Keep it Simple Stupid – On the Effect of Lower-Order Terms in BIC-Like Criteria

Teemu Roos and Yuan Zou
Helsinki Institute for Information Technology HIIT
Department of Computer Science
University of Helsinki, Finland
Email: firstname.lastname@hiit.fi

***Abstract*—We study BIC-like model selection criteria. In particular, we approximate the lower-order terms, which typically include the constant $\log \int \sqrt{\det I(\theta)}\, d\theta$, where $I(\theta)$ is the Fisher information at parameter value θ. We observe that the constant can sometimes be a huge negative number that dominates the other terms in the criterion for moderate sample sizes. At least in the case of Markov sources, including the lower-order terms in the criteria dramatically degrades model selection accuracy. A take-home lesson is to keep it simple.**

I. Introduction

Many generally applicable model selection criteria such as the Bayesian information criterion (BIC) [14] and Akaike's information criterion (AIC) [1] are large-sample asymptotic formulas. Consider, for instance, the BIC criterion

$$\mathrm{BIC}(x^n\,;\,\mathcal{M}) = \log \frac{1}{p(x^n\,;\,\hat{\theta}_{\mathcal{M}}(x^n))} + \frac{d_{\mathcal{M}}}{2} \log n, \qquad (1)$$

where $\mathcal{M}$ is a model, x^n is a data sample of size n, $\hat{\theta}_{\mathcal{M}}(x^n)$ denotes the maximum likelihood (ML) parameters, and $d_{\mathcal{M}}$ is the number of free parameters in the model. The BIC is an asymptotic approximation of the Bayesian marginal likelihood where the terms independent of n are omitted: this is done because, first, the lower-order terms depend on the prior distribution, second, their evaluation is often mathematically or computationally hard, and third, because the terms depending on n will eventually begin to dominate.

A particularly important case is obtained by using the Jeffreys prior

$$p(\theta) = \frac{\sqrt{\det I(\theta)}}{\mathrm{FII}(\mathcal{M})},$$

where $I(\theta)$ denotes the Fisher information matrix; the normalizing factor, which we call the *Fisher information integral*, is given by

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

where $\Theta_{\mathcal{M}}$ denotes the parameter space corresponding to model $\mathcal{M}$. The Jeffreys prior was originally justified by invariance arguments [6] but it has also been shown to have several minimax properties [2], [10].

Under regularity conditions on the model class, a more exact formula for the Bayesian marginal likelihood with the Jeffreys prior (of which BIC is an asymptotic version) is given by the following expression (which we decorate with frames as it will be mentioned several times in what follows):

$$\boxed{\log \frac{1}{p(x^n\,;\,\hat{\theta}(x^n))} + \frac{d}{2} \log \frac{n}{2\pi} + \log \mathrm{FII}(\mathcal{M}) + o(1). \quad (2)}$$

There are also other model selection criteria that have the same asymptotic form, for instance, Shtarkov's [15] normalized maximum likelihood (NML) criterion which is used in the recent formulations of the minimum description length (MDL) principle, see [5], [11].

It appears to be generally held that omitting the lower-order terms (terms independent of n) is bad but often unavoidable, and that whenever they are available, they should be included. Consequently, a great deal of work has been done to obtain as precise approximate formulas as possible; see, e.g., [9] and references therein. We mention that the factor 2π in the above criterion has been briefly discussed (without a definite conclusion) in [3]; see also references therein.

In this work, we study the behaviour $\mathrm{FII}(\mathcal{M})$ by means of a simple Monte Carlo approximation of the NML criterion [13]. This leads to some interesting observations: first, the value of $\log \mathrm{FII}(\mathcal{M})$, which appears in Eq. (2), can sometimes be a huge negative number that dominates all the other terms for small to moderate sample sizes, and second, quite surprisingly, the inclusion of the lower order terms in the criterion dramatically *degrades* model selection performance in the case of Markov sources where the task is to identify the correct model order.

II. A Monte Carlo Approximation

We start by introducing the normalized maximum likelihood (NML) universal model. It involves a normalizing factor that under suitable regularity conditions is asymptotically given by the terms following the first one in Eq. (2). Therefore, we can use the NML model as a means to approximate $\log \mathrm{FII}(\mathcal{M})$.

