On Model Selection, Bayesian Networks, and the Fisher Information Integral

Yuan Zou and Teemu Roos

Helsinki Institute for Information Technology HIIT
Department of Computer Science, University of Helsinki
Gustaf Hällströmin katu 2b, Helsinki, 00014, Finland
{yuan.zou,teemu.roos}@hiit.fi
www.hiit.fi/cosco/promo

Abstract. We study BIC-like model selection criteria and in particular, their refinements that include a constant term involving the Fisher information matrix. We observe that for complex Bayesian network models, the constant term is a negative number with a very large absolute value that dominates the other terms for small and moderate sample sizes. We show that including the constant term degrades model selection accuracy dramatically compared to the standard BIC criterion where the term is omitted. On the other hand, we demonstrate that exact formulas such as Bayes factors or the normalized maximum likelihood (NML), or their approximations that are not based on Taylor expansions, perform well. A conclusion is that in lack of an exact formula, one should use either BIC, which is a very rough approximation, or a very close approximation but not an approximation that is truncated after the constant term.

1 Introduction

A Bayesian network encodes joint probability distributions of a set of random variables via a directed acyclic graph (DAG). Bayesian networks with different network topologies form a lattice-like hierarchy with both nested and non-nested relations where the model complexity varies greatly. It therefore becomes imperative to regularize model complexity when learning the structure from finite data. In this paper we study BIC-like model selection criteria that can be derived via Laplace approximation, and their properties in the case of Bayesian networks. Our main focus is on complexity regularization and in particular, the lower-order terms such as the constant term, $\log \int_\Theta \sqrt{\det I(\theta)}\, d\theta$, which involves the Fisher information, $I(\theta)$. The omission of such terms in the standard BIC formula can be justified by asymptotic arguments.

An approximation of the Bayes factor (or the marginal likelihood) [5] under Jeffreys' prior, where the constant term is retained, results in a so called Fisher information approximation (FIA). We show that contrary to what might be expected, namely that a more refined approximation such as FIA should be better than a rough approximation such as BIC, FIA tends to be extremely inaccurate for small and moderate sample sizes. In particular, we observe that

for complex Bayesian network models (with thousands or tens of thousands of independent parameters), the constant term is a negative number with a very large absolute value that dominates all the other terms in FIA unless the sample size is greater than the number of parameters. The absolute value of the term grows rapidly with increasing model order, which makes the FIA criterion favor complex models unless the sample size is extremely large. Similar results have been reported for other model families such as the exponential model [9] and Markov sources [15].

In this paper, we first review the FIA approximation and discuss its relation to certain other model selection criteria. Because there is no closed form formula for the Fisher information integral under most model families, including Bayesian networks, we illustrate how to estimate it with arbitrarily fine precision using Monte Carlo techniques. We carry out model selection experiments where we highlight the complexity regularization performance by the various criteria in order to determine which of the criteria are safe and which should be avoided under given conditions.

2 The Fisher information approximation

In this section, we discuss what we call the Fisher information approximation (FIA), and relate it to other model selection criteria. First, let's consider the Bayes factor criterion before investigating asymptotic approximations. The Bayes factor measures the ratio of marginal likelihoods between competing models.

$$\mathrm{BF}_{12} = \frac{p(x^n \,;\, \mathcal{M}_1)}{p(x^n \,;\, \mathcal{M}_2)} = \frac{\int_{\Theta_{\mathcal{M}_1}} p(x^n \,;\, \theta_1, \mathcal{M}_1)\, p(\theta_1)\, d\theta_1}{\int_{\Theta_{\mathcal{M}_2}} p(x^n \,;\, \theta_2, \mathcal{M}_2)\, p(\theta_2)\, d\theta_2}, \tag{1}$$

where $p(\theta_1)$ and $p(\theta_2)$ denote the parameter priors under the two models, $\mathcal{M}_1$ and $\mathcal{M}_2$, respectively.

The marginal likelihood has a built-in, implicit penalty for model complexity, see [10]. A closed form solution for the marginal likelihood is only available for a limited set of model families when conjugate priors exist. For other model families, we usually need to resort to sampling methods such as MCMC methods [3]. Furthermore, even when an efficient formula for calculating Bayes factors is available, like in the case of Bayesian networks discussed in this work, model selection performance may be highly sensitive to the choice of the associated parameter priors [18].

2.1 Approximation of marginal likelihood

To avoid the selection of a specific prior and to obtain a more objective method for model selection, we can use asymptotic (large-sample) approximations of the Bayes factor or the marginal likelihood such as the classic BIC criterion [16]. The BIC can be obtained via Laplace approximation, which involves a Taylor expansion of the log-likelihood function around its maximum. For instance, if we

have a model $\mathcal{M}$ with $d_{\mathcal{M}}$ free parameters, jointly denoted by $\theta \in \Theta_{\mathcal{M}}$, and a data set x^n with sample size n, the Laplace approximation of the log-marginal likelihood is given by

$$\begin{aligned}\log p(x^n\,;\,\mathcal{M}) &= \log \int_{\Theta_{\mathcal{M}}} p(x^n\,;\,\theta,\mathcal{M})\,p(\theta)\,d\theta \\ &= \log p(x^n\,;\,\hat{\theta}(x^n)) + \log p(\hat{\theta}(x^n)) \\ &\quad + \frac{d_{\mathcal{M}}}{2}\log(2\pi) - \frac{1}{2}\log\det \hat{I}(\hat{\theta}(x^n)) + o(1),\end{aligned} \tag{2}$$

where $p(\theta)$ is the parameter prior, the maximum likelihood parameters are denoted by $\hat{\theta}(x^n)$, and $\hat{I}(\theta)$ is the empirical Fisher information matrix at θ. If the distributions of model $\mathcal{M}$ are independent and identically distributed (i.i.d.), by the law of large numbers, we have the average per-symbol empirical Fisher information converging to its expectation $I(\hat{\theta}(x))$:

$$n^{-1}\hat{I}(\hat{\theta}(x^n)) \to I(\hat{\theta}(x^n)),\ \text{where } I(\theta) = \mathbb{E}_\theta\,\hat{I}(\theta). \tag{3}$$

