Averaging via stacking in model-based clustering

Alessandro Casa[1], Luca Scrucca[2], Giovanna Menardi[1]
casa@stat.unipd.it, luca.scrucca@unipg.it, menardi@stat.unipd.it

[1] Dipartimento di Scienze Statistiche - Università di Padova
[2] Dipartimento di Economia - Università di Perugia

Introduction

- **Density-based clustering**: formalization of the clustering problem by assuming a probability density function underlying data generation. Two different approaches:
  - Parametric: one-to-one correspondence between clusters and unimodal components of a finite mixture model;
  - Nonparametric: clusters as the domains of attraction of the density modes.
- Regardless of the chosen paradigm the first step consists in obtaining an estimate of the density.
- Estimation in the model-based clustering framework is carried out assuming, as data generating mechanism, a finite mixture model

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

$\Theta = \{\pi_1, \ldots, \pi_{K-1}, \theta_1, \ldots, \theta_k\}$, $\pi_k > 0$, $\forall k$ and $\sum_{k=1}^{K} \pi_k = 1$.

- Often $f_k(\cdot) = \phi_k(\cdot)$ so $\theta_k = \{\mu_k, \Sigma_k\}$ and parsimony induced via eigendecomposition of Σ_k.

Model selection in model-based clustering

- Model selection involves several different choices:
  - Number of groups (via number of components K of the mixture);
  - Parametrization of the component covariance matrices Σ_k;
  - Specification of component densities $f_k(\cdot)$.

  Every combination can be seen as a separate model in the set on which selection occurs.

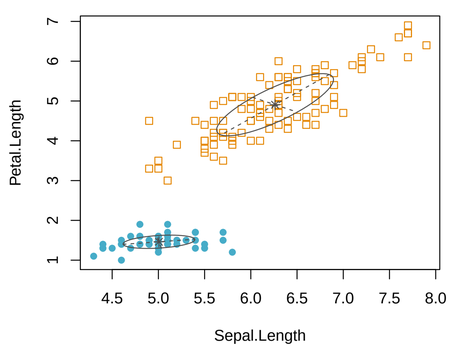
Single best model paradigm

Models corresponding to each combination are estimated and the best one among them is selected according to an information criterion (IC) e.g. BIC, ICL, and then used to obtain a partition.

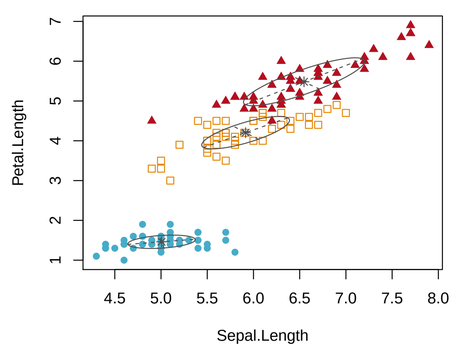
⌄

Throwing away all other models could be sub-optimal, leading to loss of useful information.
What if discarded models have IC values close to that of the selected model?

- Example: `Iris data`



`VEV2, BIC=-561.72`



`VEV3, BIC=-562.55`

Literature, solutions and surroundings

- **Possible workaround**: average a suitably chosen subset of estimated models.
- Contact points with the rationale behind *consensus clustering* approaches where several clusterings are combined:
  - Improve quality, robustness and stability;
  - Multiple views of the data.
- Two different approaches in model-based clustering based on *Bayesian Model Averaging*:
  - Wei and McNicholas (2015): average a posteriori probabilities or parameter estimates;
  - Russell et al. (2015): average over similarity matrices.

Work proposal - 1

Main issue

The meaning of the quantity to average on need to be consistent across the estimated models

⌄

We tackle the problem at its roots and choose, as a quantity to average, the density itself.

- We adapt the idea on which *stacking* is based on in an unsupervised framework (Smyth and Wolpert, 1999).
  - Stacking: way of combining base models by building a new one stacked on the top of them, learning from their predictions.
- The resulting density estimate is a convex linear combination of a subset of the fitted models

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

where $f_m(\cdot)$ are the mixture models to average, M their number and α_m the corrisponding weights.

- Critical points:
  - How do we estimate the weights?
  - How do we operationally obtain a partitions?

Work proposal - 2

- **Weights**:
  - $f_{av}(\cdot)$ is still a mixture model so α_m $(m = 1, \ldots, M)$ are estimated by maximizing the log-likelihood via EM algorithm.
  - To avoid *overfitting* we consider a *BIC-type* penalization and obtain $\hat{\alpha}_m$ by maximizing the penalized log-likelihood

$$l_p(\alpha|x) = \sum_{i=1}^{n} \log \sum_{m=1}^{M} \alpha_m f_m(x_i) - log(n) \sum_{m=1}^{M} \alpha_m \nu_m \ ,$$

  where ν_m is the number of parameters for mth model.
- **Partitions**:
  - Loss of correspondence between groups and mixture components.
  - According with nonparametric clustering formulation we explore the modality of $\hat{f}_{av}(\cdot)$ via *mean-shift* algorithm.

Some results

`Wines data`
M=5, $G_{true} = 3$

	best	pen_av
Adj Rand Index	0.830	0.964
Num groups	3	3

`DLBCL data`
M=126, $G_{true} = 5$

	best	pen_av
Adj Rand Index	0.296	0.909
Num groups	7	4

`Iris data`
M=2, $G_{true} = 3$

	best	pen_av
Adj Rand Index	0.568	0.568
Num groups	2	2

Future work

- Are there other possible penalization scheme worth considering?
- How do we choose M?
  - Occam's window built on BIC values of the fitted models;
  - Incorporate selection of M in the estimation phase with sparsity inducing penalization.

References

- Chacón, J.E. (2018). *Mixture model modal clustering*, *Advances in Data Analysis and Classification*, 1–26.
- Russell, N., Murphy, T.B. and Raftery, A.E. (2015). Bayesian model averaging in model-based clustering and density estimation, *arXiv preprint arXiv:1506.09035*.
- Smyth, P. and Wolpert, D. (1999). Linearly combining density estimators via stacking, *Machine Learning*, **36**, 59–83.
- Wei, Y. and McNicholas, P.D. (2015). Mixture model averaging for clustering, *Advances in Data Analysis and Classification*, **9**(2), 197–217.