# Quotient Normalized Maximum Likelihood Criterion for Learning Bayesian Network Structures

**Tomi Silander**[1] [1]NAVER LABS Europe

**Janne Leppä-aho**[2,4] [2]Helsinki Institute for Information Technology HIIT

**Elias Jääsaari**[2,3] [3]Department of CS, Aalto University

**Teemu Roos**[2,4] [4]Department of CS, University of Helsinki

## 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.

## 1 INTRODUCTION

Bayesian networks [Pearl, 1988] are popular models for presenting multivariate statistical dependencies that may have been induced by underlying causal mechanisms. Techniques for learning the structure of Bayesian networks from observational data have therefore been used for many tasks such as discovering cell signaling pathways from protein activity data [Sachs et al., 2002], revealing the business process structures from transaction logs [Savickas and Vasilecas, 2014] and modeling brain-region connectivity using fMRI data [Ide et al., 2014].

Learning the structure of statistical dependencies can be seen as a model selection task where each model is a different hypothesis about the conditional dependencies between sets of variables. Traditional model selection criteria such as the Akaike information criterion (AIC) [Akaike, 1973] and the Bayesian information criterion (BIC) [Schwarz, 1978] have also been used for the task, but recent comparisons have not been favorable for AIC, and BIC appears to require large sample sizes in order to identify appropriate structures [Silander et al., 2008, Liu et al., 2012]. Traditionally, the most popular criterion has been the Bayesian marginal likelihood [Heckerman, 1995] and its BDeu variant (see Section 2), but studies [Silander et al., 2007, Steck, 2008] show this criterion to be sensitive to hyperparameters and to yield undesirably complex models for small sample sizes.

The information-theoretic normalized maximum likelihood (NML) criterion [Shtarkov, 1987, Rissanen, 1996] would otherwise be a potential candidate for a good criterion, but its exact calculation is likely to be prohibitively expensive. In 2008, Silander et al. introduced a hyperparameter-free, NML inspired criterion called the factorized NML (fNML) [Silander et al., 2008] that was shown to yield good predictive models without such sensitivity problems. However, from the structure learning point of view, fNML still sometimes appears to yield overly complex models. In this paper we introduce another NML related criterion, the *quotient NML* (qNML) that yields simpler models without sacrificing predictive accuracy. Furthermore, unlike fNML, qNML is *score equivalent*, i.e., it gives equal scores to structures that encode the same independence and dependence statements. Like other common model selection criteria, qNML is also consistent.

We next briefly introduce Bayesian networks and review the BDeu and fNML criteria and then introduce the qNML criterion. We also summarize the results for 20 data sets to back up our claim that qNML yields parsimonious models with good predictive capabilities. The experiments with artificial data generated from real-world Bayesian networks demonstrate the capability of our score to quickly learn a structure close to the generating one.

Proceedings of the $21^{st}$ International Conference on Artificial Intelligence and Statistics (AISTATS) 2018, Lanzarote, Spain. PMLR: Volume 84.

## 2 BAYESIAN NETWORKS

Bayesian networks are a general way to describe the dependencies between the components of an $n$-dimensional random data vector. In this paper we only address the case in which the component $X_i$ of the data vector $X = (X_1, \dots, X_n)$ may take any of the discrete values in a set $\{1, \dots, r_i\}$. Despite denoting the values with small integers, the model will treat each $X_i$ as a categorical variable.

### 2.1 Likelihood

A Bayesian network $B = (G, \theta)$ defines a probability distribution for $X$. The component $G$ defines the structure of the model as a directed acyclic graph (DAG) that has exactly one node for each component of $X$. The structure $G = (G_1, \dots, G_n)$ defines for each variable/node $X_i$ its (possibly empty) parent set $G_i$, i.e., the nodes from which there is a directed edge to the variable $X_i$.

Given a realization $x$ of $X$, we denote the sub-vector of $x$ that consists of the values of the parents of $X_i$ in $x$ by $G_i(x)$. It is customary to enumerate all the possible sub-vectors $G_i(x)$ from 1 to $q_i = \prod_{h \in G_i} r_h$. In case $G_i$ is empty, we define $q_i = 1$ and $P(G_i(x) = 1) = 1$ for all vectors $x$.

For each variable $X_i$ there is a $q_i \times r_i$ table $\theta_i$ of parameters whose $k^{\text{th}}$ column on the $j^{\text{th}}$ row $\theta_{ij}$ defines the conditional probability $P(X_i = k \mid G_i(X) = j; \theta) = \theta_{ijk}$. With structure $G$ and parameters $\theta$, we can now express the likelihood function of the model as

$$P(x|G,\theta) = \prod_{i=1}^{n} P(x_i \mid G_i(x); \theta_i) = \prod_{i=1}^{n} \theta_{iG_i(x)x_i}. \quad (1)$$

### 2.2 Bayesian Structure Learning

Score-based Bayesian learning of Bayesian network structures evaluates the goodness of different structures $G$ using their posterior probability $P(G|D, \alpha)$, where $\alpha$ denotes the hyperparameters for the model parameters $\theta$, and $D$ is a collection of $N$ $n$-dimensional i.i.d. data vectors collected to a $N \times n$ design matrix. We use the notation $D_i$ to denote the $i^{\text{th}}$ column of the data matrix and the notation $D_V$ to denote the columns that correspond to the variable subset $V$. We also write $D_{i,G_i}$ for $D_{\{i\} \cup G_i}$ and denote the entries of the column $i$ on the rows on which the parents $G_i$ contain the value configuration number $j$ by $D_{i,G_i=j}$, $j \in \{1, \dots, q_i\}$.

