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

When is the slope heuristic useful?

Titre: Quand l'heuristique de pente est-elle utile ?

Gilles Celeux[1]

I will focus my comments on the extensive survey on the slope heuristics by Sylvain Arlot on its practical aspects. Practical activity in applied statistics consists of fitting models to data. My comments will turn around the famous aphorism of Box (1976) "All models are wrong, but some are useful."

All models are wrong... Whatever its complexity, any model does not fit the data perfectly and presents a bias. If the bias becomes small enough when the model complexity increases, standard models selection criteria such as AIC or BIC could be expected to do a good job under some conditions. But when the bias remains somewhat large for any model complexity, criteria AIC and BIC have a marked tendency to select too complex models. The claim that BIC has a tendency to underestimate the penalty of a model in practical situations, see for instance (Keribin, 2019), seems to be in contradiction with the common opinion that BIC could have a tendency to underestimate the dimension of the "true" model. The reason of this gap between two contrary behaviors of the BIC criterion could be caused by the fact that there is no "true" model in the family of models at hand. (In Keribin (2019), this tendency of BIC to overestimate the dimension of the "true" model is related to an other well identified mathematical reason: the asymptotic approximation of BIC is $O_p(1)$ and could be rough.)

..., but some are useful Obviously, when AIC and BIC choose underpenalized models, they are not useful. In order to select useful models some specific criteria have been proposed to take into account the aim of the modeller. An important example of such criteria is the ICL criterion which aims to select a useful mixture model in the model-based clustering context (Biernacki et al., 2000). Other examples are (Baudry et al., 2015; Gallopin et al., 2015). Such criteria could be actually useful but studying their theoretical properties is difficult.

When a collection of models appears to be a crude approximation of the data distribution, the slope heuristic gives a more universal and well grounded answer to the model selection problem. My claim is that when the bias of a collection of models tends to become a constant *Cst* when the complexity increases, the slope heuristic allows to choose the model of minimum variance inside the set of models with bias approximatively equal to *Cst*. Obviously the statement of this claim needs to be made precise and it requires precise conditions... But, I think it could possible to state and prove a theorem analogous to the Theorem 1 of the Arlot article under the assumption that the approximation error becomes constant as the dimension increases. [2]

[1] Inria Saclay Île-de-France

[2] I show the moon. I hope some researchers will jump to catch it.

 ISSN: 2102-6238

Many numerical applications, some of them being cited in the survey of Sylvain Arlot, support this claim and this hope of a theoretical justification. Among them, I would like to highlight the article by Rau et al. (2015) where the slope heuristic allows to select a relevant number of clusters in a Poisson mixture model to cluster RNASeq data despite the fit of a Poisson mixture to such data is questionable.

From this point of view, in the common context of the comparison of a model collection of the same family differing by their complexity, it is important to detect carefully if (and when) the chosen contrast becomes linear when the complexity of the models increases. In this respect (i) the graph of the chosen contrast as a function of the penalty, as illustrated in Figure 4 of the Arlot article is highly useful to answer this question; (ii) estimating the minimal-penalty estimator $\hat{C}$ through a robust linear regression leading to $\hat{C}_{\text{slope}}$ is more natural than the estimator $\hat{C}_{\text{jump}}$. In practice, $\hat{C}_{\text{slope}}$ is also more reliable than $\hat{C}_{\text{jump}}$ because the "unique large jump" is often decomposed into "a cascade of moderate jumps".

Finally, all these considerations show that I am a supporter of Algorithm 2...

References

Baudry, J.-P., Cardoso, M., Celeux, G., Amorim, M. J., and Sousa Ferreira, A. (2015). Enhancing the selection of a model-based clustering with external categorical variables. *Advances in Data Analysis and Classification*, 9:177–196.

Biernacki, C., Celeux, G., and Govaert, G. (2000). Assessing a mixture model for clustering with the integrated completed likelihood. *IEEE Transaction on PAMI*, 22:719–725.

Box, G. (1976). Science and statistics. *Journal of the American Statistical Association*, 71:791–799.

Gallopin, M., Celeux, G., Jaffrezic, F., and Rau, A. (2015). A model selection criterion for model-based clustering of annotated gene expression data. *Stat Appl Genet Mol Biol*, 14:413–428.

Keribin, C. (2019). A note on BIC and the slope heuristic; discussion on "minimal penalties and the slope heuristic: a survey" by Sylvain Arlot. *Journal de la Société française de statistique*, 160:2.

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.

 ISSN: 2102-6238