Then by simple manipulation, the fourth term in Eq. (2) can be approximated as

$$\frac{1}{2}\log\det \hat{I}(\hat{\theta}(x^n)) = \frac{d_{\mathcal{M}}}{2}\log n + \frac{1}{2}\log\det I(\hat{\theta}(x^n)) + o(1). \tag{4}$$

Finally, we can obtain the approximation of log marginal likelihood as

$$\begin{aligned}\log p(x^n\,;\,\mathcal{M}) &= \log p(x^n\,;\,\hat{\theta}(x^n)) - \frac{d_{\mathcal{M}}}{2}\log n \\ &\quad + \log p(\hat{\theta}(x^n)) + \frac{d_{\mathcal{M}}}{2}\log(2\pi) - \frac{1}{2}\log\det I(\hat{\theta}(x^n)) + o(1).\end{aligned} \tag{5}$$

When the sample size n increases, lower order terms that are independent of n will eventually be dominated by the terms that grow with n. Therefore, for very large sample sizes, we can omit the last four terms in Eq. (5) and change the sign to obtain the familiar BIC criterion:

$$\mathrm{BIC}(x^n\,;\,\mathcal{M}) = -\log p(x^n\,;\,\hat{\theta}_{\mathcal{M}}(x^n)) + \frac{d_{\mathcal{M}}}{2}\log n, \tag{6}$$

To get a more precise approximation, we would need to include the lower-order terms as well. However, they depend on the chosen prior. An often quoted objective choice is the Jeffreys prior. The Jeffreys prior was initially proposed to acquire an invariance property under reparameterization [4]. Later studies have shown that the Jeffreys prior also has several minimax properties [1], [11]. For example, it achieves asymptotic minimax risk for model families with smooth finite-dimensional parameters. This requirement is met in most of the cases for Bayesian networks. However, when the maximum likelihood parameters lie on the boundary of the parameter space, Jeffreys prior may fail to achieve the asymptotic minimax property. In this work, for the sake of simplicity, we assume that the necessary conditions are satisfied and ignore the boundary issues.

For further discussion on the regularity conditions and an alternative BIC-like criterion, called NIP-BIC, see [20].

The Jeffreys prior is proportional to the square root of the determinant of the Fisher information matrix:

$$p(\theta) = \mathrm{FII}(\mathcal{M})^{-1}\sqrt{\det I(\theta)}. \tag{7}$$

The normalizing term, which we call the the *Fisher information integral* (FII), is given by

$$\mathrm{FII}(\mathcal{M}) = \int_{\Theta_{\mathcal{M}}} \sqrt{\det I(\theta)}\, d\theta.$$

Plugging Eq. (7) in Eq. (5), we get the Fisher information approximation:

$$\mathrm{FIA}(x^n\,;\,\mathcal{M}) = \log p(x^n\,;\,\hat{\theta}_{\mathcal{M}}(x^n)) - \frac{d_{\mathcal{M}}}{2}\log\frac{n}{2\pi} - \log \mathrm{FII}(\mathcal{M}) + o(1). \tag{8}$$

For Bayesian networks, which is the model class studied in this work, the Jeffreys prior has been derived in [7]. Unfortunately, as the authors showed, evaluating it is NP-hard. Therefore, it is unlikely that an efficient formula for FII could be obtained for Bayesian networks. To get around this difficulty, we introduce a way to approximate FII by first linking the marginal likelihood to another model selection criterion via the FIA formula.

2.2 Approximation of normalized maximum likelihood

The FIA formula is important not only because it approximates the Bayesian marginal likelihood. It also coincides with the asymptotic form of the normalized maximum likelihood (NML) model selection criterion [17]. NML is a modern form of the minimum description length (MDL) principle, which is an information theoretic approach to select the model that has the shortest code length for describing the information in the data [2], [12].

The NML model is defined as:

$$\mathrm{NML}(x^n\,;\,\mathcal{M}) = \frac{p(x^n\,;\,\hat{\theta}_{\mathcal{M}}(x^n))}{C_n^{\mathcal{M}}}, \tag{9}$$

where the normalizing factor $C_n^{\mathcal{M}}$ is the sum of the maximum likelihoods over all potential data sets:

$$C_n^{\mathcal{M}} = \sum_{x^n} p(x^n\,;\,\hat{\theta}_{\mathcal{M}}(x^n)). \tag{10}$$

NML provides a unique solution to minimize the *worst case regret* under log loss for all possible distributions, and the constant $\log C_n^{\mathcal{M}}$ is the minimax and maximin regret, see [17, 21].

