Proceedings of Machine Learning Research vol 72, 416-427, 2018 PGM 2018

Who Learns Better Bayesian Network Structures: Constraint-Based, Score-based or Hybrid Algorithms?

Marco Scutari SCUTARI@STATS.OX.AC.UK
Department of Statistics, University of Oxford, United Kingdom

Catharina Elisabeth Graafland CATHARINA.GRAAFLAND@UNICAN.ES
José Manuel Gutiérrez MANUEL.GUTIERREZ@UNICAN.ES
Institute of Physics of Cantabria (CSIC-UC), Santander, Spain

Abstract

The literature groups algorithms to learn the structure of Bayesian networks from data in three separate classes: *constraint-based algorithms*, which use conditional independence tests to learn the dependence structure of the data; *score-based algorithms*, which use goodness-of-fit scores as objective functions to maximise; and *hybrid algorithms* that combine both approaches. Famously, Cowell (2001) showed that algorithms in the first two classes learn the same structures when the topological ordering of the network is known and we use entropy to assess conditional independence and goodness of fit.

In this paper we address the complementary question: how do these classes of algorithms perform outside of the assumptions above? We approach this question by recognising that structure learning is defined by the combination of a *statistical criterion* and an *algorithm* that determines how the criterion is applied to the data. Removing the confounding effect of different choices for the statistical criterion, we find using both simulated and real-world data that constraint-based algorithms do not appear to be more efficient or more sensitive to errors than score-based algorithms; and that hybrid algorithms are not faster or more accurate than constraint-based algorithms. This suggests that commonly held beliefs on structure learning in the literature are strongly influenced by the choice of particular statistical criteria rather than just properties of the algorithms themselves.

Keywords: Bayesian networks; structure learning; conditional independence tests; network scores; climate networks.

1. Introduction

Bayesian networks (BNs; Koller and Friedman, 2009) are a class of graphical models defined over a set of random variables $\mathbf{X} = \{X_1, \ldots, X_N\}$, each describing some quantity of interest, that are associated with the nodes of a directed acyclic graph (DAG) $\mathcal{G}$. (They are often referred to interchangeably.) The structure of the DAG, that is, the pattern of arcs in $\mathcal{G}$, encodes the independence relationships between those variables, with graphical separation in $\mathcal{G}$ implying conditional independence in probability. As a result, $\mathcal{G}$ induces the factorisation

$$\mathrm{P}(\mathbf{X} \,|\, \mathcal{G}, \Theta) = \prod_{i=1}^{N} \mathrm{P}(X_i \,|\, \Pi_{X_i}, \Theta_{X_i}), \tag{1}$$

in which the *global distribution* of $\mathbf{X}$ (with parameters Θ) decomposes in one *local distribution* for each X_i (with parameters Θ_{X_i}, $\bigcup_{X_i} \Theta_{X_i} = \Theta$) conditional on its parents Π_{X_i}. This decomposition does not uniquely identify a single BN, but groups BNs into *equivalence classes* (Chickering, 1995) of models that are probabilistically indistinguishable. All BNs in the same equivalence class have

the same underlying undirected graph and v-structures (patterns of arcs like $X_i \rightarrow X_j \leftarrow X_k$, with no arc between X_i and X_k); and each equivalence class is characterised by the completed partially-directed acyclic graph (CPDAG) that arises from the combination of these two quantities.

While in principle there are many possible choices for the distribution of $\mathbf{X}$, the literature has focused mostly on two sets of assumptions. *Discrete BNs* (Heckerman et al., 1995) assume that $X_i \,|\, \Pi_{X_i} \sim Mul(\pi_{ik\,|\,j})$, $\pi_{ik\,|\,j} = \mathrm{P}(X_i = k \,|\, \Pi_{X_i} = j)$; their parameters are the conditional probabilities of X_i given each configuration of the values of its parents. As a result, $\mathbf{X}$ is also multinomial. *Gaussian BNs* (GBNs; Geiger and Heckerman, 1994) assume that the X_i are univariate normals linked by linear dependencies to their parents: $X_i \,|\, \Pi_{X_i} \sim N(\mu_{X_i} + \Pi_{X_i}\boldsymbol{\beta}_{X_i}, \sigma^2_{X_i})$ in what is essentially a linear regression model of X_i against the Π_{X_i} with regression coefficients $\boldsymbol{\beta}_{X_i}$. Equivalently, the $X_i \,|\, \Pi_{X_i}$ can be parameterised with the partial correlations $\rho_{X_i,X_j\,|\,\Pi_{X_i}\setminus X_j}$ between X_i and each parent X_j given the rest. In both cases $\mathbf{X}$ is multivariate normal. Other distributional assumptions have seen less widespread adoption due to the lack of exact conditional inference and simple closed-form estimators (*e.g.* copulas, Elidan, 2010) or because of limitations in the DAGs they can encode (*e.g.* conditional linear Gaussian BNs, Lauritzen and Wermuth, 1989).

