Latent variable models in biology and ecology

Chapter 2: Mixtures models

Sophie Donnet. INRAE

Master 2 MathSV. January 25, 2022

Introduction

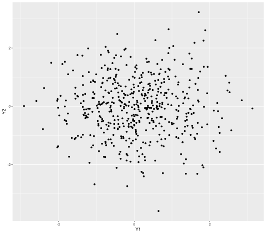
The mixture model

Statistical inference

- Mixture model: one of the most simple latent variable models
- Assumptions
  - Observations supposed to be independent,
  - Each observation arises from a given class that is unobserved
- Main goal : retrieve the class from which each observation arises
- Also referred as unsupervised classification as we do not dispose of any observation with known label.

First toy illustration

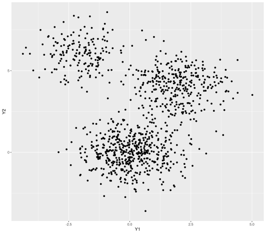
Observations described by 2 variables



Observation distribution seems easy to model with one Gaussian

First toy illustration

Observations described by 2 variables



Data are scattered and subpopulations are observed

According to the experimental design, there exists no external information about them

This is an underlying structure observed through the data

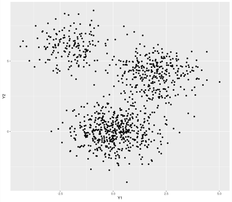
First toy illustration

Definition (Mixture model)

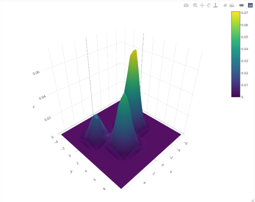
It is a probabilistic model for representing the presence of subpopulations within an overall population.

$$Y_i | Z_i = k \sim \mathcal{N}(\mu_k, \Sigma_k), \quad P(Z_i = k) = \pi_k$$

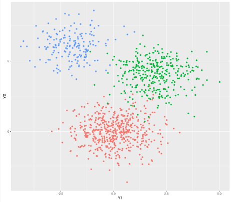
what we observe



the model



the expected results



$Z = ?$

$Z : 1 = \color{blue}{\bullet}, 2 = \color{red}{\bullet}, 3 = \color{green}{\bullet}$

$\rightarrow$ It is an unsupervised classification method

Applications in biology

Technics of clustering widely used in biology. See the Wikipedia page

Gene expression i

- To build groups of genes with related expression patterns (also known as coexpressed genes).
- Often such groups contain functionally related proteins, such as enzymes for a specific pathway, or genes that are co-regulated.
- Y_{tm} gene expression of gene at locus t in condition $m = 1, \ldots, P$ conditions.

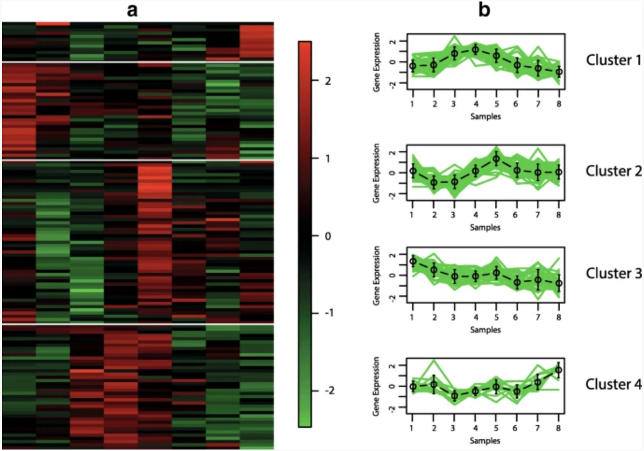


Figure from [Parraga-Alava et al., 2018]

Human genetic

- Better understand the genetic structure of populations
- Relies on the genotyping of large sets of individuals sampled in different places, environments or with different origins
- Genotype Y_{it} of a series of individuals $i \in [1, I]$ at a series of locus $t \in [1, T]$ is measured
- Aim: distinguish sub-populations.

Model without 'admixture'

Each individual i is supposed to belong to one population, labeled Z_i

$$\begin{aligned}(Z_i)_i \text{ iid} &\sim \mathcal{M}(1;\pi),\\ (Y_{it})_{i,t} \text{ indep.} \,|\, (Z_i) &\sim \mathcal{M}(1;\gamma_{Z_i t}),\end{aligned}$$

γ_{kt} is the vector of the allelic frequencies at locus t in population k

which makes explicit the fact that, if individual i belongs to population k, its genotype is generated with the allelic frequencies of its population.

Model with 'admixture'

$$
\begin{aligned}
(Y_{it})_{i,t} \text{ indep.} \,|\, (S_{it}) &\sim \mathcal{M}(1; \gamma_{S_i t}) \\
(S_{it})_{i,t} \text{ indep.} \,|\, (Q_i) &\sim \mathcal{M}(1; Q_i), \\
(Q_i)_i \text{ iid} &\sim \mathcal{D}(1; \alpha)
\end{aligned}
$$

About Q_i: individual preferential trends characterized

- Dirichlet distribution whose support is the the simplex of $\mathbb{R}^K$.
- D_i is the position of individual i in the simplex, the vertices of which correspond to fictitious individuals purely issued from each population.

Hidden variable is hence $Z = (Q, S)$.

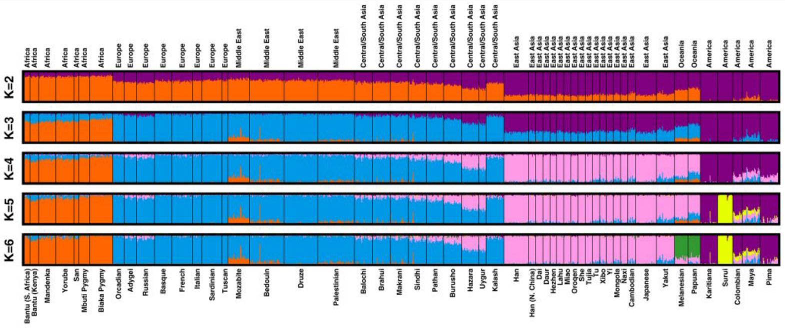
Model with 'admixture': reformulation

The model can be rewritten also after marginalization over S_{it}:

$$\begin{aligned} (Q_i)_i \text{ iid} &\sim \mathcal{D}(1;\alpha), \\ (Y_{it})_{i,t} \text{ indep.} \,|\, (Q_i) &\sim \mathcal{M}\left(1; \sum_k Q_{ik}\gamma_{kt}\right). \end{aligned}$$

The latent variable reduces then to $Z = (Q)$.

See [Pritchard et al., 2000] for more details.