As stated above, the logarithm of the NML probability shares the same asymptotic expansion as the marginal likelihood under Jeffreys prior, given by

FIA. The regularity conditions required for this to hold are discussed in [11]. Therefore, we can combine Eq. (8) with Eq. (9) and obtain an estimate of $\log \mathrm{FII}(\mathcal{M})$ by:

$$\log \mathrm{FII}(\mathcal{M}) = \log C_n^{\mathcal{M}} - \frac{d_{\mathcal{M}}}{2} \log \frac{n}{2\pi} + o(1), \tag{11}$$

However, the normalizing constant, $C_n^{\mathcal{M}}$ also lacks a closed form solution for most of model families and therefore, its value can be calculated efficiently only for a restricted set of model families such as the Bernoulli and multinomial models [6]. For other cases, one possible solution is to use factorized variants of NML [14], which approximate the formula by factorizing it as a product of locally minimax optimal models. The study in [19] proves that for Bayesian networks, the factorized NML (fNML) is asymptotically equivalent to BIC but leads to improved model selection accuracy for finite samples. In this work, we provide further evidence about the behavior of fNML.

However, instead of resorting to factorized NML variants, where no numerical guarantees about the approximation error are known, we estimate NML by Monte Carlo sampling in the same fashion as in [13]. The obtained estimates can be shown to be consistent as the number of simulated samples is increased. Hence they provide a sound approach for approximating NML and thereby also the FII constant: once we have obtained an estimate of the NML normalizing term, we deduct other terms as in Eq. (8) to approximate $\log \mathrm{FII}(\mathcal{M})$. After that, by plugging in the approximated value of $\log \mathrm{FII}(\mathcal{M})$ in Eq. (11), we can calculate FIA for any sample size without having to repeat the sampling procedure.

3 Monte Carlo approximation of NML

For Bayesian networks, there is no efficient way to compute the exact value of $\log C_n^{\mathcal{M}}$. We need to consider other approximate methods such as the Monte Carlo sampling method introduced in [13]. Based on the law of large numbers, the sample average is guaranteed to converge to the mean if the sampling size is large. By sampling m data sets $\{x_1^n, \ldots, x_m^n\}$ from distribution $q(\cdot)$, we have a consistent *importance sampling estimator* for $C_n^{\mathcal{M}}$ as:

$$\frac{1}{m} \sum_{t=1}^{m} \frac{p(x_t^n\,;\, \hat{\theta}_{\mathcal{M}}(x_t^n))}{q(x_t^n)} \xrightarrow{a.s.} C_n^{\mathcal{M}} \quad \text{as } m \to \infty. \tag{12}$$

Ideally, any proposal distribution q with full support will guarantee convergence.

However, the shape of q significantly affects the rate of convergence and the variance of the estimator. We need to choose a sampling distribution q that is similar to the target distribution. Following [13], we use the sampling distribution by drawing each set of the parameters independently from the Dirichlet distribution $\mathrm{Dir}(\frac{1}{2}, \frac{1}{2}, \ldots, \frac{1}{2})$, which results in the Krichevsky-Trofimov universal model (K-T model) [8]. It has been proved that the K-T model is asymptotically equivalent to NML as long as the parameters are not on the boundary.

4 Numerical results concerning the lower-order terms

In this section we present some properties of $\log \mathrm{FII}(\mathcal{M})$ that are important to the model selection behavior of the FIA formula.

4.1 Numerical values of $\log C_n^{\mathcal{M}}$ and $\log \mathrm{FII}(\mathcal{M})$

Firstly, for each combination of maximum indegree, number of nodes, and alphabet size, which together determine the number of parameters, we generate 100 Bayesian networks randomly. We estimate the $\log C_n^{\mathcal{M}}$ under different sample sizes to show how the $\log C_n^{\mathcal{M}}$ curve relates to the BIC curve and its upper bound. Note that while the main determinant of the model complexity, as measured by $\log C_n^{\mathcal{M}}$, is the number of parameters, these different Bayesian network models usually have somewhat different complexities. As we will see, however, the variance among networks with a fixed number of parameters is relatively small compared to the differences between networks with a different number of parameters.[1]

As an example, we show the results of Bayesian networks with $l = 20$ nodes, alphabet size $|\mathcal{X}| = 4$, and indegree (number of parents) of each node $k = 5, \ldots, 8$ subject to the acyclicity condition. All estimates of $\log C_n^{\mathcal{M}}$ under each sample size are calculated separately for 100 different Bayesian networks to obtain the mean and the standard deviation. (The variance is due to both the aforementioned differences between different model structures as well as the noise inherent to the Monte Carlo technique.)

Because $C_n^{\mathcal{M}}$ is defined as the sum of maximized likelihoods over all possible data sets, and because in the discrete case the likelihood is always at most one, a trivial upper bound for $\log C_n^{\mathcal{M}}$ is defined as

$$\log C_n^{\mathcal{M}} \leq nl \log |\mathcal{X}|. \tag{13}$$

Figure 1 shows that for small sample sizes, this upper bound tightly squeezes $\log C_n^{\mathcal{M}}$ towards zero. On the other hand, up to constant terms, $\log C_n^{\mathcal{M}}$ shares the same asymptotic form with the BIC (Eq. (6) and Eq. (11)). As the sample size increases, the slope of the $\log C_n^{\mathcal{M}}$ curve will tend to the slope of $\frac{d_{\mathcal{M}}}{2} \log n$. In terms of the graph, where the sample size is shown on a logarithmic scale, the $\log C_n^{\mathcal{M}}$ curve becomes a straight line that is parallel to the corresponding BIC curve. The difference between the curves tends to the constant $\log \mathrm{FII}(\mathcal{M}) - \frac{d_{\mathcal{M}}}{2} \log 2\pi$. The figure suggests that the constant grows rapidly as the model order is increased.

