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@cs.helsinki.fi,teemu.roos@hiit.fi

Received 19 March 2016

Abstract We study BIC-like model selection criteria and in particular, their refinements that include a constant term involving the Fisher information matrix. We perform numerical simulations that enable increasingly accurate approximation of this constant in the case of Bayesian networks. 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. For networks with a fixed number of parameters, d, the leading term in the complexity penalty, which is proportional to d, is the same. However, as we show, the constant term can vary significantly depending on the network structure even if the number of parameters is fixed. Based on our experiments, we conjecture that the distribution of the nodes' outdegree is a key factor. Furthermore, we demonstrate that the constant term can have a dramatic effect on model selection performance for small sample sizes.

Keywords Model Selection, Bayesian Networks, Fisher Information Approximation, NML, BIC.

§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, $I(\theta)$.[*1] 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) by Kass *et al.*[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 appears to grow 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 by Navarro[9] and Markov sources by Roos *et al.*[15].

In this paper, we first review the FIA approximation and discuss its relation to certain other model selection criteria. Even though there is no closed form formula for the Fisher information integral under most model families, including Bayesian networks, it can be estimated up to arbitrarily fine precision using a Monte Carlo technique.[13] Our main contributions include, first, an investigation on the effects of the network structure on the Fisher information integral. Second, we carry out model selection experiments where we highlight the complexity regularization behavior of various criteria. This leads to conclusions as to which of the criteria are safe and which should be avoided under

[*1] We denote the binary (base-2) logarithm by log and the natural logarithm by ln.

given circumstances.

§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 us 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 penalty for model complexity.[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 Approximate 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

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

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.[21] 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, refer to Ueno's work for more details.[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 *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), and dropping the $o(1)$ terms, we get the Fisher information approximation:

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

For Bayesian networks, which is the model class studied in this work, the Jeffreys prior has been derived by Kontkanen *et al.*[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 Approximations of the 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, refer to the works by Shtarkov and Xie.[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 by Rissanen[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, which approximate the formula by factorizing it as a product of locally minimax optimal models.[14] The study by Silander *et al.*[19] proves that for Bayesian networks, the factorized NML (fNML) is asymptotically equivalent to BIC, but their empirical experiments suggest that it 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 Roos[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 for a given (large) sample size, 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.

2.3 A Lower Bound on the Approximation Error

While the purpose of this paper is to explore the behavior of the Fisher information approximation numerically, the link between the FIA and NML immediately leads to a simple upper bound on the normalizing term in NML, which further leads to a theoretical observation about the approximation error

of FIA in the finite alphabet case.

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, an upper bound for $\log C_n^{\mathcal{M}}$ is obtained as

$$\log C_n^{\mathcal{M}} \leq nl \log |\mathcal{X}|, \tag{12}$$

where $|\mathcal{X}| \geq 2$ is the alphabet size, which we assume to be the same for all variables for the sake of simplicity.

This rather trivial upper bound was already pointed out by Roos[13], and it is illustrated in Fig. 1 below. Together with the fact that the leading term of $\log C_n$ agrees with that of the BIC penalty, the upper bound implies that the behavior of $\log C_n$ has two characteristics: first, for small sample sizes, it is sandwiched between zero and the linear upper bound, and second, it will eventually grow at a logarithmic rate like BIC. For complex models, the constant factor in the logarithmic term, $\frac{d_{\mathcal{M}}}{2} \log \frac{n}{2\pi}$, is so large that no approximation of the same analytic form as FIA can be accurate for both small and large sample sizes.

We can quantify the mismatch between FIA and NML in terms of the following proposition.

Proposition 2.1

Let each model be over $l \geq 1$ variables with alphabet size $|\mathcal{X}| \geq 2$.

a) The maximum discrepancy between FIA and NML has the following lower bound:

$$\max_n |\mathrm{FIA}(x^n ; \mathcal{M}) - \log \mathrm{NML}(x^n ; \mathcal{M})| \geq \eta, \tag{13}$$

where

$$\eta = \frac{d}{4} \log \left\lfloor \frac{d}{2l \ln 2 \log |\mathcal{X}|} \right\rfloor - \frac{d}{4 \ln 2}$$

b) If the number of free parameters is greater than $d > \kappa\, l \ln |\mathcal{X}|$, where $\kappa = 2(e+1) \approx 7.444$, the difference is non-zero for some sample size, i.e., $\eta > 0$.

c) For any two models, $\mathcal{M}_i$ and $\mathcal{M}_j$, on the same set of variables with $d_i > d_j > \kappa\, l \ln |\mathcal{X}|$, the respective lower bounds satisfy $\eta_i > \eta_j$.

Note that the discrepancy is a constant for all x^n with a given sample size n since the first term that depends on the actual sample is the same in both FIA and NML.

