

PAPER • OPEN ACCESS

## Improved Parameter Uniform Priors in Bayesian Network Structure Learning

To cite this article: Manxi Wang *et al* 2019 *IOP Conf. Ser.: Earth Environ. Sci.* **252** 042099

View the [article online](#) for updates and enhancements.

# Improved Parameter Uniform Priors in Bayesian Network Structure Learning

Manxi Wang<sup>1, a</sup>, Liandong Wang<sup>1, b</sup>, Zidong Wang<sup>2, c</sup>, Xiaoguang Gao<sup>2, d</sup> and Ruohai Di<sup>2, e</sup>

<sup>1</sup>State key laboratory of complex electromagnetic environment effects on electronics and information system, Luoyang, 471000, China

<sup>2</sup>Northwestern Polytechnical University, Xi'an 710000, China

<sup>a</sup>manxi\_wang@126.com, <sup>b</sup>siriusno2001@126.com, <sup>c</sup>1354560573@mail.nwpu.edu.cn,

<sup>d</sup>cxg2012@nwpu.edu.cn, <sup>e</sup>diruohai@nwpu.edu.cn

**Abstract.** Bayesian Dirichlet equivalent uniform score (BDeu) is often used in Bayesian structure learning. But it does not work well when data size is sparse because the equivalence of the prior parameter distribution isn't suit for the specific data set. To break the rules of uniform and equivalent, the paper proposes the Bayesian Dirichlet Sparse score (BDs) which change distribution of prior parameter through the all zero items in the sparse data. The circulation principle of information entropy and simulations are used to explain the reason why BDs is better than BDeu when data size is sparse. In the experiments, we also verify the stability of BDs when hyperparameters change.

## 1. Introduction

Bayesian Network (BN) which was proposed by Jude Pearl creatively is a model of Graphical Models [1]. BN expresses the causal relationship between variables through a Directed Acyclic Graph (DAG). So, in the study of BN, how to construct a network from the data is a hot topic. The algorithms of learning BN structure are divided into three categories: Independence Test Method, Score-Search Method and the hybrid Method. When the data size is complete, all of them can achieve very perfect learning results.

However, in the area of BN structure learning when data size is sparse, researchers don't make a major breakthrough. Elidan expand the data size with bootstrap sampling and use bagging to integrate the learning results from all data sets [2]. Some scholars restrict the structure learning process through adding expert constraints [3]. The expert constraints are usually expressed as the node order, the causal relationship and the existence of edges. Japanese scholar Takashi and Isozaki propose the minimum free energy principle and noisy-or-gates [4]. However, they all promote the quality of data through adding external factors to optimize the data sets but don't consider constraints in the internal data to the network structure. The prior distribution and hyperparameters can't suit the change of data size. Recently, Ueno raises a question about the uniform of prior experience, and use a non-informative Dirichlet Score to learn BN structure [5]. Steck explores the relationship between prior parameters and posterior parameters [6]. On the basis of their work, we propose a new score function which breaks the traditional uniform prior distribution in the condition of sparse data sets.



The structure of this paper is as follows: In the second section, we briefly review the process of BN structure learning. In the third section, we analyze the disadvantages of BDeu score in sparse data. In the fourth section, the principle of information entropy is used to explain the theory of improvement. Moreover, we propose a new score function BDs, which change the uniform distribution of prior parameters with the addition of data sets. In the fifth section, we explore the influence of data size to the BDeu and BDs through experiments. The simulations also prove that modified score is more stable. In the last section, we come to some conclusions and point out the direction of future research.

## 2. Bayesian Dirichlet Equivalent Uniform Score

The BDeu score function is a special case of the Dirichlet score functions. The idea of them come from the Bayes thought, which simulate the probability of all nodes in the network when the entire event does not occur by assuming an equivalent sample size (ESS). After obtaining the data, it is integrated into the previous prior distribution through the Dirichlet function to obtain the posterior distribution, and the aim of BD score is to seek the current optimal posterior distribution.

The general BD family scoring function is as follows:

$$BD(G | D) = \prod_{i=1}^n BD(X_i | Pa(X_i)) = \prod_{i=1}^n \prod_{j=1}^{q_i} \frac{\Gamma(\alpha_{ij*})}{\Gamma(\alpha_{ij*} + m_{ij*})} \prod_{k=1}^{r_i} \frac{\Gamma(\alpha_{ijk} + m_{ijk})}{\Gamma(\alpha_{ijk})} \quad (1)$$