1.1 Learning a Bayesian Network from Data

The task of learning a BN with DAG $\mathcal{G}$ and parameters Θ from a data set $\mathcal{D}$ containing n observations is performed in two steps in an inherently Bayesian fashion:

$$\underbrace{\mathrm{P}(\mathcal{G}, \Theta \,|\, \mathcal{D})}_{\text{learning}} = \underbrace{\mathrm{P}(\mathcal{G} \,|\, \mathcal{D})}_{\text{structure learning}} \cdot \underbrace{\mathrm{P}(\Theta \,|\, \mathcal{G}, \mathcal{D})}_{\text{parameter learning}}.$$

Structure learning consists in finding the DAG $\mathcal{G}$ that encodes the dependence structure of the data; *parameter learning* consists in estimating the parameters Θ given the $\mathcal{G}$ obtained from structure learning. If we assume parameters in different local distributions are independent, they can be learned in parallel for each node because (1) then implies

$$\mathrm{P}(\Theta \,|\, \mathcal{G}, \mathcal{D}) = \prod_{i=1}^{N} \mathrm{P}(\Theta_{X_i} \,|\, \Pi_{X_i}, \mathcal{D}).$$

On the other hand, structure learning is well known to be computationally challenging and several algorithms have been proposed to solve it, following one of three possible approaches: *constraint-based*, *score-based* and *hybrid*.

Constraint-based algorithms are based on the seminal work of Pearl on causal graphical models (Verma and Pearl, 1991). The most commonly used among them is the PC algorithm in its PC-Stable implementation (Colombo and Maathuis, 2014). PC-Stable first identifies which pairs of nodes (X_i, X_j) are connected by an arc, regardless of its direction. Such nodes cannot be separated by any other subset of nodes; this condition is tested heuristically by performing *conditional independence tests* with increasingly large candidate separating sets. Then the algorithm identifies the v-structures among all the pairs of non-adjacent nodes X_i and X_k with a common neighbour X_j using the separating sets found earlier; and sets the remaining arc directions using the rules from Chickering (1995) to obtain the CPDAG describing the identified equivalence class. More recent algorithms such as Grow-Shrink (Margaritis, 2003) and Inter-IAMB (Yaramakala and Margaritis, 2005) proceed along similar lines, but use faster heuristics to implement the first two steps.

Score-based algorithms represent the application of general optimisation techniques to BN structure learning. Each candidate DAG is assigned a *network score* reflecting its goodness of fit, which the algorithm then attempts to maximise. Some examples are *greedy search*, *simulated annealing* (Bouckaert, 1995) and *genetic algorithms* (Larrañaga et al., 1997); a comprehensive review of these and other approaches is provided in Russell and Norvig (2009). These heuristics can also be applied to CPDAGs, as in the case of Greedy Equivalent Search (GES; Chickering, 2002).

Finally, hybrid algorithms are based on two phases: 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. The best-known member of this family is the *Max-Min Hill Climbing* algorithm (MMHC) by Tsamardinos et al. (2006); another example was presented in our previous work (RSMAX2; Scutari et al., 2014).

1.2 Conditional Independence Tests and Network Scores

The choice of which conditional independence test or network score to use in structure learning depends mainly on the choice of the distribution of $\mathbf{X}$; and is orthogonal to the choice of algorithm. Here we provide a brief overview of those we will use in this paper, while referring the reader to Koller and Friedman (2009) for a more comprehensive treatment.

For discrete BNs, conditional independence tests are functions of the observed frequencies $\{n_{ijk}; i = 1, \ldots; R,\ j = 1, \ldots, C; k = 1, \ldots, L\}$ for any pair of variables (X, Y) given the configurations of some conditioning variables $\mathbf{Z}$. The most common is the log-likelihood ratio G^2 test

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

which is equivalent to mutual information and has an asymptotic $\chi^2_{(R-1)(C-1)L}$ distribution. For GBNs, conditional independence tests are functions of the partial correlation coefficients $\rho_{XY \mid \mathbf{Z}}$. The log-likelihood ratio (and Gaussian mutual information) test takes form

$$\mathrm{G}^2(X, Y \mid \mathbf{Z}) = n \log(1 - \rho^2_{XY \mid \mathbf{Z}}) \sim \chi^2_1; \tag{3}$$

other common options are Fisher's Z test and the exact t test for partial correlation.