Proof We start by proving part (a). Since the data-dependent first term cancels, we only need to consider the difference

$$\mathrm{FIA}(x^n\,;\,\mathcal{M}) - \log \mathrm{NML}(x^n\,;\,\mathcal{M}) = \frac{d}{2}\log\frac{n}{2\pi} + \log \mathrm{FII}(\mathcal{M}) - \log C_n.$$

We discuss two cases depending on the value of $\log \mathrm{FII}(\mathcal{M})$.

Case I: Assume that $\log \mathrm{FII}(\mathcal{M}) \geq -\eta + \frac{d}{2}\log 2\pi$. By (12), the difference is greater than or equal to

$$\frac{d}{2}\log\frac{n}{2\pi} + \log \mathrm{FII}(\mathcal{M}) - nl\log|\mathcal{X}|. \tag{14}$$

The difference (14) can be maximized by setting its derivative with respect to n to zero, which may give a non-integer solution. Let n' be the greatest integer than is less than or equal to the root of the derivative:

$$n' = \left\lfloor \frac{d}{2l\ln 2\log|\mathcal{X}|} \right\rfloor \leq \frac{d}{2l\ln 2\log|\mathcal{X}|}.$$

Plugging n' into the formula for the difference, and applying the assumption in Case I, we then obtain

$$\begin{aligned}
&\frac{d}{2}\log\frac{n'}{2\pi} + \log \mathrm{FII}(\mathcal{M}) - n'l\log|\mathcal{X}| \\
&\geq \frac{d}{2}\log\left\lfloor \frac{d}{2l\ln 2\log|\mathcal{X}|} \right\rfloor - \frac{d}{2}\log 2\pi - \eta + \frac{d}{2}\log 2\pi - \frac{d}{2\ln 2} \\
&= \frac{d}{2}\log\left\lfloor \frac{d}{2l\ln 2\log|\mathcal{X}|} \right\rfloor - \frac{d}{4}\log\left\lfloor \frac{d}{2l\ln 2\log|\mathcal{X}|} \right\rfloor + \frac{d}{4\ln 2} - \frac{d}{2\ln 2} = \eta,
\end{aligned}$$

which concludes Case I.

Case II: Assume now that $\log \mathrm{FII}(\mathcal{M}) < -\eta + \frac{d}{2}\log 2\pi$. By definition, we have $\log C_n^{\mathcal{M}} \geq 0$ for all models $\mathcal{M}$ and all $n \geq 1$. Letting the sample size be $n = 1$, we thus have

$$-\frac{d}{2}\log\frac{1}{2\pi} - \log \mathrm{FII}(\mathcal{M}) + \log C_1^{\mathcal{M}} \geq \frac{d}{2}\log 2\pi - \log \mathrm{FII}(\mathcal{M}).$$

By the assumption in Case II, we now get

$$\max_n \left| \frac{d}{2} \log \frac{n}{2\pi} + \log \mathrm{FII}(\mathcal{M}) - \log C_n^{\mathcal{M}} \right| \geq \frac{d}{2} \log 2\pi - \log \mathrm{FII}(\mathcal{M}) > \eta.$$

Combining the two cases, inequality (13) holds for all values of $\log \mathrm{FII}(\mathcal{M})$, so part (a) is complete.

Part (b) follows from the assumption that $d > \kappa\, l \ln |\mathcal{X}|$ by direct manipulation of the expression for η. Likewise, part (c) follows by noting that as long as the first term in η, which is increasing in d, dominates the second term, which is decreasing in d, the sum increases. ■

§3 Numerical Values of 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. We use Monte Carlo sampling to approximate the NML normalizer $\log C_n^{\mathcal{M}}$ for Bayesian networks.

3.1 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 by Roos[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:

$$\frac{1}{m} \sum_{t=1}^{m} \frac{p(x_t^n\,;\, \hat{\theta}_{\mathcal{M}}(x_t^n))}{q(x_t^n)} \xrightarrow{a.s.} C_n^{\mathcal{M}} \quad \text{as } m \to \infty. \tag{15}$$

In principle, 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 Roos[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.

3.2 Numerical Values of $\log \mathrm{FII}(\mathcal{M})$

We first study the numerical values of the Fisher information integral, followed by a numerical evaluation of the accuracy of the Fisher information approximation.

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.

As an example, we show the results of Bayesian networks with $l = 20$ nodes, alphabet size $|\mathcal{X}| = 4$, and indegree of each node $k = 3, 4, 5, 6$ 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 variance. (The variance is due to both the aforementioned differences between different model structures as well as the noise inherent to the Monte Carlo technique.)

