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

# Discussion on "Minimal penalties and the slope heuristic: a survey" by Sylvain Arlot

**Titre:** Discussion sur "Pénalités minimales et heuristique de pente" par Sylvian Arlot

Adrien Saumard[1]

I would like to begin this note with some sincere compliments to Sylvain Arlot for a most valuable survey. I know that this ambitious project started some years ago and I deeply thank his author for having had the courage to put an end to this quite huge amount of work. Let me also thank the Editor Gilles Celeux for giving the opportunity to publish this work in the best conditions and having raised a discussion around it.

I have no doubt that this survey will contribute to promote the whole lines of research connected to minimal penalties and optimal penalties design heuristics, from the most theoretical aspects, to the methodological ones and to their usage in applications as well. All these facets are treated with great details at one place, which is very rare, but highly precious. Indeed, it emphasizes the unity of the subject and the connections between existing works and approaches.

## 1. On the usage of the slope heuristics

With a slight abuse of simplification, I would emphasize that the slope heuristics and more general penalty calibration methods are mostly interesting when (V-fold) cross-validation - or other resampling based methods - is inefficient, or difficult to implement, or impossible. This occurs in two major domains of statistics and machine learning (and of course in a variety of other statistical tasks): clustering - that is unsupervised classification - and time series analysis.

Because of the lack of labels, a standard cross-validation of the classification performance is impossible in clustering. As reviewed in the article, there is already a great deal of work on the behavior of the slope heuristics for the clustering task, especially in the model based approach. Empirical evidence of the benefits of using penalty calibration algorithms in these settings is now clear. I would argue however that we have rather little theoretical understanding of the problem. The main obstacle is that mixture models are (highly) non-linear. So one is tempted to use some classical chaining arguments, that are inefficient when tackling the optimality of the slope heuristics (for instance), since some constants are automatically lost in the estimates. The use of chaining estimates should in fact at least be indirect in order to preserve some sharp theoretical expressions related to the excess risks and general representation formulas for the latter quantities such as in Navarro and Saumard (2017) might help in this case.

[1] Crest-Ensai, Université Bretagne-Loire

A refined analysis of the geometry of mixture models would be very welcome in my opinion, and could have an impact on the understanding of the nature of the clustering task in general. Remarkable contributions that may be related to this direction are (Gassiat and van Handel, 2013, 2014; Heinrich and Kahn, 2018).

Concerning time series analysis, which is in my point of view a most natural domain of application of penalty design heuristics, precise investigations are rather lacking. I would push towards the analysis of some classical models situations, that have a great impact for instance on econometry, such as model selection for auto-regressive processes, especially ARMA-type processes, and also for GARCH-type volatility models. Indeed, cross-validation is typically hard to implement in these settings, especially when the residuals are correlated, since for instance the technique of blocking depends on some hyper-parameters that are difficult to tune.

A way to start in this direction of research is to look for refinements in previous analysis of model selection in such frameworks. Relevant references are (but not limited to) related to early works of Fabienne Comte and co-authors (Baraud et al., 2001; Comte and Rozenholc, 2002; Comte and Genon-Catalot, 2006; Comte et al., 2008, 2010).

## 2. A conjecture in the binary classification setting

Grant the notations of the survey and set $\hat{\mathscr{R}}_n(t) = 1/n \sum_{i=1}^n 1_{\{Y_i \neq t(X_i)\}}$ for binary valued random variables $Y_i$ and function $t$ , thus corresponding to the binary classification setting, with i.i.d. sample $(X_i, Y_i), i = 1, \ldots, n$. Assume that we have a polynomial collection of models, with polynomial covering numbers. More precisely, assume that there exists $A_m, V_m > 0$ such that,

$$N(m, L_2(P_n), \varepsilon) \leq \left( \frac{A_m \|F_m\|_{L_2(P_n)}}{\varepsilon} \right)^{V_m}, \varepsilon > 0,$$

where $N(m, L_2(P_n), \varepsilon)$ is the minimal number of $L_2(P_n)$-balls covering the model $m$ and $F_m$ is a measurable envelope of $m$.

