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

On the significance of a linear relationship in density estimation with mixture models

Titre: Sur la portée d'une relation linéaire pour l'estimation de densité par le modèle de mélange

Jean-Patrick Baudry[1]

The author claims in this interesting and important survey that "minimal-penalty algorithms indeed work" for the task of "estimating the number of clusters for (unsupervised) clustering", particularly with the model-based clustering approach (Section 8.3.2). Indeed, partial theoretical results and numerical experiments support this claim.

However, ongoing research with Gilles Celeux on this very topic suggests that the observed (almost) linear relationship between the empirical risk and the number of parameters when the number of Gaussian components is overestimated can be explained by the fact that any extra component fitted has to be a degenerate component. Softwares prevent pure degeneracy by imposing more or less explicitly a minimal variance for the Gaussian components. Interestingly, numerical experiments confirm that the observed slope can be "chosen" by tuning this minimal variance.

The good point for the slope heuristics in this situation is that it can be shown that it should still perform well for a much wider spectrum of minimal variances than the BIC criterion, which underpenalizes when the minimal variance is too small. This illustrates the interest of the data-driven quality of the slope heuristics.

However the slope heuristics can also get into trouble: it can overpenalize when the minimal variance is set very small.

Ongoing work focuses on understanding how this phenomenon really relates to the usual slope heuristics phenomenon... or whether and to what extent the observed linear relationship actually reflects in this case something different from the slope heuristics assumptions (see Conclusion of Section 7.1 of the survey).

This would be an illustration that unsurprisingly, if the user should be worried about applying the slope heuristics when no linear relationship is observed (Conclusion of Section 7.1 again), they should not blindly be reassured on the possibility of applying the slope heuristics when a linear relationship is actually observed.

[1] Laboratoire de Probabilités, Statistique et Modélisation
Sorbonne Université