If the sample size is small, the sum of lower-order terms may be a very important part that should not be ignored. For example, Fig. 1 shows that for Bayesian networks with 20 nodes, alphabet size $|\mathcal{X}| = 4$ and maximum indegree $k = 6$, when the sample size is $n = 1000$, the sum of lower terms amount to a

[1] An interesting line of future research will be to zoom in into the differences in model complexity within classes of networks with a fixed number of parameters by the techniques we use here.

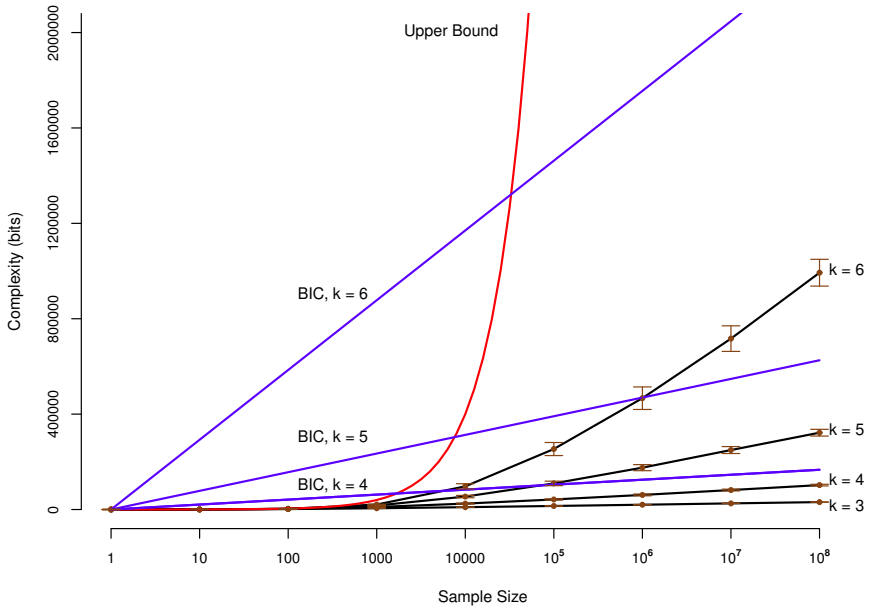

Fig. 1: Estimates of $\log C_n^{\mathcal{M}}$ by Monte Carlo sampling for Bayesian networks with $l = 20$ nodes and alphabet size $|\mathcal{X}| = 4$, labeled by the model complexity $k = \{3, \ldots, 6\}$, as a function of sample size $n = 1, 10, \ldots, 10^8$ (in log-scale). The black lines connect the mean values and the whiskers indicate standard deviation over 100 random repetitions. The red curve shows the upper bound $nl \log |\mathcal{X}|$. The straight blue lines are BIC complexity penalties over different k.

number less than $-800,000$. This is because $\log C_n^{\mathcal{M}}$ is restricted by its upper bound to almost zero but the the term $\frac{d_{\mathcal{M}}}{2} \log n$ is larger than $800,000$.

4.2 Accuracy of FIA for small sample sizes

Secondly, we look into the accuracy of FIA as an approximation of $\log C_n^{\mathcal{M}}$ when the sample size is small. Here we estimate $\log C_n^{\mathcal{M}}$ by the Monte Carlo sampling method for both small and large sample sizes. We show the estimated values for a set of nested Bayesian networks of 20 nodes. The models are nested in the sense that simpler (less edges) Bayesian networks are obtained by removing edges from a complex ($k = 8$), randomly generated Bayesian network. We simulate $m = 100$ data sets in each case and take the average to estimate the $\log C_n^{\mathcal{M}}$ value. On the other hand, we also estimate the constant term $\log \mathrm{FII}(\mathcal{M})$ (by Eq. (11)) for the same networks using a sample size of 10^9 to make sure that the term $o(1)$ becomes negligible, and plug in the resulting constant into the FIA formula for the smaller sample sizes. Table 1 lists related quantities for Bayesian networks

with 20 nodes and alphabet size $|\mathcal{X}| \in \{2, 4\}$, when sample sizes are 10^3 or 10^5 and maximum indegrees are from one to eight.

Based on Table 1, a significant observation is that when the model is very complex, for instance, when $|\mathcal{X}| = 4$ and $k \geq 6$, the $\log \mathrm{FII}(\mathcal{M})$ is a negative number with very large absolute value (less than -10^6). However, the absolute values of the term $\frac{d_{\mathcal{M}}}{2} \log \frac{n}{2\pi}$, as shown in the third row of Table 1 are much smaller than $\log \mathrm{FII}(\mathcal{M})$ for small sample sizes. Therefore, the term $\frac{d_{\mathcal{M}}}{2} \log \frac{n}{2\pi}$ is dominated by $\log \mathrm{FII}(\mathcal{M})$, which results in negative values of the sum. For example, as shown in the fourth row of Table 1, for sample size $n = 10^3$, this is the case for alphabet size $|\mathcal{X}| = 4$, with maximum indegree $k \geq 4$; and for alphabet size $|\mathcal{X}| = 2$, with maximum indegree $k = 8$. When the sample size increases to $n = 10^5$, for some simpler networks like $|\mathcal{X}| = 2$, and $k \leq 5$, the values of $\log C_n^{\mathcal{M}}$ and the sum are fairly close to each other. But for the most complex networks when $|\mathcal{X}| = 4$ and $k \geq 7$, sample sizes as large as 10^5 are still far from enough to even make the sum positive. The more complex the model, the larger sample size that we need to get sensible complexity penalties.

