 1.0, 2.0, 5.0 (large samples).
2. We learn $\mathcal{G}$ using (BIC, $\mathrm{G}^2_{\mathrm{BIC}}$), and $(\log \mathrm{BDeu}, \log \mathrm{BF})$ as well for discrete BNs.
3. We measure the accuracy of the learned DAGs using SHD$/|A|$ [10] from the reference BN; and we measure the speed of the learning algorithms with the number of calls to the statistical criterion.

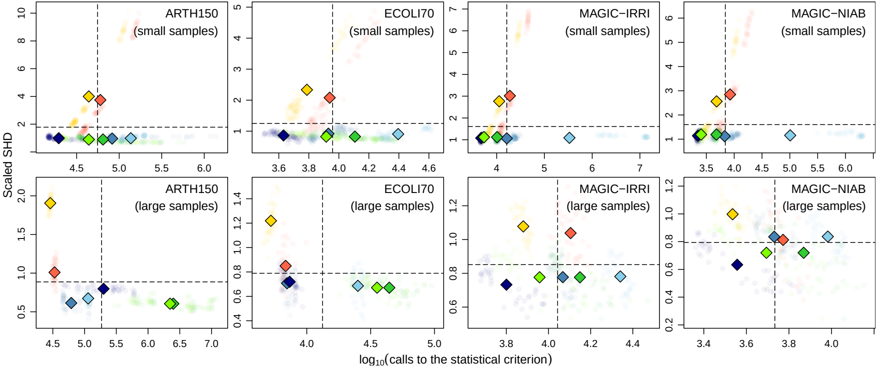
Discrete Bayesian Networks (Large Samples)



Discrete Bayesian Networks (Small Samples)



Gaussian Bayesian Networks



Overall Conclusions

Discrete networks:

- score-based algorithms often have higher SHDs for small samples;
- hybrid and constraint-based algorithms have comparable SHDs;
- constraint-based algorithms have better SHD than score-based algorithms for small sample sizes in 7/10 BNs, but it decreases more slowly as n increases for all BNs;
- simulated annealing is consistently slower; tabu search is always fast and accurate in large samples, for 6/10 BNs in small samples.

Gaussian networks:

- tabu search and simulated annealing have larger SHDs than constraint-based or hybrid algorithms for most samples;
- hybrid and constraint-based algorithms have roughly the same SHD for all sample sizes.

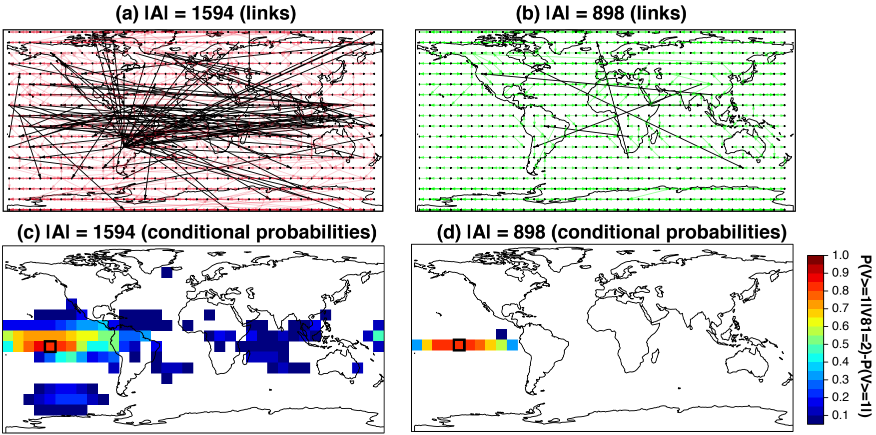
Real-World Climate Data...

Climate networks aim to analyse the complex spatial structure of climate data: spatial dependence among nearby locations, but also long-range large-scale oscillation patterns over distant regions in the world, known as teleconnections [11], such as the El Niño Southern Oscillation (ENSO) [12].

We confirm the results above using NCEP/NCAR monthly surface temperature data on a global 10°-resolution grid between 1981 and 2010. This gives sample size $n = 30 \times 12 = 360$ and variables $N = 18 \times 36 = 648$, which we model with a Gaussian Bayesian network.

