Supplementary Material for the article *Quotient Normalized Maximum Likelihood Criterion for Learning Bayesian Network Structures*

A Regularity proof

A.1 Preliminaries

We start by recalling the definition of regularity (Suzuki, 2017):

Definition 1. *Assume $H_N(X \mid Y') \leq H_N(X \mid Y)$, where $Y' \subset Y$. We say that the scoring function $Q_N(\cdot \mid \cdot)$ 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 $H_N(X \mid Y)$ refers to the empirical conditional entropy based on N samples of variables X and Y.

Let X be a categorical random variable with r possible values. Let U denote a possible parent set with q different combinations of values for the variables, and V a parent set with m different configurations. Assume that we have observed N samples of (X, U, V) (denoted by x_N, u_N and v_N) and $H_N(X \mid U) \leq H_N(X \mid U \cup V)$ holds.

Recall the definition of the qNML score:

$$\begin{aligned} Q_N^{qnml}(X \mid U) &= \log P(x_n \mid \hat{\theta}_{X|U}\ ) - (reg(N, rq) - reg(N, q)) \\ &= \log P(x_N \mid \hat{\theta}_{X|U}\ ) - \log \frac{C(N, rq)}{C(N, q)}, \end{aligned}$$

where $C(N, r)$ is the normalizing constant the of the NML distribution for a categorical variable with r possible values and sample size N and $\hat{\theta}_{X|U}$

denotes the maximum likelihood parameters of the conditional distribution of X given U which are computed from the data (x_N, u_N).

In order to prove the regularity, we need the following three lemmas:

Lemma 1. *We can write $C(N,k)$ as a polynomial of k, formally*

$$C(N,k) = \sum_{j=1}^{N} a_j k^j,$$

where $a_j > 0$.

Lemma 2. *Assume $H_N(X \mid Y') \leq H_N(X \mid Y)$, where $Y' \subset Y$. Now $\log P(x_N \mid \hat{\theta}_{X|Y}\ ) = \log P(x_N \mid \hat{\theta}_{X|Y'}\ )$.*

Lemma 3. *Let $r \in \mathbb{N}, r \geq 2$. The function $k \mapsto \frac{C(N,rk)}{C(N,k)}$ is increasing for every $k \geq 2$.*

We present the proofs of these lemmas in Section A.3.

A.2 The main proof

Theorem 1. *qNML score is regular.*

Proof. We want to show that

$$Q_N^{qnml}(X \mid U) \geq Q_N^{qnml}(X \mid U \cup V).$$

assuming $H_N(X \mid U) \leq H_N(X \mid U \cup V)$. Using the entropy assumption and Lemma 2 implies that the maximized likelihood terms are equal. In order to prove the claim, it suffices to study the penalty terms, and we want to show that

$$\begin{aligned} -\left(reg(N,rq) - reg(N,q)\right) &\geq -\left(reg(N,rqm) - reg(N,qm)\right) \\ \log \frac{C(N,rq)}{C(N,q)} &\leq \log \frac{C(N,rqm)}{C(N,qm)}. \end{aligned}$$

This holds, since logarithm is an increasing function, and $q \leq qm$, so we can apply Lemma 3 to conclude the proof.

□

A.3 Proofs of lemmas

Lemma 1. *$C(N,k)$ can be written as a polynomial of k, formally*

$$C(N,k) = \sum_{j=1}^{N} a_j k^j,$$

where $a_j > 0$.

Proof. Mononen and Myllymäki (2008) derive the following representation for the normalizing constant

$$C(N,k) = \sum_{l=0}^{N-1} \frac{(N-1)^{\underline{l}} k^{\overline{l+1}}}{N^{l+1}\ l!},$$

where $x^{\underline{l}}$ and $x^{\bar{l}}$ denote falling and rising factorials, respectively.

We utilize the fact that the rising factorial can be represented as polynomial using unsigned Stirling numbers of the first kind (see Adamchik (1997), for instance)

$$\begin{aligned}
C(N,k) &= \sum_{l=0}^{N-1} \frac{(N-1)^{\underline{l}} k^{\overline{l+1}}}{N^{l+1}\ l!} \\
&= \sum_{l=0}^{N-1} b_l\ k^{\overline{l+1}} \\
&= \sum_{l=0}^{N-1} b_l \left(\sum_{j=1}^{l+1} |s(l+1,j)|\ k^j \right) \\
&= \sum_{l=0}^{N-1} \left(\sum_{j=1}^{N} b_l\ |s(l+1,j)|\ k^j \right) \\
&= \sum_{j=1}^{N} \left(\sum_{l=0}^{N-1} b_l\ |s(l+1,j)|\ k^j \right) \\
&= \sum_{j=1}^{N} \left(\sum_{l=0}^{N-1} b_l\ |s(l+1,j)| \right) k^j \\
&= \sum_{j=1}^{N} a_j k^j,
\end{aligned}$$

where $s(i,j)$ denotes the (signed) Stirling number of the first kind and

$$a_j = \left(\sum_{l=0}^{N-1} \frac{(N-1)^{\underline{l}}}{N^{l+1} l!} \, |s(l+1,j)| \right),$$

