

---

# Sequential and Factorized NML models

---

**Tomi Silander   Teemu Roos   Petri Myllymäki**  
Helsinki Institute for Information Technology HIIT

## 1 INTRODUCTION

Bayesian networks are among most popular model classes for discrete vector-valued i.i.d data. Currently the most popular model selection criterion for Bayesian networks follows Bayesian paradigm. However, this method has recently been reported to be very sensitive to the choice of prior hyper-parameters [1]. On the other hand, the general model selection criteria, AIC [2] and BIC [3], are derived through asymptotics and their behavior is suboptimal for small sample sizes.

This extended abstract is based on an unpublished manuscript [4] in which we introduce a new effective scoring criterion for learning Bayesian network structures, the factorized normalized maximum likelihood (fNML). This score features no tunable parameters thus avoiding the sensitivity problems of Bayesian scores. It also has a probabilistic interpretation which yields a natural way to use the selected model for predicting future data.

## 2 BAYESIAN NETWORKS

Bayesian network defines a joint probability distribution for an  $n$ -dimensional data vector  $X = (X_1, \dots, X_n)$ , where each  $X_i$  may have  $r_i$  different values which, without loss of generality, can be denoted as  $\{1, \dots, r_i\}$ .

### 2.1 Model class

A Bayesian network consists of a directed acyclic graph (DAG)  $G$  and a set of conditional probability distributions. We specify the DAG with a vector  $G = (G_1, \dots, G_n)$  of parent sets, so that  $G_i \subset \{X_1, \dots, X_n\}$  denotes the parents of variable  $X_i$ , i.e., the variables from which there is an arc to  $X_i$ . Each parent set  $G_i$  has  $q_i$  ( $q_i = \prod_{X_p \in G_i} r_p$ ) possible values that are the possible value combinations of the variables belonging to  $G_i$ . We assume an enumeration of these values and denote the fact that  $G_i$  holds the  $j^{\text{th}}$  value combination simply by  $G_i = j$ .

The conditional probability distributions  $P(X_i | G_i)$  are determined by a set of parameters,  $\Theta$ , via the equation

$$P(X_i = k | G_i = j, \Theta) = \theta_{ijk}.$$

We denote the set of parameters associated with variable  $X_i$  by  $\Theta_i$ . Given a Bayesian network  $(G, \Theta)$  the joint distribution can be factorized as

$$P(x | G, \Theta) = \prod_{i=1}^n P(x_i | G_i, \Theta_i) = \prod_{i=1}^n \theta_{iG_i x_i}. \quad (1)$$

## 2.2 Data

To learn the Bayesian network structures, we assume data  $D$  of  $N$  i.i.d instantiations of the vector  $X$ , i.e., an  $N \times n$  data matrix without missing values. We select columns of the data matrix  $D$  by subscripting it with a corresponding variable index or a variable set.

Since the rows  $D$  are assumed to be i.i.d, the probability of a data matrix can be calculated by just taking the product of the row probabilities. Combining equal terms yields

$$P(D | G, \Theta) = \prod_{i=1}^n \prod_{j=1}^{q_i} \prod_{k=1}^{r_i} \theta_{ijk}^{N_{ijk}}, \quad (2)$$

where  $N_{ijk}$  denotes number of rows in which  $X_i = k$  and its parents contain the  $j^{\text{th}}$  value combination.

For a given structure  $G$ , the maximum likelihood parameters are simply the relative frequencies found in the data:  $\hat{\theta}_{ijk} = \frac{N_{ijk}}{\sum_{k'} N_{ijk'}}$ . Setting parameters  $\hat{\theta}_{ijk}$  to their maximum likelihood values for data  $D$ , gives the maximized likelihood  $P(X | G, \hat{\Theta}(D))$ . In the following, we denote the value  $P(D | G, \hat{\Theta}(D))$  by  $\hat{P}(D | G)$ <sup>1</sup>.

## 3 Model selection

The number of possible Bayesian network models for  $n$  variables is super exponential, and the model selection task has been shown to be NP-hard for practically all model selection criteria such as AIC, BIC, and marginal likelihood [5]. However, all popular Bayesian network selection criteria  $S(G, D)$  feature a convenient *decomposability*

$$\text{SCORE}(G, D) = \sum_{i=1}^n S(D_i, D_{G_i}) \quad (3)$$

that makes implementing a heuristic search for models easier [6].

Many popular scoring functions avoid overfitting by balancing the fit to the data and the complexity of the model. A common form of this idea can be expressed as

$$\text{SCORE}(G, D) = \log \hat{P}(D | G) - \Delta(D, G), \quad (4)$$

where  $\Delta(D, G)$  is a complexity penalty. For example,  $\Delta^{\text{BIC}} = \sum_i \frac{q_i(r_i-1)}{2} \log N$ , and  $\Delta^{\text{AIC}} = \sum_i q_i(r_i - 1)$ .