Due to the properties discussed above, the model selection by FIA fails under several conditions. For example, with $|\mathcal{X}| = 2$ and sample size $n = 10^3$, the FIA penalty for Bayesian networks with maximum indegree $k = 6$ is larger than for $k = 7$. Because the simpler network is a subset of the more complex one, the maximum likelihood value for the network with $k = 7$ is always higher or equal to that for the model with $k = 7$. Therefore, the FIA criterion will select the Bayesian network with $k = 7$ rather than the one with $k = 6$, *no matter what the data are*. For sample size $n = 10^5$ the problem does not occur when the alphabet size of $|\mathcal{X}| = 2$ but with $|\mathcal{X}| = 4$, the same problem occurs for $k \geq 7$ even with sample size $n = 10^5$. The rule of thumb that one should have more samples than there are free parameters in the model seems to hold quite well in these situations.

The above observations underline the importance of paying attention to the potential problems due to the $o(1)$ terms involved in the approximations for small and moderate sample sizes. Curiously enough, the BIC formula, which is based on omitting all $O(1)$ terms does not have a similar problem; we will return to this issue below.

5 Model Selection Simulations

In the above, we already made some remarks on the likely consequences of the the identified properties of FIA to model selection performance. In this section, we perform a set of simulation experiments to investigate them in detail. We focus in particular on complexity regularization in Bayesian networks. We consider networks with $l = 20$ and $l = 40$ discrete-valued nodes. The alphabet size of each node is varied to be $|\mathcal{X}| = 2$ or $|\mathcal{X}| = 4$.

In each simulation, we restrict the model comparison to a set of eight network topologies that are obtained by constructing a random DAG with each node's indegree $k = 8$ (subject to the acyclicity condition) and removing edges from

Table 1: The $\log C_n^{\mathcal{M}}$ estimates based on FIA (the fourth row) or Monte Carlo sampling (the fifth row), the Fisher information integral $\log \mathrm{FII}$ and the higher order term $\frac{d}{2}\log\frac{n}{2\pi}$ for Bayesian networks of $k = \{1,\ldots,8\}$, alphabet size $|\mathcal{X}| = \{2,4\}$ with number of nodes $l = 20$ and sample size $n \in \{10^3, 10^5\}$. Values that are based on Monte Carlo approximation are reported with four significant digits

$|\mathcal{X}| = \mathbf{2}, \mathbf{n} = \mathbf{10^3}$

k	1	2	3	4	5	6	7	8
$\log \mathrm{FII}$	-22.88	-37.57	-96.27	-349.9	-1004	-2565	-6488	-14330
$d_{\mathcal{M}}$	39	75	143	271	511	959	1791	3327
$\frac{d_{\mathcal{M}}}{2}\log\frac{n}{2\pi}$	142.6	274.3	523.0	991.1	1869	3507	6550	12167
sum	119.8	236.7	426.7	641.2	864.1	941.7	61.45**	-2163*
$\log C_n$	179.5	298.9	481.2	711.0	1092	1565	2056	2698

$|\mathcal{X}| = \mathbf{2},\ \mathbf{n} = \mathbf{10^5}$

k	1	2	3	4	5	6	7	8
$\log \mathrm{FII}$	-22.88	-37.57	-96.27	-349.9	-1004	-2565	-6488	-14330
$d_{\mathcal{M}}$	39	75	143	271	511	959	1791	3327
$\frac{d_{\mathcal{M}}}{2}\log\frac{n}{2\pi}$	272.2	523.4	998.0	1891	3566	6693	12500	23219
sum	249.3	485.9	901.7	1541	2562	4128	6011	8889
$\log C_n$	308.0	542.4	941.8	1545	2608	4204	6390	10270

$|\mathcal{X}| = \mathbf{4},\ \mathbf{n} = \mathbf{10^3}$

k	1	2	3	4	5	6	7	8
$\log \mathrm{FII}$	-86.96	-1123	-8211	-48710	-239000	-1135000	-5105000	-21230000
$d_{\mathcal{M}}$	231	879	3327	12543	47103	176127	655359	2424831
$\frac{d_{\mathcal{M}}}{2}\log\frac{n}{2\pi}$	844.8	3215	12167	45872	172263	644122	2396742	8867956
sum	757.8	2092	3956	-2840*	-66720*	-490700*	-2709000*	-12360000*
$\log C_n$	832.4	2289	5522	10300	16880	21070	23050	24500

$|\mathcal{X}| = \mathbf{4},\ \mathbf{n} = \mathbf{10^5}$

k	1	2	3	4	5	6	7	8
$\log \mathrm{FII}$	-86.96	-1123	-8211	-48710	-239000	-1135000	-5105000	-21230000
$d_{\mathcal{M}}$	231	879	3327	12543	47103	176127	655359	2424831
$\frac{d_{\mathcal{M}}}{2}\log\frac{n}{2\pi}$	1612	6135	23219	87539	328735	1229203	4573798	16923071
sum	1525	5012	15010	38830	89750	94330	-531500*	-4308000*
$\log C_n$	1582	5059	15310	4137	112500	261100	494000	858900

