The ABC of Model Selection: AIC, BIC and the New CIC

Carlos C. Rodríguez

Department of Mathematics and Statistics
The University at Albany, SUNY
Albany, NY

Abstract. The geometric theory of ignorance [1] suggests new criteria for model selection. One example is to choose model M minimizing,

$$\mathrm{CIC} = -\sum_{i=1}^{N} \log \hat{p}(x_i) + \frac{d}{2} \log \frac{N}{2\pi} + \log V + \frac{\pi R}{N \log(d+1)}$$

where $(x_1, \ldots, x_N)$ is a sample of N iid observations, $\hat{p} \in M$ is the mle, $d = \dim(M)$ is the dimension of the model M, $V = \mathrm{Vol(M)}$ is its information volume and $R = \mathrm{Ricci}(M)$ is the Ricci scalar evaluated at the mle. I study the performance of CIC for the problem of segmentation of bit streams defined as follows: Find n from N iid samples of a complete dag of n bits. The CIC criterion outperforms AIC and BIC by orders of magnitude when $n > 3$ and it is just better for the cases $n = 2, 3$.

Keywords: Geometric Theory of Ignorance, Information Geometry, Model Selection, Segmentation of Bitstreams, Statistical Ignorance.
PACS: 02.50.Tt,02.40.ky

INTRODUCTION

Consider the following decision problem: Given a finite sequence of bits, $x = b_1 b_2 \ldots b_k$ choose one M among competing statistical models (i.e. explanations for x) $M_1, M_2, \ldots$. For example M_j may explain x as $N = k/j$ independent chunks of j bits generated by a graphical model of j binary variables with a given structure but unspecified parameters. We allow models M_j to be of different dimensions for both the data and the parameter spaces for different values of j. This is a standard decision problem requiring a loss function and a prior for its solution. The CIC formula defined in the abstract is an approximation to the bayes rule for 01-loss and uniform priors. By "uniform priors" we mean that the bits are generated by first choosing M uniformly at random among the available M_j's, followed by a random choice of a probability distribution $p \in M$ and finally producing $x = x_1 x_2 \ldots x_N$ as a random sample of size N from p.

The first three terms of CIC are easy to obtain. Under 01-loss the bayes action is the mode of the posterior distribution and we only need to search for the model M with highest posterior probability $P(M|x)$. By bayes theorem,

$$P(M|x) = \frac{P(M)}{P(x)} \int_M p(x) \, \frac{dV}{V}$$

where dV is the information volume element in M and $V = \mathrm{Vol}(M)$ is the total volume of M. Taking logs, noticing that $(P(M)/P(x)) \propto 1$ and using a parameterization $M \leftrightarrow \Omega \subset R^d$ we obtain,

$$\log P(M|x_1,\ldots,x_N) = \log \int_\Omega e^{NL_N(\theta)}\, dV(\theta) - \log V + C$$

where in the θ parameterization the volume element $dV = \sqrt{\det I(\theta)}\, d\theta$ and the average log likelihood $L_N = \frac{1}{N}\sum_{i=1}^N \log p(x_i|\theta)$. $I(\theta)$ is the Fisher information matrix at θ. Expanding L_N about the mle $\hat{\theta}$, noticing that $\nabla L_N(\hat{\theta}) = 0$ and that by the LLN (Law of Large Numbers) $-\nabla\nabla L_N(\hat{\theta}) \to I(\hat{\theta})$ we can write,

$$N\, L_N(\theta) = N\, L_N(\hat{\theta}) - \frac{N}{2}(\theta-\hat{\theta})^T I(\hat{\theta})(\theta-\hat{\theta}) + o(N|\theta-\hat{\theta}|^2)$$

Thus, the bayes action (as $N \to \infty$) is the model that maximizes

$$NL_N(\hat{\theta}) + \log\left(\frac{2\pi}{N}\right)^{d/2} - \log V$$

where the second term is obtained by noticing that for large N, by the mean value theorem for integrals and the formulas for dV and the normalizing constant of a d-dim gaussian,

$$\int e^{-\frac{N}{2}(\theta-\hat{\theta})^T I(\hat{\theta})(\theta-\hat{\theta})}\, dV = |I(\hat{\theta})|^{1/2} \left|\frac{2\pi}{N} I^{-1}(\hat{\theta})\right|^{1/2}$$

The first three terms of *CIC* are then just a simple consequence of the large sample properties of mle's. The last term involving the Ricci scalar R at the mle was obtained semi-empirically by simulation guided by the more rigorous analysis in [1].

