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

The EM algorithm for GMM

Let's now consider the EM algorithm for GMM

E step: $E\left[p(X,Z)|\theta^*\right] = \sum_i \sum_k E\left[z_{ik}|x,\theta\right] \log\left(\pi_k \phi(x_i;\theta_k)\right)$

t_{ik}^*

through Bayes' theorem

$t_{ik} = \dfrac{\pi_k^* p(x_i;\theta_k^*)}{p(x)}$

M step: $\max_\theta \prod_{i=1}^{n} \sum_{k=1}^{K} t_{ik}^* \pi_k \phi(x_i, \mu_k, \Sigma_k)$

$= (\pi_k, \mu_k, \Sigma_k)_k$

The EM algorithm for GMM

E step: $t_{ik}^{*} = P[z|x;\theta^{*}) \overset{\text{Bayes rule}}{=} \frac{\pi_k^{*}\,\phi(x_i;\mu_k^{*},\Sigma_k^{*})}{p(x)} = \frac{\pi_k^{*}\,\phi(x_i;\mu_k^{*},\Sigma_k^{*})}{\sum_{k=1}^{K}\pi_k^{*}\,\phi(x_i;\mu_k^{*},\Sigma_k^{*})}$

M step:

$$\pi_k^{*(\text{new})} = \frac{\sum_{i=1}^{n} t_{ik}^{*}}{n}$$

$$\mu_k^{*(\text{new})} = \frac{\sum_{i=1}^{n} t_{ik}^{*}\, x_i}{\sum_{i=1}^{n} t_{ik}^{*}}$$

$$\Sigma_k^{*(\text{new})} = \frac{1}{n}\sum_{i=1}^{n} t_{ik}^{*}\left(x_i-\mu_k^{*}\right)^{t}\left(x_i-\mu_k^{*}\right)$$

Deeper in the GMM and EM algorithm:

Let's recall that the EM algorithm introduces a latent variable (Z) to encode the group memberships, and then maximizes iteratively the Expectation of the complete data likelihood (of the pair (X, Z)). - Start with a given value of θ^*

- E step: compute $E[\ell(X, Z|\theta)|\theta^*]$
- M step: maximize $E[$ ——— $]$ according to θ to get a new value for θ^*.

Let's first focus on the complete-data log-likelihood:

$$\ell(X, Z \mid \theta) = \sum_{i=1}^{n} \sum_{h=1}^{K} z_{ih} \log\left(\pi_h \, \phi(x_i \mid \theta_h)\right)$$

$$\log\left(L(X, Z \mid \theta)\right) = \prod_i \log \sum_h p(z_i \mid x_i, \theta) \cdot \phi(x_i, \theta_h)$$

Note: we can switch the log and the $\sum$ because for one ind. x_i it can belong to only one group. $\log \sum = \sum z_{ih} \log(\cdot)$

$$= \sum_i \sum_h z_{ih} \log\left(p(z_i \mid x_i, \theta) \times \phi(x_i; \theta_h)\right)$$

$$= \sum_i \sum_h z_{ih} \log\left(\pi_h \, \phi(x_i; \theta_h)\right)$$

Note: here Z is assumed to be a vector of binary values encoding the group membership:

$$z_i = (0,0,1,0) \Rightarrow x_i \text{ belongs to cluster 3}$$

$$E\left[\ell(X,Z|\theta)\,|\,\theta^*\right] = \sum_i \sum_k \underbrace{E\left[z_{ik}\,|\,x_i,\theta^*\right]}_{t_{ik}} \log\left(\pi_k \phi(x_i;\theta_k)\right)$$

$$t_{ik} = E\left[z_{ik}\,|\,x_i,\theta^*\right] = P\left(z_{ik}=1\,|\,x_i,\theta^*\right)$$

This quantity can be easily computed with a table of the distribution ϕ

$$\overset{\text{Bayes}}{=} \frac{P(z_{ik}=1\,|\,\theta^*) \times P(x_i\,|\,z_{ik}=1,\theta^*)}{P(x_i\,|\,\theta^*)} = \frac{\pi_k^* \phi(x_i;\theta_k^*)}{\quad}$$

