Model-based Statistical Learning

Pr. Charles BOUVEYRON

Professor of Statistics
Chair of the Institut 3IA Côte d'Azur
Université Côte d'Azur & Inria

charles.bouveyron@univ-cotedazur.fr
@cbouveyron

Parsimonious models for GMM

In many situations, it may be useful to consider more constrained models:

Full GMM : $\theta = \{\pi_k, \mu_k, \Sigma_k\}$

Com GMM : $\theta = \{\pi_k, \mu_k, \Sigma\}$

diag GMM : $\theta = \{\pi_k, \mu_k, \sigma_k^2 I_p\}$

iso GMM : $\theta = \{\pi_k, \mu_k, \sigma^2 I_p\}$

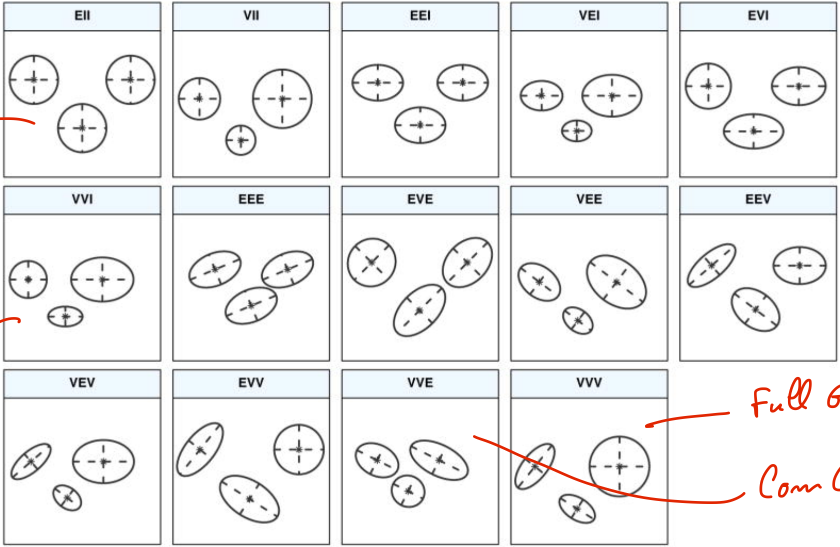
⋮

k-means : $\theta = \{\frac{1}{K}, \mu_k, \sigma^2 I_p\}$

Parsimonious models for GMM

k-means

diag GMM



+ the choice of K

Full GMM.

Com GMM.

Figure: The parsimonious models of Mclust.

How to choose between models?

In practice, when facing some data set, we need to pick the appropriate number K of components and the best GMM model to fit on data.

Model selection

The roots of model selection can be found in Bayesian statistical theory.

Let's first consider a set of models to test:

$\{\mathcal{M}_1, \mathcal{M}_2, \dots, \mathcal{M}_G\}$ and associated to prior probabilities $p(\mathcal{M}_g)$ (in practice we prefer to use $p(\mathcal{M}_g) = p \; \forall g$)

The idea of model selection is to evaluate a specific quantity: $p(\mathcal{M}_g | X)$.

Thanks to the Bayes theorem:

$$p(\mathcal{M}_g \mid X) \propto p(X \mid \mathcal{M}_g)\, p(\mathcal{M}_g) \qquad (\star).$$

To compute this quantity, we have to evaluate it for all possible combinations of the parameters of $\mathcal{M}_g$

$$p(X \mid \mathcal{M}_g) = \int_{\theta_g} p(X \mid \mathcal{M}_g, \theta_g) \cdot p(\theta_g \mid \mathcal{M}_g)\, d\theta_g .$$

Rmq: due to the high-dimensional integration over θ_g, this computation is very often intractable!

Note: The quantity $p(X \mid \mathcal{M}_g)$ is called the integrated likelihood or the evidence or the marginal likelihood.

In practice, we prefer to avoid the computation problem by approximating the integrated likelihood:

Then:

$$\mathcal{M}^* = \arg\max_{\mathcal{M}_g} \tilde{p}(X \mid \mathcal{M}_g).$$

Among the possible approximation of the integrated likelihood, the most popular one is BIC: Bayesian Information Criterion, Schwarz, 1978

$$\log p(X \mid \mathcal{M}_g) \simeq \underbrace{\log p(X \mid \mathcal{M}_g, \hat{\theta}_{g ML}) - \frac{\nu(\mathcal{M}_g)}{2} \log(n)}_{\text{BIC criterion.}}$$

where $\nu(\mathcal{M}_g)$ is the number of free parameters in the model $\mathcal{M}_g$, and n is the number of observations.

For information, BIC is both an asymptotic approximation of the marginal likelihood ($\text{BIC} \xrightarrow[n \to \infty]{} p(X \mid \mathcal{M}_g)$) and a second order Taylor expansion of the logarithm of the integrand, around its maximum $\hat{\theta}_g^{ML}$.