As for network scores, the Bayesian Information criterion

$$\mathrm{BIC}(\mathcal{G}; \mathcal{D}) = \sum_{i=1}^{N} \left[\log \mathrm{P}(X_i \mid \Pi_{X_i}) - \frac{|\Theta_{X_i}|}{2} \log n \right], \tag{4}$$

is a common choice for both discrete BNs and GBNs, as it provides a simple approximation to $\log \mathrm{P}(\mathcal{G} \mid \mathcal{D})$ that does not depend on any hyperparameter. $\log \mathrm{P}(\mathcal{G} \mid \mathcal{D})$ is also available in closed form for both discrete BNs (Heckerman et al., 1995) and GBNs (Geiger and Heckerman, 1994).

2. Performance as a Combination of Tests, Scores and Algorithms

As it may be apparent from Sections 1.1 and 1.2, we take the view that the algorithms and the statistical criteria they use are separate and complementary in determining the overall behaviour of structure learning. Cowell (2001) followed the same reasoning when showing that constraint-based and score-based algorithms can select identical discrete BNs. He noticed that the G^2 test

in (2) 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 $\mathbf{Z} = \Pi_X$. 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.

In this paper we will extend that investigation by addressing the following questions:

Q1 *Which of constraint-based and score-based algorithms provide the most accurate structural reconstruction, after accounting for the effect of the choice of statistical criteria?*

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?*

More precisely, we will drop the assumption that the topological ordering is known and we will compare the performance of different classes of algorithms outside of their equivalence conditions for both discrete BNs and GBNs. We choose questions **Q1**, **Q2** and **Q3** because they are most common among practitioners (*e.g.* Cugnata et al., 2016) and researchers (*e.g.,* Tsamardinos et al., 2006; Koller and Friedman, 2009). Overall, there is a general view in the references above and in the literature that score-based algorithms are less sensitive to individual errors of the statistical criteria, and thus more accurate, because they can reverse earlier decisions; and that hybrid algorithms are faster and more accurate than both score-based and constraint-based algorithms. These differences have been found to be more pronounced at small sample sizes. Furthermore, score-based algorithms have been found to scale less well to high-dimensional data.

An important limitation we find in these studies is the confounding between the choice of the algorithms and that of the statistical criteria, which makes it impossible to assess the merits inherently attributable to the algorithms themselves. Therefore, similarly to Cowell (2001), we construct matching scores and independence tests to make algorithms directly comparable. Consider two DAGs $\mathcal{G}^+$ and $\mathcal{G}^-$ which differ by a single arc $X_j \to X_i$. In a score-based approach, we can compare them using BIC from (4) and select $\mathcal{G}^+$ over $\mathcal{G}^-$ if

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

which is equivalent to testing the conditional independence of X_i and X_j given Π_{X_i} using the G^2 test from (2) or (3), just with a different significance threshold than the appropriate $\chi^2_{1-\alpha}$ quantile. 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, we will also construct a test from graph posterior probabilities using Bayes factors,

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

to confirm our conclusions with a second set of matching criteria.

3. Simulation Study

We address **Q1**, **Q2** and **Q3** with a simulation study based on reference BNs from the Bayesian network repository (Scutari, 2012), whose conclusions will then be confirmed using real-world

climate data in Section 4. Both will be implemented using the *bnlearn* (Scutari, 2010) and *catnet* (Balov and Salzman, 2017) R packages and TETRAD (Landsheer, 2010).

We assess three constraint-based algorithms (PC, GS, Inter-IAMB), two score-based algorithms (tabu search, simulated annealing for BIC, GES for $\log \mathrm{P}(\mathcal{G} \mid \mathcal{D})$) and two hybrid algorithms (MMHC, RSMAX2) on the networks in Table 1. For each BN:

1. We generate 20 samples of size $n/|\Theta| = 0.1, 0.2, 0.5, 1.0, 2.0$, and 5.0 to allow for meaningful comparisons between different BNs.

2. We learn $\mathcal{G}$ using (BIC, $\mathrm{G}^2_{\mathrm{BIC}}$), and ($\log \mathrm{P}(\mathcal{G} \mid \mathcal{D})$, $\log \mathrm{BF}$) as well for discrete BNs. For the latter we use the BDeu score (Heckerman et al., 1995) with a prior probability of inclusion of $1/(N-1)$ for each parent of each node, which is the default in TETRAD.

3. We measure the accuracy of the learned DAGs using the Structural Hamming Distance (SHD; Tsamardinos et al., 2006) from the reference BN scaled by the number of arcs $|A|$ of that BN (lower is better); and we measure the speed of the learning algorithms with the number of calls to the statistical criterion.

