MBC^2 - Lightning Talk Session
September 5–7, 2018

Averaging via stacking in model-based clustering

Alessandro Casa Luca Scrucca Giovanna Menardi

Università di Padova
casa@stat.unipd.it

Università di Perugia
luca.scrucca@unipg.it

Università di Padova
menardi@stat.unipd.it

Framework

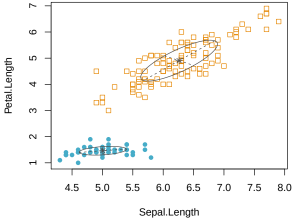
- Model selection is a crucial step in the framework of model-based clustering;
- It involves the choices of:
  - Number of clusters;
  - Parametrization of component covariance matrices;
  - Component densities.

Single best model paradigm

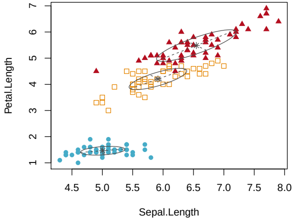
The best model among the fitted ones is chosen, according to information criteria (e.g. BIC, ICL) and used for subsequent steps.

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