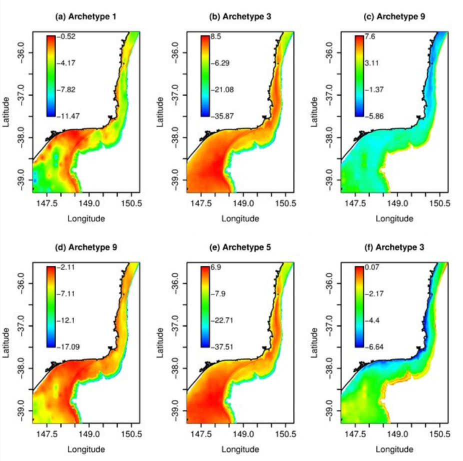
Population origine of series of human genomes with varying number of groups K. Each column corresponds to an individual. Each individual is represented by a thin vertical line partitioned into K colored segments that represent the fractions of the individual's genome estimated to belong to the K clusters. From [Rosenberg, 2011].

Plant and animal ecology i

To describe and to make spatial and temporal comparisons of communities (assemblages) of organisms in heterogeneous environments.

- Y_{is}: abundancy of species i at location s.
- Not the same repartitions with respect to species.

Plant and animal ecology ii



[Dunstan et al., 2013]

Definition

- Let $(Y_i)_{i=1,\ldots,n}$ be independent variables
- For each individual i assumes the existence an unknown (or latent) label Z_i that can take a finite number of values among $[1, K]$.
- The distribution of Y_i depends on the value Z_i.

Definition

An independent K mixture model is defined as follows: $\forall i = 1, \ldots, n$

$$\begin{array}{ccll} \mathrm{P}(Z_i = k) & = & \pi_k, & (i.i.d) \\ Y_i | (Z_i = k) & \sim_{i.i.d} & \mathcal{F}_k = \mathcal{F}(\gamma_k), & \end{array} \tag{1}$$

where $\sum_{k=1}^{K} \pi = 1$.

Let $f_k(\cdot) = f(\cdot; \gamma_k)$ be the pdf of distribution of $\mathcal{F}(\gamma_k)$.

Alternative fomulations

- $Y_i|(Z_i = k) \sim \mathcal{F}(\gamma_k)$ is equivalent to $Y_i|Z_i \sim \mathcal{F}(\gamma_{Z_i})$
- Let $Z_{ik} = \mathbf{1}_{\{Z_i = k\}}$

$$(Z_{ik})_{k=1,\ldots,K} \sim \mathcal{M}(1, \boldsymbol{\pi})$$

where $\mathcal{M}$ is the multinomial distribution $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_K)$

About the mixture proportions

- π_k = proportion of the population k
- Sometimes called prior probabilities although this denomination may be misleading in a non-Bayesian context.
- Also often refereed to as the proportions of the mixture.

About the emission distribution

- Conditionally on $\{Z_i = k\}$, Y_i has a parametric distribution $\mathcal{F}_k = \mathcal{F}(\gamma_k)$ with probability distribution function (pdf) $f_k(\cdot) = f(\cdot; \gamma_k)$.
- $\mathcal{F}_k$ is called the emission distribution in class k
- It describes how observed data arising from class k are emitted.
- f_k is called the emission pdf.

Other formulation

Useful notations

- $\mathbf{Z} = (Z_1, \ldots, Z_n)$
- $\mathbf{Y} = (Y_1, \ldots, Y_n)$
- $\boldsymbol{\pi} = (\pi_k)_{k=1,\ldots,K}$
- $\boldsymbol{\gamma} = (\gamma_k)_{k=1,\ldots,K}$
- $\theta = (\boldsymbol{\pi}, \boldsymbol{\gamma})$

Conditional distributions

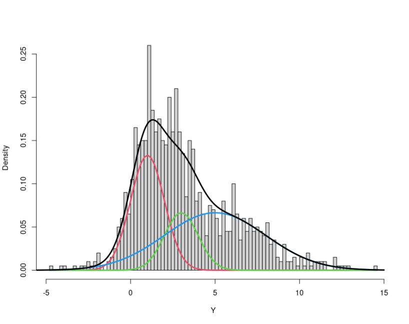
$$\begin{array}{rclcl} p_\theta(\mathbf{Z}) & = & \prod_{i=1}^n \pi_{Z_i} & = & \prod_{i=1}^n \prod_{k=1}^K (\pi_k)^{Z_{ik}}, \\ p_\theta(\mathbf{Y}|\mathbf{Z}) & = & \prod_{i=1}^n f(Y_i, \gamma_{Z_i}) & = & \prod_{i=1}^n \prod_{k=1}^K f(Y_i, \gamma_k)^{Z_{ik}}, \end{array}$$

Marginal distribution

Marginal pdf. of Y_i is the mixture distribution

$$g(y) = \sum_{k=1}^{K} \pi_k f(y; \gamma_k).$$

Example of a mixture of $K = 3$ Gaussian distributions $\frac{1}{3}\mathcal{N}(1,1) + \frac{1}{6}\mathcal{N}(3,1) + \frac{1}{2}\mathcal{N}(5,3^2)$



Label switching

Since the (Z_i) are not observed, the model is invariant for any permutation of the labels $[1, K]$.

Therefore, the mixture model with K classes has $K!$ equivalent definitions.

Number of parameters

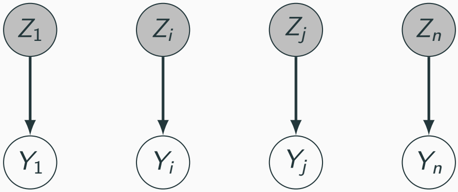
- Depends on both the dimension of the data and the number of groups
- $\sum_{k=1}^{K} \pi_k = 1$, $\boldsymbol{\pi}$ involves only $K - 1$
- About $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_K)$, its dimension is typically proportional to the number of groups K
- For $\mathcal{F}_k$: univariate Poisson distributions with respective mean γ_k , $\boldsymbol{\gamma}$ of dimension $K \Rightarrow 2K - 1$ parameters
- For $\mathcal{F}_k$: d-variate normal distributions (with respective mean vector μ_k and variance Σ_k):

$$(K - 1) + Kd + Kd(d + 1)/2 \simeq Kd^2/2$$

parameters

Dependency structures

- The (Z_i) are independent;
- the (Y_i) are independent conditionally to $\mathbf{Z} = (Z_i)_{i=1,\ldots,n}$;
- the couples $\{(Y_i, Z_i)\}_i$ are iid.



Graphical representation of a mixture model

Remarks

1. Because the $\{(Y_i, Z_i)\}_i$ are independent, we have that

$$p_\theta(Z_i|\mathbf{Y}) = p_\theta(Z_i|Y_i)$$

   which means that the information about the classification of individual i is contained in the observation Y_i.
2. Note that the variables (Y_i, Y_j) are *not* independent conditionally on the event $Z_i = Z_j$.

Two tasks

- For a fixed number of class K, estimating the parameters

$$\boldsymbol{\pi} = (\pi_1, \ldots, \pi_K), \quad \boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_K)$$

$$\theta = (\boldsymbol{\pi}, \boldsymbol{\gamma})$$

  $\Rightarrow$ (Maximum likelihood) estimation
- Would be great to obtain a classification of the observations
- Choosing the number of classes K $\Rightarrow$ Model selection

Parameter estimation

- General introduction to finite mixture models and their inference can be found in [McLachlan and Peel, 2000]
- Most popular inference method: maximum likelihood approach
- Specificity of latent variable models : the observed data $\mathbf{Y} = (Y_i)_{i=1,\dots,n}$ seen as incomplete, as the latent variables $\mathbf{Z} = (Z_i)_{i=1,\dots,n}$ are not observed
- Often referred to as incomplete data models.