The results for (BIC, $\mathrm{G}^2_{\mathrm{BIC}}$) and the discrete BNs are illustrated in Figure 1 for small samples ($n/|\Theta| < 1$) and large samples ($n/|\Theta| \geqslant 1$); results from ($\log \mathrm{P}(\mathcal{G} \mid \mathcal{D})$, $\log \mathrm{BF}$) are very similar and are not discussed separately for brevity. We find that 1) tabu search and simulated annealing have the highest SHDs for small samples, while tabu search has the lowest SHD for large samples, for 10/10 BNs; 2) the SHD of hybrid algorithms is comparable to that of constraint-based algorithms for all sample sizes and BNs; 3) the SHD of constraint-based algorithms is comparable to or better than that of score-based algorithms for small sample sizes in 7/10 BNs, but it decreases more slowly as n increases for all BNs. As for speed, while simulated annealing is consistently slower than other algorithms, tabu search is in the bottom left panel ("fast, accurate") for 10/10 BNs in large samples and for 6/10 BNs in small samples.

The corresponding results for GBNs are shown in Figure 2, and confirm that for all BNs 1) tabu search and simulated annealing have a larger SHD than constraint-based or hybrid algorithms for small samples; 2) the SHD of hybrid and constraint-based algorithms is not markedly different at different sample sizes. With the exception of simulated annealing, all algorithms have very similar SHD for all large samples. However, neither tabu search nor simulated annealing achieve lower SHD than constraint-based or hybrid algorithms regardless of the sample size.

4. Real-World Climate Data

Climate networks have recently attracted a great deal of interest due to their potential to analyse the complex spatial structure of climate data. This includes spatial dependence among nearby lo-

discrete BN	N	$\lvert A\rvert$	$\lvert\Theta\rvert$	discrete BN	N	$\lvert A\rvert$	$\lvert\Theta\rvert$	GBN	N	$\lvert A\rvert$	$\lvert\Theta\rvert$
ALARM	37	46	509	MUNIN1	186	273	15622	ARTH150	107	150	364
ANDES	223	338	1157	PATHFINDER	135	200	77155	ECOLI72	46	70	162
CHILD	20	25	230	PIGS	442	592	5618	MAGIC-IRRI	64	102	230
HAILFINDER	56	66	2656	WATER	32	66	10083	MAGIC-NIAB	44	66	154
HEPAR2	70	123	1453	WIN95PTS	76	112	574				

Table 1: Reference BNs with their numbers of nodes (N), arcs ($|A|$) and parameters ($|\Theta|$).

