Journal de la Société Française de Statistique
Vol. 160 No. 3 (2019)

A note on BIC and the slope heuristic

Titre: Discussion sur l'article de Sylvain Arlot : "Pénalités minimales et heuristique de pente"

Christine Keribin[1,2]

BIC (Schwarz et al., 1978) is a Bayesian model selection criterion relying on an asymptotic approximation of the integrated likelihood, where $O_p(1)$ are neglected. It is generally written for a parametric model $m \in \mathscr{M}$ as the following penalized maximum likelihood criterion

$$BIC(m) = -\mathscr{L}(\widehat{\theta}_m) + \frac{\log(n)}{2} D_m \tag{1}$$

where $\mathscr{L}(\widehat{\theta}_m)$ is the maximized likelihood in the model m and D_m the number of estimated parameters. The model with minimum BIC is chosen. BIC is known to be consistent in many situations when the true model belongs to a nested family (Keribin, 2000; Gassiat and Van Handel, 2013; Yang, 2005) but these asymptotic properties may not hold in practice.

In an other hand, the slope heuristic, allowing to define non asymptotic minimal and optimal penalties, can be naturally extended to frameworks where a penalty is known up to a multiplicative constant. It is the case for BIC as the optimal penalty term is known theoretically and in practice, but only asymptotically:

$$BIC(m) = -\mathscr{L}(\widehat{\theta}_m) + C\, D_m$$

The constant C can be estimated with Algorithm 2 (Arlot, 2019), as used for example by Rau et al. (2015).

We illustrate here an interesting comparison between BIC and the slope heuristic. In Keribin et al. (2019), a constrained mixture model is developed to estimate tumor genome alterations. Each single-nucleotide polymorphism (SNP) of a DNA sequence is characterized by a copy number (cn) and a bi-allele frequency (baf). It can be shown that the (baf, cn) values are located on a grid, whose frame depends on the proportion p of the normal tissue, see figure 1. Each point of the grid corresponds to a specific genomic alteration. The acquisition process gives access to the copy number through the logarithm of the R-ratio, $\text{lrr} = \alpha \log_2 \text{cn} + \beta$, where α and β are unknown constants. Thus, two noisy signals BAF and LRR are extracted from SNP arrays and segmented in S regions of assumed constant (and unknown) alteration (figure 2). The joined (BAF, LRR) observations are located around a point of a theoretical grid whose characteristics depend on the unknown proportion p of normal tissue and the experimental parameters α and β, these three values to be inferred. To reduce the variance and the size of the data, the signal

[1] Laboratoire de Mathématiques d'Orsay, Univ. Paris-Sud, CNRS, Université Paris-Saclay, 91405 Orsay, France.
E-mail: christine.keribin@math.u-psud.fr
[2] INRIA-Saclay Ile de France - Equipe CELESTE.

 ISSN: 2102-6238

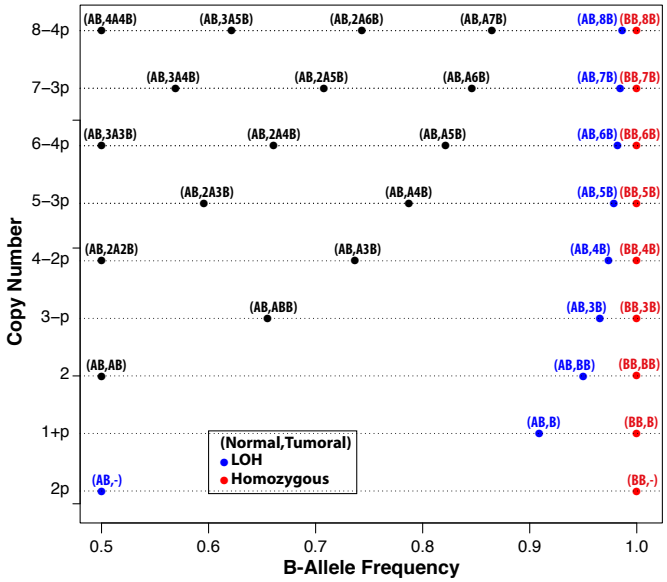

FIGURE 1. *For a given proportion p of the normal tissue, positions of the tumoral mutations in the 2D plane defined by B-Allele Frequency (*baf*) and copy number (*cn*). Mutations of germ line homozygous are in red. Mutations of germ line heterozygous are in black and blue. The blue centers characterize a loss of heterozygosity (LOH).*