The middle term of the equation (1) is established because of the decomposability of the score.  $m_{ijk}$  represents the amount of data in the sample that satisfies  $X_i = k, Pa(X_i) = j$ .  $m_{ij*}$  is the summation of  $m_{ijk}$  about all values of node  $X_i$ .  $\alpha_{ijk}$  is the prior parameter which represents the amount of the data in the ESS that satisfies  $X_i = k, Pa(X_i) = j$ , and  $\alpha_{ij*}$  is the result of  $\alpha_{ijk}$  summing k. There are many design methods for parameters prior. When assuming  $\alpha_{ijk} = 1, \alpha_i = r_i q_i$ , it represents the CH score. When assuming  $\alpha_{ijk} = 1/2, \alpha_i = r_i q_i / 2$ , it represents the BD score, and so on. By comparison, BDeu score is widely quoted by researchers. The form of parameters prior in BDeu score is defined as follows: (In some literatures,  $\alpha_{ijk}$  is also called hyperparameter)

$$\alpha_{ijk} = \frac{\alpha}{r_i q_i} \quad (2)$$

BDeu score has three characteristics. The first is the equivalence, which means that prior setting of the score is the same as the conditional probability table of the DAG. The second is uniformity. The uniform distribution is generated according to the maximum number of states that a single node owns, ensuring that the prior probabilities of all cases are same. The third is adjustability. Since  $\alpha$  is not determined, the score has a lot of freedom to control. But there are two problems in BDeu score. First, the score is very sensitive to the choice of hyperparameter, even if it is not very large (after taking 20), the learning result may be severely over-fitting. Second, in the case of sparse data, the learned network seriously deviates from the actual situation. These two problems stem from the relationship between  $\alpha$  and  $D$ .

## 3. Bayesian Dirichlet Sparse Score

The section will discuss the solution when data is sparse. We would use BDs score instead of BDeu score. The main idea of the solution is to reduce the upper bound of the parameter.

On the other words, the number of prior parameters is reduced when the data is sparse, and the decrease is relative to the previous complete prior parameters. However, if the restriction on the prior parameter is relaxed and the reduced prior parameter is used as the complete prior parameter, then the posterior parameters are modified in the same way, so that the evaluation index will be the same as when

the data is complete (There is no larger number of all zero lines). The result will be close to the true score. The implementation through the idea of the entropy is as follows:

The entropy of random variables is defined as:

$$H(X) = -\sum_X P(X) \log P(X) \quad (3)$$

Initially, the maximal likelihood estimation is  $P(X) = p_{ijk} = n_{ijk} / n_{ij}$ .

There are two graphs  $G^+$  and  $G^-$ , and  $G^+ = G^- \cup \{X_i \rightarrow X_l\}$ . If the data item is not spare for the both networks (there are no amounts of all zero lines), the score of graph will be better when we add the edge to the graph in the follow situations according to the maximum principle of entropy:

$$H(X_i | Pa(X_i)) = -\sum_{j=1}^{q_i} \sum_{k=1}^{r_i} p_{ijk} \log p_{ijk} \quad (4)$$

$$H(X_i | Pa(X_i) \cup X_l) > H(X_i | Pa(X_i)) \quad (5)$$

If the data item is sparse for the networks, the form of entropy should be:

$$H(X_i | Pa(X_i)) = -\sum_{j:n_{ij}>0} \sum_{k=1}^{r_i} p_{ijk} \log p_{ijk} \quad (6)$$

And we replace the maximum likelihood estimation with the posterior estimation. The theorem should be established [8]:

Theorem 1: In the Bayesian network, if the prior parameter  $\alpha$  of uniform Dirichlet distribution is known, the condition entropy  $H(\bullet)$  of  $X_i | Pa(X_i)$  should be:

$$H(X_i | Pa(X_i); \alpha) = -\sum_{j:n_{ij}>0} \sum_{k=1}^{r_i} p_{ijk}^{\alpha_{ijk}} \log p_{ijk}^{\alpha_{ijk}}, p_{ijk}^{\alpha_{ijk}} = \frac{\alpha_{ijk} + n_{ijk}}{r_i \alpha_{ijk} + n_{ij}} \quad (7)$$

And when  $\alpha < \beta$ , there will be  $H(X_i | Pa(X_i); \alpha) < H(X_i | Pa(X_i); \beta)$ .