Let $\mathcal{M} = \{p(\cdot\,;\,\theta_{\mathcal{M}}) : \theta_{\mathcal{M}} \in \Theta_{\mathcal{M}}\}$ be a model class, i.e., a set of probability distributions indexed by parameter(s) $\theta_{\mathcal{M}}$. Given a model class, the NML universal model is defined in terms of the maximized likelihood, $p(x^n\,;\,\hat{\theta}_{\mathcal{M}}(x^n)) = \max_{\theta_{\mathcal{M}}} p(x^n\,;\,\theta_{\mathcal{M}})$:

$$\mathrm{NML}(x^n\,;\,\mathcal{M}) = \frac{p(x^n\,;\,\hat{\theta}_{\mathcal{M}}(x^n))}{C_n^{\mathcal{M}}}, \qquad (3)$$

where the normalizing factor $C_n^{\mathcal{M}}$ is given by

$$C_n^{\mathcal{M}} = \sum_{x^n} p(x^n\,;\,\hat{\theta}_{\mathcal{M}}(x^n)). \tag{4}$$

The NML model and the normalizing factor $C_n^{\mathcal{M}}$ have several interesting properties. The NML model is the unique minimax optimal distribution in the sense that it minimizes the *worst case regret* when the loss is measured using logarithmic[1] loss (code length),

$$\max_{x^n} \mathcal{R}(q(\cdot), x^n) = \max_{x^n} \log \frac{p(x^n\,;\,\hat{\theta}_{\mathcal{M}}(x^n))}{q(x^n)},$$

over all choices of the distribution $q(\cdot)$. The worst case regret of NML is in fact a constant for all x^n and it is given by $\log C_n^{\mathcal{M}}$. The latter is also the minimax and maximin regret [17], which makes the quantity interesting in its own right.

Under regularity conditions on the model class, the negative logarithm (ideal code length) of the NML distribution is given by Eq. (2) [10]. Therefore, recalling the definition of NML, Eq. (3), we have

$$\log \mathrm{FII}(\mathcal{M}) = \log C_n^{\mathcal{M}} - \frac{d}{2} \log \frac{n}{2\pi} + o(1). \tag{5}$$

Unfortunately, computing the actual value of the normalizing factor $C_n^{\mathcal{M}}$ tends to be infeasible. An important exception is the linear-time algorithm for the multinomial (i.i.d.) case [7]. However, from [13], we have a straightforward Monte Carlo approximation that consists of sampling m data sets $x_1^n, \ldots, x_m^n$ from distribution $q(\cdot)$, which acts as a proxy for NML, and using the *importance sampling estimator*

$$\frac{1}{m} \sum_{t=1}^{m} \frac{p(x_t^n\,;\,\hat{\theta}_{\mathcal{M}}(x_t^n))}{q(x_t^n)} \xrightarrow{a.s.} C_n^{\mathcal{M}} \quad \text{as } m \to \infty. \tag{6}$$

Almost sure consistency as $m \to \infty$ holds for all finite alphabets.

Intuitively, the idea in the above importance sampling estimator is to consider $|\mathcal{X}|^{-n} C_n^{\mathcal{M}}$, where $|\mathcal{X}|$ denotes the cardinality of the source alphabet, as an average (rather than a sum) over all data sets, of which an approximation can be obtained by sampling random data sets. The case $q(x_t^n) = |\mathcal{X}|^{-n}$ gives the simple (uniform) sample average. Drawing the random data sets from a non-uniform q does not affect the mean of the estimator (6) but significantly reduces its variance if q is appropriately chosen; for more details, see [13].

Finally, we remark that the definition of $C_n^{\mathcal{M}}$ as a sum over all possible data sets immediately yields the following uniform bound

$$\log C_n^{\mathcal{M}} \leq n \log |\mathcal{X}|, \tag{7}$$

that follows from the obvious fact that the sum of maximized likelihoods over all data sets cannot be greater than the number of data sets. While this upper bound may appear trivial, it can be used to deduce some properties of the lower-order terms in BIC-like criteria, as we will see in the next section.

III. Markov sources

Markov sources are an important model class that has been extensively studied from both Bayesian and information theoretic perspectives. The main difficulty in obtaining precise results concerning the asymptotic behavior of the marginal likelihood and related quantities is that the regularity conditions employed in many of the known results do not hold at the boundaries of the parameter space. If the parameters are restricted to compact subsets in the interior of the parameter space (i.e., away from the boundaries) the minimax optimal coding rate is given by Eq. (2) which is also the asymptotic code length by the mixture with Jeffreys prior as well as NML, see [16] and references therein.