$a_j > 0$ for all j. On the second row, we denoted $b_l = (N-1)^{\underline{l}}/(N^{l+1}l!)$. On the row 4, we used the property of Stirling numbers: $s(i,j) = 0$ for all $j > i$. □

Lemma 2. *Assume $H_N(X \mid Y') \leq H_N(X \mid Y)$, where $Y' \subset Y$. Now $\log P(x_N \mid \hat{\theta}_{X|Y}\ ) = \log P(x_N \mid \hat{\theta}_{X|Y'})$.*

Proof. We can write the logarithm of the maximized likelihood, $\log P(x_N \mid \hat{\theta}_{X|Y}\ )$, as follows (Koller and Friedman, 2009)

$$\begin{aligned} \log P(x_N \mid \hat{\theta}_{X|Y}\ ) &= -N \left(H_N(X) - I_N(X;Y) \right) \\ &= -N \ H_N(X \mid Y), \end{aligned}$$

where $I_N(\cdot;\cdot)$ is the empirical mutual information. This implies that the assumption

$$H_N(X \mid Y') \leq H_N(X \mid Y)$$

is equivalent to

$$\log P(x_N \mid \hat{\theta}_{X|Y}\ ) \geq \log P(x_N \mid \hat{\theta}_{X|Y'}\ ).$$

Actually we must have the equality holding in the above expression, since

$$H_N(X \mid Y') < H_N(X \mid Y)$$

would imply that

$$I_N(X; Z \mid Y) < 0,$$

where $Z = Y \setminus Y'$, which is impossible. □

Lemma 3. *Let $r \in \mathbb{N}, r \geq 2$. The function $k \mapsto \frac{C(N,rk)}{C(N,k)}$ is increasing for every $k \geq 2$.*

Proof. Lemma 1 lets us to write

$$C(N,k) = \sum_{j=1}^{N} a_j k^j \tag{1}$$

and, similarly,

$$C(N, rk) = \sum_{j=1}^{N} a_j r^j k^j. \tag{2}$$

From this, it it easy see that the derivative of the quotient, $d/dk(C(N, rk)/C(N, k))$, will be a ratio of two polynomials of k. Our goal is to show that the polynomial in the numerator has positive coefficients, which will guarantee the positivity of derivative for every $k > 0$, and thus imply that the original function is increasing (polynomial in the denominator is squared and non-zero for $k > 0$, so it can be ignored).

Derivatives of (1) and (2) are obtained easily:

$$\begin{aligned}\frac{d}{dk}C(N, k) &= \sum_{j=1}^{N} j a_j k^{j-1} \\ &= \sum_{j=0}^{N-1} (j+1) a_{j+1} k^j\end{aligned}$$

and

$$\begin{aligned}\frac{d}{dk}C(N, rk) &= \sum_{j=1}^{N} j a_j r^j k^{j-1} \\ &= \sum_{j=0}^{N-1} (j+1) a_{j+1} r^{j+1} k^j.\end{aligned}$$

Consider next the products found in the derivative of the quotient. We obtain

$$\begin{aligned}\left(\frac{d}{dk}C(N, rk)\right) C(N, k) &= \left(\sum_{j=0}^{N-1} (j+1) a_{j+1} r^{j+1} k^j\right) \left(\sum_{l=1}^{N} a_l k^l\right) \\ &= \sum_{i=1}^{2N-1} \left(\sum_{j+l=i} (j+1) a_{j+1} r^{j+1} a_l\right) k^i\end{aligned}$$

and

$$\left(\frac{d}{dk}C(N,k)\right)C(N,rk) = \left(\sum_{j=0}^{N-1}(j+1)a_{j+1}k^j\right)\left(\sum_{l=1}^{N}a_l r^l k^l\right)$$
$$= \sum_{i=1}^{2N-1}\left(\sum_{j+l=i}(j+1)a_{j+1}a_l r^l\right)k^i.$$