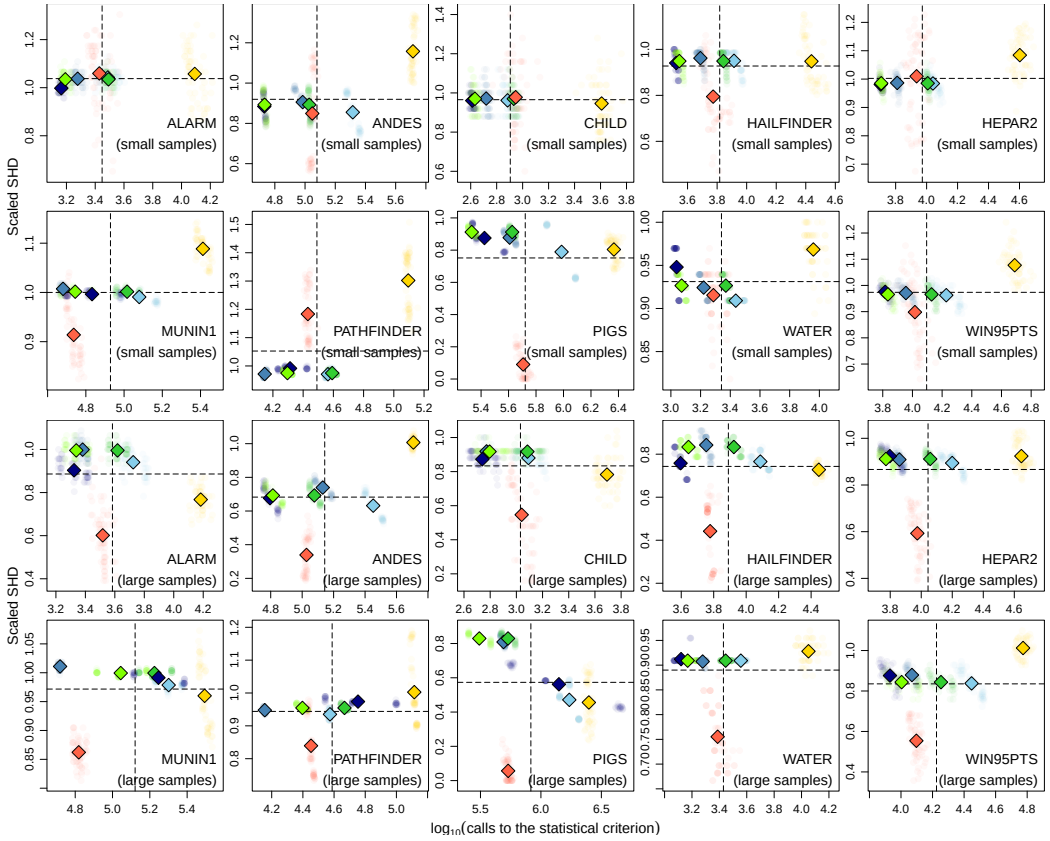

Figure 1: Scaled SHD versus speed for GS (blue), Inter-IAMB (sky blue), PC (navy blue), MMHC (green), RSMAX2 (lime green), tabu search (red) and simulated annealing (gold) and (BIC, G^2_{BIC}) for the discrete BNs. Shaded points correspond to individual simulations, while diamonds are algorithm averages. The four quadrants in each panel correspond to "fast, inaccurate" (top left), "slow, inaccurate" (top right), "slow, accurate" (bottom right) and "fast, accurate" (bottom, left) algorithms with respect to the overall mean performance in the panel.

cations, but also long-range spatial dependencies connecting distant regions in the world, known as *teleconnections* (Tsonis et al., 2008). These teleconnections represent large-scale oscillation patterns—such as the El Niño Southern Oscillation (ENSO)—which modulate the synchronous behaviour of distant regions (Yamasaki et al., 2008). The most popular climate network models are *complex networks* (Tsonis et al., 2006), which are easy to build since they are based on pairwise correlations (arcs are established between pairs of stations with correlations over a given threshold) and provide topological information in the network structure (*e.g.* highly connected regions). Bayesian networks have been proposed as an alternative methodology for climate networks that can model both marginal and conditional dependence structures and that allows probabilistic inference (Cano et al., 2004). However, learning such networks is computationally demanding and choosing an appropriate structure learning algorithm is crucial. Here we consider an illustrative case study

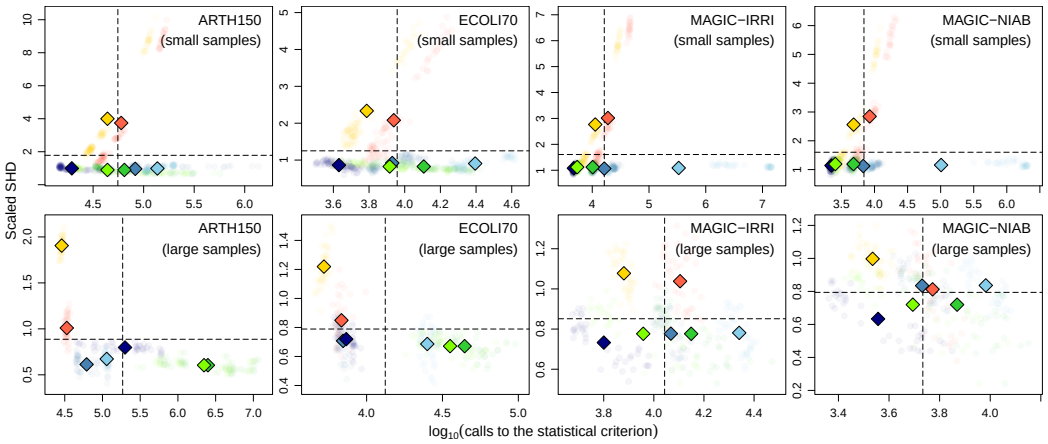

Figure 2: Scaled SHD versus speed using (BIC, G^2_{BIC}) for GBNs, formatted as in Figure 1.

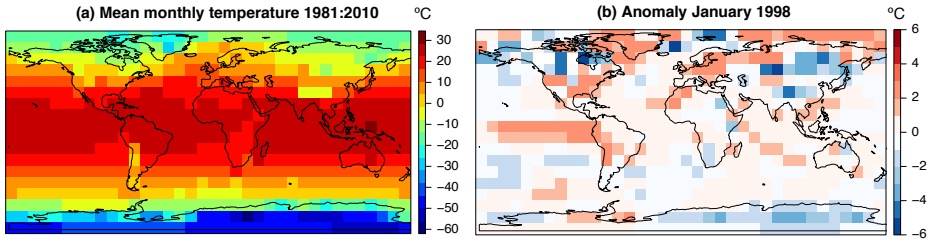

Figure 3: (a) Global mean temperature for 1981 to 2010 on a global 10° grid from the NCEP reanalysis. (b) Anomaly for January 1998 (strong El Niño episode).

modelling global surface temperature and we reassess the performance of the different learning methods we used in Section 3.