TESTING CIC

To test the performance of CIC as a criterion for model selection we compared it with AIC (the "An" Information Criterion of Akaike [2]) and with BIC (the Bayesian Information Criterion of Schwarz [3]). All three criteria search, among a list of possible models for the data, for the one minimizing the AIC, BIC or CIC expressions defined by,

$$\mathrm{AIC} = -N\, L_N(\hat{\theta}) + d \tag{1}$$

$$\mathrm{BIC} = -N\, L_N(\hat{\theta}) + \frac{d}{2}\log\, N \tag{2}$$

$$\mathrm{CIC} = -N\, L_N(\hat{\theta}) + \frac{d}{2}\log\frac{N}{2\pi} + \log V + \frac{\pi R}{N\log(d+1)} \tag{3}$$

A Note About the ABC ICs

AIC

Initially, Akaike justified AIC by maximum entropy. He asked: If there are competing models with a, possibly different, number of free parameters; How should the sample x^N be used to choose one of them?. He reasoned: If we knew the true distribution t then we could take the model that maximizes the entropy relative to this t. Translation: solve $\arg\min_M I(t:M)$ where $I(t:M)$ is the Kullback distance from t to M. Let $p_M \in M$ be the I-projection of t onto M, i.e. $I(t:M) = I(t:p_M)$. Thus, AIC attempts to find,

$$\begin{aligned} M_a &= \arg\min_M I(t:p_M) = \arg\min_M -\int t(x)\log p_M(x)\,dx \\ &= \arg\min_M -E_t \log\, p_M(X) \end{aligned}$$

The problem is that t and p_M are unknown and need to be estimated from the data x^N. By the LLN and the consistency of the mle (i.e. $\hat{p} \to p_M$) we have,

$$-E_t\, \log\, p_M(X) \approx -E_t\, \log\, \hat{p}(X) \approx -\frac{1}{N}\sum_{i=1}^{N} \hat{p}(X_i)$$

but this naïve mle estimate is biased for finite N. It could then be argued that after collecting the data x^N, it should be the term in the middle, $-E_t \log \hat{p}$, the one quantifying the loss and not the original, $-E_t \log p_M$. The asymptotic bias (with respect to $-E_t \log \hat{p}$) can be obtained by the following (tricky) considerations. By the (generalized) Pythagoras theorem (see [4]) we have:

$$I(t:p_M) + I(p_M:\hat{p}) = I(t:\hat{p})$$

and rearranging terms we get,

$$E_t \log \frac{\hat{p}(X)}{p_M(X)} = -I(p_M:\hat{p})$$

Thus,

$$2NE_t \log \frac{\hat{p}(X)}{p_M(X)} \approx -||\sqrt{N}(\theta_0 - \hat{\theta})||_0^2 \approx -\chi_d^2$$

where θ_0 is the parameter associated to p_M and we have used the consistency and asymptotic normality of the mle (i.e., $\hat{\theta} \to \theta_0$ and $\sqrt{N}(\hat{\theta} - \theta_0) \to N(0, I^{-1}(\theta_0))$, to arrive at the asymptotic Chi-square with d degrees of freedom. With this we can write:

$$\lim_{N\to\infty} \qquad E\left\{ -\sum_{i=1}^{N} \log\, \hat{p}(X_i) + N\, E_t\, \log\, \hat{p}(X) \right\} =$$

$$\lim_{N\to\infty} E\left\{-\sum_{i=1}^{N}\log\frac{\hat{p}(X_i)}{p_M(X_i)}+N\,E_t\,\log\frac{\hat{p}(X)}{p_M(X)}\right\}=-\frac{d}{2}-\frac{d}{2}=-d.$$

Where we have used the fact that twice the sum involving the likelihood ratio (line above) converges (in law) to a Chi-square with d degrees of freedom. Hence, AIC is just the naïve mle corrected to be asymptotically unbiased as an estimator of $-E_t \log \hat{p}$. These arguments try to justify the definition of AIC in (1). Nevertheless, I don't find AIC defensible as a general criterion for model selection.

BIC and CIC

The BIC of Schwarz can be obtained by following the derivation for the first three terms of CIC used in the introduction of this paper. However, instead of using the uniform prior on M use a fix arbitrary positive prior on M and neglect the terms of order $N^0 = 1$ to arrive at (2). The problem with BIC is that the neglected terms, involving the volume and the curvature of M can become the leading terms. In fact that is the case for the important case of multinomial models studied in this paper.