According to maximum entropy principle, adding the edge performs better than no adding the edge in following situation:

$$H(X_i | Pa(X_i) \cup X_l; \alpha) > H(X_i | Pa(X_i); \alpha) \quad (8)$$

When  $\alpha$  is very small and  $\alpha_{ijk} \rightarrow 0$ , the posterior estimation would convergence to the maximum likelihood estimation:

$$\begin{aligned} H(X_i | Pa(X_i) \cup X_l; \alpha) &\approx H(X_i | Pa(X_i) \cup X_l) \\ H(X_i | Pa(X_i); \alpha) &\approx H(X_i | Pa(X_i)) \end{aligned} \quad (9)$$

This is consistent with previous conclusions as for now. But if we consider that the data is not always sparse for the  $G^-$  and  $G^+$  which means that all the father nodes can be observed in the  $G^-$  but there are only  $\hat{q}_i$  father nodes can be observed in the  $G^+$  (data items for the others  $q_i - \hat{q}_i$  father nodes is all zero lines), we can adjust some conclusion according to theorem 1:

$$H(X_i | Pa(X_i) \cup X_l; \alpha \hat{q}_i / q_i) < H(X_i | Pa(X_i) \cup X_l; \alpha) \quad (10)$$

Although the data is sparse, to stay the network structure that adding the edge performs better than no adding, the following conclusion should be established:

$$H(X_i | Pa(X_i) \cup X_i; \alpha \hat{q}_i / q_i) > H(X_i | Pa(X_i); \alpha) \quad (11)$$

The requirement of the equation is hyperparameter  $\alpha$  should be updated while adding the edge:

$$\alpha = \alpha \hat{q}_i / q_i \quad (12)$$

According to equation 8 and equation 12, the interaction of information suggests that adding the edge is a better choice even though the data item is sparse. So far, we have accomplished the entropy 'convergence' from sparse data to complete data.

Then apply the algorithm that updating the hyperparameter  $\alpha$  to the BDeu score function and we can get BDs score function:

$$\alpha_{ijk} = \begin{cases} \alpha / (r_i \hat{q}_i) & n_{ij} > 0 \\ 0 & otherwise \end{cases} \quad (13)$$

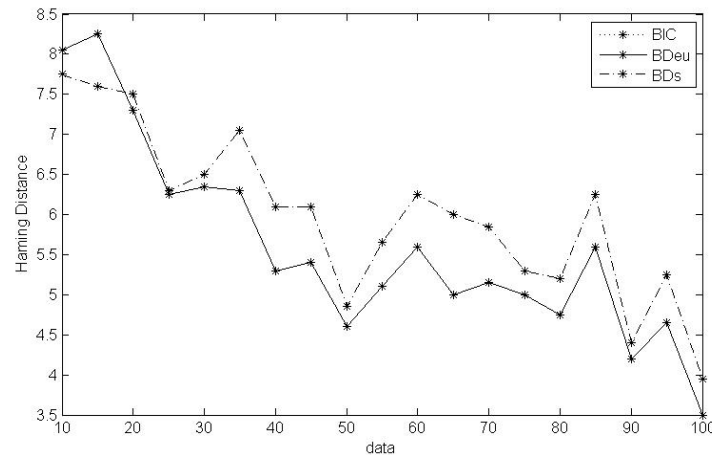
$$BDs(X_i | Pa(X_i); \alpha) = \prod_{j: n_{ij} > 0} \left( \frac{\Gamma(r_i \alpha_{ijk})}{\Gamma(r_i \alpha_{ijk} + n_{ij})} \right) \prod_{k=1}^{r_i} \frac{\Gamma(\alpha_{ijk} + n_{ijk})}{\Gamma(\alpha_{ijk})}$$

The thought of replacing the prior parameter distribution break the concept that the prior setting for definite network is always equivalence in the traditional score function. For different data sets, BDs score function may produce different prior parameters distribution which could include some constraints that is source from data.

## 4. Simulations

### 4.1. The influence of data size.

We will test the effect of BDs score and BDeu score in different data size. The standard network used to simulate is Asia Network. Set hyperparameter  $\alpha$  to 1 stably. Then we compute both score functions 20 times for every data size which is generated per 5 times from 10 to 100. The final score is the average of 20 scores, which may eliminate the error of single trial and make the result more general. Also, we choose BIC score as contrast. The experience result is as figure1:



**Figure 1.** Compare of three score function (data size: 10-5-100)

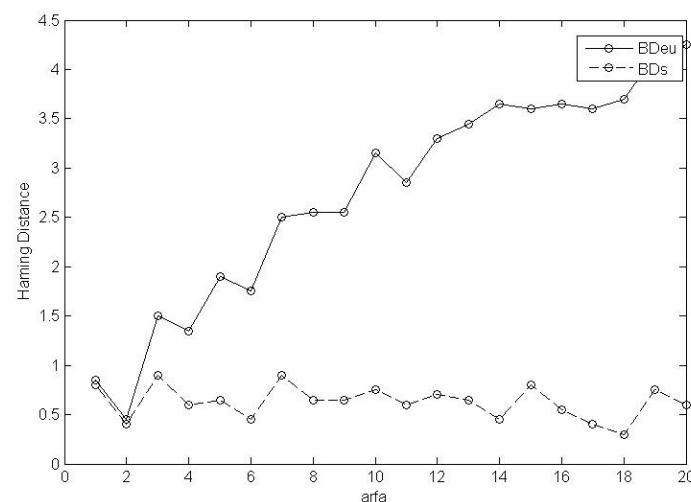
The horizontal axis of fig 1 stands for the changes of data size. The vertical axis stands for the changes of Hamming Distance (HD) that is a widely accepted function to estimate the difference between the standard network and learned network. The definition of HD is as follow:

