Better than the best? Blending models and approaches in density-based clustering

Working Group on Statistical Learning

Alessandro Casa

Joint work with: Luca Scrucca & Giovanna Menardi

School of Mathematics and Statistics
University College Dublin

alessandro.casa@ucd.ie

3rd February 2020

Mixture modelling and clustering

- Model-based clustering offers a statistically sound formalization of the clustering problem
- Data are assumed to come from a finite mixture of K components

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

  - $\Theta = (\pi_1, \ldots, \pi_{K-1}, \theta_1, \ldots, \theta_K)$ with $\pi_k > 0,\ \forall k = 1, \ldots, K$ and $\sum_k \pi_k = 1$
  - Often $\varphi_k(\cdot) = \phi_k(\cdot)$ than $\theta_k = \{\mu_k, \Sigma_k\}$ with parsimony induced by eigen-decomposition $\Sigma_k = \lambda_k A_k D_k A_k^T$

KEY IDEA

One-to-one correspondence between clusters and components of the mixture

Model Selection in MBC

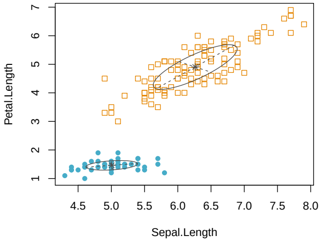
- Model selection step is essential in order to choose a model giving a good density estimate and partition of the data. Possible choices required are:
  - Number of clusters K
  - Parametrizations of Σ_k
  - Specification for component densities
- **Single-best model paradigm**
  Several models are fitted $\longrightarrow$ best one is chosen according to information criteria and used to obtain a partition

Is this the best thing we can do?

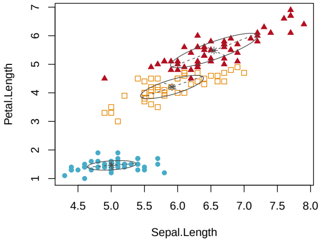
We are throwing away possibly useful models (i.e. info) and neglecting selection-related uncertainty

A new and unexpected example

- As a motivation think about the `Iris data`



`VEV2, BIC=-561.72`



`VEV3, BIC=-562.55`

- **Our aim**: propose a model averaging approach in the model-based clustering framework
    - Improve stability and robustness
    - Account for the selection step
    - More informative partitions

Model averaging in MBC

- Averaging approaches have been scarcely pursued in an unsupervised framework with respect to supervised counterpart
- **Main challenge**: an *invariant quantity*, having the same meaning across models, to average on is needed
  $\longrightarrow$ not straightforward in model-based clustering
- Relevant works in parametric clustering based on BMA:
  - Wei & McNicholas (2015) average *a posteriori* probabilities after a merging step
  - Russell *et al.* (2015) average similarity matrices obtaining partitions via hierarchical clustering

Work proposal

- **Idea**: recast the problem as a density estimation one
  $\longrightarrow$ the density is chosen as the invariant quantity
- New density estimate is a convex linear combination of a subset of the fitted models

$$\tilde{f}(x;\alpha) = \sum_{m=1}^{M} \alpha_m f_m(x|\hat{\Theta})$$

  where $f_m(\cdot)$ are the models to average, M is their number and α_m the corresponding weights
- Criticalities:
  - How to estimate the weights
  - How to choose M
  - How to operationally obtain a partition

Weights estimation

- $\tilde{f}(\cdot)$ is a mixture itself so $\alpha_m, m = 1, \ldots, M$ are estimated maximizing the log-likelihood via EM algorithm
- Overfitting issue: complex models with larger number of components will weight more in the combination
- Proposed solution: obtain $\hat{\alpha} = \{\hat{\alpha}_1, \ldots, \hat{\alpha}_M\}$ by maximizing a penalized log-likelihood defined as

$$\ell_p(\alpha|\mathcal{X}) = \sum_{i=1}^{n} \log \sum_{m=1}^{M} \alpha_m f_m(x_i) - \lambda \sum_{m=1}^{M} \alpha_m \nu_m$$

with $\mathcal{X} = \{x_i\}_{1 \leq i \leq n}$ the observed sample with $x_i \in \mathbb{R}^p$ and ν_m the number of parameters for the mth model

How do we select λ ?

Penalization strength

- The strength of the penalization plays a key role in choosing which model will have a role in the ensamble
- Different strategies have been explored
  CV-based:
  - split iteratively the dataset $\mathcal{X}$ in test and training sets
  - for λ's in a reasonable grid compute $\tilde{f}(x_{\text{test}}|x_{\text{train}})$
  - compute the *test log-likelihood*

$$\ell_{\text{test}} = \sum_{x \in \mathcal{X}_{\text{test}}} \log \tilde{f}(x|x_{\text{train}})$$

  then selct $\lambda_{\text{CV}} = \arg\max\ \ell_{\text{test}}(\lambda)$

  IC-based: BIC and AIC-type pe nalizations respectively lead to $\lambda_{\text{BIC}} = \log n/2$ and $\lambda_{\text{AIC}} = 1$

Choosing ensamble size

- A huge number of model could have been estimated
  $\longrightarrow$ choosing the ones entering in the ensamble could have a strong impact
- Possible strategies:
  - Subjective selection based on some prior knowledge
  - Build an *Occam's window* based on some quantity evaluating the goodness of the fitted models, e.g. using differences in BIC values
  - Set a large M and let the penalization to do the job for us

Small intermezzo - Density-based

- Model-based approach is widely known but what about **density based clustering**?
  $\longrightarrow$ generally speaking provides partitions by linking the concept of cluster to features of the density underlying the data
- Developed following two different (diverging?) roads:
  - Model-based approach (or parametric)
  - **Modal approach** (or nonparametric)

    clusters correspond to the domains of attraction of the modes of the density.
    Operationally requires:
    - a density estimate
    - a method to locate the modal regions of the density