Jacquet and Szpankowski [9] obtain an expression for the NML code length that applies uniformly over the full parameter space. Their formula is similar to Eq. (2) with the exception of the constant term which they give in terms of an integral that, however, can be evaluated in closed form only in the simplest case of first order Markov sources on a binary alphabet.[2] Takeuchi, Kawabata and Barron [16] show that the NML code length is uniformly bounded from above by Eq. (2). In contrast, the Bayesian marginal likelihood with Jeffreys prior is *not* asymptotically minimax optimal as it fails to achieve the optimal code length, Eq. (2), for sequences for which the maximum likelihood parameters are on the boundary of the parameter space.

For the sake of simplicity, we ignore the issues with the parameter space boundary in the following, and assume that the necessary regularity conditions hold for both the mixture with Jeffreys prior and NML. More careful treatment of the boundaries will be a topic for future research.

IV. The Fisher Information Integral FII($\mathcal{M}$)

We now use the Monte Carlo approximation of the NML normalizer, $C_n^{\mathcal{M}}$, for Markov sources to illustrate some features of the Fisher information integral FII($\mathcal{M}$) via Eq. (5). In all what follows, we use alphabet sizes $|\mathcal{X}| = 2$ and $|\mathcal{X}| = 4$, which is sufficient to prove our point.

First of all, we reproduce Fig. 1 from [13]. It clearly shows how the simple upper bound, Eq. (7), tightly squeezes $\log C_n^{\mathcal{M}}$ down towards zero for the shown models (Markov order $k = 1, \ldots, 5$) and small sample sizes. Since $\log C_n^{\mathcal{M}}$ and the corresponding BIC complexity penalty $\frac{d_{\mathcal{M}}}{2} \log n$ asymptotically differ by a constant, they have the same asymptotic slope when the sample size is plotted on a logarithmic scale (as in Fig. 1). As one can see in the figure for $k = 5$, the constant appears to be quite large even for relatively simple models.

[1] Throughout, we denote base 2 logarithms by log which corresponds to measuring code length in bits.

[2] The constant corresponding to $\log \mathrm{FII}(\mathcal{M})$ in [9] (see their Thm. 3) is $\log A_2^1 = \log 16G \approx 3.873$, where G is the Catalan constant.We note that a similar result is obtained by Giurcaneanu et al. [4] who, however, give a slightly different constant, $\log 8G \approx 2.873$. It would be interesting to investigate whether this is due to a different definition of the constant—Jacquet and Szpankowski mainly study actual codes where codeword lengths are integers—or something else. Our Monte Carlo approximation converges to 2.873.

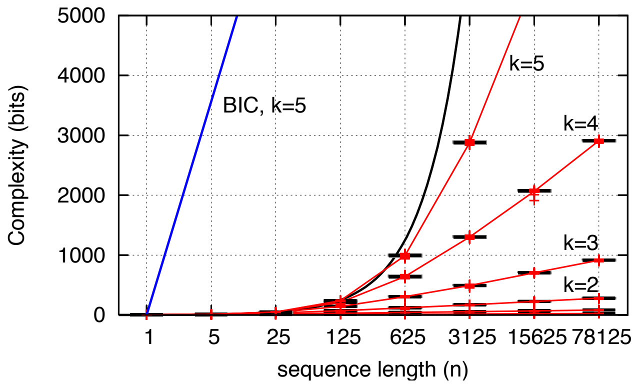

Fig. 1. Estimates of $\log C_n^{\mathcal{M}}$ for alphabet size $|\mathcal{X}| = 4$ and Markov orders $k = 1, \ldots, 5$ as a function of the sample size n. The red lines connect estimates of $\log C_n^{\mathcal{M}}$ for each value of k. The solid black curve shows the upper bound $n \log |\mathcal{X}|$, and the blue line shows the BIC complexity penalty for order $k = 5$. Note the log-scale for n. Figure source: [13]

To investigate the behavior of $\log \mathrm{FII}(\mathcal{M})$, we plug in the estimated $\log C_n^{\mathcal{M}}$ values in Eq. (5) for models of order $1, \ldots, 10$; see the second columns in Tables I and II (the estimated values are given twice in both tables as they are independent of the sample size). In the tables, values that are based on Monte Carlo approximation are reported with four significant digits to emphasize the fact that they come without a precise guarantee about their accuracy. We estimated the $\log \mathrm{FII}(\mathcal{M})$ values at sample size $n = 10^7$ (except for the three largest models where we used $n = 10^9$) for which we found that Eq. (5) seemed to have converged well enough, and used $m \geq 1000$ Monte Carlo samples in each case; see [13] for discussion about the convergence rate of the Monte Carlo estimator.[3]