Likelihoods

Definition

The observed data log-likelihood is the marginal log-likelihood of the observed variables $\mathbf{Y}$:

$$\log p_\theta(\mathbf{Y}).$$

The complete data log-likelihood is the joint log-likelihood of the observed $\mathbf{Y}$ and latent $\mathbf{Z}$ variables:

$$\log p_\theta(\mathbf{Y}, \mathbf{Z}).$$

Expression of the likelihoods

Proposition (Likelihoods)

For the mixture model (1), the log-likelihood is

$$\log p_\theta(\mathbf{Y}) = \sum_{i=1}^{n} \log \left[\sum_{k=1}^{K} \pi_k f(Y_i; \gamma_k) \right],$$

and, denoting $Z_{ik} = \mathbf{1}_{\{Z_i = k\}}$, the complete log-likelihood is

$$\log p_\theta(\mathbf{Y}, \mathbf{Z}) = \sum_{i=1}^{n} \sum_{k=1}^{K} Z_{ik} \left[\log \pi_k + \log f(Y_i; \gamma_k) \right].$$

Proof

The dependency structure described in previously ensures that

$$
\begin{aligned}
\log p_\theta(\mathbf{Y}) &= \sum_{i=1}^{n} \log p_\theta(Y_i) = \sum_{i=1}^{n} \log g(Y_i) \\
\text{and} \qquad \log p_\theta(\mathbf{Y}, \mathbf{Z}) &= \sum_{i=1}^{n} \log p_\theta(Y_i, Z_i) \\
&= \sum_{i=1}^{n} \left[\log p_\theta(Z_i) + \log p_\theta(Y_i|Z_i)\right].
\end{aligned}
$$

Remark: $\log p_\theta(Y_i)$ not easy to optimize

About the EM algorithm

- First proposed by [Dempster et al., 1977] for a large class of incomplete data models, including mixture models.
- Based on a decomposition of the incomplete data likelihood.

Proposition (Decomposition of the log-likelihood)

For any θ and θ'

$$\log p_\theta(\mathbf{Y}) = \mathbb{E}_{\theta'}\left[\log p_\theta(\mathbf{Y}, \mathbf{Z}) | Y\right] - \mathbb{E}_{\theta'}\left[\log p_\theta(\mathbf{Z}|\mathbf{Y}) | \mathbf{Y}\right].$$

Proof

It suffices to develop

$$\mathbb{E}_{\theta'}\left[\log p_\theta(\mathbf{Z}|\mathbf{Y})|\mathbf{Y}\right] = \mathbb{E}_{\theta'}\left[\log p_\theta(\mathbf{Y},\mathbf{Z}) - \log p_\theta(\mathbf{Y})|\mathbf{Y}\right]$$

reminding that $\mathbb{E}_{\theta'}\left[\log p_\theta(\mathbf{Y})|\mathbf{Y}\right] = \log p_\theta(Y)$.

Remarks

1. Decomposition of Slide 36 is convenient bacause makes a connexion between $\log p_\theta(\mathbf{Y})$ (often intractable) and $\log p_\theta(\mathbf{Y}, \mathbf{Z})$ (generally more manageable).
2. if $\theta' = \theta$, the second term is the entropy of the latent variables $\mathbf{Z}$ given the observed $\mathbf{Y}$:

$$\mathcal{H}[p_\theta(\mathbf{Z}|\mathbf{Y})] := -\mathbb{E}_\theta[\log p_\theta(\mathbf{Z}|\mathbf{Y})|\mathbf{Y}]$$

EM Algorithm

$$\widehat{\theta} = \arg\max_{\theta} \log p_{\theta}(\mathbf{Y}).$$

Algorithm (EM)

Repeat until convergence:

Expectation step *(E-step) given the current estimate θ^h of θ, compute $p_{\theta^h}(\mathbf{Z}|\mathbf{Y})$, or at least all the quantities needed to compute $\mathbb{E}_{\theta^h}[\log p_{\theta}(\mathbf{Y}, \mathbf{Z})|\mathbf{Y}]$;*

Maximization step *(M-step) update the estimate of θ as*

$$\theta^{h+1} = \arg\max_{\theta} \mathbb{E}_{\theta^h}[\log p_{\theta}(\mathbf{Y}, \mathbf{Z})|\mathbf{Y}].$$

Property

Proposition ([Dempster et al., 1977])

The log-likelihood of the observed data $\log p_\theta(\mathbf{Y})$ *increases at each step:*

$$\log p_{\theta^{h+1}}(\mathbf{Y}) \geq \log p_{\theta^h}(\mathbf{Y}).$$

Proof i

Because $\theta^{h+1} = \arg\max_\theta \mathbb{E}_{\theta^h}[\log p_\theta(\mathbf{Y}, \mathbf{Z})|\mathbf{Y}]$, we have

$$
\begin{aligned}
0 &\leq \mathbb{E}_{\theta^h}[\log p_{\theta^{h+1}}(\mathbf{Y}, \mathbf{Z})|\mathbf{Y}] - \mathbb{E}_{\theta^h}[\log p_{\theta^h}(\mathbf{Y}, \mathbf{Z})|\mathbf{Y}] && (2) \\
&= \mathbb{E}_{\theta^h}\left[\log \frac{p_{\theta^{h+1}}(\mathbf{Y}, \mathbf{Z})}{p_{\theta^h}(\mathbf{Y}, \mathbf{Z})}|\mathbf{Y}\right] && (3) \\
&\leq \log \mathbb{E}_{\theta^h}\left[\frac{p_{\theta^{h+1}}(\mathbf{Y}, \mathbf{Z})}{p_{\theta^h}(\mathbf{Y}, \mathbf{Z})}|\mathbf{Y}\right] && (4)
\end{aligned}
$$

by Jensen's inequality.