In the M step, the expected complete data log likelihood is maximised over θ to get the new θ^{new}:

$$\max_{\theta} E\left[\ell(x,z|\theta)\,\middle|\,\theta^{old}\right] \Leftrightarrow \max_{\theta} \sum_i \sum_h t_{ih} \log\left(\pi_h \phi(x_i;\theta_h)\right)$$

where $\theta = (\pi_1, \dots, \pi_K, \theta_1, \dots, \theta_K)$

For the Gaussian mixture model:

$$\phi(x_i;\theta_h) = \phi(x_i;\mu_h,\Sigma_h) = \frac{1}{(2\pi)^{p/2}} |\Sigma_h|^{-\frac{1}{2}} \exp\left(-\frac{1}{2}(x_i-\mu_h)^t \Sigma_h^{-1}(x_i-\mu_h)\right)$$

where $\theta = (\pi_1, \dots, \pi_K, \mu_1, \dots, \mu_K, \Sigma_1, \dots, \Sigma_K)$

Let's now do the calculations to get update values for θ^*, meaning π_k^*, μ_k^* and Σ_k^* :

Exo 1 : Write down the $Q(\theta) = E[\ell(X, Z|\theta)|\theta^*]$ for the GMM

$$Q(\theta) = \sum_i \sum_k t_{ik} \left[\log(\pi_k) + \log\left(\phi(x_i|\theta_k)\right) \right]$$

$$= \sum_i \sum_k t_{ik} \left[\log \pi_k - \frac{p}{2}\log(2\pi) - \frac{1}{2}\log(|\Sigma_k|) - \frac{1}{2}(x_i - \mu_k)^t \Sigma_k^{-1} (x_i - \mu_k) \right]$$

Exo 2 : maximisation over μ_k

$$\frac{\partial Q(\theta)}{\partial \mu_k} = 0 \Rightarrow \mu_k^*$$

$$\frac{\partial}{\partial \mu_k} Q(\theta) = \sum_k \left[\sum_i t_{ik} \left[\log \pi_k - \frac{p}{2} \log(2\pi) - \frac{1}{2} \log |\Sigma_k| \right] - \frac{1}{2} \sum_i t_{ik} (x_i - \mu_k)^t \Sigma_k^{-1} (x_i - \mu_k) \right]$$

$$= \frac{\partial}{\partial \mu_k} \left[-\frac{1}{2} \sum_i t_{ik} (x_i - \mu_k) \Sigma_k^{-1} (x_i - \mu_k) \right]$$

$$= -\frac{1}{2} \sum_i t_{ik} \times 2 \times \Sigma_k^{-1} (x_i - \mu_k)$$

$$\frac{\partial Q(\theta)}{\partial \mu_k} = 0 \iff \sum_i t_{ik} \Sigma_k^{-1} (x_i - \mu_k) = 0$$

$$\iff \sum_i t_{ik} (x_i - \mu_k) = 0$$

$$\iff \sum_i t_{ik} x_i = \sum_i t_{ik} \mu_k$$

$$\iff \mu_k^* = \frac{\sum_i t_{ik} x_i}{\sum_i t_{ik}} = \frac{1}{n_k} \sum_{i=1}^{n} t_{ik} x_i$$

Exo 3 : max over Σ_k :

$$\frac{\partial Q}{\partial \Sigma_k} = \frac{\partial}{\partial \Sigma_k}\left[-\frac{1}{2}\sum_{i=1}^{m} t_{ik}\left(\log|\Sigma_k| + (x_i-\mu_k)^t \Sigma_k^{-1}(x_i-\mu_k)\right)\right]$$

$$= \frac{\partial}{\partial \Sigma_2}\left(-\frac{1}{2}\sum t_{ik}\left(\log|\Sigma_2| + \operatorname{tr}\left((x_i-\mu_k)^t \Sigma_2^{-1}(x_i-\mu_k)\right)\right)\right)$$

$$= \operatorname{tr}\left(\Sigma_k^{-1}(x_i-\mu_k)(x_i-\mu_k)^t\right)$$

$$= \frac{\partial}{\partial \Sigma_2}\left[-\frac{1}{2}\sum_i t_{ik}\log|\Sigma_k| - \frac{1}{2}\operatorname{tr}\left(\Sigma_k^{-1}\sum_i t_{ik}(x_i-\mu_2)(x_i-\mu_2)^t\right)\right] = -\frac{1}{2}\sum_i t_{ik}\log|\Sigma_2| - \frac{1}{2}\operatorname{tr}\left(\Sigma_k^{-1}S_k\right)$$

$$= -\frac{1}{2}\left(\sum t_{ik}\left(\Sigma_k^{-1}\right)^t - \left(\Sigma_k^{-1}S_k\Sigma_k^{-1}\right)^t\right)$$

$$= -\frac{1}{2}\left(\sum_{i=1}^{m} t_{ik}\,\Sigma_k^{-1} - \Sigma_k^{-1}S_k\Sigma_k^{-1}\right) \overset{?}{=} 0$$