Remark 1 (Very large absolute values): The most significant observation is that in the case where $|\mathcal{X}| = 4$, the $\log \mathrm{FII}(\mathcal{M})$ values quickly become very large in absolute value as the model order is increased. The tables also give the second term in Eq. (2), $\frac{d}{2} \log \frac{n}{2\pi}$. Comparing the second and the third columns, we notice that $\log \mathrm{FII}(\mathcal{M})$ dominates

[3] For the 0th order (i.i.d.) models, the estimates match those obtained using an exact formula [7].

TABLE I

ESTIMATED VALUES OF $\log \mathrm{FII}(\mathcal{M})$ AND OTHER RELATED QUANTITIES FOR MARKOV SOURCES OF ORDER $k = 0, \ldots, 10$ ON SOURCE ALPHABET OF SIZE $|\mathcal{X}| = 2$ WITH SAMPLE SIZES $n \in \{10^4, 10^5\}$.

$|\mathcal{X}| = \mathbf{2}, \mathbf{n} = \mathbf{10^4}$

k	$\log \mathrm{FII}$ *	$\frac{d}{2} \log \frac{n}{2\pi}$	sum *	$\log C_n$ *
0	1.647	5.318	6.965	6.977
1	2.873	10.64	13.51	13.53
2	3.514	21.27	24.78	24.82
3	−2.008	42.54	40.54	44.75
4	−5.748	85.09	79.34	79.87
5	−29.96	170.2	140.2	141.6
6	−95.80	340.4	244.6	248.6
7	−260.4	680.7	420.3	431.2
8	−659.8	1361	701.6	736.2
9	−1602	2723	1121	1230
10	−3735	5446	1711	1979

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

k	$\log \mathrm{FII}$ *	$\frac{d}{2} \log \frac{n}{2\pi}$	sum *	$\log C_n$ *
0	1.647	6.979	8.626	8.633
1	2.873	13.96	16.83	16.84
2	3.514	27.92	31.43	31.43
3	−2.008	55.83	53.83	57.95
4	−5.748	111.7	105.9	106.2
5	−29.96	223.3	193.4	193.2
6	−95.80	446.7	350.9	352.4
7	−260.4	893.3	632.9	638.0
8	−659.8	1787	1127	1139
9	−1602	3573	1971	2013
10	−3735	7147	3412	3517

*) Monte Carlo estimates (4 significant digits)

TABLE II

ESTIMATED VALUES OF $\log \mathrm{FII}(\mathcal{M})$ AND OTHER RELATED QUANTITIES FOR MARKOV SOURCES OF ORDER $k = 0, \ldots, 10$ ON SOURCE ALPHABET OF SIZE $|\mathcal{X}| = 4$ WITH SAMPLE SIZES $n \in \{10^4, 10^5\}$.

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

k	$\log \mathrm{FII}$ *	$\frac{d}{2} \log \frac{n}{2\pi}$	sum *	$\log C_n$ *
0	3.293	15.95	19.25	19.27
1	1.876	63.82	65.69	65.95
2	−44.81	255.3	210.5	212
3	−380.7	1021	640.4	656
4	−2319	4084	1765	1888
5	−12 420	16 337	3917	4870
6	−62 050	65 349	3299	10 370
7	−296 900	261 396	−35 500	16 100
8	−1 392 000	1 045 583	−346 400	18 900
9	−6 350 000	4 182 330	−2 168 000	19 800
10	−28 490 000	16 729 322	−11 760 000	19 950

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

k	$\log \mathrm{FII}$ *	$\frac{d}{2} \log \frac{n}{2\pi}$	sum *	$\log C_n$ *
0	3.293	20.94	24.23	24.25
1	1.876	83.75	85.62	85.7
2	−44.81	335	290.2	291
3	−380.7	1340	959.3	961
4	−2319	5360	3041	3070
5	−12 420	21 440	9020	9320
6	−62 050	85 759	23 710	25 990
7	−296 900	343 035	46 140	64 070
8	−1 392 000	1 372 141	−19 860	123 900
9	−6 350 000	5 488 566	−861 400	172 200
10	−28 494 000	21 954 263	−6 539 000	192 400

*) Monte Carlo estimates (4 significant digits)

the other term for higher order models, so that their sum (the fourth column in the tables) suddenly dips as the model order exceeds a certain point.