4.1 Data and Methods

We use monthly surface temperature values on a global 10°-resolution (approx. 1000 km) regular grid for a representative climatic period (1981 to 2010), as provided by the NCEP/NCAR reanalysis[1]. Figure 3 shows the mean temperature (climatology) for the whole period as well as the anomaly (difference from the mean climatological values) for a particular date (January 1998, from a strong El Niño episode with high tropical Pacific temperatures).

The surface temperature at each gridpoint is assumed to be normally distributed; hence we construct GBNs from the data in which nodes represent the (anomaly of) surface temperature at different gridpoint and arcs represent spatial dependencies. Thus, we define X_i as the monthly anomaly value of the temperature at location i for a period of 30 years ($n = 30 \times 12 = 360$). The anomaly value is obtained by removing the mean annual cycle (*i.e.* the 30-year mean monthly

1. `https://www.esrl.noaa.gov/psd/data/gridded/data.ncep.reanalysis.html`

values) from the raw data. The location of a gridpoint i is defined by its latitude and longitude. Hence the node set $\mathbf{X}$ in the corresponding network is characterised as $\mathbf{X} = \{X_1, \ldots, X_N\}$ with $N = 18 \times 36 = 648$.

Similarly to Section 3, we assess two constraint-based algorithms (PC, GS), two score-based algorithms (tabu search and hill climbing, HC) and one hybrid algorithm (MMHC). Note, however, that in this case the sample size is fixed to what was considered a "small sample" even for a DAG with no arcs: $n/|\Theta| \leqslant 360/(648 \times 2) = 0.28$.

In order to construct an appropriate pair of matching criteria, we introduce an extended version of BIC in which we introduce a regularisation parameter γ that penalises the number of parameters. We refer to this score as BIC_γ, with $\mathrm{BIC}_\gamma = \mathrm{BIC}$ if $\gamma = 0$, defined as

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

From BIC_γ we then construct the corresponding independence test $\mathrm{G}^2_{\mathrm{BIC}_\gamma}$ as follows:

$$\mathrm{BIC}_\gamma(\mathcal{G}^+; \mathcal{D}) > \mathrm{BIC}_\gamma(\mathcal{G}^-; D) \Rightarrow 2 \log \frac{\mathrm{P}(X_i \,|\, \Pi_{X_i} \cup \{X_j\})}{\mathrm{P}(X_i \,|\, \Pi_{X_i})} > (|\Theta^{\mathcal{G}^+}_{X_i}| - |\Theta^{\mathcal{G}^-}_{X_i}|)(2\gamma \log N + \log n).$$

The additional regularisation in $\mathrm{G}^2_{\mathrm{BIC}_\gamma}$ is required to make constraint-based algorithms reliable; using $\mathrm{G}^2_{\mathrm{BIC}}$ with climate data often results in learning graphs that are not valid CPDAGs. For all algorithms (BIC_γ, $\mathrm{G}^2_{\mathrm{BIC}_\gamma}$) allow us to obtain graphs of comparable size $n/|\Theta|$. We have chosen to scale γ with the factor $2|\Theta^{\mathcal{G}^+}_{X_i}| - |\Theta^{\mathcal{G}^-}_{X_i}| \log p$ as in the EBIC score from Chen and Chen (2012) due to its effectiveness in feature selection. We refer to the range of γs in which an algorithm can return directed acyclic graphs (DAGs) as the *parameter range* of the algorithm.

Motivated by the above, we proceed as in Section 3 but with the following changes:

1. We generate 5 permutations of the order of the variables in the data to cancel local preferences in the learning algorithms (see *e.g.* Colombo and Maathuis, 2014).

2. From each permutation, we learn $\mathcal{G}$ using (BIC, $\mathrm{G}^2_{\mathrm{BIC}}$) as well as ($\mathrm{BIC}_\gamma$, $\mathrm{G}^2_{\mathrm{BIC}_\gamma}$) for different values of $\gamma \in (0, 50]$.

3. Since we do not have a "true" model to use as a reference, we measure the accuracy of learned BNs along the parameter range of the algorithm by their log-likelihood. We also analyse the long-distance arcs (teleconnections) established by the DAGs and assess their suitability for probabilistic inference by testing the conditional probabilities obtained when introducing some El Niño related evidence.