It is common to assume the uniform prior for structures, in which case the objective function for structure learning is reduced to the marginal likelihood $P(D|G, \alpha)$. If the model parameters $\theta_{ij}$ are further assumed to be independently Dirichlet distributed only depending on $i$ and $G_i$ and the data $D$ is assumed to have no missing values, the marginal likelihood can be decomposed as

$$\begin{aligned} &P(D|G,\alpha) \\ &\quad = \prod_{i=1}^{n} \prod_{j=1}^{q_i} P(D_{i,G_i=j}; \alpha) \\ &\quad = \prod_{i=1}^{n} \prod_{j=1}^{q_i} \int P(D_{i,G_i=j} | \theta_{ij}) P(\theta_{ij}; \alpha) d\theta_{ij}. \quad (2) \end{aligned}$$

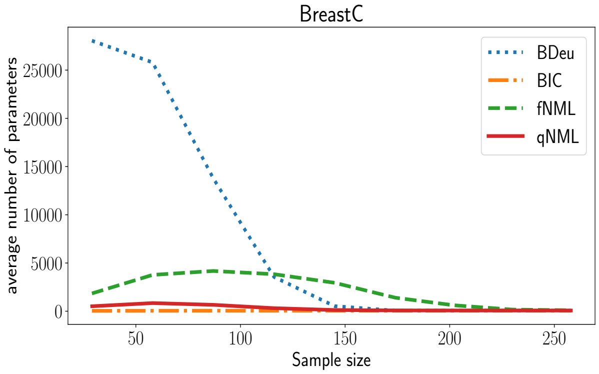

Figure 1: Number of parameters in a breast cancer model as a function of sample size for different model selection criteria.

In coding terms this means that each data column $D_i$ is first partitioned based on the values in columns $G_i$, and each part is then coded using a Bayesian mixture that can be expressed in closed form [Buntine, 1991, Heckerman et al., 1995].

### 2.3 Problems, Solutions and Problems

Finding satisfactory Dirichlet hyperparameters for the Bayesian mixture above has, however, turned out to be problematic. Early on, one of the desiderata for a good model selection criterion was that it is score equivalent, i.e., it would yield equal scores for essentially equivalent models [Verma and Pearl, 1991]. For example, the score for the structure $X_1 \rightarrow X_2$ should be the same as the score for the model $X_2 \rightarrow X_1$ since they both correspond to the hypothesis that variables $X_1$ and $X_2$ are statistically dependent on each other. It can be shown [Heckerman et al., 1995] that to achieve this, not all the hyperparameters $\alpha$ are possible and for practical reasons Buntine [Buntine, 1991] suggested a so-called BDeu score with just one hyperparameter $\alpha \in R_{++}$ so that $\theta_{ij\cdot} \sim Dir(\frac{\alpha}{q_i r_i}, \dots, \frac{\alpha}{q_i r_i})$. However, it soon turned out that the BDeu score was very sensitive to the selection of this hyperparameter [Silander et al., 2007] and that for small sample sizes this method detects spurious correlations [Steck, 2008] leading to models with suspiciously many parameters.

Recently, Suzuki [Suzuki, 2017] discussed the theoretical properties of the BDeu score and showed that in certain settings BDeu has the tendency to add more and more parent variables for a child node even though the empirical conditional entropy of the child given the parents has already reached zero. In more detail, assume that in our data $D$ the values of $X_i$ are completely determined by variables in a set $Z$, so that the empirical entropy $H_N(X_i|Z)$ is zero. Now, if we can further find one or more variables, denoted by $Y$, whose values are determined completely by the variables in $Z$, then BDeu will prefer the set $Z \cup Y$ over $Z$ alone as the parents of $X_i$. Suzuki argues that this kind of behavior violates *regularity* in model selection as the more complex model is preferred over a simpler one even though it does not fit the data any better. The phenomenon seems to stem from the way the hyperparameters for the Dirichlet distribution are chosen in BDeu as using Jeffreys' prior, $\theta_{ijk} \sim Dir(\frac{1}{2}, \dots, \frac{1}{2})$, does not suffer from this anomaly. However, using Jeffreys' prior causes marginal likelihood score not to be score equivalent. In Section 3.4, we will give the formal definition of regularity and state that qNML is regular. In addition, we provide a proof of regularity for fNML criterion, which has not appeared in the literature before. The detailed proofs can be found in Appendix B in the Supplementary Material.

A natural solution to avoid parameter sensitivity of BDeu would be to use a normalized maximum likelihood (NML) criterion [Shtarkov, 1987, Rissanen, 1996], i.e., to find the structure $G$ that maximizes

$$P_{NML}(D;G) = \frac{P(D|\hat{\theta}(D;G))}{\sum_{D'} P(D'|\hat{\theta}(D';G))}, \tag{3}$$

where $\hat{\theta}$ denotes the (easy to find) maximum likelihood parameters and the sum in the denominator goes over all the possible $N \times n$ data matrices. This information-theoretic NML criterion can be justified from the minimum description length point of view [Rissanen, 1978, Grünwald, 2007]. It has been shown to be robust with respect to different data generating mechanisms where a good choice of prior is challenging, see [Eggeling et al., 2014, Määttä et al., 2016]. While it is easy to see that the NML criterion satisfies the requirement of giving equal scores to equal structures, the normalizing constant renders the computation infeasible.