Remark 2 (Complex models always win): To get some perspective on the magnitude of the dip, notice that for the used alphabet size $|\mathcal{X}| = 4$, the first part of the marginal likelihood approximation, Eq. (2), is bounded between 0 and $2n$ since each observation can always be encoded using $\log |\mathcal{X}| = 2$ bits in the worst case. In other words, the higher order models will *always* be chosen by the criterion (2) for sample sizes at least up to 10^5. No matter what the observed data!

Remark 3 (Exact criterion still works): To see that the above failure (highest order model is always chosen) is not inherent to the used criterion (NML or the Bayesian marginal likelihood), we give the values of the normalizing constant $\log C_n$ in the last column of the tables. It is interesting to see how the sum of the penalty terms in Eq. (2) (fourth column) is rather a good approximation of $\log C_n$ for alphabet size $|\mathcal{X}| = 2$. However, in the $|\mathcal{X}| = 4$ case, the approximation completely breaks down at model orders $k > 5$.

From a practical point of view, it appears that the approximation should only be used for sample sizes that are unrealistically large in most application domains or for models so simple that they can be considered trivial. As an example, for alphabet size $|\mathcal{X}| = 4$ and sample size $n = 10^4$, the approximation assigns a *larger* penalty (3917 bits) to the model of order $k = 5$ than to the model of order $k = 6$ (3299 bits) even though the former is clearly a subset of the latter![4] The same "inverted complexity penalization" phenomenon occurs for all models with $k > 5$ at sample size $n = 10^4$ and for all models with $k > 7$ at sample size $n = 10^5$.

V. A Model Selection Experiment

To illustrate the practical significance of the above observations, we conduct a simulation experiment where we compare the model selection performance of criteria of varying level of approximation. We include BIC as well as the criterion based on the Fisher information approximation in Eq. (2) to gauge the effect the lower-order terms. We also include a recent sequential variant of NML, called the sequential NML (sNML) [12], which is much easier to compute than the exact NML criterion. To implement the Fisher information approximation, we need to resort to Monte Carlo approximation as explained above. We also tried AIC, which gave results that were in most cases slightly better than BIC but clearly worse than SNML (results not shown for the sake of clarity).

Lastly, we also include a theoretical Bayes factor criterion based on the Bayesian marginal likelihood where the used prior matches the one used for generating the simulated data sequences. Such a criterion is clearly not available in practice because the prior generating the data—the "true" prior, if such a thing exists at all—is not known. Including it in the experiment provides a way to get an idea of the hardness of the model selection problem in each case, and a yardstick to which the practical criteria can be compared.

Data was generated by picking a model order between $0, \ldots, 10$ and drawing all model parameters independently from Dirichlet distribution $\mathrm{Dir}(\frac{1}{2}, \frac{1}{2})$ in the binary alphabet case and $\mathrm{Dir}(\frac{1}{2}, \frac{1}{2}, \frac{1}{2}, \frac{1}{2})$ in the case $|\mathcal{X}| = 4$.

Figures 2–3 show the results. The main observation is that the Fisher information approximation fails except in the binary alphabet case with sample sizes $n \geq 10^4$. In such cases all criteria achieve nearly perfect accuracy (with the exception of BIC for orders $k \geq 9$). Note in particular that the results for order $k = 10$ where the Fisher information approximation achieves 100 % accuracy are due to the phenomenon mentioned in Remark 2: the criterion simply lead to the choice of the most complex available model which in this case just happened to be the correct one.

The well-known under-fitting tendency of BIC is also evident in that BIC works better than the other criteria only for model orders $k = 1$ and $k = 2$. For alphabet size $|\mathcal{X}| = 4$, BIC completely fails except for low model orders and large samples.

VI. Discussion: Possible Issues and Future Work

Based on the above observations, we argue that including the lower-order terms in BIC-like model selection criteria is harmful. This may or may not come as a surprise: On one hand, more refined approximations of criteria tend to behave more similarly to the exact versions than coarse approximations, which would suggest that including the lower-order terms is beneficial. On the other hand, some of the models we use (Markov sources of order 1–10) have an exorbitant number of parameters (e.g., with alphabet size $|\mathcal{X}| = 4$, order $k = 8$, the number of parameters is $d = 196608$) which suggests that any approximation even at sample size $n = 10^5 < d$ is likely to be inaccurate. Still, the fact that using BIC, which is a very coarse approximation, works much better than the more refined Fisher information approximation is somewhat surprising. In any case, we emphasize the necessity to be careful when applying *any* approximation (such as BIC) when in problems with large d and small n.