Figure 1 shows that for small sample sizes, the upper bound in Eq. (12) 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 the 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 the lower-order terms amounts to a number less than $-800,000$. This is because $\log C_n^{\mathcal{M}}$ is restricted

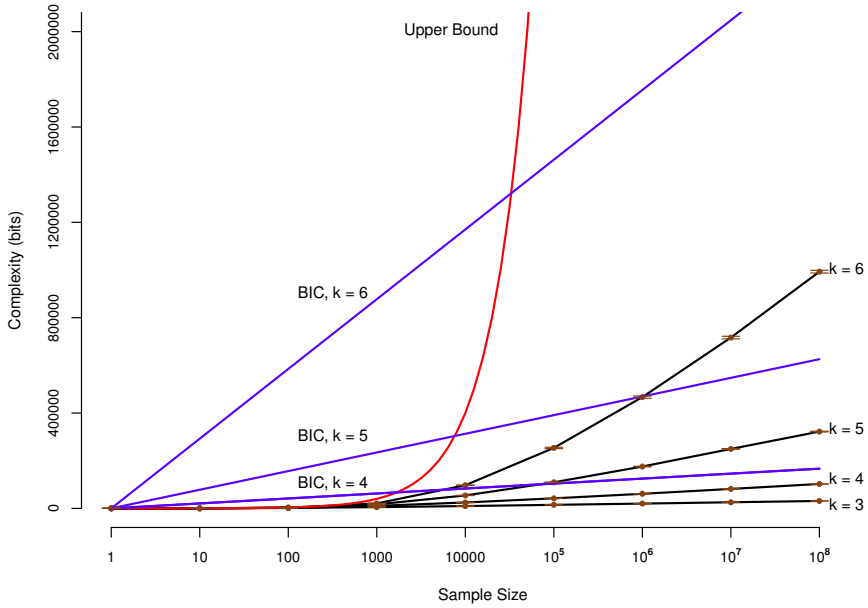


Fig. 1 Estimates of $\log C_n^{\mathcal{M}}$ by Monte Carlo sampling for Bayesian networks with $l = 20$ nodes and alphabet size $|\mathcal{X}| = 4$, labeled by the model complexity (indegree) $k = \{3, \ldots, 6\}$, as a function of sample size $n = 1, 10, \ldots, 10^8$ (in log-scale). The black lines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

by its upper bound to be relatively close to zero but the term $\frac{d_{\mathcal{M}}}{2} \log n$ is larger than $800,000$.

3.3 Accuracy of FIA with Small Samples

We next 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

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 with indegree $k = \{1,\ldots,8\}$, 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	41370	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

based on omitting all $O(1)$ terms does not have a similar problem; we will return to this point in the model selection comparisons in Sec. 6.

§4 Beyond the Number of Parameters

In this section, we focus on how the network structures influence the $\log \mathrm{FII}(\mathcal{M})$ term and result in different FIA penalties. In particular, we illustrate how the numerical values of $\log \mathrm{FII}(\mathcal{M})$ differ when the number of parameters in the model is fixed. We design three sets of networks with different characteristics. In Set 1, we compare networks where groups of five nodes are linked either as a sequence or a tree structure with increasing branching factor. This leads to an observation that a star-like structure where a group of five nodes are linked to all other nodes has smaller complexity than other structures where the outdegree distribution is more uniform.

In Set 2, we have a number of binary trees with a fixed outdegree (two). The trees differ in terms of how balanced they are. Here we observe that all binary trees have very similar values of the constant term, $\log \mathrm{FII}(\mathcal{M})$. Last, we compare three distinct categories of networks in Set 3. These include two types of grid-like networks, a star-like network, a second order Markov chain, and a hybrid between a star and a chain. The outcome of this experiment agrees with the observation in the first set, namely that a large maximum outdegree, such as in the star-like structure, appears to lead to small values of the constant term. However, there is no clear difference in the values of the constant term between a second order Markov chain and the grid structures.

The three sets of networks are depicted in Fig. 2. They are described in more detail in the following three subsections, together with observations made by evaluating the constants. The $\log \mathrm{FII}(\mathcal{M})$ estimates were obtained by the above Monte Carlo technique by generating $m = 100$ random data sets for each network. We use the sample size 10^9, which we found to be well sufficient to guarantee convergence of the constant term $\log \mathrm{FII}(\mathcal{M})$ in all cases below.