Hamming Distance = loss edges + reversion edges + over edges

The difference between networks would get bigger and learned network is worse accurate with the HD is bigger. From the above figure we can see, there is no prior structure in BDeu score, which leads to the score of BIC and BDeu are same and the lines of them are also coincident. Compared with them in the condition of sparse data size, BDs score has the smaller HD and better learning effect if the data size is varied from 10 to 30. But BDeu score is better than BDs score when data size is during 30 and 100. It is because that the nature of BDs score is to remove the all zero lines in the sparse data and then remove the corresponding hyperparameter (setting them to zeros) which means there is not a such combination of the node and the father node and the edge is not defaulted in the prior experience. So, the edge will not be learned in the condition of sparse data.

#### 4.2. The influence of hyperparameter.

Next, we consider the influence of hyperparameter  $\alpha$  to the BDs and BDeu. Like last experiment, we choose Asia Network, select a data size of 1000 and set  $\alpha$  vary from 1 to 20. To ensure the accuracy of results, every  $\alpha$  should be simulated 20 times to get the mean.



**Figure 2.** Compare two score functions for different hyperparameter in Asia Network

From the analysis of BDeu in the section 3 we can know it is very sensitive to the selection of hyperparameter. When  $\alpha$  is not suitable, the algorithm will pretend to over-fitting. But the prior parameters in BDs that change the form of hyperparameters always obey the parameters distribution in real network. So, no such wrong edges against the prior experience would be add to network during the training process. On the other hand, BDs could avoid the over-fitting effectively. Figure 2 is the best evidence. When  $\alpha$  is increasing, HD of BDeu is rapidly raising while HD of BDs is always stable.

## 5. Conclusion

The paper design an improved Bayesian Structure learning algorithm for uniform prior parameters. For the problem that in BDeu score function the prior parameter selection method is always based on the uniform distribution of data whether the data is sparse or not, we come up with the BDs score function of which the prior parameters are decided by the all zero lines in the sparse data. BDs score break the equivalence of prior parameters and make them suitable for different data set. We also explain the theory why BDs always perform better than BDeu through the principle of entropy and explore the relationship between two score functions. To prove the superiority of BDs score when data is sparse, we simulate

the HD of different score functions in the condition of different data size. Finally, we explore the influence of hyperparameter  $\alpha$  to BDs score and BDeu score and further certify that BDs may adapt to the dynamic changes of hyperparameter without over-fitting.

However, duo to our limited level, the whole process is still relatively rough. There are still many problems to be discussed. Although BDs score has higher learning accuracy than BDeu score, the improvement of learning accuracy rate varies with the size of data set for different networks. So how to choose a best hyperparameter  $\alpha$  to suit the given data size for networks with different complexity is still a direction worth to research.

## References

- [1] J Pearl. Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference. Series in Representation and Reasoning [M]. San Mateo, CA: Morgan Kaufmann Press, 1988.
- [2] Gal Elidan. Bagged Structure Learning of Bayesian Networks [C]. Proceedings of the 14<sup>th</sup> International conference on Artificial Intelligence and Statistics, For Lauderdale, 2011: 15-25.
- [3] Borboudakis G, Tsamardinos I. Incorporating causal prior knowledge as Path-Constraints in Bayesian networks and maximal ancestral graphs [C]. Proceedings of the 29<sup>th</sup> International Conference on Machine Learning. Edinburgh, 2012: 1-8.
- [4] Takashi Isozaki. Learning causal Bayesian networks using minimum free energy principle [J]. New Generation Computing. 2012, 1 (30): 17-52.
- [5] M Ueno, M Uto. Non-informative Dirichlet score for learning Bayesian networks [C]. Proceeding of the 6<sup>th</sup> European Workshop on Probabilistic Graphical Models. 2012, 331-338.
- [6] Harald Steck. Learning the Bayesian Network Structure: Dirichlet Prior versus Data. UAI-P 2008-PG-511-518.
- [7] Z I Wen, G H Peng, Introduction to Bayesian Networks [M]. Beijing: Science Press. 2006.
- [8] Scutari M. Beyond Uniform Priors in Bayesian Network Structure Learning [J]. 2017.