Consequently, Silander et al. [Silander et al., 2008] suggested solving the BDeu parameter sensitivity problem by using the NML code for the column partitions, i.e., changing the Bayesian mixture in equation (2) to

$$P^1_{NML}(D_{i,G_i=j};G) = \frac{P(D|\hat{\theta}(D_{i,G_i=j};G))}{\sum_{D'} P(D'|\hat{\theta}(D';G))}, \tag{4}$$

where $D' \in \{1,\dots,r_i\}^{|D_{i,G_i=j}|}$. The logarithm of the denominator is often called the regret, since it indicates the extra code length needed compared to the code length obtained using the (a priori unknown) maximum likelihood parameters. The regret for $P^1_{NML}$ depends only on the length $N$ of the categorical data vector with $r$ different categorical values,

$$reg(N,r) = \log \sum_{D \in \{1,\dots,r\}^N} P(D|\hat{\theta}(D)). \tag{5}$$

While the naive computation of the regret is still prohibitive, Silander et al. approximate it efficiently using a so-called Szpankowski approximation [Kontkanen et al., 2003]:

$$\begin{aligned} reg(N,r) \approx & \frac{\sqrt{2}r\Gamma\left(\frac{r}{2}\right)}{3\sqrt{N}\Gamma\left(\frac{r-1}{2}\right)} \\ & + \left(\frac{r-1}{2}\right)\log\left(\frac{N}{2}\right) - \log\Gamma\left(\frac{r}{2}\right) + \frac{1}{2}\log(\pi) \\ & - \frac{r^2\Gamma^2\left(\frac{r}{2}\right)}{9N\Gamma^2\left(\frac{r-1}{2}\right)} + \frac{2r^3-3r^2-2r+3}{36N}. \end{aligned} \tag{6}$$

However, equation (6) is derived only for the case where $r$ is constant and $N$ grows. While with fNML it is typical that $N$ is large compared to $r$, an approximation for all ranges of $N$ and $r$ derived by Szpankowski and Weinberger [Szpankowski and Weinberger, 2012] can also be used:

$$\begin{aligned} reg(N,r) \quad \approx \quad & N\left(\log\alpha + (\alpha+2)\log C_\alpha - \frac{1}{C_\alpha}\right) \\ & -\frac{1}{2}\log\left(C_\alpha + \frac{2}{\alpha}\right), \end{aligned} \tag{7}$$

where $\alpha = \frac{r}{N}$ and $C_\alpha = \frac{1}{2} + \frac{1}{2}\sqrt{1+\frac{4}{\alpha}}$. These approximations are compared in Table 1 to the exact regret for various values of $N$ and $r$. For a constant $N$, equation (6) provides a progressively worse approximation as $r$ grows. Equation (7) on the other hand is a good approximation of the regret regardless of the ratio of $N$ and $r$. In our experiments, we will use this approximation for implementation of the qNML criterion.

fNML solves the parameter sensitivity problem and yields predictive models superior to BDeu. However, the criterion does not satisfy the property of giving the same score for models that correspond to the same dependence statements. The score equivalence is usually viewed desirable when DAGs are considered only

Table 1: Regret values for various values of $N$ and $r$.

| N | r | eq. (6) | eq. (7) | exact |
|---|---|---|---|---|
| 50 | 10 | 13.24 | 13.26 | 13.24 |
| | 100 | 62.00 | 60.01 | 60.00 |
| | 1000 | 491.63 | 153.28 | 153.28 |
| | 10000 | 25635.15 | 265.28 | 265.28 |
| 500 | 10 | 22.67 | 22.69 | 22.67 |
| | 100 | 144.10 | 144.03 | 144.03 |
| | 1000 | 624.35 | 603.93 | 603.93 |
| | 10000 | 4927.24 | 1533.38 | 1533.38 |
| 5000 | 10 | 32.74 | 32.76 | 32.74 |
| | 100 | 247.97 | 247.97 | 247.97 |
| | 1000 | 1452.51 | 1451.78 | 1451.78 |
| | 10000 | 6247.83 | 6043.16 | 6043.16 |

as models for conditional independence, without any causal interpretation. Furthermore, the learned structures are often rather complex (see Figure 1) which also hampers their interpretation. 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.

# 3 QUOTIENT NML SCORE

We will now introduce a quotient normalized maximum likelihood (qNML) criterion for learning Bayesian network structures. While equally efficient to compute as BDeu and fNML, it is free from hyperparameters, and it can be proven to give equal scores to equivalent models. Furthermore, it coincides with the actual NML score for exponentially many models. In our empirical tests it produces models featuring good predictive performance with significantly simpler structures than BDeu and fNML.

Like BDeu and fNML, qNML can be expressed as a product of $n$ terms, one for each variable, but unlike the other two, it is not based on further partitioning the corresponding data column

$$\begin{aligned} s^{qNML}(D;G) &:= \sum_{i=1}^{n} s_i^{qNML}(D;G) \qquad (8) \\ &:= \sum_{i=1}^{n} \log \frac{P^1_{NML}(D_{i,G_i};G)}{P^1_{NML}(D_{G_i};G)}. \end{aligned}$$

The trick here is to model a subset of columns as though there were no conditional independencies among the corresponding variables $S \subset X$. In this case, we can collapse the $\prod_{X_i \in S} r_i$ value configurations and consider them as values of a single variable with $\prod_{X_i \in S} r_i$ different values which can then be modeled with a one-dimensional $P^1_{NML}$ code. The $s^{qNML}$ score does not necessarily define a distribution for $D$, but it is easy to verify that it coincides with the NML score for all networks that are composed of fully connected components. The number of such networks is lower bounded by the number of nonempty partitions of a set of $n$ elements, i.e., the $n^{\text{th}}$ Bell number.