4.1 Set 1: From a Chain to a Five-Star Network

In Set 1, each network has 45 nodes corresponding to random variables with alphabet size $|\mathcal{X}| = 4$. For each network, there is a root group consisting of five nodes that have no parents, and eight child groups, each of which consists of five nodes that each have five parent nodes. Thus, all eight networks have the same number of parameters, $d_{\mathcal{M}} = 5 \times 3 + 8 \times 5 \times 4^5 \times 3 = 122895$. Network

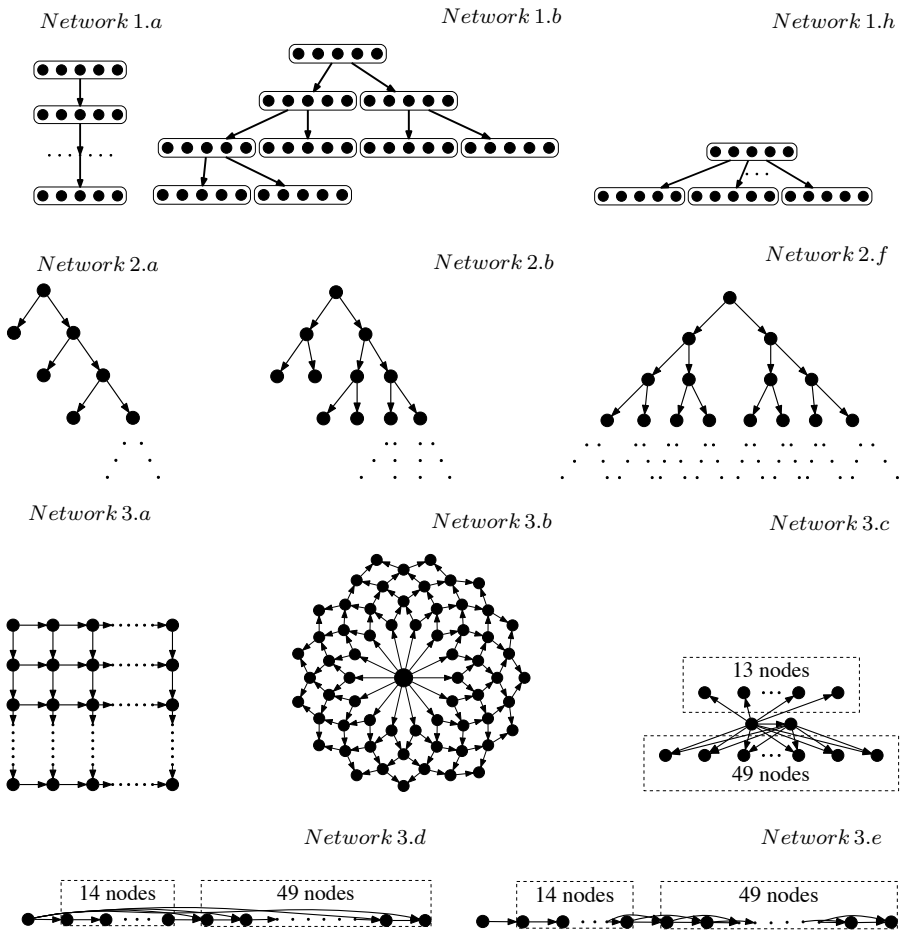
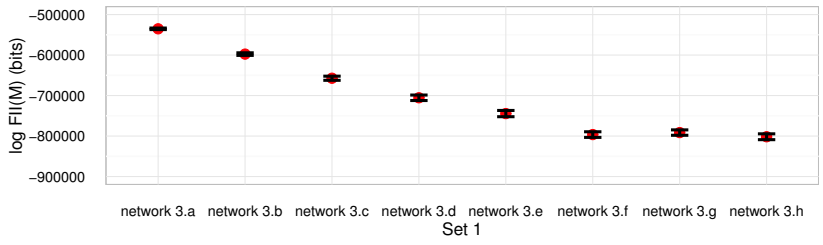
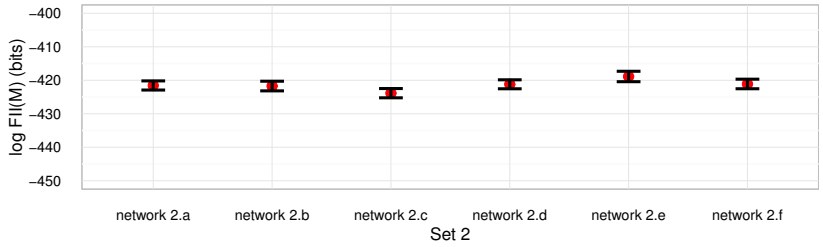