Proof ii

We further develop $\log \mathbb{E}_{\theta^h}\left[p_{\theta^{h+1}}(\mathbf{Y},\mathbf{Z}) / p_{\theta^h}(\mathbf{Y},\mathbf{Z}) \,|\mathbf{Y}\right]$ as

$$
\begin{aligned}
\log \int \frac{p_{\theta^{h+1}}(\mathbf{Y},\mathbf{Z})}{p_{\theta^h}(\mathbf{Y},\mathbf{Z})} p_{\theta^h}(\mathbf{Z}|\mathbf{Y})\,\mathrm{d}\mathbf{Z} &= \log \int \frac{p_{\theta^{h+1}}(\mathbf{Y},\mathbf{Z})}{p_{\theta^h}(\mathbf{Y},\mathbf{Z})} \frac{p_{\theta^h}(\mathbf{Y},\mathbf{Z})}{p_{\theta^h}(\mathbf{Y})}\,\mathrm{d}\mathbf{Z} && (5)\\
&= \log \left[\frac{1}{p_{\theta^h}(\mathbf{Y})} \int p_{\theta^{h+1}}(\mathbf{Y},\mathbf{Z})\,\mathrm{d}\mathbf{Z}\right] && (6)\\
&= \log \left[\frac{p_{\theta^{h+1}}(\mathbf{Y})}{p_{\theta^h}(\mathbf{Y})}\right] && (7)
\end{aligned}
$$

Finally :

$$
\log \left[\frac{p_{\theta^{h+1}}(\mathbf{Y})}{p_{\theta^h}(\mathbf{Y})}\right] \geq 0
$$

Convergence

There is no general guaranty about the convergence of the EM algorithm towards the MLE $\widehat{\theta}$. The main property is that the observed likelihood increases at each iteration step.

Although, in practice : very sensible to the initialisation point.

Illustration of the problems of convergence (I)

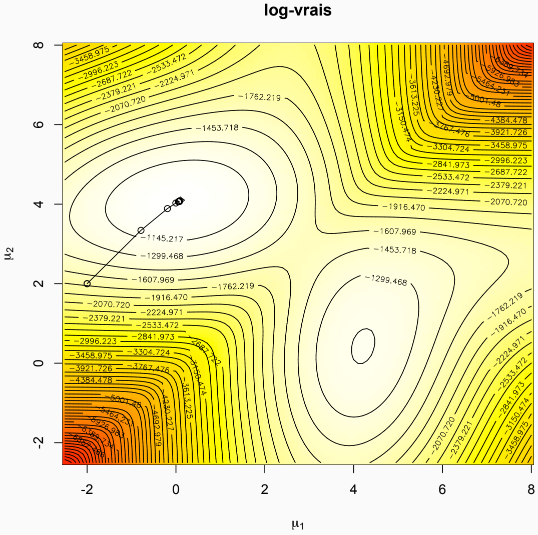
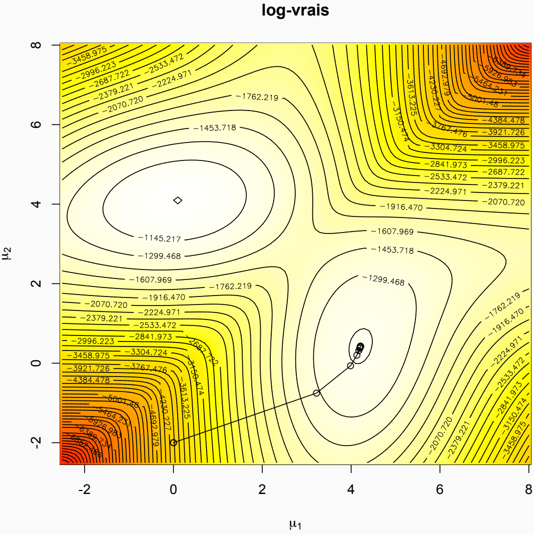
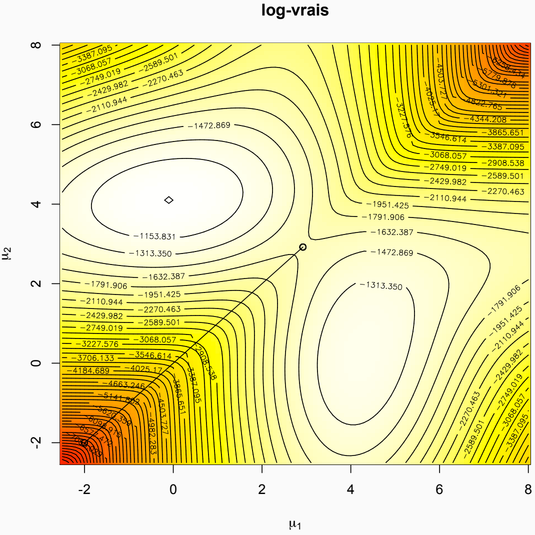


Illustration of the problems of convergence (II)




log-vrais
μ_1
μ_2
-2
0
2
4
6
8
−1153.831
−1313.350
−1472.869
−1632.387
−1791.906
−1951.425
−2110.944
−2270.463
−2589.501
−2749.019
−3068.057
−3387.095

Application for the mixture model : E step

E-step is straightforward for independent mixture models.

Proposition

In a mixture model (1), the hidden states Z_i are independent conditional on the observations:

$$p_\theta(\mathbf{Z}|\mathbf{Y}) = \prod_{i=1}^{n} p_\theta(Z_i|Y_i)$$

and, denoting $Z_{ik} = \mathbf{1}_{\{Z_i=k\}}$, the conditional distribution of each Z_i is given by

$$\tau_{ik} := P_\theta(Z_i = k|Y_i) = \mathbb{E}_\theta(Z_{ik}|Y_i) = \frac{\pi_k f_k(Y_i)}{\sum_{\ell=1}^{K} \pi_\ell f_\ell(Y_i)}.$$

Proof i

- First result is a direct consequence of Slide 29
- Second result follows from the Bayes formula

$$
\begin{aligned}
\tau_{ik} &= P_\theta(Z_i = k|Y_i) = \frac{P_\theta(Z_i = k) p_\theta(Y_i|Z_i = k)}{p_\theta(Y_i)} \\
&= \frac{P_\theta(Z_i = k) p_\theta(Y_i|Z_i = k)}{\sum_\ell P_\theta(Z_i = \ell) p_\theta(Y_i|Z_i = \ell)}.
\end{aligned}
$$

- $P_\theta(Z_i = k|Y_i) = \mathbb{E}_\theta(Z_{ik}|Y_i)$ because Z_{ik} is binary.

Proof ii

The update formula's of the τ_{ik} at the $(h+1)$-th E-step is then

$$\tau_{ik}^{h+1} = \frac{\pi_k^h f(Y_i; \gamma_k^h)}{\sum_\ell \pi_\ell^h f(Y_i; \gamma_\ell^h)}$$

where θ^h stands for the current estimate of θ resulting from the h-th M step.

Remark

Conditional probability τ_{ik} is sometimes referred to as the **posterior probability** for observation i to belong to class k (as opposed to the **prior probability** π_k).

Again this phrase is misleading in a non-Bayesian context and 'conditional probability' should be preferred.