The Simulations

We played repeatedly (100 repetitions per sample size) the standard game of generating a sample of size N from a chosen true distribution. Then, acting as if we didn't know this true distribution, we let AIC, BIC, and CIC, guess a model for the simulated data and counted the proportion of correct guesses for each criterion.

The underlying true distributions were chosen from the set $\{M_2, M_3, \ldots, M_9\}$ where M_n is the complete dag of n binary variables. The observed sequence of bits was created by concatenating a random sample of size N from M_n with random values for the parameters.

The dimension d, volume V, and scalar curvature R for the complete dag M_n were computed in [5] as,

$$d = 2^n - 1 \tag{4}$$

$$V = \frac{\pi^k}{(k-1)!}, \text{ where } k = 2^{n-1} \tag{5}$$

$$R = \frac{1}{4}d(d-1). \tag{6}$$

The simulations, are summarized in figure (1). The graphs show that CIC is orders of magnitude better than its two competitors (AIC and BIC) as a criterion for determining the size of complete bitnets. When $n = 9$ the graph seems to show only the CIC curve. The two other curves (for AIC and BIC) are in fact there but are never different from 0!. For example with 100 observations from complete bitnets of 9 bits, CIC chose the

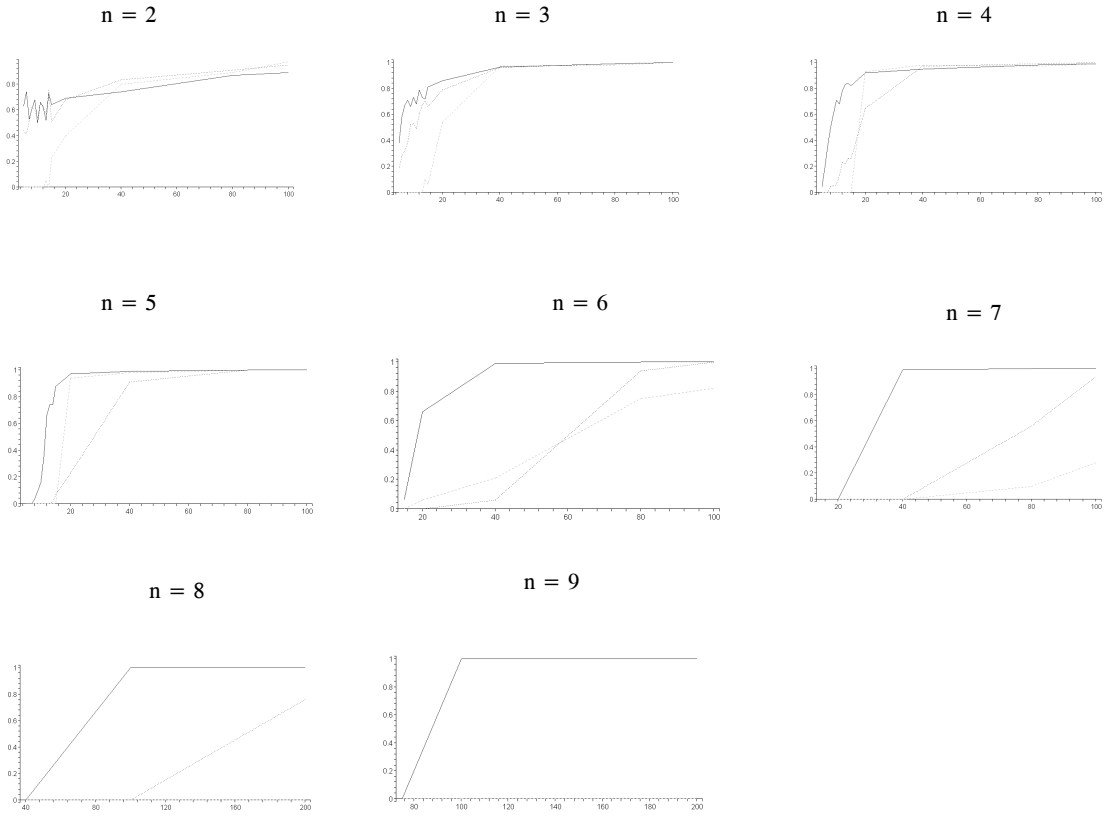

FIGURE 1. Proportion of Successes vs. Sample size. AIC(dot), BIC(dash), CIC(solid)

correct size $n = 9$, 100 times out of 100 but AIC and BIC failed all 100 times. When the sample size was increased to 200, CIC chose the correct $n = 9$ all 100 times, but AIC and BIC still failed all 100 trials. We should emphasize that each sample was chosen with random values for the parameters. Thus, the superiority of CIC over AIC and BIC appears to be independent of the actual values of the parameters of the complete bitnet. This is compatible with the homogeneous (constant curvature) geometry of complete bitnets.