In practice:

$$\mathcal{M}^* = \underset{\mathcal{M}_g}{\arg\max}\ \text{BIC}(\mathcal{M}_g).$$

Rmk: it is worth mentioning that, unfortunately, the assumptions made about the regularity of the models in BIC are not satisfied for mixture models!

$\Rightarrow$ Fortunately, BIC is behaving very well for mixture models and even when $n < +\infty$.

what we do :

$\mathcal{M}_1$ = Full GMM with $K=2$ $\xrightarrow{EM}$ $\hat{\theta}_{1MC}$ $\xrightarrow{BIC}$ $BIC(\mathcal{M}_1)$

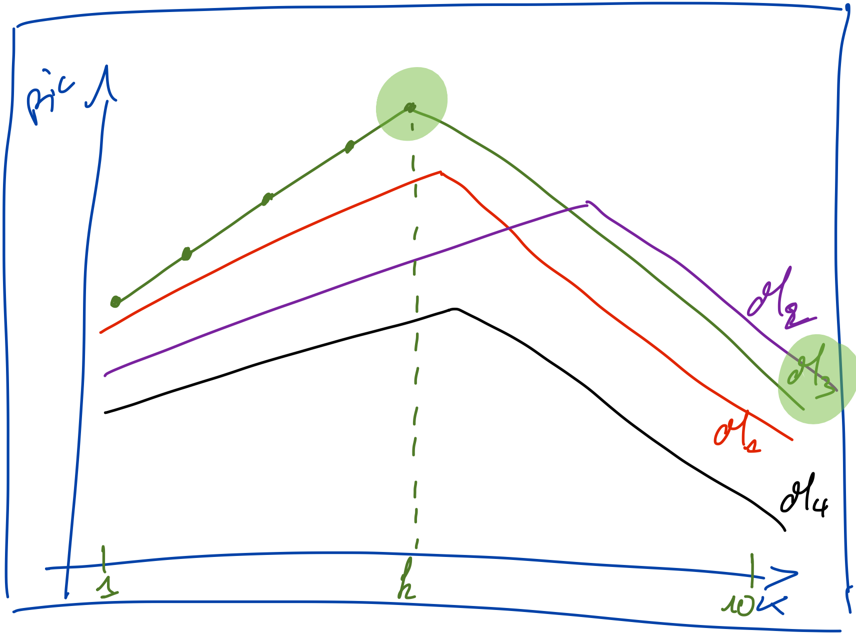
$\mathcal{M}_2$ = diag GMM — $K=2$ $\xrightarrow{EM}$ $\hat{\theta}_2$ $\longrightarrow$ $BIC(\mathcal{M}_2)$

$\mathcal{M}_3$ = Full GMM — $K=3$ $\xrightarrow{EM}$ $\hat{\theta}_3$ $\longrightarrow$ $BIC(\mathcal{M}_3)$

$\mathcal{M}_4$ = diag GMM — $K=3$ $\xrightarrow{EM}$ $\hat{\theta}_4$ $\longrightarrow$ $BIC(\mathcal{M}_4)$

and we choose $\mathcal{M}_3$ because $BIC(\mathcal{M}_3)$ is the largest one.

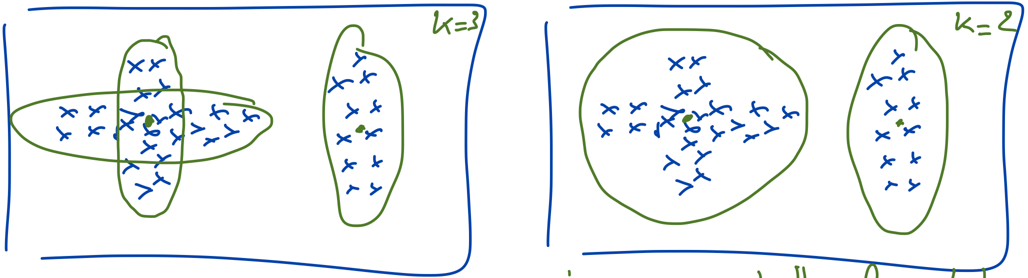
with Malust :



Rmk: BIC (and AIC which is a close criterion) can be used for model selection in all situations: regression, clustering, ...

Model selection for clustering:

The notion of mixture components and clusters may differ in some specific situations and it would be great to have a MS criterion that takes this into account.



is K=2 better for clustering?

A sound way to answer this it to come back to the model selection theory and consider the integrated complete likelihood instead of the integrated likelihood:

$$\mathcal{M}^* = \arg\max_{\mathcal{M}_g} p(X, Y \mid \mathcal{M}_g)$$

Then, it is possible to make the same approximations.

$$p(X, Y \mid \mathcal{M}_g) \simeq \log p(X \mid \hat{\theta}_g, \mathcal{M}_g) - \frac{\nu(\mathcal{M}_g)}{2} \log(n) - \frac{1}{2} \sum_{i=1}^{n} \sum_{k=1}^{K} t_{ik} \log(t_{ik}) = \underline{\underline{\mathrm{ICL}}}$$