An Experimental Study of Prior Dependence in Bayesian Network Structure Learning

ISIPTA 2019 - Ghent, Belgium

Alvaro H. C. Correia Cassio P. de Campos Linda C. van der Gaag

Department of Information and Computing Sciences - Utrecht University

4th of July 2019

Background

Bayesian Networks

Probabilistic graphical models based on a directed acyclic graph (DAG) G, where each node represents a random variable in $\mathbf{X} = \{X_1, ..., X_N\}$.

Markov property

The local distribution of each node X_i depends only on the values of its parents $\Pi_{X_i}^G$ in G.

$$P(\mathbf{X}|G) = \prod_i^N P(X_i|\Pi_{X_i}^G)$$

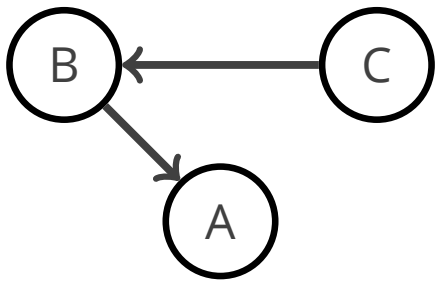
Parameters

We refer to all parameters specifying probability distributions $P(X_i|\Pi_{X_i}^G)$ as Θ_G.

Background

Bayesian Networks

A simple example of a Bayesian network with 3 variables.



G

And the joint distribution it induces.

$$P(A, B, C|G) = P(C)P(B|C)P(A|B)$$

Structure Learning

Bayesian Network Structure Learning

To learn the structure of a Bayesian network from data D.

Structure Learning

Score-Based Bayesian Network Structure Learning

To learn the structure of a Bayesian network from data D by searching for the graph G that maximises a given fitness score.

Structure Learning

BDeu Score-Based Bayesian Network Structure Learning

To learn the structure of a Bayesian network from data by searching for the graph G that maximises the posterior $P(G|D)$.

$$P(G|D) \propto P(G) \int P(D|G, \Theta_G) P(\Theta_G|G) d\Theta_G$$

Structure Learning

BDeu Score-Based Bayesian Network Structure Learning

To learn the structure of a Bayesian network from data by searching for the graph G that maximises the posterior $P(G|D)$.

$$P(G|D) \propto P(G) \int P(D|G, \Theta_G) \boxed{P(\Theta_G|G)} d\Theta_G$$

We focus on the influence of this prior on the learned structure.

BDeu

Bayesian Dirichlet equivalent uniform

With some assumptions (including complete discrete data and Dirichlet priors), we can write $P(G|D)$ in closed form and derive the BDeu score.[1]

$$\text{BDeu}(G, \alpha) = \prod_{i=1}^{n} \prod_{j=1}^{q_i} \frac{\Gamma(\alpha_{ij})}{\Gamma(\alpha_{ij} + N_{ij})} \prod_{k=1}^{r_i} \frac{\Gamma(\alpha_{ijk} + N_{ijk})}{\Gamma(\alpha_{ijk})},$$

[1] D. Heckerman, D. Geiger, and D. M. Chickering. "Learning Bayesian Networks : The Combination of Knowledge and Statistical Data". In: *Machine Learning* 20 (1995), pp. 197–243.

BDeu

$$\text{BDeu}(G, \alpha) = \prod_{i=1}^{n} \prod_{j=1}^{q_i} \frac{\Gamma(\alpha_{ij})}{\Gamma(\alpha_{ij} + N_{ij})} \prod_{k=1}^{r_i} \frac{\Gamma(\alpha_{ijk} + N_{ijk})}{\Gamma(\alpha_{ijk})},$$

where for a variable i

r_i is its arity

q_i is the number of joint instantiations of its parents

N_{ijk} is the number of observations of full instantiation ijk

$N_{ij} = \sum_k N_{ijk}$

$\alpha_{ijk} = \alpha/(r_i q_i)$ and $\alpha_{ij} = \alpha/q_i$

BDeu

$$\text{BDeu}(G, \alpha) = \prod_{i=1}^{n} \prod_{j=1}^{q_i} \frac{\Gamma(\alpha_{ij})}{\Gamma(\alpha_{ij} + N_{ij})} \prod_{k=1}^{r_i} \frac{\Gamma(\alpha_{ijk} + N_{ijk})}{\Gamma(\alpha_{ijk})},$$

where for a variable i

r_i is its arity

q_i is the number of joint instantiations of its parents

N_{ijk} is the number of observations of full instantiation ijk

$N_{ij} = \sum_k N_{ijk}$

$\alpha_{ijk} = \alpha/(r_i q_i)$ and $\alpha_{ij} = \alpha/q_i$

α is the Equivalent Sample Size (ESS)