We are now ready to prove some important properties of the qNML score.

## 3.1 qNML Is Score Equivalent

qNML yields equal scores for network structures that encode the same set of independencies. Verma and Pearl [Verma and Pearl, 1991] showed that the equivalent networks are exactly those which a) are the same when directed arcs are substituted by undirected ones and b) which have the same *V-structures*, i.e. the variable triplets $(A, B, C)$ where both $A$ and $B$ are parents of $C$, but there is no arc between $A$ and $B$ (in either direction). Later, Chickering [Chickering, 1995] showed that all the equivalent network structures, and only those structures, can be reached from each other by reversing, one by one, the so-called *covered arcs*, i.e. the arcs from node $A$ to $B$, for which $B$'s parents other than $A$ are exactly $A$'s parents ($G_B = \{A\} \cup G_A$).

We will next state this as a theorem and sketch a proof for it. A more detailed proof appears in Appendix A in the Supplementary Material.

**Theorem 1.** *Let $G$ and $G'$ be two Bayesian network structures that differ only by a single covered arc reversal, i.e., the arc from $A$ to $B$ in $G$ has been reversed in $G'$ to point from $B$ to $A$, then*

$$s^{qNML}(D;G) = s^{qNML}(D;G').$$

*Proof.* Now the scores for structures can be decomposed as $s^{qNML}(D;G) = \sum_{i=1}^{n} s_i^{qNML}(D;G)$ and $s^{qNML}(D;G') = \sum_{i=1}^{n} s_i^{qNML}(D;G')$. Since only the terms corresponding to the variables $A$ and $B$ in these sums are different, it is enough to show that the sum of these two terms are equal for $G$ and $G'$. Since we can assume the data to be fixed we lighten up the notation and write $P^1_{NML}(i, G_i) := P^1_{NML}(D_{i,G_i};G)$ and

$P^1_{NML}(G_i) := P^1_{NML}(D_{G_i}; G)$. Now

$$\begin{aligned} & s_A^{qNML}(D;G) + s_B^{qNML}(D;G) \\ &= \log \frac{P^1_{NML}(A, G_A)}{P^1_{NML}(G_A)} \frac{P^1_{NML}(B, G_B)}{P^1_{NML}(G_B)} \\ &= \log 1 \cdot \frac{P^1_{NML}(B, G_B)}{P^1_{NML}(G_A)} \\ &= \log \frac{P^1_{NML}(B, G'_B)}{P^1_{NML}(G'_A)} \frac{P^1_{NML}(A, G'_A)}{P^1_{NML}(G'_B)} \\ &= s_A^{qNML}(D;G') + s_B^{qNML}(D;G'), \end{aligned}$$

using the equations $\{A\} \cup G_A = G_B$, $\{B\} \cup G'_B = G'_A$, $\{B\} \cup G_B = \{A\} \cup G'_A$, and $G_A = G'_B$ which follow easily from the definition of covered arcs. □

### 3.2 qNML is Consistent

One important property possessed by nearly every model selection criterion is consistency. In our context, consistency means that given a data matrix with $N$ samples coming from a distribution faithful to some DAG $G$, the qNML will give the highest score to the true graph $G$ with a probability tending to one as $N$ increases. We will show this by first proving that qNML is asymptotically equivalent to the widely used BIC criterion which is known to be consistent [Schwarz, 1978, Haughton, 1988]. The outline of this proof follows a similar pattern to that in [Silander et al., 2010] where the consistency of fNML was proved.

The BIC criterion can be written as

$$\mathrm{BIC}(D;G) = \sum_{i=1}^{n} \log P(D_i \mid \hat{\theta}_{i|G_i}) - \frac{q_i(r_i-1)}{2} \log N, \tag{9}$$

where $\hat{\theta}_{i|G_i}$ denotes the maximum likelihood parameters of the conditional distribution of variable $i$ given its parents in $G$.

Since both the BIC and qNML scores are decomposable, we can focus on studying the local scores. We will next show that, asymptotically, the local qNML score equals the local BIC score. This is formulated in the following theorem:

**Theorem 2.** *Let $r_i$ and $q_i$ denote the number of possible values for variable $X_i$ and its possible configurations of parents $G_i$, respectively. As $N \to \infty$,*

$$s_i^{qNML}(D;G) = \log P(D_i \mid \hat{\theta}_{i|G_i}) - \frac{q_i(r_i-1)}{2} \log N.$$

In order to prove this, we start with the definition of qNML and write

$$\begin{aligned} s_i^{qNML}(D;G) = \log & \frac{P(D_{i,G_i} \mid \hat{\theta}_{i,G_i})}{P(D_{G_i} \mid \hat{\theta}_{G_i})} \\ & - (reg(N, q_i r_i) - reg(N, q_i)). \end{aligned} \tag{10}$$

By comparing the equations (9) and (10), we see that proving our clam boils down to showing two things: 1) the terms involving the maximized likelihoods are equal and 2) the penalty terms are asymptotically equivalent. We will formulate these as two lemmas.