The results of these simulations also agree, reinforce, and validate the findings by Eitel Lauría in [6, 7]. Lauría's Monte Carlo experiments show conclusively that by adding the approximation for $\log V$ (see [5]) to the BIC formula (2) (i.e. essentially using the first three terms of CIC) outperforms plain BIC in the much more difficult task of identifying the full structure of a bitnet of n (fixed) bits and the performance increases with the maximum number of parents in the true bitnet.

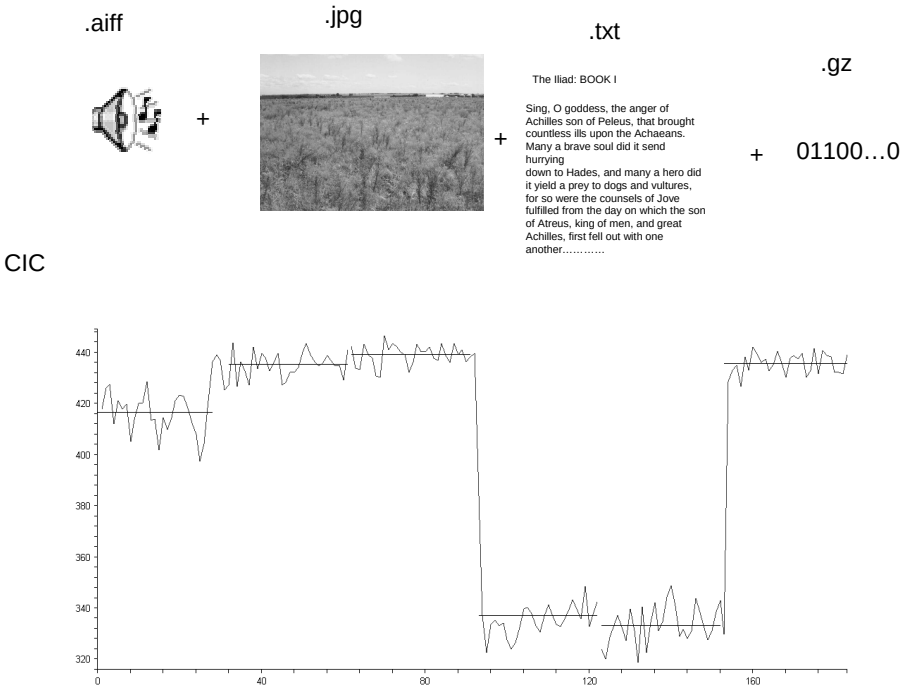

FIGURE 2. Segmentation of concatenated files by CIC

SEGMENTING REAL DATA

To demonstrate the sensitivity of CIC for recognizing model changes along a bitstream, we created a long sequence of bits by concatenating pieces of files of different types. The data x was,

$$x = 10k(.aiff) + 20k(.jpg) + 5k(.txt) + 5k(.gz)$$

i.e., ten thousand bytes from a sound file (.aiff type), followed by 20k bytes from an image file (.jpg type), followed by 5k from an ascii file (.txt type), followed by 5k from the same ascii file by compressed with gnuzip (.gz type). Figure (2) shows the value of CIC for a sliding window of 640 bits traversing the boundaries between files of different types. The horizontal lines indicate the average values of CIC for the different segments: end of sound, sky, ground, beginning of text, end of text, compressed data.

FOR MODEL SELECTION IGNORANCE IS BLISS

Take the blue pill, the story ends. You wake up in your bed and believe whatever you want to believe. You take the red pill, you stay in Wonderland, and I show you how deep the rabbit hole goes.... Remember, all I'm offering is the truth, nothing more.... Morpheus (holding out two pills): The Matrix.

Let M be a manifold of homogeneous theories, i.e. M is a standard regular, parametric statistical model. M is riemannian with the induced metric from the Hellinger distance (i.e. with Fisher information as the metric) and therefore carries a notion of volume element $dV = \sqrt{\det I(\theta)}\, d\theta$. Consider the following two ways for generating (Data,Theory):

1) The Informative Prior Way Pick Theory $p \in M$ with prior probability scalar density $\pi(p)$. Then, observe Data $x^\alpha = x_1 x_2 \dots x_\alpha$, i.e. with probability $p(x^\alpha) = p(x_1)p(x_2)\dots p(x_\alpha)$.
Here (Data,Theory) are *dependent*: $\text{Prob}(x^\alpha, p) = p(x^\alpha)\ \pi(p) \equiv 1_\pi$.

