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

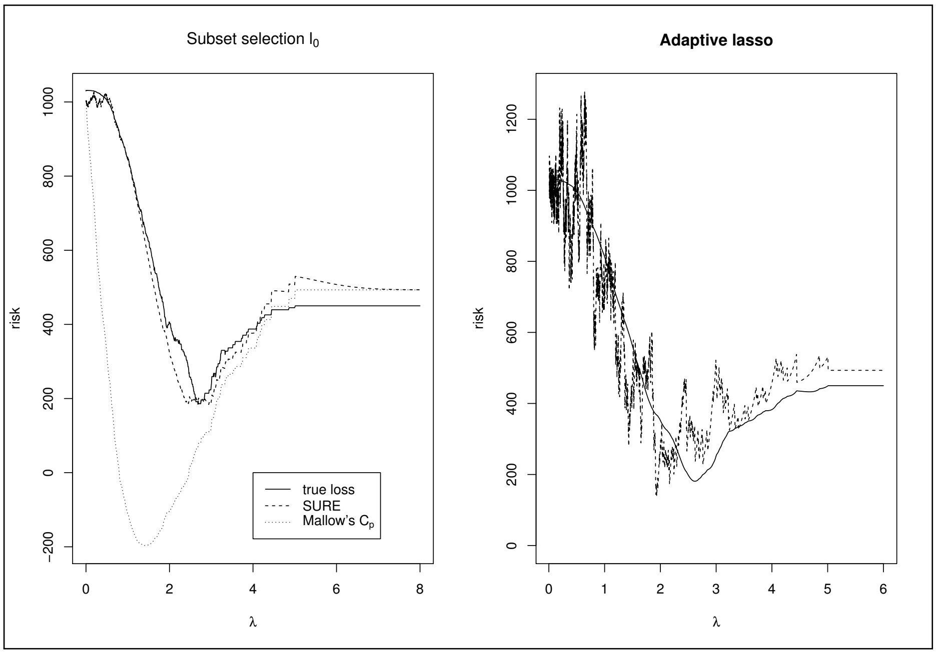
Discussion of “Minimal penalties and the slope heuristics: a survey” by Sylvain Arlot

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

Sylvain Sardy[1]

Most machine learning methods require the selection of a regularization parameter that controls the complexity and the fit of the estimated model. The learner considered here is a sequence of projections into a collection of linear subspaces $(S_m)_{m\in\mathscr{M}}$, the regularization penalizes the least squares by $C\dim(S_m)$, and the goodness-of-fit measure is the predictive risk. The *optimal* penalty is seen as the constant C that unbiasedly estimates the predictive risk. Sylvain Arlot makes a thorough theoretical and empirical survey and provides great insights of the minimal penalty and the slope heuristics that circumvent the difficult problem of estimating the noise variance σ^2.

FIGURE 1. *Example of risk estimation compared to true loss.*



Regularization methods know that bias is good, yet they often paradoxically seek an optimal

[1] Section de Mathématiques, Université de Genève

model by minimizing an *unbiased* estimate of the risk. We recommend biased estimation of the risk towards models of lower complexity. We illustrate our point with orthonormal regression $\mathbf{Y} = \mu + \varepsilon$, where μ is believed to be sparse. We consider two estimators.

A typical projection estimator is subset selection with C_p^n models of size $p \in \{0, 1, \ldots, n\}$ and a total of $|\mathscr{M}| = 2^n$ models $(S_m)_{m \in \mathscr{M}}$. Conditional on $p = \dim(\hat{S}_m)$ the best model $\hat{S}_m$ is the support of $\hat{\mu}_\varphi = \eta_\varphi^{\text{hard}}(\mathbf{Y})$ with threshold $\varphi = \sqrt{C}$ and $C = Y_{(n-p)}^2$. In that case the unbiased risk estimate formula based on Stein (1981) and Sardy (2009, Equation (12)) takes into account that the optimal model $\hat{S}_m$ is estimated. On the contrary Mallow's C_p of (9) which unbiasedness property is conditional on each S_m of size p underestimates the variance and consequently selects an over-complex model. The left plot of Figure 1 illustrates this behavior. The factor 2 in (9) is too small, which concurs with Birgé and Massart (2007).

Adaptive lasso (Zou, 2006) indexed by (λ, ν) includes best subset selection at its limit when $\nu \to \infty$. For adaptive lasso, the right plot of Figure 1 shows that, for a fixed large $\nu = 20$, the unbiased estimate of the risk (Sardy, 2012) as a function of λ has high variance on the left side of the minimum of the true loss which itself has a high negative derivative. Biasing towards smaller complexity (i.e., larger λ) would lead to an estimator with smaller risk.

These two examples suggest a slope method with a larger constant than 2. An even larger constant should be employed when the design is not fixed, the regression matrix is badly condition (ill-posed inverse problems) and σ^2 is unknown. BIC (Schwarz, 1978) and Quantile universal threshold (QUT) (Giacobino et al., 2017) lead to low complexity models. QUT is also a good competitor of the scree test to recover the number of components in principal component analysis (Josse and Sardy, 2016).

References

Birgé, L. and Massart, P. (2007). Minimal penalties for gaussian model selection. *Probability theory and related fields*, 138(1-2):33–73.

Giacobino, C., Sardy, S., Diaz Rodriguez, J., and Hengardner, N. (2017). Quantile universal threshold. *Electronic Journal of Statistics*, 11(2):4701–4722.

Josse, J. and Sardy, S. (2016). Adaptive shrinkage of singular values. *Statistics and Computing*, 26(3):715–724.

Sardy, S. (2009). Adaptive posterior mode estimation of a sparse sequence for model selection. *Scandinavian Journal of Statistics*, 36:577–601.

Sardy, S. (2012). Smooth blockwise iterative thresholding: a smooth fixed point estimator based on the likelihood's block gradient. *Journal of the American Statistical Association*, 107(498):800–813.

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, 6(2):461–464.

Stein, C. M. (1981). Estimation of the mean of a multivariate normal distribution. *The Annals of Statistics*, 9(6):1135–1151.

Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American Statistical Association*, 101(476):1418–1429.