**Conjecture:** Under the above framework, the “classical” slope heuristics - with constant 2 between the optimal and minimal penalty - is valid for a penalty shape equal to the (smallest) power $V_m$ appearing in the polynomial entropy number bound of the model $m$, if there exists a (uniform) strong margin relation over the union of the models and if the models are “nice” enough (in terms of metric entropy). If only a weak margin condition holds, then the minimal penalty phenomenon - i.e. the phase transition - is still satisfied for a penalty shape equal to $V_m$, for “nice” enough models, but the ratio of the optimal penalty over the minimal one is greater than 2 - in fact equal to $2/\alpha$ - and depends on the (highest) exponent $\alpha$ in the margin relation, $\operatorname{Var}\{\gamma(s)\text{-}\gamma(s_m)\} \leq L[\mathscr{R}(s) - \mathscr{R}(s_m)]^\alpha$, $\alpha \leq 1$, $L > 0$, where $\gamma(t)(x,y) = 1\{y \neq t(x)\}$ is the so-called binary classification contrast.

Let me explain the rationale behind this conjecture. Using (Navarro and Saumard, 2017, Proposition 6.6), we have for any $m \in \mathscr{M}$,

$$p_1(m) \in \arg\max_{C \geq 0} \left\{ \sup_{s \in m_C} \{(P_n - P)(\gamma(s_m) - \gamma(s))\} - C \right\}$$

and

$$p_2(m) = \max_{C\geq 0}\left\{\sup_{s\in m_C}\{(P_n-P)(\gamma(s_m)-\gamma(s))\}-C\right\} ,$$

with $m_C = \{s\in m : \mathscr{R}(s)-\mathscr{R}(s_m)\leq C\}$. Now, for $C$ sufficiently large (it happens that the constant in the lower bound for $C$ is actually problematic when there is existence of a strong margin relation), results in Giné and Koltchinskii (2006), see also (Koltchinskii, 2011), show that the following estimate is reasonable for "nice" enough models,

$$K_1\sigma_C\sqrt{\frac{V_m}{n}} \leq \mathbb{E}\left[\sup_{s\in m_C}\{(P_n-P)(\gamma(s_m)-\gamma(s))\}\right] \leq K_2\sigma_C\sqrt{\frac{V_m}{n}} ,$$

where $\sigma_C^2 = \sup_{s\in m_C}\operatorname{Var}\{\gamma(s)\text{-}\gamma(s_m)\}$. In some cases however, some extra log factor would be necessary (Massart and Nédélec, 2006). Then by abusively taking the latter equivalence for an equality ($K_1 = K_2 = K$) and considering that the margin relation saturates, $\sigma_C^2 \sim LC^\alpha$, easy calculations give that, with probability close to one (controlled by Talagrand's type concentration inequalities for suprema of bounded empirical processes),

$$p_1(m) \sim \left(K\sqrt{L}\frac{\alpha}{2}\right)^{\frac{2}{2-\alpha}}\left(\frac{V_m}{n}\right)^{\frac{1}{2-\alpha}} \text{ and } p_2(m) \sim \left(\frac{2}{\alpha}-1\right)p_1(m) .$$

The identity $p_1(m) = p_2(m)$ then arises if and only if the exponent in the margin relation is equal to 1. Note that actually, the particular structure of the binary contrast is not essential and the rationale should work for the selection of some "nice" enough models - and nice enough sample distribution - in a general bounded M-estimation setting.

## 3. Remarks about the over-penalization effect

Theoretical validation of the slope heuristics is based on estimates of the excess risks that are optimal to first order. As emphasized in Section 8.4 of the survey, second order effects also play a significant role in the behavior of estimator selection rules in the moderate sample size regime, as they influence what is known as the over-penalization effect.

I agree on the fact that the slope heuristics has a tendency to avoid over-fitting (from its very definition actually!) and this might be related to some kind of over-penalization. I also pointed this phenomenon in a paper about the slope heuristics in MLE, Saumard (2010, page 2), emphasizing the superiority of the slope heuristics over AIC for small to moderate sample sizes.