M-step for the mixture model

$$\theta^{h+1} = \arg\max_{\theta} \mathbb{E}_{\theta^h}[\log p_\theta(\mathbf{Y}, \mathbf{Z})|\mathbf{Y}]$$

We use Proposition on Slide 34 to get an explicit formula for this quantity

$$\begin{aligned}
\mathbb{E}_{\theta^h}[\log p_\theta(Y, Z)|Y] &= \mathbb{E}_{\theta^h}\left[\sum_{i=1}^{n}\sum_{k=1}^{K} Z_{ik}[\log \pi_k + \log f(Y_k; \gamma_k)]|\mathbf{Y}\right] \\
&= \sum_{i=1}^{n}\sum_{k=1}^{K} \mathbb{E}_{\theta^h}(Z_{ik}|Y_i)[\log \pi_k + \log f(Y_k; \gamma_k)] \\
&= \sum_{i=1}^{n}\sum_{k=1}^{K} \tau_{ik}^{h}[\log \pi_k + \log f(Y_k; \gamma_k)].
\end{aligned}$$

Has to be maximized with respect to $\theta = (\boldsymbol{\pi}, \boldsymbol{\gamma})$, the τ_{ik} being fixed

Application for the mixture model : M step (π) i

$$\pi_k^{h+1} = \frac{1}{n} \sum_{i=1}^{n} \tau_{ik}^h \tag{8}$$

Indeed:

- Using the Lagrange multiplier to take into account the constraint $\sum_{k=1}^{K} \pi_k = 1$
- $$\frac{\partial}{\partial \pi_k} \left[\sum_{i=1}^{n} \sum_{k=1}^{K} \tau_{ik}^h [\log \pi_k + \log f(Y_k; \gamma_k)] - \lambda \left(\sum_{k=1}^{K} \pi_k - 1 \right) \right] = 0$$
- Leads to $\sum_{i=1}^{n} \frac{\tau_{ik}^h}{\pi_k^{(h+1)}} - \lambda = 0$ and so $\pi_k^{(h+1)} = \frac{1}{\lambda} \sum_{i=1}^{n} \tau_{ik}^h$

Application for the mixture model : M step (π) ii

- Moreover $\sum_{k=1}^{K} \pi_k^{(h+1)} = 1$. So
  $\frac{1}{\lambda} \sum_{k=1}^{K} \sum_{i=1}^{n} \tau_{ik}^{h} = \frac{1}{\lambda} \sum_{i=1}^{n} \underbrace{\sum_{k=1}^{K} \tau_{ik}^{h}}_{=1} = n.$
- Which implies Formula (8)

Application for the mixture model : M step (γ)

- For γ : solution of this optimization problem has no general form as it strongly depends on the model at hand
- Some general formula can be derived in the case of the exponential family, as we will see in Slide 56

About the entropy

$$\mathcal{H}[p_\theta(\mathbf{Z}|\mathbf{Y})] = -\mathbb{E}_\theta[\log p_\theta(\mathbf{Z}|\mathbf{Y})|\mathbf{Y}]$$

Can be calculated using the conditional independence of the Z_i given the data $\mathbf{Y}$:

$$\begin{aligned}
\mathcal{H}[p_\theta(\mathbf{Z}|\mathbf{Y})] &= \sum_{i=1}^{n} H[p_\theta(Z_i|Y_i)] \\
&= -\sum_{i=1}^{n} \mathbb{E}_\theta[\log P(Z_i = k|Y_i)|Y_i] \qquad (9) \\
&= -\sum_{i=1}^{n} \sum_{k=1}^{K} \tau_{ik} \log \tau_{ik}.
\end{aligned}$$

Exponential family

Definition (Exponential family of distributions)

The distribution $f(;\gamma)$ belongs to exponential family with *canonical parameter* γ if

$$f(y;\gamma) = \exp[\gamma^\intercal t(y) - a(y) - b(\gamma)]$$

where $t(y)$ is the vector of the *sufficient statistics*.

Maximum likelihood for the exponential family

Two general properties that show connections between maximum likelihood estimates and moment estimates for this class of distribution.

Proposition

$b'(\gamma) = \mathbb{E}_\gamma[t(Y)]$.

Proposition

For an iid sample $(Y_1, \ldots Y_n)$, *the MLE* $\widehat{\gamma}$ *of* γ *satisfies*

$$b'(\widehat{\gamma}) = \frac{1}{n}\sum_{i=1}^{n} t(Y_i) =: \bar{t}(Y).$$

This shows that the MLE $\widehat{\gamma}$ is also the moment estimate of γ based on the mean of the sufficient statistics.

Proof in appendix slides 81 and 83.

EM for the exponential family

Proposition

If all emission distributions $\mathcal{F}_k$ belong to the exponential family with respective sufficient statistics t_k and normalizing functions a_k and b_k, the maximization in the M step results in the weighted moment estimates based on the expectation of the sufficient statistics, i.e. γ_k^{h+1} satisfies:

$$\mathbb{E}_{\gamma_k^{h+1}}[t_k(U)] = \frac{T_k^{h+1}}{N_k^{h+1}}$$

where

- $U \sim f(\cdot, \gamma_k^{h+1})$,
- $\tau_{ik}^{h+1} = \mathbb{E}_{\theta^{h+1}}[Z_{ik}|Y_i]$,
- $N_k^{h+1} = \sum_{i=1}^n \tau_{ik}^{h+1}$
- *and* $T_k^{h+1} = \sum_{i=1}^n \tau_{ik}^{h+1} t_k(Y_i)$.

Proof i

Complete-likelihood for exponential family

$$\begin{aligned}\log p_\theta(Y,Z) &= \sum_{i=1}^{n}\sum_{k=1}^{K} Z_{ik}[\log \pi_k + \log f_k(Y_i)] \\ &= \sum_{i=1}^{n}\sum_{k=1}^{K} Z_{ik}[\log \pi_k + \gamma_k^\intercal t_k(Y_i) - a_k(Y_i) - b_k(\gamma_k)]\end{aligned}$$

So conditional expectation is

$$\begin{aligned}&\mathbb{E}[\log p_\theta(Y,Z)|Y] = \\ =\ &\mathbb{E}\left[\sum_{i=1}^{n}\sum_{k=1}^{K} Z_{ik}\left[\log \pi_k - b_k(\gamma_k)\right]|Y\right] + \mathbb{E}\left[\sum_{i=1}^{n}\sum_{k=1}^{K} Z_{ik}\left[\gamma_k^\intercal t_k(Y_i) - a_k(Y_i)\right]|Y\right] \\ =\ &\sum_{k=1}^{K} N_k[\log \pi_k - b_k(\gamma_k)] + \sum_{k=1}^{K}\gamma_k^\intercal T_k - \sum_{i=1}^{n}\tau_{ik} a_k(Y_i).\end{aligned}$$

Proof ii

The derivative with respect to γ_k is null iff $b_k'(\gamma_k) = T_k/N_k$ and the result follows from the general properties of the exponential family given in Propositions slide 57.