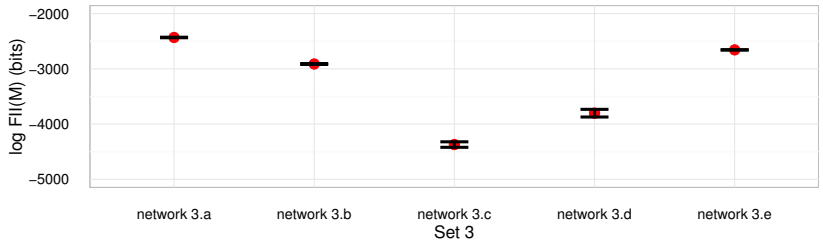
Fig. 2 Set 1: All five nodes in each parent group enclosed by a frame are parents of all five nodes in the child group. All networks have $9 \times 5 = 45$ nodes. The networks range from Network 1.a (a "five-chain") where each group is a parent of one other group, to Network 1.h (a "five-star") where the root group is a parent of all the other groups. **Set 2:** Binary trees from the "caterpillar-tree" (Network 2.a) to the balanced binary tree (Network 2.f). **Set 3:** Network 3.a is a square grid of size 8×8 and Network 3.b is a so called "polar grid" with the same number of nodes and the same number of parameters. Network 3.c is a "twin-star" with two node in the middle. Network 3.d is a "star-chain" hybrid between a first order Markov chain and a star. Network 3.e is a second order Markov chain. Note that in Networks 3.c–3.e, there is a group of 14 nodes with only one parent each, which keeps the number of parameters the same as for the other networks in Set 3.



(a)



(b)



(c)

Fig. 3 Estimates of $\log \mathrm{FII}(\mathcal{M})$ for the three sets of Bayesian network structures in Sec. 4. The red dots show the mean values and the error bars indicate the standard error of the mean over $m = 100$ random data sets.

1.*a* has a chain-like structure where the groups are parents of each other in a sequential ordering, and thus, each node except nodes in the last child group have five children, i.e., their outdegree equals five. In Network 1.*b*, each group has two child groups except those at the bottom level of the tree-like arrangement, and in Network 1.*c*, each group has three child groups, etc. At the other extreme, the groups in Network 1.*h* are organized in a star-like arrangement, which we call a "five-star" network.

Table 3a and Fig. 3a show $\log \mathrm{FII}(\mathcal{M})$ and $\log C_n^{\mathcal{M}}$ for the eight Bayesian networks. In this set of networks, we investigate how the maximum outdegree

Table 2 The normalizing term $\log C_n^{\mathcal{M}}$ and the Fisher information integral $\log \mathrm{FII}$ estimates based on Monte Carlo sampling for three sets of Bayesian networks in Sec. 4 with sample size 10^9. All values are reported with four significant digits

Network id	1.a	1.b	1.c	1.d	1.e	1.f	1.g	1.h
$\log \mathrm{FII}$	-534800	-597600	-657200	-709600	-755600	-837200	-861700	-901600
$\log C_n$	1139000	1077000	1017000	964500	918600	837000	812000	773000

(a) Set 1

Network id	2.a	2.b	2.c	2.d	2.e	2.f
$\log \mathrm{FII}$	-421.3	-418.1	-423.3	-422.5	-419.4	-419.4
$\log C_n$	20220	20220	20220	20220	20220	20220

(b) Set 2

Network id	3.a	3.b	3.c	3.d	3.e
$\log \mathrm{FII}$	-2431	-2911	-5858	-4486	-2656
$\log C_n$	31940	31460	28510	29880	31710

(c) Set 3

affects the constants. It can be seen quite clearly that $\log C_n^{\mathcal{M}}$ and $\log \mathrm{FII}(\mathcal{M})$ decrease as the maximum outdegree increases. Network $1.h$ ("five-star"), is the least complex one among Set 1.

4.2 Set 2: Binary Trees

We compare six binary trees (Networks $2.a, \ldots, 2.f$) with 127 nodes and alphabet size $|\mathcal{X}| = 4$ in Set 2. For all the binary trees in this set, there are two children for each parent and each child has exactly one parent. Therefore, all the trees have the same number of parameters. We restrict the maximum number of nodes in a layer to $\{2, 4, 8, \ldots, 64\}$ in Network $2.a$ to $2.f$, respectively. In other words, we have a set of binary trees from the least balanced (the "caterpillar-tree", $2.a$) with depth 64 to the balanced binary tree with depth seven ($2.f$).

We estimate $\log \mathrm{FII}(\mathcal{M})$ in the same way as for Set 1 in Sec. 4.1 with 100 randomly generated data sets for each binary tree and list the corresponding values in Table 3b. For all different types of trees in this data set, the values of $\log \mathrm{FII}(\mathcal{M})$ as well as $\log C_n^{\mathcal{M}}$ are almost the same. This is in line with the conjecture that the complexity term is affected by the outdegree distribution rather than, for instance, the diameter of the network (measured by the longest

path between any two nodes).