There are a couple of potential weaknesses in this work. First, the Fisher information approximation of the Bayesian mixture with Jeffreys prior, Eq. (2), does not hold for Markov sources when the parameters are not bounded away from the boundaries of the parameter space. Hence, the asymptotics may be somewhat different for strings for which the maximum likelihood parameters converge onto the boundaries. While such cases may be exceptional in many model classes, in Markov sources they in fact become more and more typical as the model order increases.

Regarding other model classes, we are unaware of how easily the break-down of the Fisher information approximation occurs, and whether it pertains to situations where the data is continuous. Hence, the generalizability of our results is presently unknown. A similar earlier result [8] suggests that that the problem may in fact be more a rule than an exception.

[4] For a similar case in the context of psychological models where a simpler model is assigned higher complexity penalty than a complex one for sample sizes up to $n = 2095$, see [8].

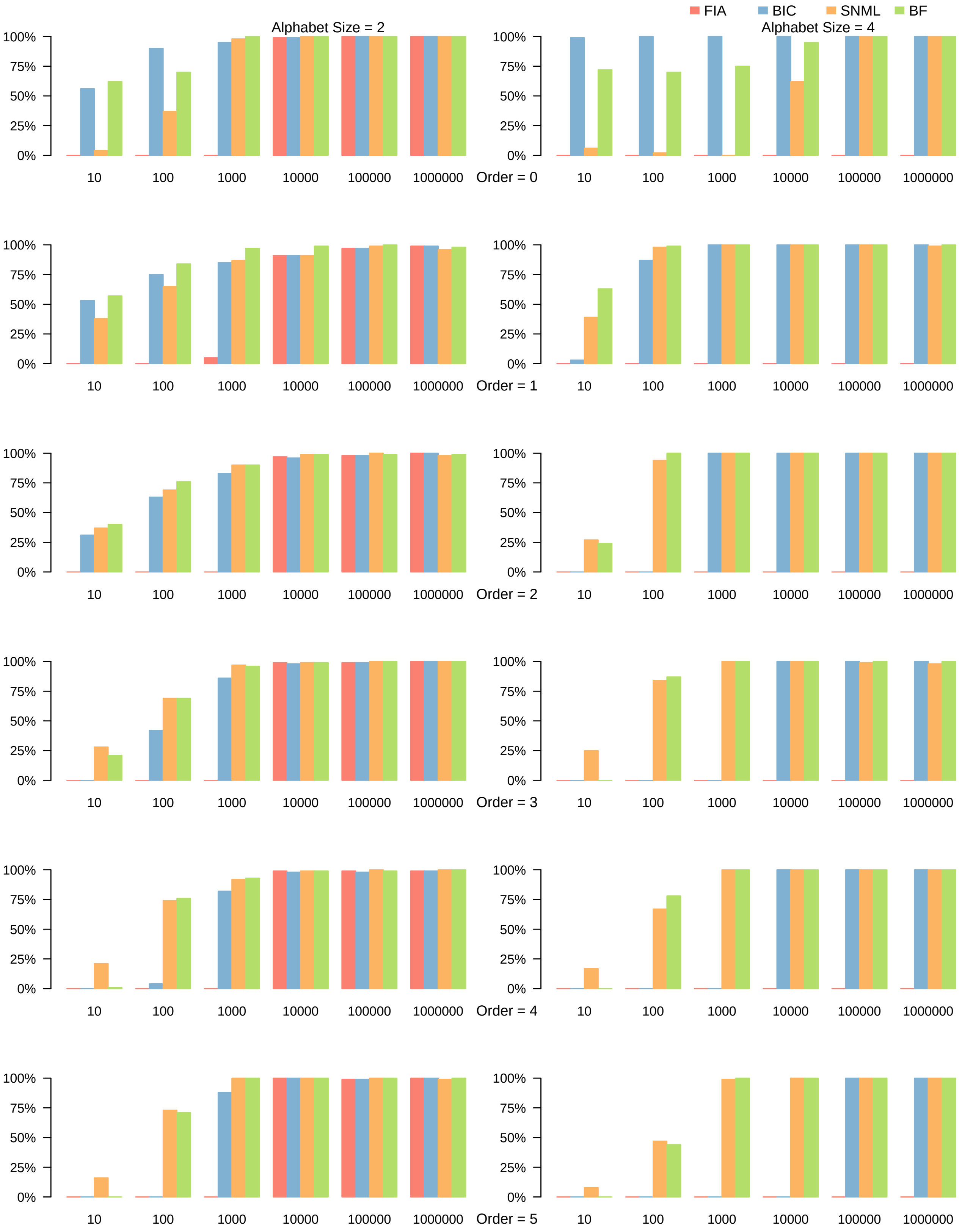

Fig. 2. Model selection experiment. Model orders $k = 0, \ldots, 5$. Bars show percentage of correctly identified model order for four different criteria as a function of sample size ($n \in \{10, 10^2, 10^3, 10^4, 10^5, 10^6\}$). Alphabet size is $\mathcal{X}| = 2$ on the left, and $|\mathcal{X}| = 4$ on the right. Criteria are FIA: Fisher information approximation, Eq. (2), BIC: Eq. (1), SNML: sequential NML [12], BF: ideal Bayes factor (Bayesian marginal likelihood with “true” prior).