Prior Dependence

Impact of the ESS in the final structure

It is well known that ESS has a large influence on the final structure.[1]

[1] T. Silander, P. Kontkaken, and P. Myllymaki. “On sensitivity of the map Bayesian network structure to the equivalent sample size parameter”. In: *Proceedings of the 23rd Conference on Uncertainty in Artificial Intelligence (UAI-07)*. 2007, pp. 360–367.

Prior Dependence

Impact of the ESS in the final structure

It is well known that ESS has a large influence on the final structure.[1]

Bayesian score

For large enough data, one should learn the same network regardless of the prior knowledge expressed via the ESS.

We investigate whether that holds for BDeu-based structure learning.

[1] Silander, Kontkaken, and Myllymaki, "On sensitivity of the map Bayesian network structure to the equivalent sample size parameter".

Prior Dependence

Impact of the ESS in the final structure

We analyse the influence of the ESS from two different angles:

Graph Complexity

How variations in the ESS affect the number of arcs in the structure.

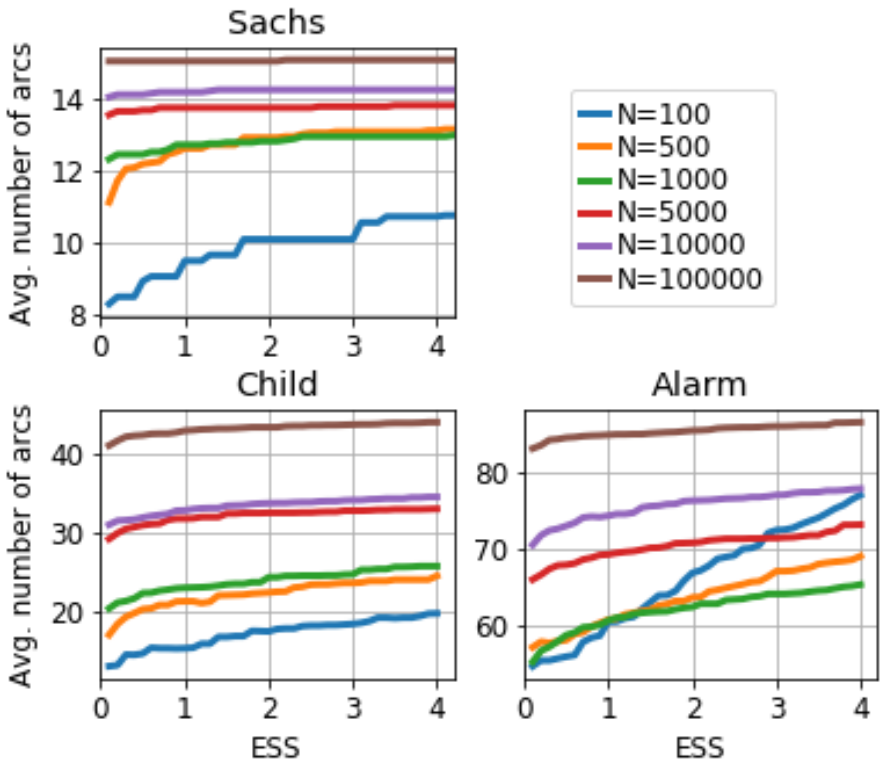
Robustness

What conditions are required for prior-independence.

Graph Complexity

Variation of the number of arcs with ESS and sample size

- The number of arcs increase with the ESS.
- Prohibitive large amount of data to achieve prior independence.



Robustness

Definition (Robust Interval)

The largest range of ESS values for which all obtained optimal structures (for each ESS) are Markov equivalent.