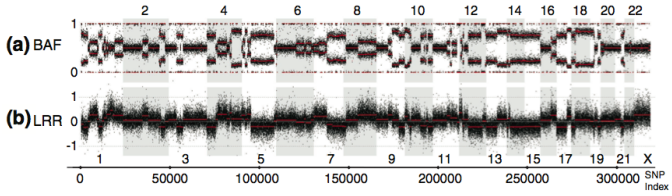

FIGURE 2. *Example of tumoral measurements, from Popova et al. (2009).*

is averaged on each segment $i = 1, \ldots, S$ of length n_i (between 500 to 1500 SNPs in a segment). Hence, knowing the theoretical mutation $k(i)$ of segment i, the conditional distribution of the observed mean signal $(\mathrm{BAF}_i, \mathrm{LRR}_i)$ is defined as a bi-variate independent Gaussian distribution:

$$\mathrm{BAF}_i(k(i)) \sim \mathcal{N}\left(\mathrm{baf}_{k(i)}(p,\alpha,\beta), \frac{\sigma^2}{n_i}\right), \quad \mathrm{LRR}_i(k(i)) \sim \mathcal{N}\left(\mathrm{lrr}_{k(i)}(p,\alpha,\beta), \frac{\eta^2}{n_i}\right) \qquad (2)$$

where $(\mathrm{baf}_{k(i)}(p,\alpha,\beta), \mathrm{lrr}_{k(i)}(p,\alpha,\beta))$ is the theoretical center of the mutation $k(i)$ of segment i. As the allocations of the observations to the mutation centers are unknown, a Gaussian mixture model is defined, whose centers are constrained to belong to a grid. The unknown parameters to be inferred are α, β, p, the variances σ^2 and η^2 and the mixing weights π. Hence, the model size D_m depends essentially on the size of the mixing weights π, defined by the number of possible centers for a maximum copy number m.

This model was applied on a real colon tumor sample, with $S = 200$ segments, leading to $n = 2S$ (homozygous or heterozygous) observations from a genome sequence of $N = 262\,000$

 ISSN: 2102-6238

SNPs. BIC largely over-estimates the maximum copy number that was known for these data. As the graph of the maximal log-likelihood against D_m clearly shows a linear trend for large values of D_m, the slope heuristic was also tested. It leads to select the correct model, with an estimated minimal $\widehat{C}_{slope} = 3.76$ greater than the BIC value $C_{BIC} = \log(2S)/2 = 3$. This may seem surprising, because BIC does not tend in general to overestimate the size of a model. However, BIC can be sensitive to bias, and the observed overestimation could be due to the rough assumptions (independence, homoscedasticity for example) that have been made.

We propose an alternative interpretation. In case of identifiable models, BIC results from the asymptotic expansion of the integrated likelihood (see Lebarbier and Mary-Huard (2006) for details):

$$\log \mathbb{P}(X|m) = \mathscr{L}(\widehat{\theta}|m) - \frac{D_m}{2}\log n \tag{3}$$

$$+ \underbrace{\log \mathbb{P}(\widehat{\theta}_m|m) + D_m\frac{\log(2\pi)}{2} - \frac{1}{2}\log|I(\widehat{\theta}_m)|}_{O_P(1)} + O_P(n^{-1/2}) \tag{4}$$

where $\mathscr{L}(\widehat{\theta}|m)$ is the maximum likelihood under model m and $|I(\widehat{\theta}_m)|$ is the determinant of the Fisher information matrix, estimated by

$$I(\widehat{\theta}_m) = -\frac{\partial^2 \mathscr{L}(\theta)/n}{\partial\theta^2}_{|\theta=\widehat{\theta}_m}.$$