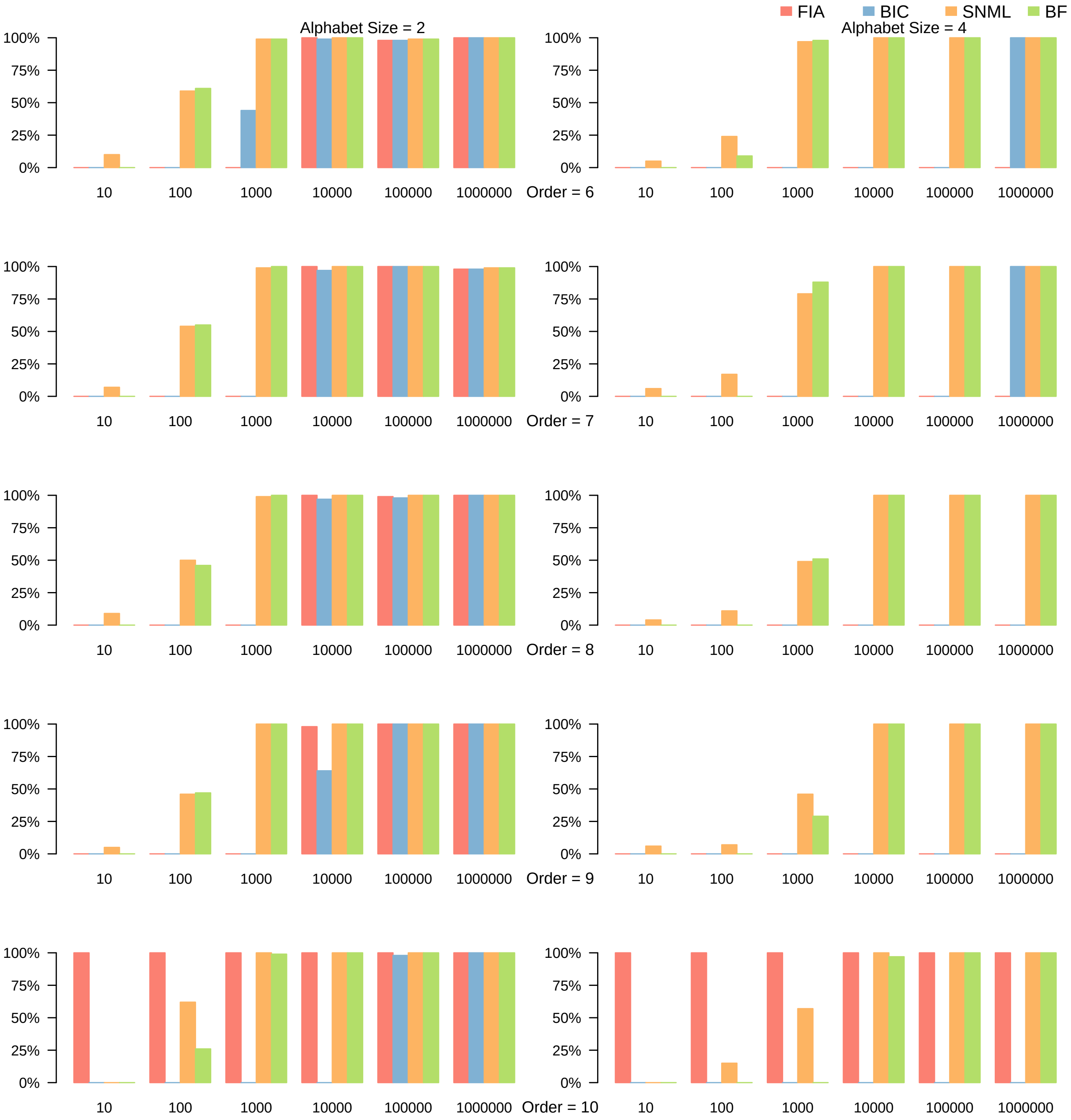

Fig. 3. Model selection experiment. (cont'd from Fig. 2): Model orders $k = 6, \ldots, 10$.

In future work, we plan to extend this work to other practically relevant model classes such as context tree models. It will also be interesting to consider whether it is possible to construct a correction term to compensate for the error in the Fisher information approximation; in other words, whether the $o(1)$ term in Eq. (2) could be broken down to obtain an accurate approximation also for small sample sizes.

VII. Conclusion

Keep it simple.

Acknowledgment

The authors that Peter Grünwald for insightful comments. This research has been supported by the Academy of Finland (Projects PRIME and MODEST and the Finnish Centre of Excellence in Computational Inference Research COIN), the Helsinki Doctoral Programme in Computer Science (HECSE), and the IST Programme of the European Community, under the PASCAL Network of Excellence, IST-2002-506778.

References

[1] H. Akaike "A new look at the statistical model identification," *IEEE Trans. Autom. Contr.*, vol. 19, no. 6, pp. 716–723, 1974.

[2] B. S. Clarke and A. R. Barron, "Jeffreys prior is asymptotically least favorable under entropy risk," *J. Statist. Planning and Inference,* vol. 41, no. 1, pp. 37–61, 1994.

[3] D. Draper, "Assessment and propagation of model uncertainty," *J. Roy. Statist. Soc. B*, vol. 57, no. 1, pp. 45–97, 1995.

[4] C. D. Giurcăneanu, D. Mihalache, M. Omarjee and M. Tetiva, "On the stochastic complexity for order-1 Markov chains and the Catalan constant," In *Proc. 16th Int. Conf. Contr. Syst. and Comp. Sci.*, pp. 114–120, Bucharest, Romania, 2007.

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