However, I would not be too optimistic on the ability of the dimension jump - or slope estimation - algorithm to produce a satisfying, close to optimal, over-penalization. Indeed, in my opinion, avoiding over-fitting does not necessary lead to a sharp over-penalization. As explained in Saumard and Navarro (2018, Section 2.3), optimal over-penalization is in fact related to a precise multiple pseudo-testing problem on a collection of "random hypotheses". So there may exist procedures that avoid over-fitting, even with small sample sizes, but that are still sub-optimal for the latter "multiple hypotheses" problem.

This objection is actually inspired by some simulations that we carried out with Fabien Navarro for the paper (Saumard and Navarro, 2018) - even if we decided not to report them for simplicity of exposition (but we may add them in a further revision!). In our density estimation setting, we observe a superiority of the slope heuristics over AIC for small sample sizes, but the performances of both algorithms are still quite poor in this regime compared to our over-penalization strategy.

Finally, I agree that proposing a completely data-driven over-penalization procedure with nearly optimal performances for small to moderate sample sizes is quite a challenge. So is the theoretical validation of the superiority of over-penalization over classical approaches.

**Aknowledgements:** I am grateful to Lionel Truquet for instructive discussions about times series models and especially about their cross-validation. I also warmly thank Sylvain Arlot, Camille Saumard and Lionel Truquet for having carefully revised this note.

## References

Baraud, Y., Comte, F., and Viennet, G. (2001). Model selection for (auto-)regression with dependent data. *ESAIM Probab. Statist.*, 5:33–49.

Comte, F., Dedecker, J., and Taupin, M. L. (2008). Adaptive density estimation for general ARCH models. *Econometric Theory*, 24(6):1628–1662.

Comte, F. and Genon-Catalot, V. (2006). Penalized projection estimator for volatility density. *Scand. J. Statist.*, 33(4):875–893.

Comte, F., Lacour, C., and Rozenholc, Y. (2010). Adaptive estimation of the dynamics of a discrete time stochastic volatility model. *J. Econometrics*, 154(1):59–73.

Comte, F. and Rozenholc, Y. (2002). Adaptive estimation of mean and volatility functions in (auto-)regressive models. *Stochastic Process. Appl.*, 97(1):111–145.

Gassiat, E. and van Handel, R. (2013). Consistent order estimation and minimal penalties. *IEEE Trans. Inform. Theory*, 59(2):1115–1128.

Gassiat, E. and van Handel, R. (2014). The local geometry of finite mixtures. *Trans. Amer. Math. Soc.*, 366(2):1047–1072.

Giné, E. and Koltchinskii, V. (2006). Concentration inequalities and asymptotic results for ratio type empirical processes. *Ann.Probab.*, 33:1143–1216.

Heinrich, P. and Kahn, J. (2018). Strong identifiability and optimal minimax rates for finite mixture estimation. *Ann. Statist.*, 46(6A):2844–2870.

Koltchinskii, V. (2011). *Oracle inequalities in empirical risk minimization and sparse recovery problems*, volume 2033 of *Lecture Notes in Mathematics*. Springer, Heidelberg. Lectures from the 38th Probability Summer School held in Saint-Flour, 2008, École d'Été de Probabilités de Saint-Flour. [Saint-Flour Probability Summer School].

Massart, P. and Nédélec, E. (2006). Risks bounds for statistical learning. *Ann.Stat.*, 34(5):2326–2366.

Navarro, F. and Saumard, A. (2017). Slope heuristics and $V$-fold model selection in heteroscedastic regression using strongly localized bases. *ESAIM Probab. Stat.*, 21:412–451.

Saumard, A. (2010). Nonasymptotic quasi-optimality of AIC and the slope heuristics in maximum likelihood estimation of density using histogram models. hal-00512310.

Saumard, A. and Navarro, F. (2018). Finite sample improvement of Akaike's Information Criterion. *arXiv preprint arXiv:1803.02078*.