### 3.1 Bayesian Dirichlet scores

The current state-of-the-art is to use marginal likelihood scoring criterion

$$S_{\text{BD}}(D_i, D_{G_i}, \bar{\alpha}) = \log \int_{\theta_i} P(D_i | D_{G_i}, \theta_i) W(\theta_i | \alpha_i) d\theta_i. \quad (5)$$

The most convenient form of this, the Bayesian Dirichlet (BD) score, uses conjugate priors in which parameter vectors  $\Theta_{ij}$  are assumed independent of each other and distributed by Dirichlet distributions so that

$$W(\theta_i | \alpha_i) = \prod_{j=1}^{q_i} P(\theta_{ij} | \alpha_{ij*}), \quad (6)$$

in which  $\theta_{ij} \sim \text{Dir}(\alpha_{ij1}, \dots, \alpha_{ijr_i})$ . With a choice of  $\alpha_{ijk} = \frac{\alpha}{q_i r_i}$  we get a family of BDeu scores popular for giving equal scores for different Bayesian network structures that encode same independence

<sup>1</sup>We often drop the dependency on  $G$  from the notation when it is clear from the context.

assumptions. The BDeu score depends only on single parameter  $\alpha$ , but recent studies show that model selection is very sensitive to it.

For predictive purposes it is natural to parameterize the model learned with the  $BD$ -score by expected parameter values

$$\theta_{ijk}^{BD} = \frac{N_{ijk} + \alpha_{ijk}}{\sum_{k'=1}^{r_i} [N_{ijk'} + \alpha_{ijk'}]}. \quad (7)$$

## 4 FACTORIZED NML

The factorized normalized maximum likelihood (fNML) score is based on the normalized maximum likelihood (NML) distribution [7, 8]

$$P_{\text{NML}}(D | \mathcal{M}) = \frac{\hat{P}(D | \mathcal{M})}{\sum_{D'} \hat{P}(D' | \mathcal{M})}, \quad (8)$$

where the normalization is over all data sets  $D'$  of a fixed size  $N$ . The log of the normalizing factor is called the *parametric complexity* or the *regret*. Evaluation of the regret is often hard due to the exponential number of terms in the sum. We propose a decomposable factorized normalized maximum likelihood criterion with a local score

$$S_{\text{NML}}(D_i, D_{G_i}) = \log P_{\text{NML}}(D_i | D_{G_i}) = \log \left( \frac{\hat{P}(D_i | D_{G_i})}{\sum_{D'_i} \hat{P}(D'_i | D_{G_i})} \right),$$

where the normalizing sum goes over all the possible  $D_i$ -column vectors of length  $N$ , i.e.,  $D'_i \in \{1, \dots, r_i\}^N$ . Using recently discovered methods for calculating the regret for a single  $r$ -ary multinomial variable [9] the fNML-criterion can be calculated as efficiently as other decomposable scores.

For predictive purposes its is natural to parameterize the model learned with the fNML-score by predictive conditional NML parameters [10]

$$\theta_{ijk} = \frac{e(N_{ijk})(N_{ijk} + 1)}{\sum_{k'=1}^{r_i} e(N_{ijk'})(N_{ijk'} + 1)}, \quad (9)$$

where  $e(n) = \binom{n+1}{n}^n$ .

Empirical tests with real data sets indicate that the fNML selection criterion performs very well in a code length sense when compared with the state of the art BDeu criterion. The predictive capabilities of the Bayesian and fNML approaches are currently under investigation.

## References

- [1] T. Silander, P. Kontkanen, and P. Myllymäki, "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*, R. Parr and L. van der Gaag, Eds. 2007, pp. 360–367, AUAI Press.
- [2] H. Akaike, "Information theory and an extension of the maximum likelihood principle," in *Proceedings of the Second International Symposium on Information Theory*, B.N. Petrox and F. Caski, Eds., Budapest, 1973, pp. 267–281, Akademiai Kiado.
- [3] G. Schwarz, "Estimating the dimension of a model," *Annals of Statistics*, vol. 6, pp. 461–464, 1978.
- [4] T. Silander, T. Roos, P. Kontkanen, and Myllymäki, "Factorized normalized maximum likelihood criterion for learning bayesian network structures," Submitted for PGM08, 2008.
- [5] D.M. Chickering, "Learning Bayesian networks is NP-Complete," in *Learning from Data: Artificial Intelligence and Statistics V*, D. Fisher and H. Lenz, Eds., pp. 121–130. Springer-Verlag, 1996.
- [6] D. Heckerman, D. Geiger, and D.M. Chickering, "Learning Bayesian networks: The combination of knowledge and statistical data," *Machine Learning*, vol. 20, no. 3, pp. 197–243, September 1995.

- [7] Yu.M. Shtarkov, "Universal sequential coding of single messages," *Problems of Information Transmission*, vol. 23, pp. 3–17, 1987.
- [8] J. Rissanen, "Fisher information and stochastic complexity," *IEEE Transactions on Information Theory*, vol. 42, no. 1, pp. 40–47, January 1996.
- [9] P. Kontkanen and P. Myllymäki, "A linear-time algorithm for computing the multinomial stochastic complexity," *Information Processing Letters*, vol. 103, no. 6, pp. 227–233, 2007.
- [10] J. Rissanen and T. Roos, "Conditional NML models," in *Information Theory and Applications Workshop (ITA-07)*, San Diego, CA, January–February 2007.