Small intermezzo - Modal clustering

- **Kernel density estimator (KDE)** is usually considered to obtain a nonparametric density estimate

$$\hat{f}_H(x) = \frac{1}{nh_1 \cdots h_p} \sum_{i=1}^{n} \left\{ \prod_{j=1}^{p} K\left(\frac{x_j - x_{ij}}{h_j} \right) \right\}$$

  - $H = \mathrm{diag}(h_1, \ldots, h_p)$ is the bandwidth matrix
  - $K(\cdot)$ is the kernel function
- How to obtain partitions?
  - Mode-searching algorithm
  - Level-set methods

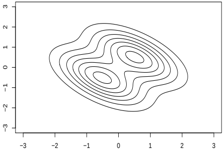
Obtaining partitions

- The obtained estimate $\tilde{f}(\cdot;\hat{\alpha})$ has a mixture structure but the correspondence clusters-components is lost
- **Idea**: blend together the two formulations of density-based clustering by shifting the concept of cluster itself
  $\longrightarrow$ partitions are obtained drawing the correspondence between groups and modal regions
- Use of gradient ascent algorithm to explore modality of $\tilde{f}(\cdot;\hat{\alpha})$

Modal EM algorithm: EM-like algorithm, exploiting the mixture structure, searching for local maxima of the density

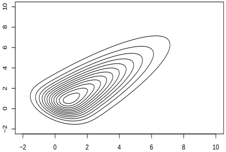
Some results - Syntethic data

Bimodal



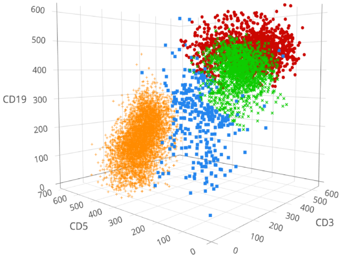
	n=500		n=5000	
	MISE	ARI	MISE	ARI
SB	0.666 (0.714)	0.677 (0.125)	0.057 (0.035)	0.680 (0.067)
SB-NP	-	0.690 (0.119)	-	0.720 (0.012)
λ_{AIC}	0.809 (0.435)	0.687 (0.063)	0.072 (0.044)	0.719 (0.013)
λ_{BIC}	0.714 (0.522)	0.683 (0.129)	0.057 (0.035)	0.720 (0.012)
λ_{CV}	0.687 (0.402)	0.688 (0.064)	0.058 (0.036)	0.720 (0.012)

Skewed unimodal



	n=500		n=5000	
	MISE	ARI	MISE	ARI
SB	0.626 (0.235)	0.000 (0.000)	0.087 (0.032)	0.000 (0.000)
SB-NP	-	0.520 (0.501)	-	0.725 (0.448)
λ_{AIC}	0.441 (0.181)	0.440 (0.498)	0.059 (0.024)	0.915 (0.280)
λ_{BIC}	0.438 (0.167)	0.705 (0.457)	0.074 (0.030)	0.965 (0.184)
λ_{CV}	0.436 (0.172)	0.440 (0.498)	0.059 (0.026)	0.850 (0.358)

Some results - Real data



	SB	SB-NP	λ_{AIC}	λ_{BIC}	λ_{CV}
ARI	0.401	0.867	0.909	0.910	0.912
$\hat{K}$	7	4	4	4	4

DLBCL data

- $d = 8$ chemical variables
  $n = 572$ olive oils
- $\hat{K}_{\text{true}} = 9$ regions of Italy
- Hierarchical structure in the data

	SB	SB-NP	λ_{AIC}	λ_{BIC}	λ_{CV}
ARI	0.782	0.792	0.902	0.892	0.902
$\hat{K}$	6	6	8	8	8

Olive Oil data

Remarks, directions and open questions

- **Post-selection inference** usually ignored in clustering
  $\longrightarrow$ this approach may be inferentially more appropriate
- Proposed as a solution to single-best model
  $\longrightarrow$ what about bootstrap replications or **initialization** issues?
- Bayesian approaches with **shrinkage priors** on the ensamble weights may be useful to select ensemble size
- **Penalization** scheme is still density estimation oriented
  $\longrightarrow$ choosing λ in a clustering-oriented fashion
- Some strong contact points with **merging** approaches and with deep Gaussian mixture models

Remarks, directions and open questions

Modal and model-based clustering are two sides of the density-based coin

- Two perspectives:
  - Solution to the single-best model paradigm in the parametric framework
  - Solution to the density estimation issues in the nonparametric framework
- Density estimator lies in the semi-parametric realm
- Blending them together we reduce their weaknesses while mixing their strengths

Some references

Check the paper out on arXiv
https://arxiv.org/pdf/1911.06726.pdf

Other references

- Chacòn, J.E. (2019) *Mixture model modal clustering*, Advances in Data Analysis and Classification, 13(2), 379–404.
- Russell, N., Murphy, T.B. & Raftery, A.E. (2015) *Bayesian model averaging in model-based clustering and density estimation*. arXiv preprint arXiv:1506:09035.
- Scrucca, L. (2016) *Identifying connected components in Gaussian finite mixture models for clustering*. Comoputational Statistics and Data Analysis, 93, 5–17.
- Smyth, P. & Wolpert, D. (1999) *Linearly combining density estimators via stacking*. Machine Learning, 36, 59–83
- Wei, Y. & McNicholas, P.D. (2015) *Mixture model averaging for clustering*. Advances in Data Analysis and Classification, 9(2), 197–217.