**Lemma 1.** *The maximized likelihood terms in equations (9) and (10) are equal:*

$$\frac{P(D_{i,G_i} \mid \hat{\theta}_{i,G_i})}{P(D_{G_i} \mid \hat{\theta}_{G_i})} = P(D_i \mid \hat{\theta}_{i|G_i}).$$

*Proof.* We can write the terms on the left side of the equation as

$$\begin{aligned} P(D_{i,G_i} \mid \hat{\theta}_{i,G_i}) &= \prod_{j,k} \left(\frac{N_{ijk}}{N}\right)^{N_{ijk}}, \text{ and} \\ P(D_{G_i} \mid \hat{\theta}_{G_i}) &= \prod_{j} \left(\frac{N_{ij}}{N}\right)^{N_{ij}}. \end{aligned}$$

Here, $N_{ijk}$ denotes the number of times we observe $X_i$ taking value $k$ when its parents are in $j^{\text{th}}$ configuration in our data matrix $D$. Also, $N_{ij} = \sum_k N_{ijk}$ (and $\sum_{k,j} N_{ijk} = N$ for all $i$). Therefore,

$$\begin{aligned} \frac{P(D_{i,G_i} \mid \hat{\theta}_{i,G_i})}{P(D_{G_i} \mid \hat{\theta}_{G_i})} &= \frac{\prod_{j,k} \left(\frac{N_{ijk}}{N}\right)^{N_{ijk}}}{\prod_j \left(\frac{N_{ij}}{N}\right)^{N_{ij}}} \\ &= \frac{\prod_{j,k} \left(\frac{N_{ijk}}{N}\right)^{N_{ijk}}}{\prod_j \prod_k \left(\frac{N_{ij}}{N}\right)^{N_{ijk}}} \\ &= P(D_i \mid \hat{\theta}_{i|G_i}). \end{aligned}$$

□

Next, we consider the difference of regrets in (10) which corresponds to the penalty term of BIC. The following lemma states that these two are asymptotically equal:

**Lemma 2.** *As $N \to \infty$,*

$$reg(N, q_i r_i) - reg(N, q_i) = \frac{q_i(r_i-1)}{2} \log N + O(1).$$

*Proof.* The regret for a single multinomial variable with $m$ categories can be written asymptotically as

$$reg(N, m) = \frac{m-1}{2} \log N + O(1). \tag{11}$$

For the more precise statement with the underlying assumptions (which are fulfilled in the multinomial case) and for the proof, we refer to [Rissanen, 1996, Grünwald, 2007]. Using this, we have

$$\begin{aligned} reg(N, q_i r_i) - reg(N, q_i) \\ &= \frac{q_i r_i - 1}{2} \log N - \frac{q_i - 1}{2} \log N + O(1) \\ &= \frac{q_i r_i - 1 - q_i + 1}{2} \log N + O(1) \\ &= \frac{q_i (r_i - 1)}{2} \log N + O(1). \end{aligned}$$

□

This concludes our proof since Lemmas 1 and 2 imply Theorem 2.

### 3.3 qNML Equals NML for Many Models

The fNML criterion can be seen as a computationally feasible approximation of the more desirable NML criterion. However, the fNML criterion equals the NML criterion only for the Bayesian network structure with no arcs. It can be shown that the qNML criterion equals the NML criterion for all the networks $G$ whose connected components are tournaments (i.e., complete directed acyclic subgraphs of $G$). These networks include the empty network, the fully connected one and many networks in between having different complexity. While the generating network is unlikely to be composed of tournament components, the result increases the plausibility that qNML is a reasonable approximation for NML in general[1].

**Theorem 3.** *If $G$ consists of $C$ connected components $(G^1, \ldots, G^C)$ with variable sets $(V^1, \ldots, V^C)$, then $\log P_{NML}(D; G) = s^{qNML}(D; G)$ for all data sets $D$.*

*Proof.* The full proof can be found in Appendix C in the Supplementary Material. The proof first shows that NML decomposes for these particular structures, so it is enough to show the equivalence for fully connected graphs. It further derives the number $a(n)$ of different $n$-node networks whose connected components are tournaments, which turns out to be the formula for OEIS sequence A000262[2]. In general this sequence grows rapidly; $1, 1, 3, 13, 73, 501, 4051, 37633, 394353, 4596553, \ldots$. □

### 3.4 qNML is Regular

Suzuki [Suzuki, 2017] defines regularity for a scoring function $Q_n(X \mid Y)$ as follows:

**Definition 1.** *Assume $H_N(X \mid Y') \leq H_N(X \mid Y)$, where $Y' \subset Y$. We say that $Q_N(\cdot \mid \cdot)$ is regular if $Q_N(X \mid Y') \geq Q_N(X \mid Y)$.*

In the definition, $N$ denotes the sample size, $X$ is some random variable, $Y$ denotes the proposed parent set for $X$, and $H_N(\cdot \mid \cdot)$ refers to the empirical conditional entropy. Suzuki [Suzuki, 2017] shows that BDeu violates this principle and demonstrates that this can cause the score to prefer more complex networks even though the data do not support this. Regular scores are also argued to be computationally more efficient when applied with branch-and-bound type algorithms for Bayesian network structure learning [Suzuki and Kawahara, 2017].

