Model selection

Model selection: How many clusters ?

The number of clusters K controls the model complexity.

Choosing K is an example of model selection.

The optimal Bayesian approach is to pick the model with the largest marginal likelihood,

$$K^* = \arg\max_k p(\mathcal{D}|K).$$

In practice,

1. Simple approximations, such as BIC, ICL can be used.
2. We can use the cross-validated likelihood as a performance measure
3. An alternative approach is to perform stochastic sampling in the space of models (MCMC)

The Laplace approximation

Taylor expansion

For functions of multiple variables $\boldsymbol{f}(\boldsymbol{z}) = \boldsymbol{f}(z_1, z_2, \cdots, z_p)$.

$$\log f(\boldsymbol{z}) \approx \log f(\boldsymbol{z}_0) + \nabla \log f(\boldsymbol{z}_0)^T (\boldsymbol{z} - \boldsymbol{z}_0) + \frac{1}{2}(\boldsymbol{z} - \boldsymbol{z}_0)^T$$

where $\boldsymbol{A}$ is the Hessian matrix of second derivatives of $\log f(\boldsymbol{z})$ at $z = z_0$.

The Laplace approximation

Gaussian approximation to a probability density defined over a set of continuous variables.

Considering the density

$$p(\boldsymbol{z}) = \frac{1}{Z} f(\boldsymbol{z})$$

The normalizing constant is

$$\begin{aligned} Z &= \int f(\boldsymbol{z}) d\boldsymbol{z} \\ &= f(\boldsymbol{z}_0) \int \exp -\frac{1}{2} (\boldsymbol{z}_{-} \boldsymbol{z}_0)^T A (\boldsymbol{z}_{-} \boldsymbol{z}_0) d\boldsymbol{z} \\ &\approx f(\boldsymbol{z}_0) \frac{(2\pi)^{p/2}}{|A|^{1/2}} \end{aligned}$$

where $\boldsymbol{z}_0$ is a mode of the distribution (where $\nabla \log f(\boldsymbol{z}_0) = \mathbf{0}$)

BIC

From the Bayes theorem the model evidence is

$$p(\mathcal{D}) = \int p(\boldsymbol{\theta}) p(D|\boldsymbol{\theta}) d\boldsymbol{\theta}$$

Using Laplace approximation in for $f(\boldsymbol{\theta}) = p(\boldsymbol{\theta}) p(D|\boldsymbol{\theta})$ in $\boldsymbol{\theta} = \boldsymbol{\theta}_{MAP}$:

$$\log p(\mathcal{D}) \approx \log p(\mathcal{D}|\boldsymbol{\theta}_{MAP}) + \underbrace{\log p(\boldsymbol{\theta}_{MAP}) + \frac{p}{2} \log(2\pi) - \frac{1}{2}}_{Occam\,factor}$$

- $\log p(\mathcal{D}|\boldsymbol{\theta}_{MAP})$ represents the log-likelihood
- the Occam factor penalizes the model complexity

Assuming a simple Gaussian prior distribution over parameters, with full rank Hessian we can further approximate the Hessian matrix by the fisher information and we get

$$\log p(\mathcal{D}) \approx \log p(\mathcal{D}|\boldsymbol{\theta}_{MAP}) - \frac{1}{2} p \log n$$

which is known a the BIC (Bayesian Information Criterion) or the Schwartz criterion (1978).

BIC for chosing the number of clusters

$$K_{BIC} = \arg\max_k \log p(\mathcal{D}|\boldsymbol{\theta}^k_{MAP}) - \frac{1}{2}p_k \log n$$

where p_k is the number of parameters of the model with k clusters and θ^k_{MAP} the MAP estimate of the model

Integrated Complete Likelihood (ICL)

$$BIC(k) = p(\mathcal{D}|\boldsymbol{\theta}^k_{MAP}) - \frac{1}{2}p_k \log n$$

$$= \mathbb{E}_{Z|X;\boldsymbol{\theta}^k_{MAP}}[\log p(X, Z; \boldsymbol{\theta}^k_{MAP})] - \mathbb{E}_{Z|X;\boldsymbol{\theta}^k_{MAP}}[\log p($$

Biernacki et al. (2000) proposed to favour clustering with high-confidence (low entropy) by removing entropy term to BIC.

ICL for chosing the number of clusters

$$K_{ICL} = \arg\max_k \log \mathbb{E}_{Z|X;\boldsymbol{\theta}^k_{MAP}}[\log p(X, Z; \boldsymbol{\theta}^k_{MAP})] - \frac{1}{2}p_k \mathrm{l}$$

Akaike information criterion

Using information theory Akaike (1974) derived an alternative criterion:

AIC

$$K_{AIC} = \arg\max_k \log p(\mathcal{D}|\boldsymbol{\theta}_k) - p_k$$

Penalized Likelihood criteria

Generally AIC chooses more complex models than BIC which chooses more complex models than ICL

$$K_{AIC} \geq K_{BIC} \geq K_{ICL}$$

Multivariate Gaussian Mixture models

Assumes K classes in proportion $\pi_1, \ldots, \pi_K$ with component densities

$$\boldsymbol{x}_i | z_i = k \sim \mathcal{N}_p(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$$

- $\boldsymbol{\mu}_k \in \mathbb{R}^p$
- $\boldsymbol{\Sigma}_k \in \mathbb{R}^{p \times p}$

Number of parameters

$$p_k = p.K + p(p-1)/2 + K - 1$$

Covariance matrix parametrization

Table 1: Parameterizations of the covariance matrix Σ_k currently available in mclust for hierarchical clustering (HC) and/or EM for multidimensional data. ('•' indicates availability).

identifier	Model	HC	EM	Distribution	Volume	Shape	Orientation
`E`		•	•	(univariate)	equal		
`V`		•	•	(univariate)	variable		
`EII`	λI	•	•	Spherical	equal	equal	NA
`VII`	λ_k I	•	•	Spherical	variable	equal	NA
`EEI`	λA		•	Diagonal	equal	equal	coordinate axes
`VEI`	$\lambda_k A$		•	Diagonal	variable	equal	coordinate axes
`EVI`	λA_k		•	Diagonal	equal	variable	coordinate axes
`VVI`	$\lambda_k A_k$		•	Diagonal	variable	variable	coordinate axes
`EEE`	λDAD^T	•	•	Ellipsoidal	equal	equal	equal
`EEV`	$\lambda D_k A D_k^T$		•	Ellipsoidal	equal	equal	variable
`VEV`	$\lambda_k D_k A D_k^T$		•	Ellipsoidal	variable	equal	variable
`VVV`	$\lambda_k D_k A_k D_k^T$	•	•	Ellipsoidal	variable	variable	variable

Relation to kmeans algorithm

Complete (Classification) log-likelihood

$$
\begin{aligned}
CL(\boldsymbol{\theta}; X, Z) &= \log \prod_i p(x_i, z_i = k; \boldsymbol{\theta}_k) \\
&= \log \prod_i \prod_k p(x_i, z_i = k; \boldsymbol{\theta}_k)^{\mathbb{I}(z_i = k)} \\
&= \sum_i \sum_k \mathbb{I}(z_i = k) \log p(x_i, z_i = k; \boldsymbol{\theta}_k)
\end{aligned}
$$

CEM algorithm

$CL(\boldsymbol{\theta}; X, Z)$ can be maximized using CEM algorithm

Principal Component Analysis