Some elements about derivative according to matrices (cf. the Matrix Cook Book)

$\operatorname{tr}(AB) = \operatorname{tr}(BA)$

$\operatorname{tr}(\lambda) = \lambda$

$\frac{\partial}{\partial A}\log(|A|) = \left(A^{-1}\right)^t$

$\frac{\partial}{\partial A}\operatorname{trace}(A^{-1}B) = -\left(A^{-1}BA^{-1}\right)^t$

$$\Leftrightarrow \quad n_k \Sigma_k^{-1} = \Sigma_k^{-1} S_k \Sigma_k^{-1}$$

$$\Leftrightarrow \quad n_k \Sigma_k^{-1} \Sigma_k = \Sigma_k^{-1} S_k \Sigma_k^{-1} \Sigma_k$$

$$\Leftrightarrow \quad n_k = \Sigma_k^{-1} S_k \quad \Leftrightarrow \quad \Sigma_k n_k = \Sigma_k \Sigma_k^{-1} S_k$$

$$\Leftrightarrow \quad \Sigma_k n_k = S_k$$

$$\Leftrightarrow \quad \Sigma_k = \frac{S_k}{n_k} = \frac{\sum_i t_{ik} (x_i - \mu_k)(x_i - \mu_k)'}{\sum t_{ik}}$$

Exo 4 : Max $Q(\theta)$ according to π_k given that $\sum_k \pi_k = 1$

$$\mathcal{L}(\theta) = Q(\theta) - \omega \left(\sum_k \pi_k - 1 \right)$$

$$\frac{\partial \mathcal{L}}{\partial \pi_k} = \frac{\sum_i t_{ik}}{\pi_k} - \omega \quad (1) \quad \Leftrightarrow$$

$$\frac{\partial \mathcal{L}}{\partial \pi_k} = 0 \iff \frac{u_k}{\pi_k} = w \iff w\pi_k = u_k$$

and we know that $\sum_k \pi_k = 1$

$$\sum_k w\pi_k = \sum_k u_k \iff w = n$$

Coming back to (1) $\Rightarrow$ $\frac{u_k}{\pi_k} = n$

$$\iff \pi_k^* = \frac{u_k}{n}$$

Parsimonious models for GMM

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

- because we have a limited number of obs. and fitting a (full) GMM requires to estimate a lot of parameters → parsimonious models are better
- we may have extra informations about the data (eg. the variables are independant → $\Sigma_k = \sigma_k I_p$) and we can encode this information as a constraint on the model
- ...

