Quotient Normalized Maximum Likelihood Criterion for Learning Bayesian Network Structures

Tomi Silander[1], Janne Leppä-aho[2], Elias Jääsaari[2], and Teemu Roos[2]

[1] NAVER LABS Europe, France
[2] HIIT / Department of Computer Science, University of Helsinki, Finland

Abstract

We introduce an information theoretic criterion for Bayesian network structure learning which we call quotient normalized maximum likelihood (qNML). In contrast to the closely related factorized normalized maximum likelihood criterion, qNML satisfies the property of score equivalence. It is also decomposable and completely free of adjustable hyperparameters. For practical computations, we identify a remarkably accurate approximation proposed earlier by Szpankowski and Weinberger. Experiments on both simulated and real data demonstrate that the new criterion leads to parsimonious models with good predictive accuracy.

Background: Structure Learning of Bayesian Networks

Bayesian Networks

- Provide a compact way to represent a joint distribution over a random vector $X = (X_1, \ldots, X_d)$.
- Consist of:
  1. A Directed acyclic graph G which encodes the dependencies between the components of X.
  2. Parameters $\theta = (\theta_1, \ldots, \theta_d)$, where θ_i denotes the parameters of the conditional distribution of X_i given its parents X_{G_i}.
- Decomposition: $P(X \mid G, \theta) = \prod_{i=1}^{d} P(X_i \mid X_{G_i}, \theta_i)$

Structure Learning

Data: Each X_i is a categorical variable. We observe n independent samples of X which are collected in a data matrix D of size $n \times d$.

Goal: We consider a *score-based* approach and seek a graph G that maximizes a scoring function which evaluates how well a given graph fits the observed data.

- Some scoring functions: BIC, BDeu and fNML.

Scoring Functions

BDeu Bayesian marginal likelihood based on Dirichlet priors. Depends on a single hyperparameter $\alpha > 0$ called equivalent sample size.

BIC Maximized log-likelihood with penalty $\frac{k}{2} \log n$, where k is the number of free parameters in the network.

fNML Factorized Normalized Maximum Likelihood. Maximized log-likelihood with penalty defined via *regret* functions.

Quotient Normalized Maximum Likelihood Criterion

Motivation

BDeu
- Is very sensitive to the choice of the hyperparameter α [2].
- Is not *regular* [4] (can be shown in certain situations to favour too complex models over simpler)

fNML
- Is not score equivalent: the graphs expressing the same independence statements are not scored equally.
- Learned structures are often rather complex, which hampers their interpretation.

BIC
- Appears to require large sample sizes in order to identify appropriate structures [3, 1].

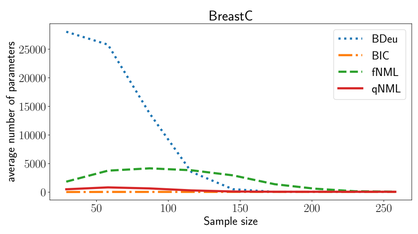

Figure : Number of parameters as a function of sample size for Breast Cancer (UCI) data.

The quest for a model selection criterion that would yield **more parsimonious**, **easier to interpret**, but still **predictive** Bayesian networks structures is one of the main motivations for this work.

qNML: Idea

- Following the *MDL-principle*, we would like to pick a model G maximizing

$$\log P_{NML}(D; G) = \underbrace{\log P(D \mid \hat{\theta}, G)}_{\text{Maximized log-likelihood}} \underbrace{- \log \sum_{D'} P(D' \mid \hat{\theta}, G)}_{:= \text{ regret. } \textbf{Intractable!}},$$

- However, for **a single categorical data vector**, regret and one dimensional P^1_{NML}-code can be computed.
- **Trick**. Recall factorization w.r.t. graph G:

$$P(D \mid G) = \prod_{i=1}^{d} P(D_i \mid D_{G_i}) = \prod_{i=1}^{d} \frac{P(D_{i,G_i})}{P(D_{G_i})}$$

- Assume there are **no independencies** among X_{G_i}. Now, $D_{i,G_i} =$

$$\begin{array}{c|cc} D_i & D_{G_i} & \\ \hline 1 & 1 & 1 \\ 0 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & 1 \end{array} \rightarrow \begin{pmatrix} 0 \\ 1 \\ 2 \\ 0 \end{pmatrix}$$

- We can treat D_{i,G_i} and D_{G_i} as values of single variables!
- $P(D_{i,G_i})$ and $P(D_{i,G_i})$ can be modeled using P^1_{NML}-code.

qNML: Definition & Properties

Definition of qNML score:

$$s^{qNML} := \sum_{i=1}^{d} \log \frac{P^1_{NML}(D_{i,G_i})}{P^1_{NML}(D_{G_i})},$$

$\log P^1_{NML}(D) = \log P(D \mid \hat{\theta}) - reg(n, r)$.