Remarks

- $\frac{T_k^{h+1}}{N_k^{h+1}}$ is an empirical weighted moment of the Y_i
- So the estimate of γ_k resulting from Proposition slide 57 is a moment-type estimate
- Depending on the form of $\mathbb{E}_{\gamma_k}[t_k(U)]$ as a function of γ_k, this estimate can have a close form or not

Expression for some popular models

- Poisson mixture: $\mathcal{F}_k = \mathcal{P}(\gamma_k)$:

$$\widehat{\gamma}_k = \frac{1}{N_k} \sum_{i=1}^{n} \tau_{ik} Y_i.$$

- Gaussian mixture: $\mathcal{F}_k = \mathcal{N}(\mu_k, \sigma_k^2)$:

$$\widehat{\mu}_k = \frac{1}{N_k} \sum_{i=1}^{n} \tau_{ik} Y_i, \qquad \widehat{\sigma}_k^2 = \frac{1}{N_k} \sum_{i=1}^{n} \tau_{ik} (Y_i - \widehat{\mu}_k)^2.$$

- Multinomial mixture: $\mathcal{F}_k = \mathcal{M}(1; \gamma_k)$, denoting $Y_{ia} = \mathbf{1}_{\{Y_i = a\}}$:

$$\widehat{\gamma}_{ka} = \frac{1}{N_k} \sum_{i=1}^{n} \tau_{ik} Y_{ia}.$$

Fisher information and asymptotic variance of the ML

Asymptotic variance of the maximum likelihood estimate

$$\widehat{\theta} = (\widehat{\boldsymbol{\pi}}, \widehat{\gamma})$$

is provided by the Fisher information matrix I by

$$\mathbb{V}_\infty(\widehat{\theta}) = I_\theta^{-1}$$

where

$$\begin{aligned} S_\theta(\mathbf{Y}) &= \partial_\theta \log p_\theta(\mathbf{Y}) \\ I_\theta &= \mathbb{E}[S_\theta(\mathbf{Y}) S_\theta(\mathbf{Y})^\intercal] = -\mathbb{E}_{\mathbf{Y}}\left[\partial^2_{\theta^2} \log p_\theta(\mathbf{Y})\right]. \end{aligned}$$

Problem: Evaluation of $S'_\theta(\mathbf{Y}) = \partial^2_{\theta^2} \log p_\theta(\mathbf{Y})$ because $p_\theta(\mathbf{Y})$ is a sum.

[Louis, 1982] provides a convenient way to compute the Hessian matrix

$$S'_\theta(\mathbf{Y}) = \partial^2_{\theta^2} \log p_\theta(\mathbf{Y}),$$

which only uses by-products of the EM algorithm.

Proposition ([Louis, 1982])

$$\begin{aligned} S'_\theta(\mathbf{Y}) &= \mathbb{E}[S'_\theta(\mathbf{Y},\mathbf{Z})|\mathbf{Y}] + \mathbb{E}\left[S_\theta(\mathbf{Y},\mathbf{Z})S_\theta(\mathbf{Y},\mathbf{Z})^\intercal|\mathbf{Y}\right] \\ &\quad -\mathbb{E}[S_\theta(\mathbf{Y},\mathbf{Z})|\mathbf{Y}]\mathbb{E}[S_\theta(\mathbf{Y},\mathbf{Z})|\mathbf{Y}]^\intercal. \end{aligned}$$

Proof is given in Appendix on Slide 84.

Louis's formula ii

Two main interests:

- Involve the complete likelihood and can, most of the times, be easily computed (see example in Apppendix Slide 88)
- Last term null when $\theta = \widehat{\theta} = \arg\max \log p_\theta(\mathbf{Y})$.
  Indeed (see the proof Slide 84)

$$\mathbb{E}[S_\theta(\mathbf{Y},\mathbf{Z})|\mathbf{Y}] = S_\theta(\mathbf{Y}) = \frac{p'_\theta(\mathbf{Y})}{p_\theta(\mathbf{Y})}$$

  which is equal to 0 for $\theta = \widehat{\theta}$ since $p'_\theta(\mathbf{Y})|_{\widehat{\theta}} = 0$.

How many states?

- K is not known general
- A model with $K-1$ classes is nested in a model with K classes : the likelihood increases as well
- Likelihood not a relevant criterion to estimate K
- Dimension of the parameter θ increases with K.

Penalized likelihood criteria

Penalized likelihood criterion

- Let $\widehat{\theta}_K$ be the maximum likelihood estimate of θ for a model with K components:
$$\widehat{\theta}_K = \arg\max_{\theta \in \Theta_K} \log p_\theta(\mathbf{Y})$$
where Θ_K : parameter space for a K-mixture model
- Penalized likelihood estimate of K:
$$\widehat{K} = \arg\max_K \left(\log p_{\widehat{\theta}_K}(Y) - \mathrm{pen}(K) \right).$$

Bayesian information criterion

- Most commonly used criterion [Schwarz, 1978]
- Originally defined in a Bayesian framework

Three levels of hierarchy:

1. a prior distribution $p(K)$ for the number of components;
2. a conditional distribution $p(\theta|K)$ for the parameter θ given the number of components;
3. a likelihood $p_\theta(\mathbf{Y})$ which corresponds to the conditional distribution of the observations $\mathbf{Y}$ given the parameters: $p_\theta(\mathbf{Y}) = p(\mathbf{Y}|\theta, K)$.

Posterior probability of K

- Model selection problem relies on conditional distribution of K given the observations:

$$p(K|\mathbf{Y}) = \frac{p(\mathbf{Y}, K)}{p(\mathbf{Y})} = \frac{p(K)p(\mathbf{Y}|K)}{p(\mathbf{Y})}.$$

- Ideally, one would choose

$$\widehat{K} \quad = \quad \arg\max_K \; p(K|\mathbf{Y}) \; = \; \arg\max_K \; (\log p(K) + \log p(\mathbf{Y}|K))$$

- But $\log p(Y|K) = \log \int p(\mathbf{Y}|\theta, K)p(\theta|K) \, \mathrm{d}\theta$
  - Difficult to evaluate
  - Laplace approximation

Laplace approximation

Proposition (Laplace approximation)

Under regularity conditions,

$$\log p(\mathbf{Y}|K) = \log p_{\widehat{\theta}_K}(\mathbf{Y}) - \frac{d_K}{2}\log n + \mathcal{O}_n(1).$$

where d_K denotes the number of independent parameters in a model with K components.

- Detailed proof: [Lebarbier and Mary-Huard, 2004], together with precise comparative study between BIC and another popular model selection criterion: AIC.
- The term $\log p(K)$ remains fix when n grows large: neglected

BIC

Definition

$$\widehat{K}_{BIC} = \arg\max_{K} \left(\log p_{\widehat{\theta}_K}(\mathbf{Y}) - \frac{d_K}{2} \log n \right).$$

From BIC to Integrated Complete Likelihood (ICL)

Using Proposition 36

$$\log p_{\widehat{\theta_K}}(\mathbf{Y}) = \mathbb{E}_{\widehat{\theta_K}}\left[\log p_{\widehat{\theta_K}}(\mathbf{Y},\mathbf{Z})|\mathbf{Y}\right] - \underbrace{\mathbb{E}_{\widehat{\theta_K}}\left[\log p_{\widehat{\theta_K}}(\mathbf{Z}|\mathbf{Y})|\mathbf{Y}\right]}_{(1)}$$

- (1): entropy of the classification distribution
- Entropy is small when the observations are classified with reasonable confidence.
- [Biernacki et al., 2000]: account for the classification uncertainty in the selection of K
- Penalize value of K with large entropy