Parsimonious models for GMM

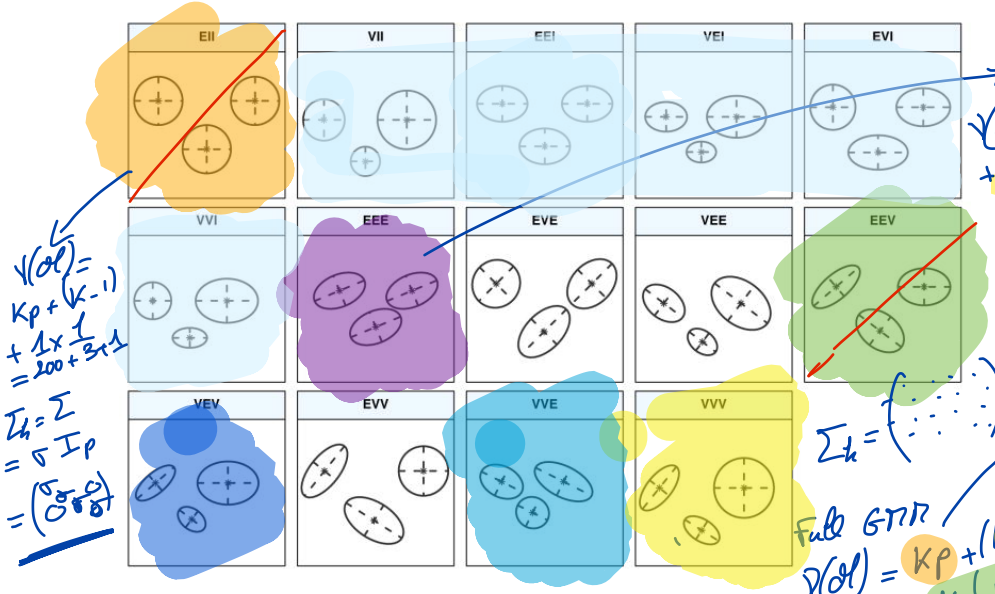
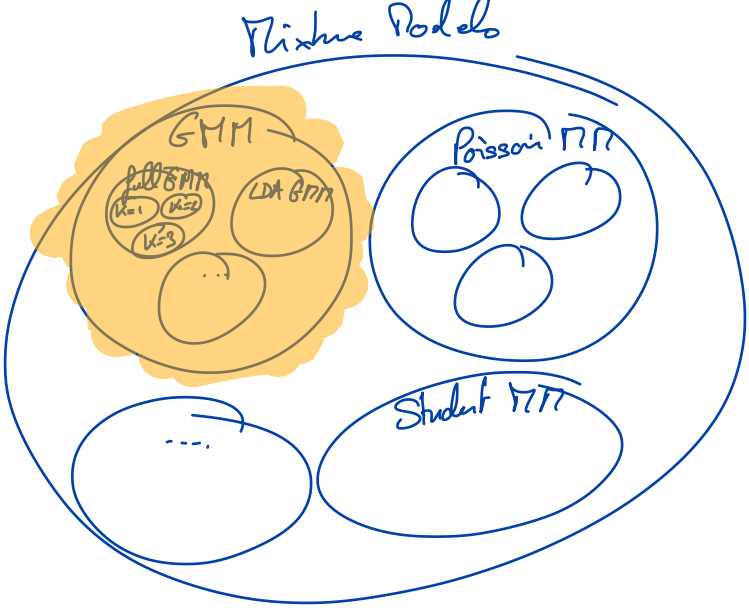


Figure: The parsimonious models of Mclust.

The interest here is to exploit the "catalog" of mixture models to find the most appropriate one and rely on model selection to make the choice

⚠ the list of models that you can test is very large due to the combination of pdf family + constraints + nb of components



$\Rightarrow$ Model selection selection is here to pick the most appropriate model for the data.

Parsimonious models for GMM

Such models are available in most softwares:

→ Mclust package for R

↳ a selection of constrained models (using the VVV → EEE nomenclature)

↳ BIC is used for model selection.

→ (R)Mixmod (Matlab + C++ + Python)

→ flexmix, ...

How to choose between models?

↳ we rely on model selection criteria allow to pick a good model in a list of candidates.

- → AIC
- → BIC
- → ICL
- → Slope heuristic
- → ...

	$\mathcal{M}_1$	$\mathcal{M}_2$		$\hat{\mathcal{M}}$*	$\mathcal{M}_G$
Crit	-200	-120	-10	-5	-215

Model selection

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

Let us first consider a list of candidate models $\{\mathcal{M}_1, \mathcal{M}_2, \ldots, \mathcal{M}_G\}$ and associated prior probabilities $p(\mathcal{M}_g)$. In practice, we can assume that all prior probabilities are equal. $p(\mathcal{M}_g) = p \ \forall g$.

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

Model selection

~~The roots of model selection can be found in Bayesian statistic theory:~~

thanks to the Bayes' theorem, we can write:

$$p(\mathcal{M}_g \mid X) \propto p(X \mid \mathcal{M}_g)\, p(\mathcal{M}_g) \qquad (1)$$

when the models have unknown parameters, the law of total probability allows to calculate it by integrating out the model parameters:

$$p(X \mid \mathcal{M}_g) = \int p(X \mid \mathcal{M}_g, \Theta_g)\, p(\Theta_g \mid \mathcal{M}_g)\, d\Theta_g$$

Rmq: this integration is usually very difficult or impossible to do!

Model selection

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

In the Bayesian framework, the knowledge of the integrated likelihood allows to do model selection by picking the model with the highest posterior proba:

$$\mathcal{M}^* = \underset{\mathcal{M}_g}{\operatorname{argmax}}\ p(\mathcal{M}_g|X)$$
$$= \underset{\mathcal{M}_g}{\operatorname{argmax}}\ p(X|\mathcal{M}_g).$$

Note that, when comparing two specific models $\mathcal{M}_1$ and $\mathcal{M}_2$, the ratio $B_{12} = \frac{p(X|\mathcal{M}_1)}{p(X|\mathcal{M}_2)}$ is called the Bayes' factor.