- D is understood as n observations on single categorical variable with r possible values.
- Accurate approximation for $reg(n, r)$ from [5]:

$$reg(n, r) \approx n \left(\log \beta + (\beta + 2) \log C_\beta - \frac{1}{C_\beta} \right) - \frac{1}{2} \log \left(C_\beta + \frac{2}{\beta} \right),$$

where $\beta = \frac{r}{n}$ and $C_\beta = \frac{1}{2} + \frac{1}{2}\sqrt{1 + \frac{4}{\beta}}$.

qNML score is:
- **hyper-parameter free**
- **consistent**
- **score equivalent**
- **regular**
- **it equals NML for many models.**

Experimental Results

Structure Learning

- Data generated from real world DAG structures.
- 5 networks, $d = 6, 5, 11, 5$ and 8, edges: $6, 4, 17, 4$ and 8.
- Networks learned using exact DP-algorithm.
- We measure Structural Hamming Distance (SHD).
- Average ranks (over 5 networks and 1000 tests) also shown.

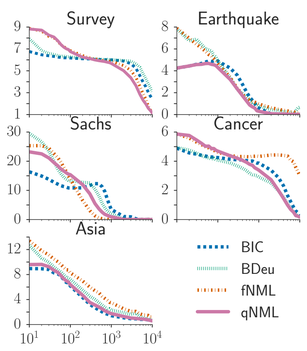

Figure : SHD as a function of sample size.

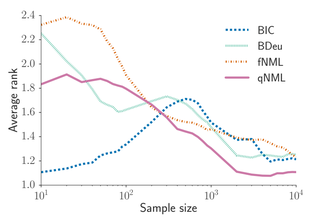

Figure : Average ranks for each score.

qNML:
- obtains the best ranking when $n > 300$.
- never has the worst average ranking.

Prediction

- 20 UCI data sets split to train and test sets.
- Used training fractions: $10\%, 20\%, \ldots, 90\%$.
- BIC, fNML and qNML use sequential predictive (sNML) parameters.
- BDeu uses Bayesian predictive parameter (BPP) values with $\alpha = 1$.
- Scores are ranked from 1 to 4 according to test set log-loss.

Table: Average predictive performance rank over different sample sizes for different scores. **Best**, worst.

Data	n	BDeu	BIC	fNML	qNML
PostOpe	90	2.79	**1.20**	3.06	2.94
Iris	150	2.82	2.37	**2.27**	2.54
Wine	178	3.23	**1.88**	2.67	2.22
Glass	214	3.61	3.09	**1.42**	1.88
Thyroid	215	2.55	3.21	**1.80**	2.44
HeartSt	270	3.12	**1.39**	3.12	2.37
BreastC	286	3.09	**1.41**	2.97	2.53
HeartHu	294	3.18	**1.66**	2.90	2.27
HeartCl	303	3.46	**1.38**	2.99	2.17
Ecoli	336	3.20	3.53	**1.24**	2.04
Liver	345	3.17	2.39	2.69	**1.75**
Balance	625	3.35	1.91	**1.59**	3.16
BcWisco	699	3.06	2.03	2.89	**2.02**
Diabete	768	2.91	2.70	2.68	**1.71**
TicTacT	958	3.44	2.71	**1.31**	2.53
Yeast	1484	2.60	3.76	**1.55**	2.10
Abalone	4177	2.60	3.64	**1.04**	2.72
PageBlo	5473	2.24	3.61	**1.31**	2.83
Adult	32561	3.23	3.77	**1.00**	2.00
Shuttle	58000	**1.44**	3.78	1.56	3.22
Mean		2.95	2.57	**2.10**	2.37
SD		0.49	0.90	0.76	**0.43**

- fNML obtains on average the best rank with qNML as the runner-up.
- qNML has the lowest standard deviation.

We also studied the number of parameters in learned models when using 10% samples for structure learning.

- qNML almost always yields more parsimonious models than fNML, but more complex than BIC.

References

[1] Liu, Z. et al. (2012). Empirical evaluation of scoring functions for Bayesian network model selection. *BMC Bioinformatics*, 13(15):S14.
[2] Silander et al. (2007). On sensitivity of the MAP Bayesian network structure to the equivalent sample size parameter. In *Proceedings of UAI–07*.
[3] Silander et al. (2008). Factorized normalized maximum likelihood criterion for learning Bayesian network structures. In *Proceedings of PGM-08*.
[4] Suzuki, J. (2017). A theoretical analysis of the BDeu scores in Bayesian network structure learning. *Behaviormetrika*, 44(1):97–116.
[5] Szpankowski, W. and Weinberger, M. J. (2012). Minimax pointwise redundancy for memoryless models over large alphabets. *IEEE Transactions on Information Theory*, 58(7):4094–4104.

Acknowledgements

J.L., E.J. and T.R. were supported by the Academy of Finland (COIN CoE and Project TENSORML). J.L. was supported by the DoCS doctoral programme of the University of Helsinki.