*) $\log C_n^{\mathcal{M}}$ approximations by FIA with negative values
**) $\log C_n^{\mathcal{M}}$ approximations by FIA with a changing order

it to obtain DAGs with maximum indegrees $k = 7, \ldots, 1$. Such a comparison is admittedly atypical since most practical scenarios involve several possible network topologies with the same maximum indegree, whereas we only consider one topology for each value of k. We adopt the present methodology for the purpose of highlighting the complexity regularization aspect and in order to be able to estimate the FII term accurately for each individual Baysian network model.[2]

Within each group of Bayesian networks, we compare FIA with other model selection criteria of varying levels of approximation, including BIC [16], and fNML [19]. To obtain a measure of the ideal performance, we also include the Bayes factor based on the "true" prior. In practice, the true prior is obviously not known in advance, and therefore, the Bayes factor criterion should be taken simply as a yardstick against which to compare the other methods. The effect of using different priors in Bayes factors has been studied in [18].

We perform the comparison for sample sizes $10, 100, \ldots, 10^6$. For each sample size we draw 100 random data sets from the true network, and apply the different criteria to select one of the eight possible network structures. We show the results as percentages of correctly identified models in Figs 2 and 3. For the Bayesian networks with alphabet size $|\mathcal{X}| = 2$ (for both $l = 20$ and $l = 40$), sample size 10^4 is enough for FIA to achieve nearly 100% accuracy. But for the cases when $|\mathcal{X}| = 4$, FIA needs $n \geq 10^6$ to achieve good performance. Most of the failures are caused by selecting the most complex models with maximum indegree $k = 8$: see the bottom panels of each figure to verify that when the true model is $k = 8$, FIA achieves 100% accuracy just because it always favors the most complex model available unless the sample size is large enough to avoid the reversed complexity penalty phenomenon discussed in the previous section.

On the contrary, the BIC criterion works better than FIA except when the true model is the most complex one. Its accuracy decreases when the maximum indegree of the true model increases. For networks with $|\mathcal{X}| = 4$ and $k = 8$, the BIC criterion fails even when the sample size reaches 10^6. Based on Table 1, we can see that BIC puts unnecessary large penalties to complex models. Therefore, it tends to select simple models. On the other hand, we note that the fNML criterion performs almost as well as the Bayes factor criterion with the true prior.

6 Conclusions

The simulation experiment verifies that whenever the sample size is not sufficient, the FIA model selection criterion is unreliable for Bayesian network model selection. We emphasize that none of the above suggests that NML or Bayes factors have similar issues for small sample sizes. Indeed, the experiments also show that another kind of (non-asymptotic) approximation of NML, the fNML criterion, behaves almost as well as Bayes factor with the true prior. A remarkable

[2] Unlike in the numerical studies in the previous section, here we want to take into account the fine-grained differences between FII values between different Bayesian network models with a fixed number of parameters.