In particular, if $B_{12} > 1$, the model $\mathcal{M}_1$ should be preferred. And, if $B_{12} > 100$, the model $\mathcal{M}_1$ as a strong evidence against $\mathcal{M}_2$ (Jeffreys, 1961)

In practice, when considering the frequentist case, this framework is also useful because the integrated likelihood can be approximated.

$$\mathcal{M}^* = \underset{\mathcal{M}_g}{\operatorname{argmax}} \; \tilde{p}(X \mid \mathcal{M}_g)$$

In particular, the BIC (Bayesian Information Criterion), proposed by Schwarz in 1978, is an approximation of the integrated likelihood:

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

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

BIC is an asymptotic approximation of the integrated likelihood based on a second order Taylor expansion of the logarithm of the integrand, around its maximum $\hat{\theta}_g$.

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

Fortunately, BIC is behaving very well for mixture models in practice!

Model selection: AIC, BIC, ICL

Classical model selection tools can be used easily:

- AIC and BIC are two criteria that can be used in large numbers of situations (regression, mixture models, ...)
- ICL (integrated classification likelihood) was proposed for the specific case of clustering.

The idea of ICL is the same as BIC but instead of approximating $p(X|\mathcal{M}_g)$ it focuses on the integrated classification likelihood $p(X, Z|\mathcal{M}_g)$

where Z is the latent variable that encodes the group memberships.

$$p(X \mid \mathcal{M}_g) \longrightarrow p(X, Z \mid \mathcal{M}_g)$$

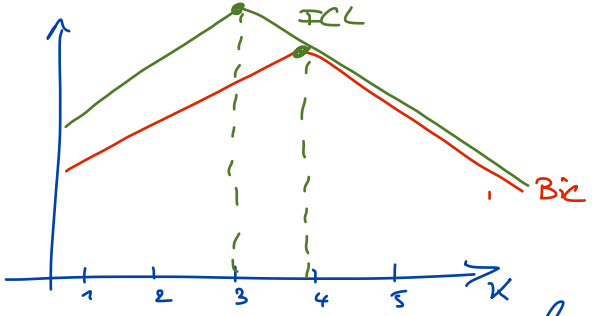
$p(X \mid \mathcal{M}_g)$ ↑ BIC

$p(X, Z \mid \mathcal{M}_g)$ ↑ ICL.

ICL then uses the same approximations of BIC and we end up with the following approximation:

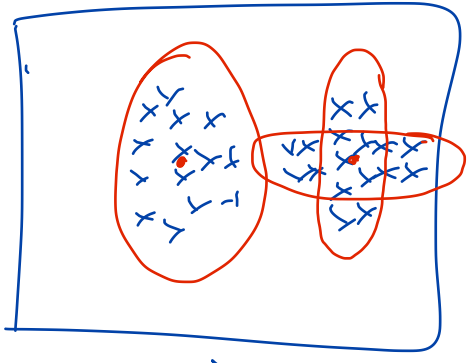
$$\log p(X, Z \mid \mathcal{M}_g) \simeq \log p(X \mid \hat{\theta}_g, \mathcal{M}_g) - \frac{\nu(\mathcal{M}_g)}{2} \log(n) - \frac{1}{2} \sum_{i=1}^{n} \sum_{k=1}^{K} \hat{z}_{ik} \log(\hat{z}_{ik}) \quad \Big\} \text{ ICL.}$$

Rmq: it clearly appears that $ICL = BIC - \frac{1}{2}\sum_{i}\sum_{k} z_{ik} \log(z_{ik})$ and that ICL penalizes more the likelihood compared to BIC.



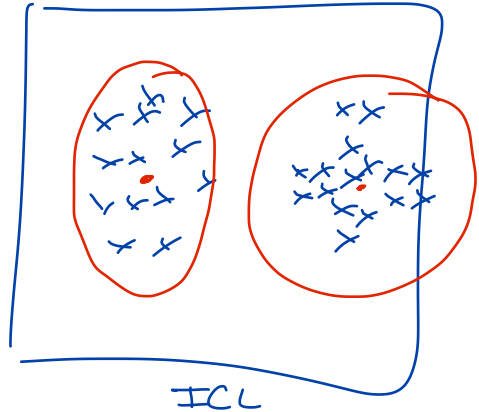
Rmq: the objectives of BIC and ICL are slightly different:

- BIC aims to find a good fit of the data
- ICL aims to find a good model to cluster the data.



BIC

⇒ GMM with 3 comp.



ICL

⇒ GMM with 2 comp.