2) The Ignorant Prior Way Pick Theory $p \in M$ uniformly at random, i.e. with constant scalar density (assume M of finite volume). Then, observe Data $x^\alpha = x_1 x_2 \dots x_\alpha$ from the *true* distribution, i.e. with probability $t(x^\alpha) = t(x_1)t(x_2)\dots t(x_\alpha)$.
Here (Data,Theory) are *independent*: $\text{Prob}(x^\alpha, p) = t(x^\alpha)\ \omega(p)$.

Ignorance is self-similar

Notice that in the ignorant way above, data is assumed to come from the true (t) distribution. If you happen to know this t then you have complete knowledge; That's all there is to know about the distribution of the data and you have arrived at the true theory of everything. Enjoy!

If, on the other hand, all you know about t is the manifold M of guesses, then the *ignorant* generative model (2, above) preserves that prior state of knowledge a posteriori. The prior state of indifference (uniform $\omega(p)$) about the elements of M does not change after observing the data x^α, for all $\alpha > 0$. The posterior is the same as the prior since Data and Theory are independent.

LIPREM: The Red Pill

The notion of ignorance sketched above, produces π^* as LIPREM: Least Informative Prior RElative to M. Where,

$$\pi^* = \arg\min_\pi \text{Dist}(1_\pi : 2)$$

where "Dist" is any statistically meaningful measure of separation between the joint distribution of (Data,Theory) specified by 1_π and by the ignorant way 2 above. The class of all the statistically meaningful notions of separation between unnormalized probability distributions can be shown to be generated by the δ-*information deviations* I_δ (see [1] and the references there), where $\delta \in [0,1]$. If we let,

$$\mathscr{A}^*(M) = I_\delta(1_{\pi^*} : 2) \tag{7}$$

to be the total information in M then,

$$M^* = \arg\max_M I_\delta(1_{\pi^*} : 2) \tag{8}$$

is the minimax ignorant (or maximin informative) model. The CIC is just an estimate of $\mathscr{A}^*$ for $\delta = 0, \alpha = N, t = \hat{p}$ (the mle) that keeps the first terms of an asymptotic expansion in $\alpha = N$ (see [1]). LIPREM and its extensions produce all the statistically meaningful actions for model selection. The standard maximum posterior probability model is just one special case.

CONCLUSION: MORE GEOMETRY

It is natural to decompose AIC, BIC and CIC as the sum of two terms. The term providing the fit of the data to the model (common to all the three criteria) plus the rest. That rest is obviously a penalty on the complexity of the model. In retrospect, it is to be expected that the complexity of a model M should involve some (or all?) of its geometric and topological invariants like: dimension, volume and curvature, as CIC does. But we need to keep in mind that CIC, like AIC and BIC, is only an approximation. It would be much better to be able to show that useful models spring from the optimization of a global topological quantity, like the total (or mean?) scalar curvature of M. In fact, we already know that that is precisely the case in classical physics. I would like to show that that is also the case for the whole of inference.

REFERENCES

1. C. Rodríguez, A geometric theory of ignorance, Tech. rep., SUNY Albany, Dept. of Mathematics, http://omega.albany.edu:8008/ignorance (2003).
2. H. Akaike, *IEEE Transactions on Automatic Control*, pp. 716–723 (1974).
3. G. Schwarz, *Annals of Statistcs*, **6**, 461–464 (1978).
4. S.-i. Amari, *Differential-Geometrical Methods in Statistics*, vol. 28 of *Lecture Notes in Statistics*, Springer-Verlag, 1985.
5. C. Rodríguez, "The Volume of Bitnets," in *Maximum Entropy and Bayesian Methods*, edited by R. Fischer, R. Preuss, and U. von Toussaint, 2004, vol. 735 of *AIP Conf. Proc.*, pp. 555–564, http://omega.albany.edu:8008/bitnets.
6. E. Lauría, *Learning structure and parameters of Bayesian Belief Networks*, Ph.D. thesis, The University at Albany, SUNY. School of Information Science (2003), http://omega.albany.edu:8008/bitnets/references/.
7. E. Lauría, "Learning the structure of a Bayesian network," in *Maximum Entropy and Bayesian Methods*, AIP Conf. Proc., 2005, these Proceedings.