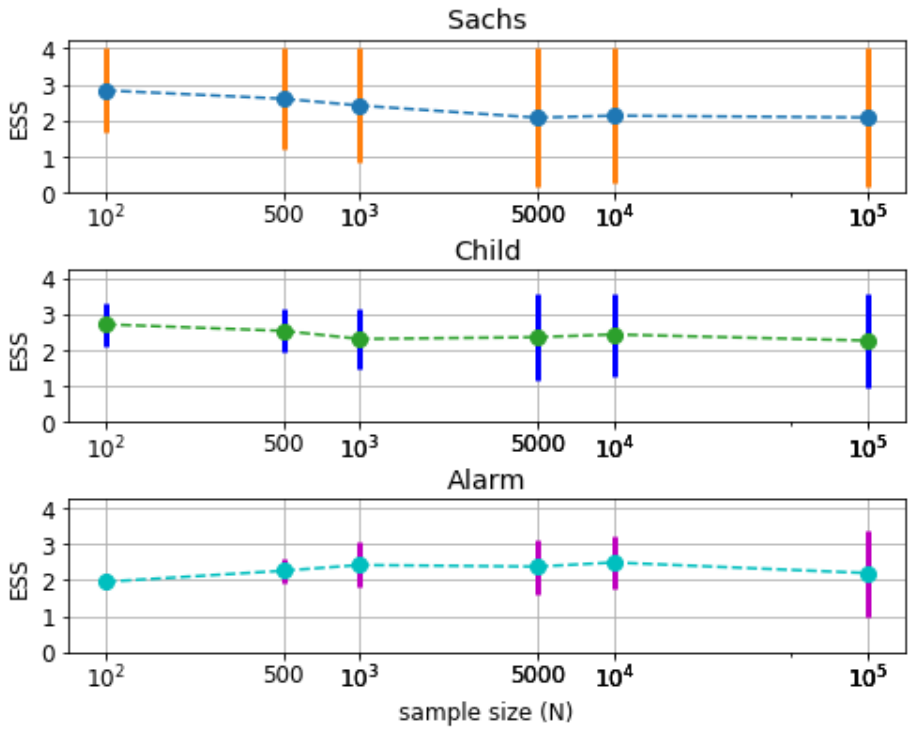
$$RI := \underset{[\alpha_1,\alpha_2]}{\arg\max}\{|\alpha_2 - \alpha_1| : G^*(\alpha') \equiv G^*(\alpha''), \forall \alpha', \alpha'' \in [\alpha_1, \alpha_2]\},$$

where $G^*(\alpha) = \arg\max_G$ *BDeu*(G, α) *is the optimal graph for a given ESS, and* $\equiv$ *denotes Markov equivalence.*

Robust Interval

Variation of the RI with ESS and sample size

- The RI increases with the sample size.
- Prohibitive large amount of data to cover small ESS range.



UCI Datasets

RI for datasets with no known distribution

- Insufficient data to cover small ESS range.
- Similar observations for graphs learned with and without ordering constraints.

Dataset	n	N	RIo	RIf
car	7	1728	(0.1, 4.0)	(0.4, 4.0)
glass	8	214	(1.3, 2.3)	(0.3, 4.0)
spambase	8	4601	(1.2, 4.0)	(1.7, 4.0)
diabetes	9	768	(0.2, 1.7)	(1.6, 4.0)
nursery	9	12960	(1.4, 2.9)	(1.4, 4.0)
breast-cancer	10	286	(1.9, 4.0)	(2.2, 4.0)
tic-tac-toe	10	958	(1.8, 2.1)	(1.7, 2.2)
cmc	10	1473	(1.7, 2.9)	(0.8, 2.8)
heart-h	12	294	(0.8, 1.6)	(2.2, 2.9)
vowel	14	990	(0.6, 1.8)	(1.9, 4.0)
zoo	17	101	(0.6, 1.3)	(0.9, 2.1)
vote	17	435	(0.8, 1.8)	(2.3, 3.1)
segment	17	2310	(1.5, 2.9)	(2.3, 4.0)
primary-tumor	18	339	(1.1, 1.5)	(3.1, 3.5)
vehicle	19	846	(0.9, 1.7)	(3.3, 4.0)

Available data is likely insufficient to avoid prior dependence in BDeu-based Structure Learning.