By analyzing the penalty term of the qNML scoring function, one can prove the following statement:

**Theorem 4.** *qNML score is regular.*

*Proof.* The proof is given in Appendix B in the Supplementary Material. □

As fNML criterion differs from qNML only by how the penalty term is defined, we obtain the following result with little extra work:

**Theorem 5.** *fNML score is regular.*

*Proof.* The proof is given in Appendix B in the Supplementary Material. □

Suzuki [Suzuki, 2017] independently introduces a Bayesian Dirichlet quotient (BDq) score that can also be shown to be score equivalent and regular. However, like BDeu, this score features a single hyperparameter $\alpha$, and our initial studies suggest that BDq is also very sensitive to this hyperparameter (see Appendix D in the Supplementary Material), the issue that was one of the main motivations to develop a parameter-free model selection criterion like qNML.

# 4 EXPERIMENTAL RESULTS

We empirically compare the capacity of qNML to that of BIC, BDeu ($\alpha = 1$) and fNML in identifying the data generating structures, and producing models that are predictive and parsimonious. It seems that none of the criteria uniformly outperform the others in all these desirable aspects of model selection criteria.

[1] A claim that is naturally subject for further study.

[2] https://oeis.org/A000262

### 4.1 Finding Generating Structure

In our first experiment, we took five widely used benchmark Bayesian networks[3], sampled data from them, and tried to learn the generating structure with the different scoring functions using various sample sizes. We used the following networks: Asia ($n = 5$, 8 arcs), Cancer ($n = 5$, 4 arcs), Earthquake ($n = 5$, 4 arcs), Sachs ($n = 11$, 17 arcs) and Survey ($n = 6$, 6 arcs). These networks were picked in order to use the dynamic programming based exact structure learning [Silander and Myllymäki, 2006] which limited the number $n$ of variables to less than 20. We measured the quality of the learned structures using structural Hamming distance (SHD) [Tsamardinos et al., 2006].

Figure 2 shows SHDs for all the scoring criteria for each network. Sample sizes range from 10 to 10000 and the shown results are averages computed from 1000 repetitions. None of the scores dominates in all settings considered. BIC fares well when the sample size is small as it tends to produce a nearly empty graph which is a good answer in terms of SHD when the generating networks are relatively sparse. qNML obtains strong results in the Earthquake and Asia networks, being the best or the second best with all the sample sizes considered.

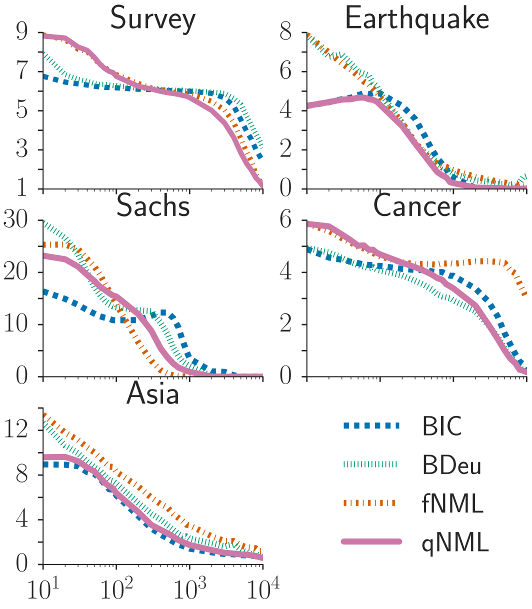

Figure 2: Sample size versus SHD with data generated from real world DAGs.

Figure 3 summarizes the SHD results in all networks by showing the average rank for each score. The ranking was done by giving the score with the lowest SHD rank 1 and the worst one rank 4. In case of ties, the methods with the same SHD were given the same rank. The shown results are averages computed from 5000 values (5 networks, 1000 repetitions). From this, we can see that qNML never has the worst average ranking, and it has the best ranking with sample sizes greater than 300. This suggests that qNML is overall a safe choice in structure learning, especially with moderate and large sample sizes.

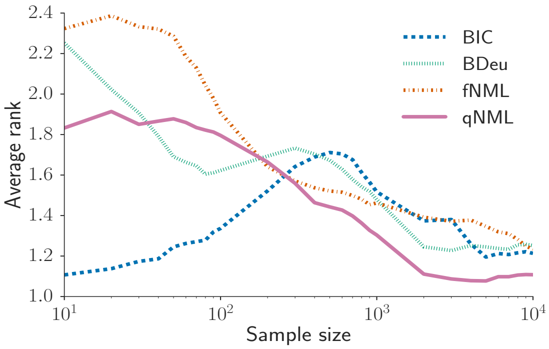

Figure 3: Average ranks for the scoring functions in structure learning experiments.

### 4.2 Prediction and Parsimony

To empirically compare the model selection criteria, we took 20 UCI data sets [Lichman, 2013] and ran train and test experiments for all of them. To better compare the performance over different sample sizes, we picked different fractions $(10\%, 20\%, \ldots, 90\%)$ of the data sets as training data and the rest as the test data. This was done for 1000 different permutations of each data set. The training was conducted using the dynamic programming based exact structure learning algorithm.

When predicting $P(d_{test}|D_{train}, S, \theta)$ with structures $S$ learned by the BDeu score, we used the Bayesian predictive parameter values (BPP) $\theta_{ijk} \propto N_{ijk} + \frac{1}{r_i q_i}$. In the spirit of keeping the scores hyperparameter-free, for structures learned by the other model selection criteria, we used the sequential predictive NML (sNML) parametrization $\theta_{ijk} \propto e(N_{ijk})(N_{ijk} + 1)$, where $e(n) = (\frac{n+1}{n})^n$ as suggested in [Rissanen and Roos, 2007].