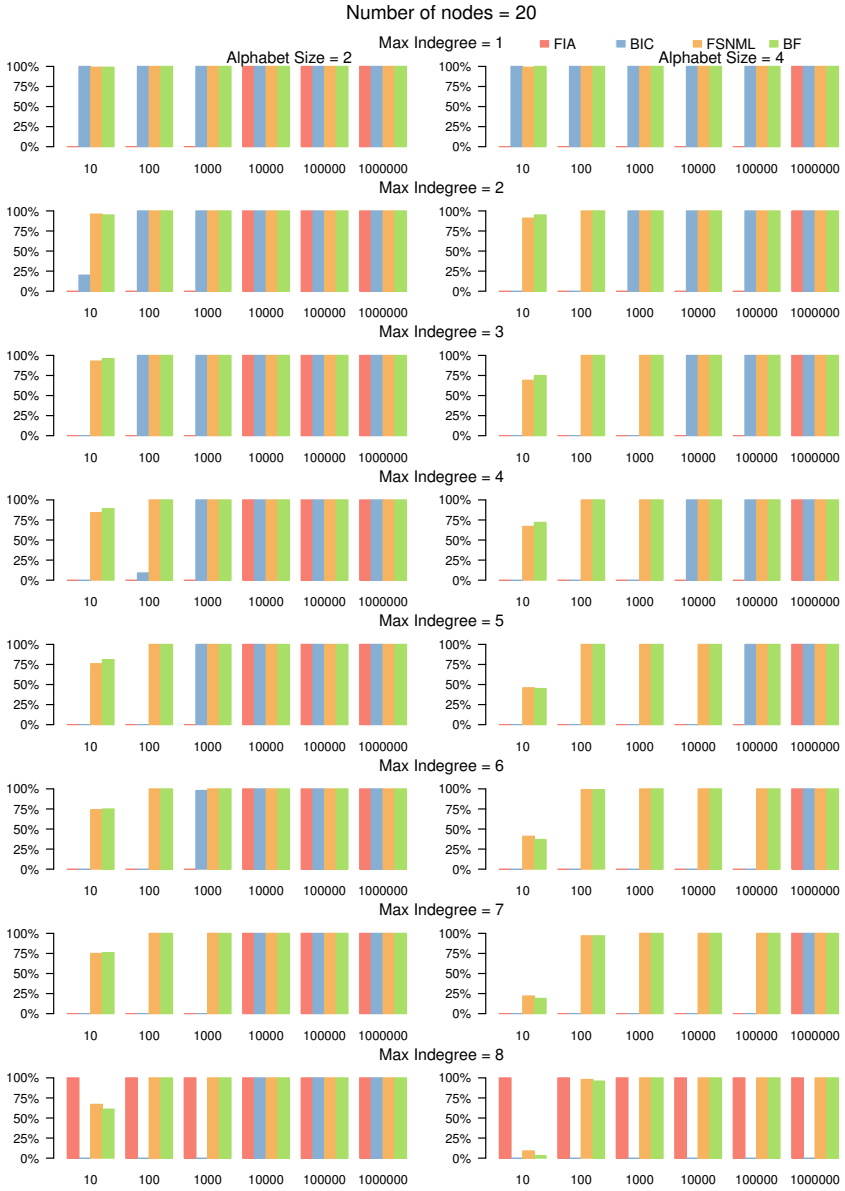

Fig. 2: Model selection experiments for selecting Bayesian networks with 20 nodes and maximum indegree $k = \{1, \ldots, 8\}$. Bars show percentages of correctly identified models by four different criteria as a function of sample size $n = \{10, 10^2, \ldots, 10^6\}$. For the left plots, we have alphabet size $|\mathcal{X}| = 2$, and for the right ones we have $|\mathcal{X}| = 4$. Four criteria, from left to right at each sample size, are: FIA (Fisher information approximation) by Eq. (8), BIC by Eq. (6), fsNML (factorized sequential NML) [19], and BF (Bayes factor with "true" prior).

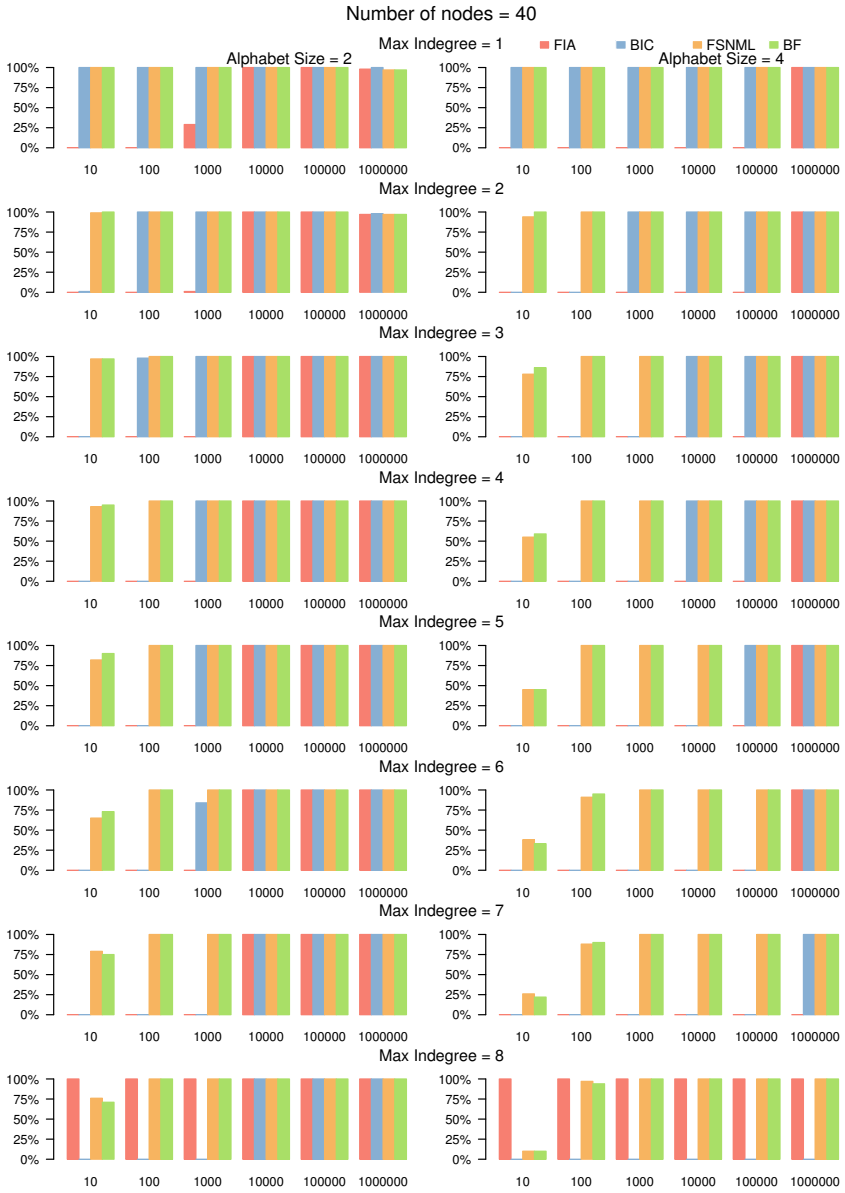

Fig. 3: Model selection experiments with the same settings for Bayesian networks with 40 nodes. (cont'd from Fig. 2)

fact is that a very rough approximation (of the Bayes factor as well as the NML), namely the classic BIC criterion where all $O(1)$ terms are ignored, was in our experiments actually never worse and often much better than the FIA criterion where the asymptotic formula is truncated only at the $o(1)$ term.