4.2 Results

Figure 4 shows the performance (speed, performance and number of arcs) of various structure learning algorithms as a function of γ, using the same colours as in Figure 1 (with the exception of hill climbing, which is new in this figure and it is shown in orange). Figure 5 (a-b) shows the resulting graphs for the representative network from MMHC in Figure 4c and a comparable intermediate

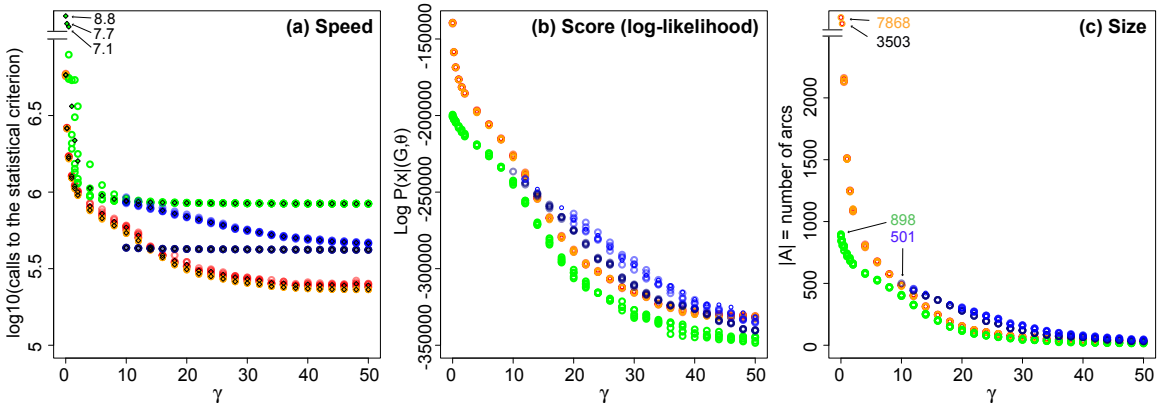

Figure 4: (a) Speed, (b) performance (log-likelihood), (c) number of arcs for different values of γ, learned by GS (blue), PC (navy), MMHC (green), tabu search (red) and HC (orange). Note that orange results are on top of red ones in some cases. For clarity panel (a) includes the mean of the 5 realisation results for each γ. Labelled points in (a) have means returned by MMHC for $\gamma \in \{0, 0.2, 0.5\}$ that are in speed-range higher than 7.0. Labelled points in (c) represent the biggest networks of tabu for $\gamma \in \{0, 0.2\}$ and the biggest networks found by MMHC and PC (to be analysed in Figure 5).

network of tabu search. This figure also compares the suitability of the learned BNs for probabilistic inference by propagating an El Niño-like evidence ($V_{81} = 2$, *i.e.* warm temperatures in the corresponding gridbox in tropical Pacific).

Constraint-based GS and PC produce BNs with the highest log-likelihood in the high parameter penalisation region ($\gamma \geq 10$). However, they do not produce valid DAGs for low parameter penalisation ($\gamma < 10$), yielding a maximum number of 501 arcs (smaller than the number of nodes) with no large arcs representing teleconnections when $\gamma \geq 10$. MMHC exhibits the poorest log-likelihood values and produces a maximum number of 898 arcs, including only a few teleconnections (Figure 5b). The absence of a sufficient number of teleconnections makes both unsuitable for propagating evidence (Figure 5d). Therefore, tabu search and HC (with almost identical results) produce the best results, with large networks (with over 2500 arcs for $\gamma \leq 0.2$) and high likelihood values. In this case, even intermediate networks (with around 1500 arcs) include a large number of teleconnections and allow propagating evidences with realistic results (Figures 5a and c).

Finally, we find that score-based algorithms are faster than both hybrid and constraint-based algorithms. The difference in speed is relatively amplified compared to MMHC for $\gamma \in \{0, 0.2, 0.5, 1, 1.5, 2\}$ accounting for the fact that in this region the score-based algorithms return DAGs containing more edges than MMHC for the same γ.

5. Conclusions

In this paper we revisited the problem of assessing different classes of BN structure learning algorithms; we improved over existing comparisons of learning accuracy and speed in the literature by removing the confounding effect of different choices of statistical criteria. Interestingly, we found

