An Experimental Study of Prior Dependence in Bayesian Network Structure Learning

Alvaro H. C. Correia A.H.CHAIMCORREIA@UU.NL
Cassio P. de Campos C.DECAMPOS@UU.NL
Linda C. van der Gaag L.C.VANDERGAAG@UU.NL
Department of Information and Computing Sciences, Utrecht University, The Netherlands

Abstract

The Bayesian Dirichlet equivalent uniform (BDeu) function is a popular score to evaluate the goodness of a Bayesian network structure given complete categorical data. Despite its interesting properties, such as likelihood equivalence, it does require a prior expressed via a user-defined parameter known as Equivalent Sample Size (ESS), which significantly affects the final structure. We study conditions to obtain prior independence in BDeu-based structure learning. We show in experiments that the amount of data needed to render the learning robust to different ESS values is prohibitively large, even in big data times.

Keywords: Robustness, Bayesian Networks, Structure Learning, BDeu

1. Introduction

Bayesian networks are a class of probabilistic graphical models based on a Directed Acyclic Graph (DAG) G that defines a factorisation of the joint probability distribution over a set of variables $X = \{x_1, ..., x_i, ..., x_n\}$. One can learn the DAG (also called structure) G from complete categorical data D via the popular Bayesian Dirichlet equivalent uniform (BDeu) score function [3, 5], which aims at finding a *maximum a posteriori* (MAP) G that maximises $P(G|D)$ (under uniform prior for G). Under these assumptions, the BDeu score for G is defined by the marginal likelihood of the data D given G and the ESS (denoted here by $\alpha > 0$):

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

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 with instantiation jk of its parents and itself, and $N_{ij} = \sum_k N_{ijk}$. Finally, $\alpha_{ijk} = \alpha/(r_i q_i)$ and $\alpha_{ij} = \alpha/q_i$.

Structure learning with BDeu requires the definition of the Dirichlet parameters, which is done through $\alpha > 0$, roughly expressing the strength of our prior belief. However, there is no consensus on what value an 'uninformative' α should take and several studies have focused on measuring the influence of α on the final structure [8, 9, 11], analysing the asymptotic behaviour of the BDeu for $\alpha \to 0$ and $\alpha \to \infty$ [13, 14, 16, 17], or finding the optimal α [11, 13].

Nonetheless, to the best of our knowledge, no work has directly addressed the robustness of BDeu-based structure learning to variations of the ESS. By robustness we mean prior-independence, i.e., for large enough data, one should expect the structure learning algorithm to produce the same network regardless of the prior knowledge expressed via the ESS. As we show in the experiments, even for a small number of variables the amount of data required to achieve such robustness is prohibitively large. That suggests the prior on the BDeu function might be too strong for some real-world applications, where other scores (or some variation of the BDeu score) might be more adequate.

2. Experiments

We conducted experiments with three known Bayesian networks [2, 7, 12] and 16 datasets from the UCI Machine Learning repository [6] to study the influence of the ESS. In all experiments, we assumed complete categorical data (we discretised continuous variables into two categories by their median values, when needed). As we wanted to study the intrinsic behaviour of BDeu-based structure learning, and not the particularities of a given approximate solver, we focused on exact solutions. For that, there are multiple exact solvers [4, 10, 18], and we used GOBNILP [1], which finds the optimal graph via integer linear programming.

In the experiments in Figure 1 and 2, we assumed a given ordering of the variables, i.e., for any nodes X and Y, if X precedes Y in the ordering, then an arc between X and Y (if it exists) must be directed from X to Y. That restriction considerably reduces the search space and allows us to consider larger sample sizes, while still guaranteeing an exact solution. Conversely, the UCI datasets are small enough that we could gather exact results both with and without order constraints.

Graph Complexity. The ESS can be interpreted as a regularizer on the structure of the Bayesian network. Hence, we start by investigating how it affects the total number of arcs (parents) in the network. In particular, we are interested

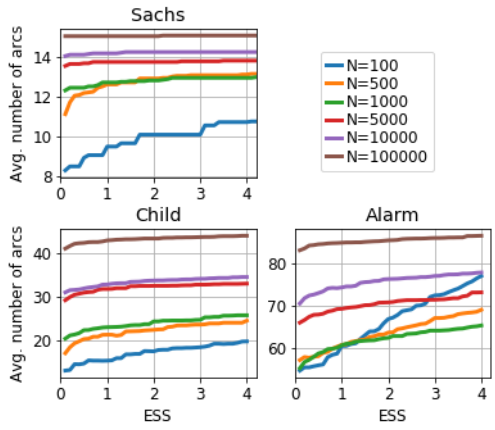

Figure 1: Average number of arcs as a function of the ESS for different N. Scale of the graphs varies.

in the interplay between the *sample size* (N) and the ESS in determining the complexity of the network.