Subtracting these two expression yields

$$\left(\frac{d}{dk}C(N,rk)\right)C(N,k) - \left(\frac{d}{dk}C(N,k)\right)C(N,rk)$$
$$= \sum_{i=1}^{2N-1}\left(\sum_{j+l=i}(j+1)a_{j+1}r^{j+1}a_l\right)k^i - \sum_{i=1}^{2N-1}\left(\sum_{j+l=i}(j+1)a_{j+1}a_l r^l\right)k^i$$
$$= \sum_{i=1}^{2N-1}\left(\sum_{j+l=i}(j+1)a_{j+1}a_l(r^{j+1}-r^l)\right)k^i$$

which is the polynomial in the numerator of the derivative of $C(N,rk)/C(N,k)$. Next, we study the coefficient of k^i, if $i \leq N$

$$\sum_{j+l=i}(j+1)a_{j+1}a_l(r^{j+1}-r^l) = \sum_{l=1}^{i}(i-l+1)a_{i-l+1}a_l(r^{i-l+1}-r^l)$$
$$= \sum_{l=1}^{i}(i-l+1)c_l$$
$$= \sum_{k=1}^{\lfloor i/2\rfloor}(i-k+1)c_k + (i-(i-k+1)+1)c_{i-k+1}$$
$$= \sum_{k=1}^{\lfloor i/2\rfloor}(i-k+1)c_k + kc_{i-k+1}$$
$$= \sum_{k=1}^{\lfloor i/2\rfloor}(i-k+1)c_k - kc_k$$
$$= \sum_{k=1}^{\lfloor i/2\rfloor}(i-2k+1)c_k.$$

On the first row, we re-wrote sum using only one running index. On the second row we denoted $c_l = a_{i-l+1}a_l(r^{i-l+1} - r^l)$. On the third row, we re-arranged the sum so that we are summing over pairs of terms of the original sum: the first and the last term, the second and the second to last, and so on. This resulting sum has $\lfloor i/2 \rfloor$ terms. We have to use the floor-function since if i is odd, there exists an index l' in the original sum such that $r^{i-l'+1} - r^{l'} = 0$. On the fifth row, we make use of the identity $c_k = -c_{i-k+1}$ which is straightforward to verify. From the last row, we can observe that every term of the sum is positive since $i - 2k + 1$ and $r^{i-k+1} - r^k$ are both positive if $k \leq (i+1)/2$ which holds since k ranges from 1 to $\lfloor i/2 \rfloor$.

Let us now consider the situation where $n < i \leq 2N - 1$. We start with the special case where $i = 2N - 1$. Then, we have only one term in the sum

$$\begin{aligned}\sum_{j+l=i}(j+1)a_{j+1}a_l(r^{j+1} - r^l) &= \sum_{l=N}^{N}(2N-1-l+1)a_{2N-1-l+1}a_l(r^{2N-1-l+1} - r^l) \\ &= Na_Na_N(r^N - r^N) \\ &= 0.\end{aligned}$$

Now, let $N < i < 2N - 1$, we follow a similar procedure as before to manip-

ulate the sum

$$\begin{aligned}
\sum_{j+l=i}(j+1)a_{j+1}a_l(r^{j+1}-r^l) &= \sum_{l=i-N+1}^{N}(i-l+1)a_{i-l+1}a_l(r^{i-l+1}-r^l)\\
&= \sum_{l=i-N+1}^{N}(i-l+1)c_l\\
&= \sum_{k=1}^{2N-i}(N-k+1)c_{i-N+k}\\
&= \sum_{k=1}^{\lfloor N-i/2\rfloor}(N-k+1)c_{i-N+k}\\
&+ (N-(2N-i-k+1)+1)c_{i-N+(2N-i-k+1)}\\
&= \sum_{k=1}^{\lfloor N-i/2\rfloor}(N-k+1)c_{i-N+k}+(i-N+k)c_{N-k+1}\\
&= \sum_{k=1}^{\lfloor N-i/2\rfloor}(N-k+1)c_{i-N+k}-(i-N+k)c_{i-N+k}\\
&= \sum_{k=1}^{\lfloor N-i/2\rfloor}(N-k+1-(i-N+k))c_{i-N+k}\\
&= \sum_{k=1}^{\lfloor N-i/2\rfloor}(2N-i-2k+1)c_{i-N+k}.
\end{aligned}$$

It is now easy to verify that $(2N-i-2k+1)$ and c_{i-N+k} are positive if $k \leq N-(i-1)/2$ which holds since k ranges from 1 to $\lfloor N-i/2\rfloor$. The floor function is again used when we sum over pairs of terms since if i is odd there is zero-term. Since all the coefficients are non-negative and the $k \geq 2$, the derivative is positive. This implies that the original function is increasing. □

References

Adamchik, V. (1997). On stirling numbers and euler sums. *Journal of Computational and Applied Mathematics*, 79(1):119 – 130.

Koller, D. and Friedman, N. (2009). *Probabilistic Graphical Models: Principles and Techniques*. MIT Press.

Mononen, T. and Myllymäki, P. (2008). On the multinomial stochastic complexity and its connection to the birthday problem. In *Proceedings of the International Conference on Information Theory and Statistical Learning*, Las Vegas, NV.

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