Comparing FIA penalties with $\log C_n^{\mathcal{M}}$ makes it clear that the $o(1)$ term in Eq. (8) is also an essential part when the sample size is small, which leads to huge differences between the FIA penalty and $\log C_n^{\mathcal{M}}$. Similar results are also reported in the early work in [9] for an exponential model and in [15] for Markov sources. Based on the simulation experiment, we suggest that including the constant term alone may actually be dangerous, and in case useful asymptotic formulas are sought after, one should consider more refined approximations that also include $o(1)$ terms. As a rule of thumb, situations where a FIA type approximation can be considered "safe" seem to be those where the sample size exceeds the number of parameters in any of the models being compared.

It is important to note that the goal of this study was not to evaluate the model selection performance of a criterion where the constant FII term is obtained by Monte Carlo techniques. Such a criterion may not be very practical since for complex networks, the sample size at which the $o(1)$ term becomes negligible can be enormous, and drawing a sufficient number of random data sets from each of the candidate models would be time consuming. Instead, we wanted to illustrate the performance of the FIA criterion, independently of the method by which the FII term is obtained. In other words, we wanted to find out whether evaluating the FII term via an approximate analytic formula, for example, would lead to a useful model selection criterion. The answer turns out to be negative unless the model complexity is severly restricted or the sample size is extremely large. Hence, studying analytic approximations without paying close attention to the $o(1)$ terms is likely to be of limited interest.

In the future, it will be interesting to extend the scope of this study to other model classes such as generalized linear models with continuous parameters to see if the problem of FIA for small sample sizes also applies to them. To address the small sample issues related to FIA, we may also try to analytically break down the $o(1)$ term to obtain more reliable approximations. A closer study for the performance of FIA and related model selection criteria in general can then be done in these two directions.

Acknowledgments

The authors thank the anonymous reviewers for insightful comments and suggestions. This work was funded in part by the Academy of Finland (Centre-of-Excellence COIN).

References

1. Clarke, B. S., Barron, A. R.: Jeffreys prior is asymptotically least favorable under entropy risk. J. Stat. Plan. Inference 41(1), 37–61 (1994)

2. Grünwald, P. D.: The Minimum Description Length Principle. MIT Press (2007)
3. Han, C., Carlin, B. P.: Markov chain Monte Carlo methods for computing Bayes factors. J. Am. Statist. Assoc. 96(455), 1122–1132 (2001)
4. Jeffreys, H.: An invariant form for the prior probability in estimation problems. J. Roy. Statist. Soc. A. 186(1007), 453–461 (1946)
5. Kass, R.E., Raftery, A.E. : Bayes factors. J. Am. Statist. Assoc. 90(430), 773–795 (1995)
6. Kontkanen, P., Myllymäki P.: A linear-time algorithm for computing the multinomial stochastic complexity. Inform. Process. Lett. 103(6), 227–233 (2007)
7. Kontkanen, P., Myllymäki P., Silander, T., Tirri, H., Grünwald, P.: On predictive distributions and Bayesian networks. Stat. Comput. 10, 39–54 (2000).
8. Krichevsky, R., Trofimov, V: The performance of universal coding. IEEE Trans. Inf. Theory 27(2), 199–207 (1981)
9. Navarro, D.: A note on the applied use of MDL approximations. Neural Comput. 16(9), 1763–1768 (2004)
10. Rasmussen, C. E., Ghahramani, Z.: Occam's razor. In: Leen, T., Dietterich, T., Tresp, V. (eds.) Advances in Neural Information Processing Systems, pp. 294–300 (2001)
11. Rissanen, J.: Fisher information and stochastic complexity. IEEE Trans. Inf. Theory 42(1), 40–47 (1996)
12. Rissanen, J.: Information and Complexity in Statistical Modeling. Springer (2007)
13. Roos, T.: Monte Carlo estimation of minimax regret with an application to MDL model selection. In: Proc. IEEE Information Theory Workshop, pp. 284–288. IEEE Press, (2008)
14. Roos, T., Rissanen, J.: On sequentially normalized maximum likelihood models. In: Rissanen, J., Liski, E., Tabus, I., Myllymäki, P., Kontoyiannis, I., Heikkonen, J. (eds.), Proc. Workshop on Information Theoretic Methods in Science and Engineering (WITMSE-08), Tampere, Finland (2008)
15. Roos, T., Zou, Y.: Keep it simple stupid — On the effect of lower-order terms in BIC-like criteria. In: Information Theory and Applications Workshop (ITA), pp. 1–7. IEEE Press, 2013
16. Schwarz, G.: Estimating the dimension of a model. Ann. Statist. 6, 461–464 (1978)
17. Shtarkov, Y. M.: Universal sequential coding of single messages. Probl. Inform. Transm. 23(3), 3–17 (1987)
18. Silander, T., Roos, T., Kontkanen, P., Myllymäki, P.: Factorized normalized maximum likelihood criterion for learning Bayesian network structures. In: Jaeger, M. and Nielsen, T. D. (eds.) Proc. 4th European Workshop on Probabilistic Graphical Models (PGM-08), pp. 257–272 (2008)
19. Silander, T., Roos, T., Myllymäki, P.: Learning locally minimax optimal Bayesian networks. Int. J. Approx. Reason 51(5), 544–557 (2010)
20. Ueno, M.: Robust learning Bayesian networks for prior belief. In: Cozman, F.G. and Pfeffer, A. (eds.), Proc. Uncertainty in Artificial Intelligence (UAI-2011), pp. 698–707, Barcelona, Spain (2011)
21. Xie, Q., Barron, A. R.: Asymptotic minimax regret for data compression, gambling, and prediction. IEEE Trans. Inform. Theory 46(2), 431–445 (2000)