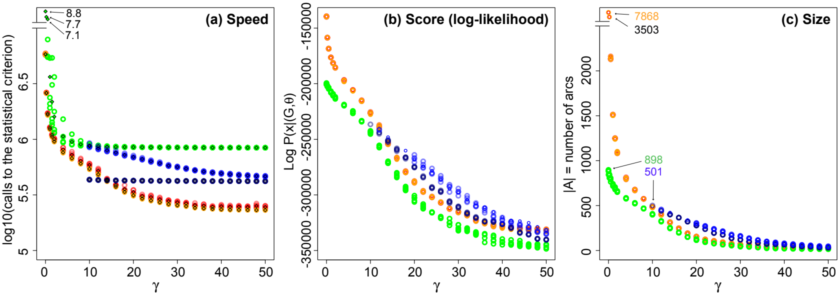
The sample would count as a “small sample” in the simulation study.

… Gives Networks that Look Like This…



We want to find teleconnections, so we are looking forward to learning networks like that on the left more than that on the right because the latter only encodes short-range perturbations.

... and Agree with the Simulation Study



- Constraint-based algorithms produce BNs with the highest log-likelihood, hybrid have the worst log-likelihood values and includes only a few teleconnections;
- score-based algorithms produce high-likelihood networks with a large number of teleconnections that allow propagating evidences with realistic results.
- score-based algorithms are faster than both hybrid and constraint-based algorithms.

Conclusions

We assessed the three classes of BN structure learning algorithms, removing the confounding effect of different choices of statistical criteria.

Interestingly, we found that:

Q1 constraint-based algorithms are more accurate than score-based algorithms for small sample sizes;

Q2 that they are as accurate as hybrid algorithms;

Q3 and that tabu search, as a score-based algorithm, is faster than constraint-based algorithms more often than not.

This in contrast with the general view in the literature that score-based algorithms are less sensitive to individual errors and more accurate than constraint-based algorithms; and that hybrid algorithms are faster and more accurate than both. More so at small sample sizes. Also, score-based algorithms are supposed to scale less well to high-dimensional data.

Thanks!

References

References I

D. M. Chickering.
Optimal Structure Identification With Greedy Search.
Journal of Machine Learning Research, 3:507–554, 2002.

D. Colombo and M. H. Maathuis.
Order-Independent Constraint-Based Causal Structure Learning.
Journal of Machine Learning Research, 15:3921–3962, 2014.

R. Cowell.
Conditions Under Which Conditional Independence and Scoring Methods Lead to Identical Selection of Bayesian Network Models.
In *Proceedings of the 17th Conference on Uncertainty in Artificial Intelligence*, pages 91–97, 2001.

D. Geiger and D. Heckerman.
Learning Gaussian Networks.
In *Proceedings of the 10th Conference on Uncertainty in Artificial Intelligence*, pages 235–243, 1994.

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.

References II

D. Margaritis.
Learning Bayesian Network Model Structure from Data.
PhD thesis, School of Computer Science, Carnegie-Mellon University, Pittsburgh, PA, May 2003.

S. J. Russell and P. Norvig.
Artificial Intelligence: A Modern Approach.
Prentice Hall, 3rd edition, 2009.

M. Scutari.
Bayesian Network Repository.
http://www.bnlearn.com/bnrepository, 2012.

M. Scutari, P. Howell, D. J. Balding, and I. Mackay.
Multiple Quantitative Trait Analysis Using Bayesian Networks.
Genetics, 198(1):129–137, 2014.

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.

References III

A. A. Tsonis, K. L. Swanson, and G. Wang.
On the Role of Atmospheric Teleconnections in Climate.
Journal of Climate, 21(12):2990–3001, 2008.

K. Yamasaki, A. Gozolchiani, and S. Havlin.
Climate Networks around the Globe are Significantly Affected by El Niño.
Phys. Rev. Lett., 100:228501, 2008.

S. Yaramakala and D. Margaritis.
Speculative Markov Blanket Discovery for Optimal Feature Selection.
In *ICDM '05: Proceedings of the Fifth IEEE International Conference on Data Mining*, pages 809–812. IEEE Computer Society, 2005.