4.3 Set 3: Grids, Chains and Stars

In this set of Bayesian networks, we compare three categories of networks: two different grids, a star-like network, a second order Markov chain, and a hybrid between a chain and a star. We show the structures of networks in Fig. 2, networks $\{3.a, \ldots, 3.e\}$. The total number of nodes in each network is 64 and as for the above networks, the alphabet size is $|\mathcal{X}| = 4$. The numbers of parameters for all networks are the same as well. Networks $3.a$ and $3.b$ are a square grid and a so called polar grid, respectively. Network $3.c$ ("twin-star") has two nodes in the middle that are parents to all the other nodes except a group of 13 nodes, whose only parent is one of the middle nodes. This is to ensure that the number of parameters is the same as in the grid structures. Networks $3.d$ and $3.e$ contain a sequence where each node, except the first, is a child of the previous node. In Network $3.d$, the last 49 nodes have also the root node as their second parent, while in Network $3.e$, the last 49 nodes have a second order Markov chain structure, where node i has nodes $i-1$ and $i-2$ as parents. The group of 14 nodes shown separately in the diagrams have only one parent in order to guarantee the same number of parameters for all networks in Set 3. Network $3.d$ is a hybrid between a chain and a star (a "star-chain") since it contains a first order Markov chain as well as a star component as subgraphs.

We show the estimated values of $\log \mathrm{FII}(\mathcal{M})$ in Table 3c and Fig. 3c. It is quite apparent that the least complex models are the twin-star ($3.c$) and the star-chain ($3.d$). These two network include one or more nodes with large outdegree whereas Networks $3.a$, $3.b$ and $3.e$ have a more uniform outdegree distribution. There is a slight difference between the complexity of polar grid network ($3.b$) and Networks $3.a$ and $3.e$ — the polar grid is less complex than the other two — which can be explained by noting that the central node in the polar grid has outdegree 14, which is more than the maximum degree in the square grid (2) or the Markov chain (2) but less than that of the twin-star (maximum outdegree 62) or the star-chain (63).

§5 Model Selection Simulations

In the above, we already made some remarks on the likely consequences of 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 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 by Schwarz[16], and fNML by Silander *et al.*[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 by Silander *et al.*[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 4 and 5. 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.

[*2] Based on the observations in Sec. 4, which make it clear that Bayesian networks with a fixed number of parameters can have large differences in FII values, we evaluate the constants for individual networks instead of using the same complexity penalty for all networks with a fixed number of parameters.