ICL

Definition (ICL)

$$
\begin{aligned}
\widehat{K}_{ICL} &= \arg\max_{K} \left(\log p_{\widehat{\theta_K}}(Y) - \mathcal{H}[p_{\widehat{\theta_K}}(\mathbf{Z}|\mathbf{Y})] - \frac{d_K}{2} \log n \right) \\
&= \arg\max_{K} \left(\mathbb{E}_{\widehat{\theta_K}} \left[\log p_{\widehat{\theta_K}}(\mathbf{Y}, \mathbf{Z}) | \mathbf{Y} \right] - \frac{d_K}{2} \log n \right)
\end{aligned}
$$

Unsupervised classification

- Often the main aim when using a mixture model.
- Maximum likelihood inference provides estimates of θ
- By-product of EM : conditional distribution of the hidden classes $\mathbf{Z}$ conditional to the observed data $\mathbf{Y}$

Soft classification

$$\tau_{ik} = P_{\widehat{\theta}}(Z_i = k|\mathbf{Y})$$

- Gives a measure of the confidence with which an observation could be classified into a given group
- Uncertainty of the classification summarized by:

$$\mathcal{H}[p_{\widehat{\theta}}(Z_i|\mathbf{Y})] = \mathcal{H}[p_{\widehat{\theta}}(Z_i|Y_i)] = -\sum_{k=1}^{K} \tau_{ik} \log \tau_{ik}.$$

Sometimes referred to as the *classification uncertainty*

- Entropy of the whole conditional distribution of Z given Y: sum of all the individual's uncertainties

Hard classification

When observations need to be classified into groups, the most common rule is the 'maximum a posteriori' (MAP) rule.

Definition

The MAP classification rule is given by:

$$\widehat{\mathbf{Z}} = \arg\max_{z} p_\theta(\mathbf{Z} = z|\mathbf{Y}).$$

- The MAP rule can be applied to each observation label Z_i as

$$\widehat{Z}_i = \arg\max_{k} \tau_{ik}$$

- In the case of mixture, equivalent:

$$\widehat{\mathbf{Z}} = \arg\max_{\mathbf{z}} p_\theta(\mathbf{Z} = \mathbf{z}|\mathbf{Y}) = (\widehat{Z}_i)_i$$

since the Z_i are independent conditionally on $\mathbf{Y}$.

Conclusion

- Idea really simple.
- Example of R package : mixtools
- Used in many context, even for complexe data. The emission distribution has to be adapted.
- Next chapter : Hidden Markov Models

References

Biernacki, C., Celeux, G., and Govaert, G. (2000).
Assessing a mixture model for clustering with the integrated completed likelihood.
IEEE Transactions on Pattern Analysis and Machine Intelligence, 22(7):719–725.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977).
Maximum likelihood from incomplete data via the EM algorithm.
Jr. R. Stat. Soc. B, 39:1–38.

Dunstan, P. K., Foster, S. D., Hui, F. K. C., and Warton, D. I. (2013).
Finite Mixture of Regression Modeling for High-Dimensional Count and Biomass Data in Ecology.
Journal of Agricultural, Biological, and Environmental Statistics, 18(3):357–375.

Lebarbier, E. and Mary-Huard, T. (2004).
Le critère bic : fondements théoriques et interprétation.
147.

Louis, T. A. (1982).
Finding the observed information matrix when using the EM algorithm.
J. Royal Statist. Society Series B, 44:226–233.

McLachlan, G. J. and Peel, D. (2000).
Finite mixture models.
Wiley Series in Probability and Statistics, New York.

Parraga-Alava, J., Dorn, M., and Inostroza-Ponta, M. (2018).
A multi-objective gene clustering algorithm guided by apriori biological knowledge with intensification and diversification strategies.
BioData Mining, 11(1):16.

Pritchard, J. K., Stephens, M., and Donnelly, P. (2000).
Inference of population structure using multilocus genotype data.
Genetics, 155(2):945–959.

Rosenberg, N. A. (2011).

Proposition

$b'(\gamma) = \mathbb{E}_\gamma[t(Y)]$.

Remind that the moment generating function of V

$$m(z) = \mathbb{E}[e^{z^\intercal V}]$$

with $m'(0) = \mathbb{E}(V)$

For the exponential family, consider the moment generating function of the sufficient statistics

$$\begin{aligned} m(z) &:= \mathbb{E}[e^{z^\intercal t(Y)}] = \int e^{z^\intercal t(y)} f_\gamma(y)\, \mathrm{d}y \\ &= \int \exp[(z+\gamma)^\intercal t(y) - a(y) - b(\gamma)]\, \mathrm{d}y. \end{aligned}$$

Appendix. Properties of the exponential family ii

Because f_γ is a pdf, $e^{b(\gamma)}$ is a normalizing constant:

$$\begin{aligned} \int \exp[\gamma^\intercal t(y) - a(y)]\,\mathrm{d}y &= e^{b(\gamma)} \\ \Rightarrow \qquad \int \exp[(z+\gamma)^\intercal t(y) - a(y)]\,\mathrm{d}y &= e^{b(z+\gamma)} \end{aligned}$$

so

$$m(z) = e^{-b(\gamma)} \int \exp[(z+\gamma)^\intercal t(y) - a(y)]\,\mathrm{d}y = e^{b(z+\gamma)-b(\gamma)}.$$

The result follows from the fact that
$m'(z) = b'(\gamma + z) \Rightarrow m'(0) = b'(\gamma)$. ◂ Go back to the course

Proposition

For an iid sample $(Y_1, \dots Y_n)$, the MLE $\widehat{\gamma}$ of γ satisfies

$$b'(\widehat{\gamma}) = \frac{1}{n}\sum_{i=1}^{n} t(Y_i) =: \bar{t}(Y).$$

Take the derivative of the log-likelihood

$$\sum_i \log p(Y_i; \gamma) = \sum_i [\gamma^\intercal t(Y_i) - a(Y_i)] - nb(\gamma)$$

with respect to γ.

◂ Go back to the course

Proposition ([Louis, 1982])

$$\begin{aligned} S'_\theta(Y) &= \mathbb{E}[S'_\theta(Y,Z)|Y] + \mathbb{E}\left[S_\theta(Y,Z)S_\theta(Y,Z)^\intercal|Y\right] \\ &\quad - \mathbb{E}[S_\theta(Y,Z)|Y]\mathbb{E}[S_\theta(Y,Z)|Y]^\intercal. \end{aligned}$$

Demonstration

Recalling that

$$\log p_\theta(Y) = \log\left[\sum_z p_\theta(Y,z)\right],$$

we have

$$
\begin{aligned}
S_\theta(Y) &= p'_\theta(Y)\,/\,p_\theta(Y) = \sum_z p'_\theta(Y,z)\,/\,p_\theta(Y) \\
&= \sum_z \frac{p'_\theta(Y,z)}{p_\theta(Y,z)} p_\theta(Y,z)\,/\,p_\theta(Y) = \sum_z \frac{p'_\theta(Y,z)}{p_\theta(Y,z)} p_\theta(z|Y) \\
&= \mathbb{E}\left[\frac{\partial}{\partial\theta}\log p_\theta(Y,z)\right] = \mathbb{E}[S_\theta(Y,Z)|Y]. \qquad (10)
\end{aligned}
$$