Standard *BIC* is defined by neglecting terms less than order $O_P(1)$, that is, all terms in (4). Taking the opposite gives (1). As noticed in Remark 2 of Lebarbier and Mary-Huard (2006), if the error in $O_P(n^{-1/2})$ in (4) is negligible when n tends to infinity, the error in $O_P(1)$ resulting of the Laplace approximation can disturb the choice of the final model even if the two terms in (3) are preponderant when n is large. Moreover, following Kass and Wasserman (1995), the sample size n should be the rate at which the Hessian matrix of the log-likelihood function grows; thus n becomes the number of data values contributing to the summation that appears in the formula of the Hessian. In our model, both the number of observations and the Hessian are difficult to determine. Remember that each observation i is itself the average of the signal on a DNA segment with n_i SNPs, and coming from the constrained unknown center $(\mathrm{baf}_{k(i)}, \mathrm{lrr}_{k(i)})$. If all the segments were of the same length, say a mean length $n_i = \ell = N/S$, we assume that $I(\widehat{\theta}_m)$ could be written as $\tilde{I}(\widehat{\theta}_m)/\ell$, so that the penalty constant including $O_P(1)$ terms in (4) would not be negligible and $\tilde{C}_{BIC} = C_{BIC} + (\log(\ell) - \log(2\pi))/2 = 5.6$. With this adapted definition, BIC selects the correct model. This would be consistent with Raftery (1995) who stressed on the fact that n should be the actual number of individuals rather than the number of cases or cells. For logistic regression for example, it should be the number of individuals, and not the number of grouped data. Hence in our case, BIC penalty term would be $\check{C}_{BIC} = \log(2N)/2 = 6.6$. Both penalty terms $\tilde{C}_{BIC}$ and $\check{C}_{BIC}$ are between the minimal $\widehat{C}_{slope}$ and optimal $2\widehat{C}_{slope}$ values coming from the slope heuristic.

In conclusion, we claim that using the slope heuristic when the maximum likelihood presents a linear trend for large values of D_m is a robust way to perform model selection. In anyways, when C_{BIC} is less than the minimal $\widehat{C}_{slope}$, BIC should not be used; at least the way it is approximated

 ISSN: 2102-6238

should be reconsidered. Moreover, as some rough assumptions have been made to define the model (such as independence and homoscedasticity), this example shows a case where the slope heuristic can be used with some model bias. One could hence conjecture that the slope heuristic can be justified when a bias exists and remains constant for the large models.

References

Arlot, S. (2019). Minimal penalties and the slope heuristics: a survey. *arXiv preprint arXiv:1901.07277*.

Gassiat, E. and Van Handel, R. (2013). Consistent order estimation and minimal penalties. *IEEE Transactions on Information Theory*, 59(2):1115–1128.

Kass, R. E. and Wasserman, L. (1995). A reference bayesian test for nested hypotheses and its relationship to the schwarz criterion. *Journal of the american statistical association*, 90(431):928–934.

Keribin, C. (2000). Consistent estimation of the order of mixture models. *Sankhyā: The Indian Journal of Statistics, Series A*, pages 49–66.

Keribin, C., Liu, L., Popova, T., and Rozenholc, Y. (2019). A mixture model to characterize genomic alterations of tumors. *Journal de la SFdS*, 160(1):130–148.

Lebarbier, É. and Mary-Huard, T. (2006). Une introduction au critère bic: fondements théoriques et interprétation. *Journal de la Société française de statistique*, 147(1):39–57.

Popova, T., Manié, E., Stoppa-Lyonnet, D., Rigaill, G., Barillot, E., Stern, M. H., et al. (2009). Genome alteration print (GAP): a tool to visualize and mine complex cancer genomic profiles obtained by SNP arrays. *Genome Biol*, 10(11):R128–R128.

Raftery, A. E. (1995). Bayesian model selection in social research. *Sociological methodology*, 25:111–164.

Rau, A., Maugis-Rabusseau, C., Martin-Magniette, M.-L., and Celeux, G. (2015). Co-expression analysis of high-throughput transcriptome sequencing data with poisson mixture models. *Bioinformatics*, 31(9):1420–1427.

Schwarz, G. et al. (1978). Estimating the dimension of a model. *The annals of statistics*, 6(2):461–464.

Yang, Y. (2005). Can the strengths of AIC and BIC be shared? a conflict between model indentification and regression estimation. *Biometrika*, 92(4):937–950.

 ISSN: 2102-6238