On the selection uncertainty in parametric clustering

European Conference on Data Analysis

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

5th July 2018

Selection uncertainty and statistical scandals

- Model selection is ubiquitous in modern statistical analysis and applications;
- Selection preceeds inference and these steps are considered as separated $\longrightarrow$ selected model treated as fixed;

Model selection is data-dependent $\longrightarrow$ *we are neglecting a source of uncertainty possibly ending up with anti-conservative statements*

- Possible workarounds:
  - Data splitting;
  - Model averaging estimators;
  - Use a corrected estimators.

Aim and contribution

- **Density-based clustering**: definition of cluster linked to features of the density underlying the data:
  - Parametric: clusters as unimodal components within an appropriate finite mixture model;
  - Nonparametric: clusters as domains of attraction of the density modes.
- Model selection tools required to choose among different models for the true density function;
- **Aim**: propose a model averaging approach accounting for the selection step in model-based clustering.

Model-based clustering

- Data comes from a finite mixture of K components (corresponding to the groups):

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

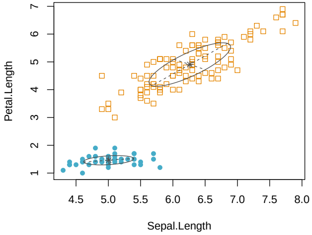
  - $\Theta = \{\pi_1, \dots, \pi_{K-1}, \theta_1, \dots, \theta_k\}$ with $\pi_k > 0$ and $\sum_{k=1}^{K} \pi_k = 1$;
  - Often $f_k(\cdot) = \phi_k(\cdot)$ hence $\theta_k = \{\mu_k, \Sigma_k\}$;
  - Parsimony is induced by considering eigenvalue decomposition $\Sigma_k = \lambda_k A_k D_k A_k^T$;
- Selection step in model-based clustering involves choices of:
  - Number of clusters (through number of components);
  - Parametrization of the component covariance matrices Σ_k;
  - Component densities.

Selection in model-based clustering

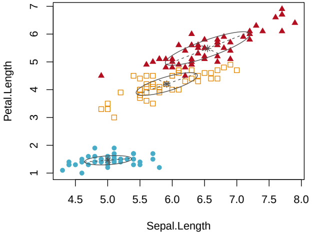
- **Single-best model paradigm**
  Several models are fitted $\longrightarrow$ best one then chosen according to information criteria (e.g. BIC, ICL) and used to obtain a partition;

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 averaging in model-based clustering

- Model averaging accounts for model uncertainty combining parameter estimates across a set of competing models;
- **Problem**: the quantity to average should have the same meaning in each estimated model;
- Two different *Bayesian Model Averaging* (BMA) approaches in model-based clustering literature:
  - Wei & McNicholas (2015) average *a posteriori* probabilities or parameter estimates $\longrightarrow$ same number of components is needed;
  - Russell et al. (2015) average over similarity matrices $\longrightarrow$ need to resort to distance-based algorithm to obtain a partition.

Work proposal

- **Idea**: choose as quantity to be averaged the density itself tackling the problem at its roots;
- 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 mixture models to average, M is their number and α_m the corresponding weights;
- Criticalities:
  - How to estimate the weights;
  - How to operationally obtain a partition.

Choosing weights

- $f_{av}(\cdot)$ is a mixture model itself so $\alpha_m, m = 1, \dots, M$ can be estimated maximizing the log-likelihood via EM algorithm;
- Overfitting issue: complex models with larger number of components weight more in the combination;
- **Proposed solution**: consider a *BIC-type* penalization and obtain $\hat{\alpha} = \{\hat{\alpha}_1, \dots, \hat{\alpha}_m\}$ 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 and $\{x_i\}_{i=1}^{n}$ is the sample.

Obtaining partition

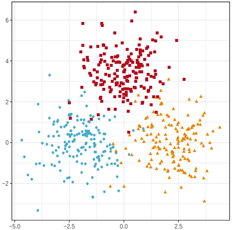
- Averaging process implies the loss of the correspondence between components and clusters
  $\longrightarrow$ final partition cannot be obtain in the usual way;
- Clusters are obtained as domain of attractions of the modes of the fitted density in a nonparametric fashion;
- Use of gradient ascent algorithm to explore modality of $\hat{f}_{av}(\cdot)$: *mean shift* specifically adjusted for mixture densities (Chacon, 2018).

Results - Simulated data

Three component Gaussian mixture

	best	pen_av
NWS, $n = 100$	0.8755	0.8664
NWS, $n = 500$	0.8796	0.8797
WS, $n = 100$	0.9918	0.9924
WS, $n = 500$	0.9909	0.9910

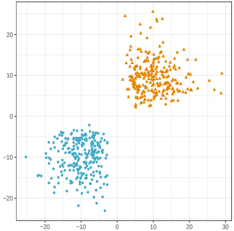
Mean Adjusted Rand Index



Skewed components mixture

	best	pen_av
$n = 100$	0.9548	0.9805
$n = 500$	0.6253	0.9936
$n = 1000$	0.5579	0.9821
$n = 5000$	0.3175	0.9547

Mean Adjusted Rand index



Results - Real data

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

Open issues and future work

- We introduce a viable and flexible alternative to BMA approaches in order to overcome single best model limitations in model-based clustering framework;

- **Open questions**:
  - Should we consider other penalitazion schemes (e.g. inspired by other IC)?
  - How do we choose M?
    - Occam's window built on BIC values of fitted models;
    - Alternatives to EM algorithm in order to incorporate selection of M in the estimation process;

Relevant 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.