We sampled data from three known networks, Sachs ($n = 11$) [7], Child ($n = 20$) [12] and Alarm ($n = 37$) [2] and learned the structure while varying the sample size (N) and the ESS. We repeated that process for 30 different orderings and reported the average number of arcs (across orderings) in Figure 1. The results indicate the graph increases in complexity with the ESS. Indeed, the number of arcs is expected to grow *almost* monotonically to the maximum (complete graph) for large values of α [11, 13, 14, 16, 17].

A more interesting analysis that received little attention in the literature is how the complexity of the graph varies with the sample size. Naturally, one would expect that, for large datasets, the prior would have little effect on the learned network. That is what we observe for the Sachs network in Figure 1: the number of arcs remains constant across all ESS values for $N \geq 10^5$. However, considering Sachs contains only 11 variables, that is an extremely high number of data points to guarantee robustness over a relatively small range of ESS values. For the other networks, no amount of data ensured prior-independence. The number of arcs increased with the ESS, and providing more data points did not alleviate this trend significantly.

In a statistical sense, the ESS is not a typical Dirichlet prior because it also defines the number of parameters in the model. It expresses a trade-off between regularisation and complexity [14], which increases with the ESS and with N, as shown in the experiments. That trade-off partially explains why it is hard to avoid prior-dependence in BDeu-based structure learning.

Robustness To study prior-independence in BDeu-based structure learning, one needs a metric that captures the influence of the ESS on the final structure.

Definition 1 (Robust Interval) *is defined by the largest range of ESS values for which all obtained optimal structures (for each ESS) are Markov equivalent.*

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

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

Intuitively, the larger the Robust Interval (RI), the more prior-independent the learning algorithm (for a given dataset). We do not distinguish structures representing the same set of conditional independence statements (Markov equivalent), as they encode the same 'information' and have the same BDeu score [5]. The advantage of the RI against other metrics, such as structural Hamming distance (SHD) [15], is that it does not require a gold standard network and also signals a safe range of ESS values over which the influence of the prior is mitigated.

Note that the RI is only meaningful if reported for non-complete graphs. For $\alpha \to \infty$, the learned structure tends to a complete graph [16], and it follows that, for large enough α, the structure is 'infinitely' robust but overfitted, which is an uninteresting result. In the experiments, we computed the RI by finding the optimal structure with α covering the range $(0.1, 4.0)$ in increments of 0.1. By using increments we do not obtain the true RI but a conservative estimate: the true RI is either smaller (due to unobserved variations in-between increments) or at most 0.2 larger.

We report the results in Figure 2, where for each dataset and for each pair (α, N), we see the average RI of 30 randomly sampled orderings.

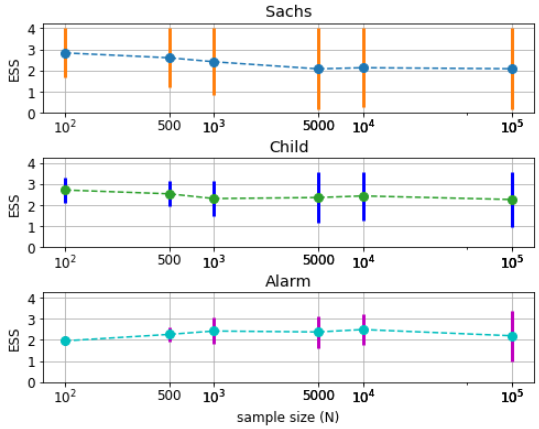

Figure 2: Robust Interval for Sachs [7], Child [12] and Alarm [2] in function of the sample size (N).

In a Bayesian framework, we want the prior to become less relevant in determining the final structure of the network as we gather more data. This in turn should result

in larger robust intervals. In Figure 2, we see the BDeu does comply with that requirement to some extent, as the robust interval does increase with N. However, the amount of data required to cover the small ESS range we analysed is already prohibitively large even for a small number of variables. That supports recent studies that claim the BDeu is not fit for sparse data [8, 9], but also alerts us that almost every real-world dataset might be too sparse for the BDeu.