For each train/test sample, we ranked the predictive performance of the models learned by the four different scores (rank 1 being the best and 4 the worst). Table 2 features the average rank for different data sets, the average being taken over 1000 different train/test

[3] Bayesian Network Repository: `http://www.bnlearn.com/bnrepository/`

Table 2: Average predictive performance rank over different sample sizes for different model selection criteria in 20 different data sets.

| Data | N | BDeu | BIC | fNML | qNML |
|---|---|---|---|---|---|
| | | BPP | sNML | sNML | sNML |
| 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 |

samples for each 9 sample sizes. BIC's bias for simplicity makes it often win (written bold in the table) with small sample sizes, but it performs worst (underlined) for the larger sample sizes (for the same reason), while fNML seems to be good for large sample sizes. The striking feature about the qNML is its robustness. It is usually between BIC and fNML for all the sample sizes making it a "safe choice". This can be quantified if we further average the columns of Table 2, yielding the average ranks of $2.95, 2.57, 2.10$, and $2.37$, with standard deviations $0.49, 0.90, 0.76$, and $0.43$. While fNML achieves on average the best rank, the runner-up qNML has the lowest standard deviation.

Figure 1 shows how fNML still sometimes behaves strangely in terms of model complexity as measured by the number of parameters in the model. qNML, instead, appears to yield more parsimonious models. To study the concern of fNML producing too complex models for small sample sizes, we studied the number of parameters in models produced by different scores when using 10% of each data set for structure learning.

Looking at the number of parameters for the same 20 data sets again features BIC's preference for simple models (Table 3). qNML usually (19/20) yields more parsimonious models than fNML that selects the most complex model for 7 out of 20 data sets.

Table 3: Average number of parameters in models for different scores in 20 different data sets.

| Data | N | BDeu | BIC | fNML | qNML |
|---|---|---|---|---|---|
| Iris | 15 | 37 | **23** | 33 | 29 |
| PostOpe | 18 | 1217 | **19** | 397 | 146 |
| Ecoli | 34 | 182 | **31** | 162 | 77 |
| Liver | 35 | 45 | **15** | 61 | 24 |
| Wine | 36 | 16521 | **70** | 807 | 205 |
| Glass | 44 | 1677 | **48** | 506 | 97 |
| Thyroid | 44 | 40 | **23** | 66 | 28 |
| HeartSt | 54 | 16861 | **44** | 1110 | 256 |
| BreastC | 58 | 25797 | **49** | 3767 | 844 |
| HeartHu | 60 | 1634 | **43** | 792 | 90 |
| HeartCl | 62 | 34381 | **47** | 1433 | 404 |
| BcWisco | 70 | 4630 | **42** | 603 | 89 |
| Diabete | 77 | 39 | **22** | 216 | 34 |
| TicTacT | 96 | 13701 | **25** | 1969 | 767 |
| Balance | 126 | **20** | 24 | 49 | 611 |
| Yeast | 149 | 71 | **31** | 265 | 75 |
| Abalone | 418 | 91 | **46** | 150 | 63 |
| PageBlo | 548 | 703 | **45** | 380 | 56 |
| Shuttle | 5800 | 535 | **99** | 717 | 130 |
| Adult | 6513 | 699 | **479** | 1555 | 945 |

The graphs for different sample sizes for both predictive accuracy and the number of parameters can be found in Appendix E in the Supplementary Material.

## 5 CONCLUSION

We have presented qNML, a new model selection criterion for learning structures of Bayesian networks. While being competitive in predictive terms, it often yields significantly simpler models than other common model selection criteria other than BIC that has a very strong bias for simplicity. The computational cost of qNML equals the cost of the current state-of-the-art criteria. The criterion also gives equal scores for models that encode the same independence hypotheses about the joint probability distribution. qNML also coincides with the NML criterion for many models. In our experiments, the qNML criterion appears as a safe choice for a model selection criterion that balances parsimony, predictive capability and the ability to quickly converge to the generating model.

### 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.

## References

[Akaike, 1973] Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In Petrox, B. and Caski, F., editors, *Proceedings of the Second International Symposium on Information Theory*, pages 267–281, Budapest. Akademiai Kiado.

[Buntine, 1991] Buntine, W. (1991). Theory refinement on Bayesian networks. In D'Ambrosio, B., Smets, P., and Bonissone, P., editors, *Proceedings of the Seventh Conference on Uncertainty in Artificial Intelligence*, pages 52–60. Morgan Kaufmann Publishers.

[Chickering, 1995] Chickering, D. M. (1995). A Transformational Characterization of Equivalent Bayesian Network Structures. In *UAI '95: Proceedings of the Eleventh Annual Conference on Uncertainty in Artificial Intelligence*, pages 87–98. Morgan Kaufmann.

[Eggeling et al., 2014] Eggeling, R., Roos, T., Myllymäki, P., Grosse, I., et al. (2014). Robust learning of inhomogeneous PMMs. In Kaski, S. and Corander, J., editors, *Proceedings of the Seventeenth International Conference on Artificial Intelligence and Statistics*, pages 229–237. PMLR.

[Grünwald, 2007] Grünwald, P. (2007). *The Minimum Description Length Principle*. MIT Press.