[6] H. Jeffreys, "An invariant form for the prior probability in estimation problems," *J. Roy. Statist. Soc. A,* vol. 186, no. 1007, pp. 453–461, 1946.

[7] P. Kontkanen and P. Myllymäki, "A linear-time algorithm for computing the multinomial stochastic complexity," *Inform. Process. Lett.*, vol. 103, no. 6, pp. 227–233, 2007.

[8] D. Navarro, "A note on the applied use of MDL approximations," *Neural Comput.*, vol. 16, no. 9, pp. 1763–1768, 2004.

[9] P. Jacquet and W. Szpankowski, "Markov types and minimax redundancy for Markov sources," *IEEE Trans Inform. Theory* vol. 50, no. 7, pp. 1393–1402, 2004.

[10] J. Rissanen, "Fisher information and stochastic complexity," *IEEE Trans Inform. Theory* vol. 42, no. 1, pp. 40–47,1996.

[11] J. Rissanen, *Information and Complexity in Statistical Modeling*, Springer, 2007.

[12] J. Rissanen and T. Roos, "Conditional NML universal models," in *Proc. 2007 Information Theory and Applications Workshop*, IEEE Press, pp, 337–341, 2007.

[13] T. Roos, "Monte Carlo estimation of minimax regret with an application to MDL model selection," in *Proc. 2008 IEEE Inform. Theory Workshop*, IEEE Press, pp. 284–288, 2008.

[14] G. Schwarz, "Estimating the dimension of a model," *Ann. Statist.*, vol. 6, pp. 461–464, 1978.

[15] Y. M. Shtarkov, "Universal sequential coding of single messages," *Probl. Inform. Transm.*, vol. 23, no. 3, pp. 3–17, 1987.

[16] J.-I. Takeuchi, T. Kawabata and A. R. Barron, "Properties of Jeffreys mixture for Markov sources," *IEEE Trans. Inform. Theory*, vol. 59, no. 1, pp. 438–457, 2013.

[17] Q. Xie and A. R. Barron, "Asymptotic minimax regret for data compression, gambling, and prediction," *IEEE Trans. Inform. Theory*, vol. 46, no. 2, pp. 431–445, 2000.