Because the second derivative of $\log f$ is

$$
(\log f)'' = \frac{f''}{f} - \left(\frac{f'}{f}\right)\left(\frac{f'}{f}\right)^{\mathsf{T}} \qquad (11)
$$

the second derivative of $\log p_\theta(Y)$ is

Appendix. Asymptotic variance iii

$$
\begin{aligned}
S'_\theta(Y) &= \frac{\partial^2}{\partial\theta^2} \log p_\theta(Y) \\
&= \frac{p''_\theta(Y)}{p_\theta(Y)} - \left[\frac{p'_\theta(Y)}{p_\theta(Y)}\right]\left[\frac{p'_\theta(Y)}{p_\theta(Y)}\right]^\intercal \\
&= \frac{\sum_z p''_\theta(Y,z)}{p_\theta(Y)} - \mathbb{E}[S_\theta(Y,Z)|Y]\mathbb{E}[S_\theta(Y,Z)|Y]^\intercal .
\end{aligned}
$$

The same trick as in (10) can be combines with (11) for the first term to get

$$\begin{aligned}
\frac{\sum_z p''_\theta(Y,z)}{p_\theta(Y)} &= \sum_z \left[\frac{p''_\theta(Y,z)}{p_\theta(Y,z)} - \left(\frac{p'_\theta(Y,z)}{p_\theta(Y,z)}\right)\left(\frac{p'_\theta(Y,z)}{p_\theta(Y,z)}\right)^\intercal + \left(\frac{p'_\theta(Y,z)}{p_\theta(Y,z)}\right)\left(\frac{p'_\theta(Y,z)}{p_\theta(Y,z)}\right)^\intercal\right] \\
&\qquad \times \underbrace{\frac{p_\theta(Y,z)}{p_\theta(Y)}}_{=p_\theta(Z|\mathbf{Y})} \\
&= \sum_z \left[\frac{\partial^2}{\partial\theta^2}\log p_\theta(Y,z) + \left(\frac{p'_\theta(Y,z)}{p_\theta(Y,z)}\right)\left(\frac{p'_\theta(Y,z)}{p_\theta(Y,z)}\right)^\intercal\right] p_\theta(z|Y) \\
&= \mathbb{E}[S'_\theta(Y,Z)|Y] + \mathbb{E}\left[S_\theta(Y,Z)S_\theta(Y,Z)^\intercal|Y\right]
\end{aligned}$$

which completes the proof.

Go back to the course

Appendix 2: Assymptotic variance for the Poisson emission distribution i

- Mixture model (1) where $F(\gamma_k) = \mathcal{P}(\gamma_k)$.
- Complete log-likelihood

$$\log p_\theta(Y, Z) = \sum_{i,k} Z_{ik} \left[\log \pi_k - \gamma_k + Y_i \log \gamma_k - \log(Y_i!)\right]$$

  where $\pi_K = 1 - \sum_{k<K} \pi_k$.
- First derivatives
  - $\partial_{\pi_k} \log p_\theta(Y, Z) = \frac{\sum_{i=1}^n Z_{ik}}{\pi_k} - \frac{\sum_{i=1}^n Z_{iK}}{\pi_K}$
  - $\partial_{\gamma_k} \log p_\theta(Y, Z) = -\sum_{i=1}^n Z_{ik} + \frac{\sum_{i=1}^n Z_{ik} Y_i}{\gamma_k}$

Appendix 2: Assymptotic variance for the Poisson emission distribution ii

- Second derivatives:
  - $\partial^2_{\pi_k^2} \log p_\theta(Y, Z) = -\frac{\sum_{i=1}^n Z_{ik}}{\pi_k^2} + \frac{\sum_{i=1}^n Z_{iK}}{\pi_K^2}$,
  - $\partial^2_{\pi_k, \pi_\ell} \log p_\theta(Y, Z) = \frac{\sum_{i=1}^n Z_{iK}}{\pi_K^2}$
  - $\partial^2_{\gamma_k^2} \log p_\theta(Y, Z) = -\frac{\sum_{i=1}^n Z_{ik} Y_i}{\gamma_k^2}$,
  - $\partial^2_{\gamma_k, \gamma_\ell} \log p_\theta(Y, Z) = 0$.

The first term of Prop slide 64 requires the calculation of the following moments, denoting here $\mathbb{E}^Y(\cdot) = \mathbb{E}(\cdot | Y)$:

$$\begin{array}{rcl} \mathbb{E}^Y\left(\sum_{i=1}^n Z_{ik}\right) & = & \sum_{i=1}^n \tau_{ik} =: N_k, \\ \mathbb{E}^Y\left(\sum_{i=1}^n Z_{ik} Y_i\right) & = & \sum_{i=1}^n \tau_{ik} Y_i =: S_k. \end{array}$$

The second term requires these of

Appendix 2: Assymptotic variance for the Poisson emission distribution iii

$$
\begin{aligned}
\mathbb{E}^Y\left[\left(\sum_{i=1}^n Z_{ik}\right)\left(\sum_{i=1}^n Z_{i\ell}\right)\right] &= \mathbb{E}^Y\left(\sum_{i=1}^n Z_{ik}Z_{i\ell} + \sum_{i\neq j} Z_{ik}Z_{j\ell}\right) \\
&= \sum_{i=1}^n \mathbb{E}^Y(Z_{ik}Z_{i\ell}) \\
&\quad + \sum_{i\neq j} \mathbb{E}^Y(Z_{ik})\mathbb{E}^Y(Z_{j\ell}) \\
&=^* \sum_{i=1}^n \mathbf{1}_{\{k=\ell\}}\tau_{ik} + \sum_{i\neq j}\tau_{ik}\tau_{j\ell} \\
&= \mathbf{1}_{\{k=\ell\}}N_k + N_k N_\ell - \sum_{i=1}^n \tau_{ik}\tau_{i\ell}, \\
\mathbb{E}\left[\left(\sum_{i=1}^n Z_{ik}Y_i\right)\left(\sum_{i=1}^n Z_{i\ell}\right)\right] &= \mathbf{1}_{\{k=\ell\}}S_k + S_k N_\ell - \sum_{i=1}^n Y_i\tau_{ik}\tau_{i\ell}, \\
\mathbb{E}^Y\left[\left(\sum_{i=1}^n Z_{ik}Y_i\right)\left(\sum_{i=1}^n Z_{i\ell}Y_i\right)\right] &= \mathbf{1}_{\{k=\ell\}}Q_k + S_k S_\ell - \sum_{i=1}^n Y_i^2\tau_{ik}\tau_{i\ell},
\end{aligned}
$$

where $Q_k = \sum_{i=1}^n Y_i^2\tau_{ik}$ and * because $Z_{ik}Z_{i\ell} = 0$ if $k \neq \ell$.

◂ Go back to the course