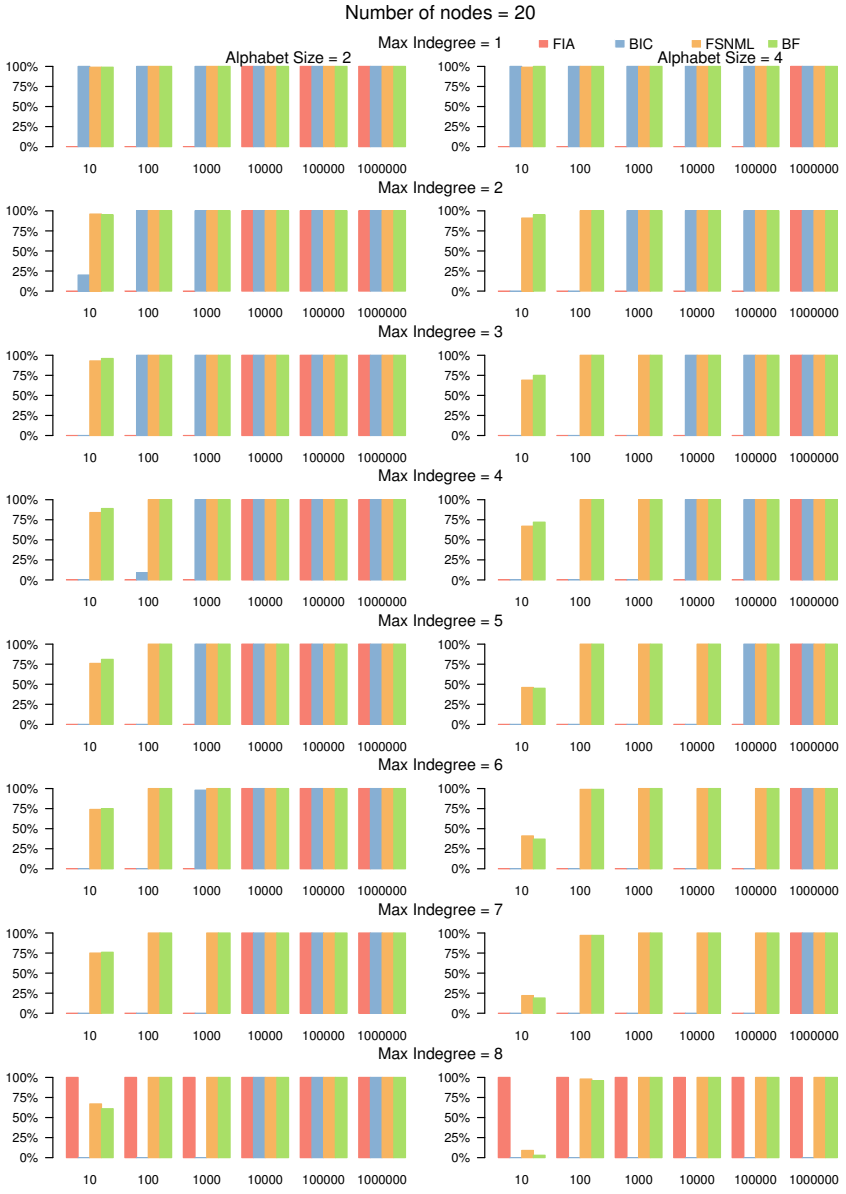


Fig. 4 Model selection experiments for selecting Bayesian networks with 20 nodes and maximum indegree $k = \{1, \ldots, 8\}$. Bars show percentages of correctly identified models by four different criteria as a function of sample size $n = \{10, 10^2, \ldots, 10^6\}$. For the left plots, we have alphabet 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) by Silander *et al.*[19], and BF (Bayes factor with "true" prior).

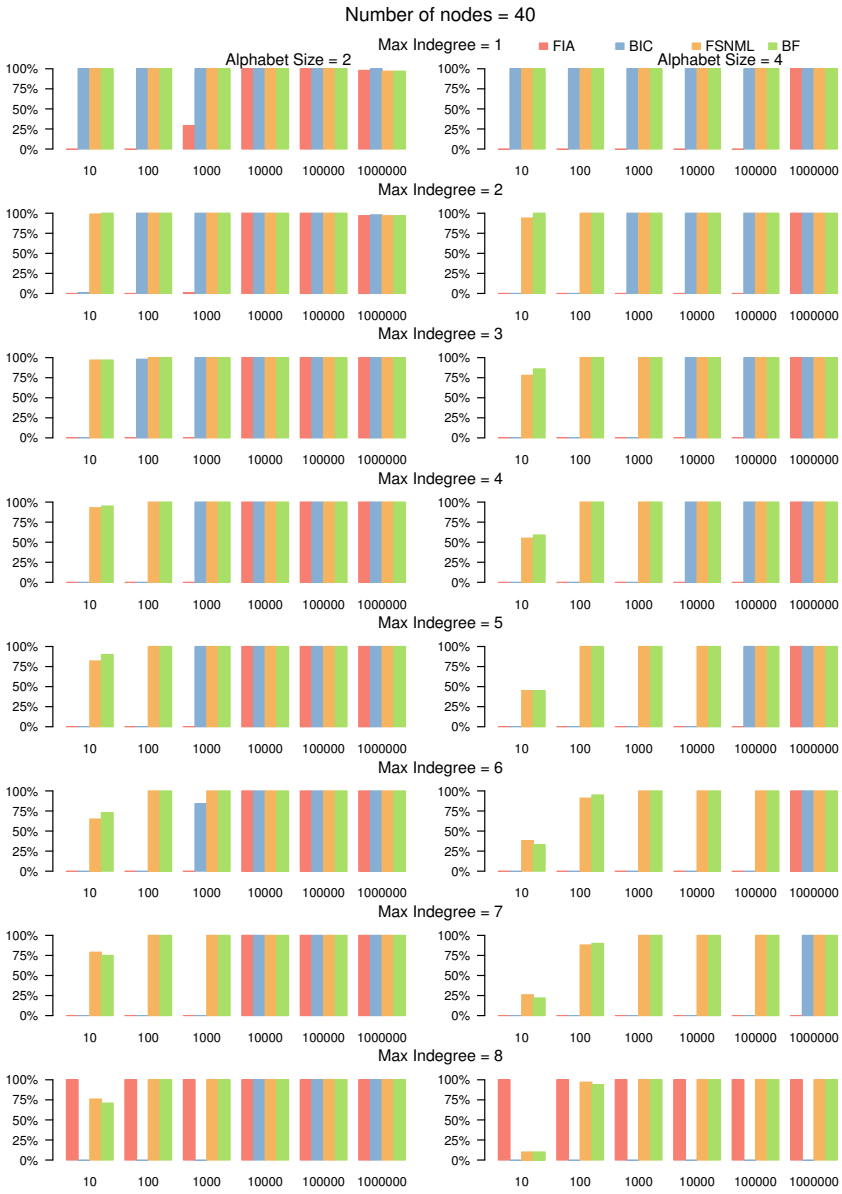


Fig. 5 Model selection experiments with the same settings for Bayesian networks with 40 nodes. (cont'd from Fig. 4)

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

In this study, we used a Monte Carlo approach to evaluate the constant term $\log \mathrm{FII}(\mathcal{M})$ in the complexity of Bayesian network models. The main contributions are *i*) the constant term can be very large compared to the asymptotically leading term that is proportional to the number of parameters; *ii*) the constant term can be used to investigate the differences in the complexity of individual networks beyond the number of parameters.