[Haughton, 1988] Haughton, D. (1988). On the choice of a model to fit data from an exponential family. *Annals of Statistics*, 16(1):342–355.

[Heckerman, 1995] Heckerman, D. (1995). A Bayesian approach to learning causal network. In Besnard, P. and Hanks, S., editors, *Uncertainty in Artificial Intelligence: Proceedings of the Eleventh Conference*, pages 285–295, Montreal, Canada.

[Heckerman et al., 1995] Heckerman, D., Geiger, D., and Chickering, D. (1995). Learning Bayesian networks: The combination of knowledge and statistical data. *Machine Learning*, 20(3):197–243.

[Ide et al., 2014] Ide, J. S., Zhang, S., and Li, C. S. (2014). Bayesian network models in brain functional connectivity analysis. *International Journal of Approximate Reasoning*, 56(1 Pt 1).

[Kontkanen et al., 2003] Kontkanen, P., Buntine, W., Myllymäki, P., Rissanen, J., and Tirri, H. (2003). Efficient computation of stochastic complexity. In Bishop, C. and Frey, B., editors, *Proceedings of the Ninth International Conference on Artificial Intelligence and Statistics*, pages 233–238. Society for Artificial Intelligence and Statistics.

[Lichman, 2013] Lichman, M. (2013). UCI machine learning repository.

[Liu et al., 2012] Liu, Z., Malone, B., and Yuan, C. (2012). Empirical evaluation of scoring functions for Bayesian network model selection. *BMC Bioinformatics*, 13(15):S14.

[Määttä et al., 2016] Määttä, J., Schmidt, D. F., and Roos, T. (2016). Subset selection in linear regression using sequentially normalized least squares: Asymptotic theory. *Scandinavian Journal of Statistics*, 43(2):382–395.

[Pearl, 1988] Pearl, J. (1988). *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann Publishers, San Mateo, CA.

[Rissanen, 1978] Rissanen, J. (1978). Modeling by shortest data description. *Automatica*, 14:445–471.

[Rissanen, 1996] Rissanen, J. (1996). Fisher information and stochastic complexity. *IEEE Transactions on Information Theory*, 42(1):40–47.

[Rissanen and Roos, 2007] Rissanen, J. and Roos, T. (2007). Conditional NML models. In *Proceedings of the Information Theory and Applications Workshop (ITA-07)*, San Diego, CA.

[Sachs et al., 2002] Sachs, K., Gifford, D., Jaakkola, T., Sorger, P., and Lauffenburger, D. A. (2002). Bayesian network approach to cell signaling pathway modeling. *Science's STKE*, 2002(148):38–38.

[Savickas and Vasilecas, 2014] Savickas, T. and Vasilecas, O. (2014). Bayesian belief network application in process mining. In *Proceedings of the 15th International Conference on Computer Systems and Technologies*, CompSysTech '14, pages 226–233, New York, NY, USA. ACM.

[Schwarz, 1978] Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics*, 6:461–464.

[Shtarkov, 1987] Shtarkov, Y. (1987). Universal sequential coding of single messages. *Problems of Information Transmission*, 23:3–17.

[Silander et al., 2007] Silander, T., Kontkanen, P., and Myllymäki, P. (2007). On sensitivity of the MAP Bayesian network structure to the equivalent sample size parameter. In Parr, R. and van der Gaag, L., editors, *Proceedings of the 23rd Conference on Uncertainty in Artificial Intelligence (UAI–07)*, pages 360–367. AUAI Press.

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

[Silander et al., 2008] Silander, T., Roos, T., Kontkanen, P., and Myllymäki, P. (2008). Factorized normalized maximum likelihood criterion for learning Bayesian network structures. In *Proceedings of the 4th European Workshop on Probabilistic Graphical Models (PGM-08)*, pages 257–264, Hirtshals, Denmark.

[Silander et al., 2010] Silander, T., Roos, T., and Myllymäki, P. (2010). Learning locally minimax optimal Bayesian networks. *International Journal of Approximate Reasoning*, 51(5):544 – 557.

[Steck, 2008] Steck, H. (2008). Learning the Bayesian network structure: Dirichlet prior vs data. In McAllester, D. A. and Myllymäki, P., editors, *Proceedings of the 24th Annual Conference on Uncertainty in Artificial Intelligence (UAI–08)*, pages 511–518. AUAI Press.

[Suzuki, 2017] Suzuki, J. (2017). A theoretical analysis of the BDeu scores in Bayesian network structure learning. *Behaviormetrika*, 44(1):97–116.

[Suzuki and Kawahara, 2017] Suzuki, J. and Kawahara, J. (2017). Branch and bound for regular Bayesian network structure learning. The 33rd Conference on Uncertainty in Artificial Intelligence, UAI 2017, August 11-15, 2017, Sydney, Australia.

[Szpankowski and Weinberger, 2012] 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.

[Tsamardinos et al., 2006] Tsamardinos, I., Brown, L. E., and Aliferis, C. F. (2006). The max-min hill-climbing Bayesian network structure learning algorithm. *Machine Learning*, 65(1):31–78.

[Verma and Pearl, 1991] Verma, T. and Pearl, J. (1991). Equivalence and synthesis of causal models. In *Proceedings of the Sixth Annual Conference on Uncertainty in Artificial Intelligence (UAI-90)*, pages 255–270, New York, NY, USA. Elsevier Science Inc.