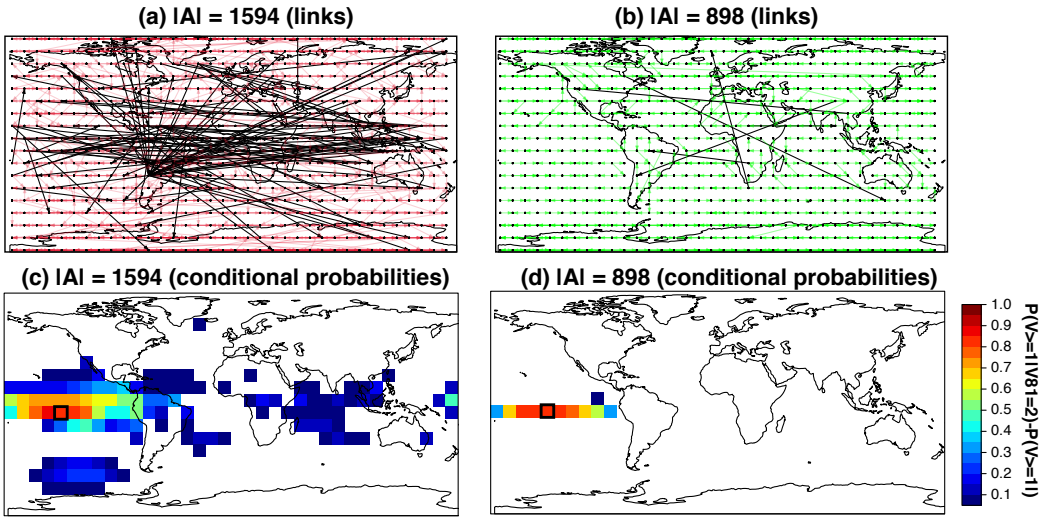

Figure 5: DAGs learned by (a) tabu search (intermediate) and (b) MMHC with $\gamma = 0$. Teleconnections are shown in black. (c) and (d) show the differences of the conditional and marginal probabilities obtained with both Bayesian networks after propagation of $V_{81} = 2$ (denoted with a black box), simulating El Niño conditions.

that constraint-based algorithms are more accurate than score-based algorithms for small sample sizes (**Q1**); and that they are as accurate as hybrid algorithms (**Q2**). We also found that tabu search, as a score-based algorithm, is faster than constraint-based algorithms more often than not (**Q3**). For climate data we found that score-based algorithms produce the largest networks allowing good propagation of evidence. These results, which we confirmed on both simulated data and real-world climate data, are intended to provide guidance for additional studies; we do not exclude the existence of other sources of confounding, such as tuning parameters, which should be further investigated.

Acknowledgments

CEG and JMG were supported by the project MULTI-SDM (CGL2015-66583-R, MINECO/FEDER).

References

N. Balov and P. Salzman. *catnet: Categorical Bayesian Network Inference*, 2017. R package version 1.15.3.

R. R. Bouckaert. *Bayesian Belief Networks: from Construction to Inference*. PhD thesis, Utrecht University, The Netherlands, 1995.

R. Cano, C. Sordo, and J. M. Gutiérrez. Applications of Bayesian Networks in Meteorology. In J. A. Gámez, S. Moral, and A. Salmerón, editors, *Advances in Bayesian Networks*, pages 309–

328. Springer, 2004.

J. Chen and Z. Chen. Extended BIC For Small-n-Large-p Sparse GLM. *Statistica Sinica*, 22(2): 555–574, 2012.

D. M. Chickering. A Transformational Characterization of Equivalent Bayesian Network Structures. In P. Besnard and S. Hanks, editors, *Proceedings of the 11th Conference on Uncertainty in Artificial Intelligence*, pages 87–98. Morgan Kaufmann, 1995.

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.

F. Cugnata, R. S. Kenett, and S. Salini. Bayesian Networks in Survey Data: Robustness and Sensitivity Issues. *Journal of Quality Technology*, 4(3):253–264, 2016.

G. Elidan. Copula Bayesian Networks. In J. D. Lafferty, C. K. I. Williams, J. Shawe-Taylor, R. S. Zemel, and A. Culotta, editors, *Advances in Neural Information Processing Systems 23*, pages 559–567, 2010.

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.

D. Koller and N. Friedman. *Probabilistic Graphical Models: Principles and Techniques*. MIT Press, 2009.

J. A. Landsheer. The Specification of Causal Models with Tetrad IV: A Review. *Structural Equation Modeling*, 17(4):703–711, 2010.

P. Larrañaga, B. Sierra, M. J. Gallego, M. J. Michelena, and J. M. Picaza. Learning Bayesian Networks by Genetic Algorithms: A Case Study in the Prediction of Survival in Malignant Skin Melanoma. In *Proceedings of the 6th Conference on Artificial Intelligence in Medicine in Europe (AIME'97)*, pages 261–272. Springer, 1997.

S. L. Lauritzen and N. Wermuth. Graphical Models for Associations Between Variables, Some of which are Qualitative and Some Quantitative. *The Annals of Statistics*, 17(1):31–57, 1989.

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. Learning Bayesian Networks with the bnlearn R Package. *Journal of Statistical Software*, 35(3):1–22, 2010.

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.

A. A. Tsonis, K. L. Swanson, and P. J. Roebber. What Do Networks Have to Do with Climate? *Bulletin of the American Meteorological Society*, 87(5):585–595, 2006.

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.

T. S. Verma and J. Pearl. Equivalence and Synthesis of Causal Models. *Uncertainty in Artificial Intelligence*, 6:255–268, 1991.

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.