Averaging via stacking in model-based clustering

Advanced Statistics for Physics Discovery

Alessandro Casa[1]

Luca Scrucca[2] and Giovanna Menardi[1]

Università degli Studi di Padova[1]
Università degli Studi di Perugia[2]

casa@stat.unipd.it

24th September 2018

Framework

- **Model-based clustering**, data come from a finite mixture of K components (corresponding to the groups):

$$f(x|\Theta) = \sum_{k=1}^{K} \pi_k f_k(x|\theta_k) \ ;$$

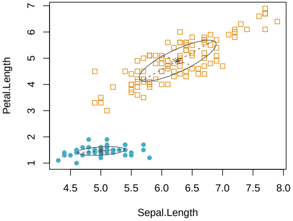
- Model selection is a crucial step in this framework involving the choices of:
    - Number of clusters;
    - Parametrization of component covariance matrices;
    - Component densities.

Single best model paradigm

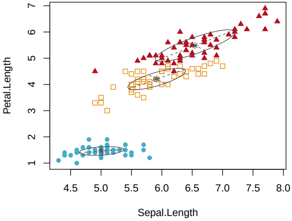
The best model among the fitted ones is chosen, according to information criteria (e.g. BIC, ICL) and used for subsequent steps.

Problem

- What if discarded models have IC values close to the one of the selected model?
- Example: `Iris data`



VEV2, BIC=-561.72



VEV3, BIC=-562.55

- Model selection-related uncertainty is neglected, possibly useful models are thrown away.

Proposal

- **Idea**: average densities of fitted models to improve robustness and stability of clustering solutions;
- Resulting estimate is a convex linear combination of a subset of fitted models

$$f_{av}(x) = \sum_{m=1}^{M} \alpha_m f_m(x|\hat{\Theta}_m) \;;$$

- **Issues**:
  - *Weights*
    $f_{av}(\cdot)$ is still a mixture model $\longrightarrow$ α_m estimated via EM, maximizing a BIC-penalized log-likelihood;
  - *Partitions*
    correspondence components-clusters is lost $\longrightarrow$ explore modality of $f_{av}(\cdot)$ via mean-shift algorithm.