Concerning the first of the above contributions, it turned out that the constant $\log \mathrm{FII}(\mathcal{M})$ tends to make the Fisher information approximation of marginal likelihood or the NML criterion break down when the sample size is less than the number of parameters. In terms of the second contribution, our experiments suggest a conjecture that the outdegree distribution of the nodes in a Bayesian network structure is a key factor in determining the value of $\log \mathrm{FII}(\mathcal{M})$. The difference in $\log \mathrm{FII}(\mathcal{M})$ of networks with the same number of parameters can be non-trivial and have interesting consequences on the appropriate complexity penalization.

Our model selection experiment further indicates that while the FIA model selection criterion may be unreliable when applied to complex Bayesian network models, the NML criterion and Bayes factors are nevertheless reliable and applicable even 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 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 by Navarro [9] for an exponential model and by Roos *et al.* [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.

Acknowledgment An earlier version of this paper was presented at the Second Workshop on Advanced Methodologies for Bayesian Networks (AMBN 2015) in Yokohama. The authors thank the anonymous reviewers for insightful comments and suggestions and the organizers of AMBN-2015 for their invitation to submit this work to this special issue. This work was funded in

part by the Academy of Finland (Centre-of-Excellence COIN).

References

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

2) Grünwald, P. D., *The Minimum Description Length Principle*, MIT Press, 2007.

3) Han, C. and Carlin, B. P., "Markov chain Monte Carlo methods for computing Bayes factors", *J. Am. Statist. Assoc. 96, 455*, pp. 1122–1132, 2001.

4) Jeffreys, H., "An invariant form for the prior probability in estimation problems", *J. Roy. Statist. Soc. A. 186, 1007*, pp. 453–461, 1946.

5) Kass, R.E. and Raftery, A.E., "Bayes factors", *J. Am. Statist. Assoc., 90, 430*, pp. 773–795, 1995.

6) Kontkanen, P. and Myllymäki P., "A linear-time algorithm for computing the multinomial stochastic complexity", *Inform. Process. Lett., 103, 6*, pp. 227–233, 2007.

7) Kontkanen, P., Myllymäki P., Silander, T., Tirri, H. and Grünwald, P., "On predictive distributions and Bayesian networks", *Stat. Comput. 10*, pp. 39–54, 2000.

8) Krichevsky, R. and Trofimov, V, "The performance of universal coding", *IEEE Trans. Inf. Theory, 27, 2*, pp. 199–207, 1981.

9) Navarro, D., "A note on the applied use of MDL approximations", *Neural Comput. 16, 9*, pp. 1763–1768, 2004.

10) Rasmussen, C. E. and Ghahramani, Z., "Occam's razor", in *Advances in Neural Information Processing Systems* (Leen, T., Dietterich T. and Tresp, V.), pp. 294–300, 2001.

11) Rissanen, J., "Fisher information and stochastic complexity", *IEEE Trans. Inf. Theory 42, 1*, pp. 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. and Rissanen, J., "On sequentially normalized maximum likelihood models", in *Proc. Workshop on Information Theoretic Methods in Science and Engineering (WITMSE-08)* (Rissanen, J., Liski, E., Tabus, I., Myllymäki, P., Kontoyiannis, I. and Heikkonen, J.), Tampere, Finland, 2008.

15) Roos, T. and 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*, pp. 461–464, 1978.

17) Shtarkov, Y. M., "Universal sequential coding of single messages", *Probl. Inform. Transm. 23, 3*, pp. 3–17, 1987.

18) Silander, T., Roos, T., Kontkanen, P. and Myllymäki, P., "Factorized normalized maximum likelihood criterion for learning Bayesian network structures", in *Proc. 4th European Workshop on Probabilistic Graphical Models (PGM-08)* (Jaeger, M. and Nielsen, T. D.), pp. 257–272, 2008.

19) Silander, T., Roos, T. and Myllymäki, P., "Learning locally minimax optimal Bayesian networks", *Int. J. Approx. Reason 51, 5*, pp. 544–557, 2010.

20) Ueno, M., "Robust learning Bayesian networks for prior belief", in *Proc. Uncertainty in Artificial Intelligence (UAI-2011)* (Cozman, F.G. and Pfeffer, A.), pp. 698–707, Barcelona, Spain, 2011.

21) Xie, Q. and Barron, A. R., "Asymptotic minimax regret for data compression, gambling, and prediction", *IEEE Trans. Inform. Theory 46, 2*, pp. 431–445, 2000.