Table 1: Largest ESS range yielding the same structure (RI) for UCI datasets. N and n are the number of samples and variables, RIo the average RI of 10 orderings, and RIf the RI without order constraint.

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)

We did the same analysis for 16 UCI datasets [6]. In these experiments, we computed both the average RI of 10 different orderings (RIo) and the RI with no constraint on the ordering (RIf). Again, in Table 1, we see that except for the *car* dataset, none of them had enough data to guarantee robustness of the BDeu-based structure learning with $\alpha \in (0.1, 4.0)$. Interestingly, the size of the interval did not change significantly between solutions with and without a order constraint, but the RI stabilised at slightly higher ESS values when no ordering was given.

The RI can also be seen as an indication of a safe interval at which the influence of the prior is minimal. However, for more than half of the datasets, the robust interval did not includes the canonical $\alpha = 1$. That contrasts with previous studies suggesting the influence of the ESS on the learned structure is minimised when it is set to one [16].

All in all, the results support previous research in confirming the BDeu is highly sensitive to the ESS. That is crucial when one wants to study the graph per se but may also impact the predictive power of the models. Therefore, one must be aware and accept the large influence the prior may have when using BDeu, since the amount of data will likely be insufficient to avoid prior dependence. Future work will extend the analysis to parameter learning to investigate further the overall impact of the ESS on the learned models.

References

[1] M. Bartlett and J. Cussens. Advances in Bayesian Network Learning using Integer Programming. *Proceedings of the 29th Conference on Uncertainty in Artificial Intelligence (UAI-13)*, pages 182–191, 2013.

[2] 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. Springer-Verlag, 1989.

[3] W. Buntine. Theory Refinement on Bayesian Networks. *Proceedings of the 7th Conference on Uncertainty in Artificial Intelligence (UAI-91)*, pages 52–60, 1991.

[4] C. P de Campos and Q. Ji. Efficient structure learning of Bayesian networks using constraints. *Journal of Machine Learning Research*, 12:663–689, 2011.

[5] D. Heckerman, D. Geiger, and D. M. Chickering. Learning Bayesian Networks : The Combination of Knowledge and Statistical Data. *Machine Learning*, 20:197–243, 1995.

[6] M. Lichman. UCI machine learning repository, 2013. URL http://archive.ics.uci.edu/ml.

[7] K. Sachs, O. Perez, D. Pe, D. Lauffenburger, and G. Nolan. Causal protein-signaling networks derived from multiparameter single-cell data. *Science*, 308 (5721):523–529, 2005.

[8] M. Scutari. An Empirical-Bayes Score for Discrete Bayesian Networks. *Journal of Machine Learning Research (Proceedings Track, PGM 2016)*, 52:438–448, 2016.

[9] M. Scutari. Dirichlet bayesian network scores and the maximum relative entropy principle. *Behaviormetrika*, 45(2):337–362, Oct 2018.

[10] T. Silander and P. Myllymäki. A simple approach for finding the globally optimal Bayesian network structure. In *Proceedings of the 22nd Conference on Uncertainty in Artificial Intelligence (UAI-06)*, pages 445–452, 2006.

[11] 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)*, pages 360–367, 2007.

[12] D. J. Spiegelhalter, A. P. Dawid, S. L. Lauritzen, and R. G. Cowell. Bayesian Analysis in Expert Systems. *Statistical Science*, 8(3):219–247, 1993.

[13] H. Steck. Learning the Bayesian Network Structure: Dirichlet Prior versus Data. In *Proceedings of the 24th Conference on Uncertainty in Artificial Intelligence (UAI-08)*, pages 511–518, 2008.

[14] H. Steck and T. S. Jaakkola. On the Dirichlet Prior and Bayesian Regularization. *Advances in Neural Information Processing Systems*, pages 713–720, 2003.

[15] 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.

[16] M. Ueno. Learning networks determined by the ratio of prior and data. *Proceedings of the 26th Conference on Uncertainty in Artificial Intelligence (UAI-10)*, pages 598–605, 2010.

[17] M. Ueno. Robust learning Bayesian networks for prior belief. *Proceedings of the 27th Conference on Uncertainty in Artificial Intelligence (UAI-11)*, pages 698–707, 2011.

[18] C. Yuan and B. Malone. An improved admissible heuristic for learning optimal Bayesian networks. In *Proceedings of the 28th Conference on Uncertainty in Artificial Intelligence (UAI-12)*, pages 924–933, Catalina Island, CA, 2012.