Notes de cours. Latent variable models and their inference via EM: applications in ecology

Sophie Donnet, Pierre Gloaguen, Stéphane Robin

January 15, 2025

Contents

Chapter 1

Introduction

Latent variable models are known to be stochastic models that relates the observable variables, namely Y, to a set of unobserved random variables Z, called latent or hidden variables. The hidden variables may have a physical meaning with respect to the observed phenomena, as it is the case in the Hidden Markov Models, where the hidden variable is an indicator of the sate of the system at each time-step. However, in other cases, the latent variables have no physical reality and are used to enrich a simpler model to fit the data, as it is the case in mixture models where the latent variables are used to classify the observations.

Latent variable models are widely used in ecology. See for instance Peyrard and Gimenez [2022].

In any case, the presence of hidden variable in the model formulation makes the inference much more difficult. Indeed, the likelihood of the observations Y involves an integration over the latent variables Z. As a consequence, the maximisation of the likelihood may be much more complex or the computation of the likelihood itself may be tricky.

The Expectation-Maximisation algorithm proposed by Dempster et al. [1977] is an iterative method to find the maximum likelihood of models involving latent variables. It has been successfully used in multiple contexts, making it a central tool in statistical inference. However, few models allow to apply it in its original version and several extensions have been developed.

In this book, we propose to present a collection of latent variable models classically used in ecology but we classify the models according to the type of EM algorithm required for model inference. The second chapter introduces the classical EM model and presents the mixture model which is typically the case where the EM algorithm can be applied directly. The third chapter is dedicated to models where the E-step of the EM algorithm requires a greater effort. The fourth and fifth chapter tackle the case where the E-step requires an approximation, respectively deterministic and stochastic.

Spatial models. The analysis of spatially organized data is a subject by itself, for which a huge amount of specific models, theory and methods have been proposed [see, e.g. Cressie, 2015, for a fairly comprehensive view]. An important feature of spatially structured data is that observations made at neighbour sites are expected to be correlated and the spatial statistics precisely aim at accounting for such a dependency structure.

Chapter 2

Latent variable models and the EM algorithm

Contents

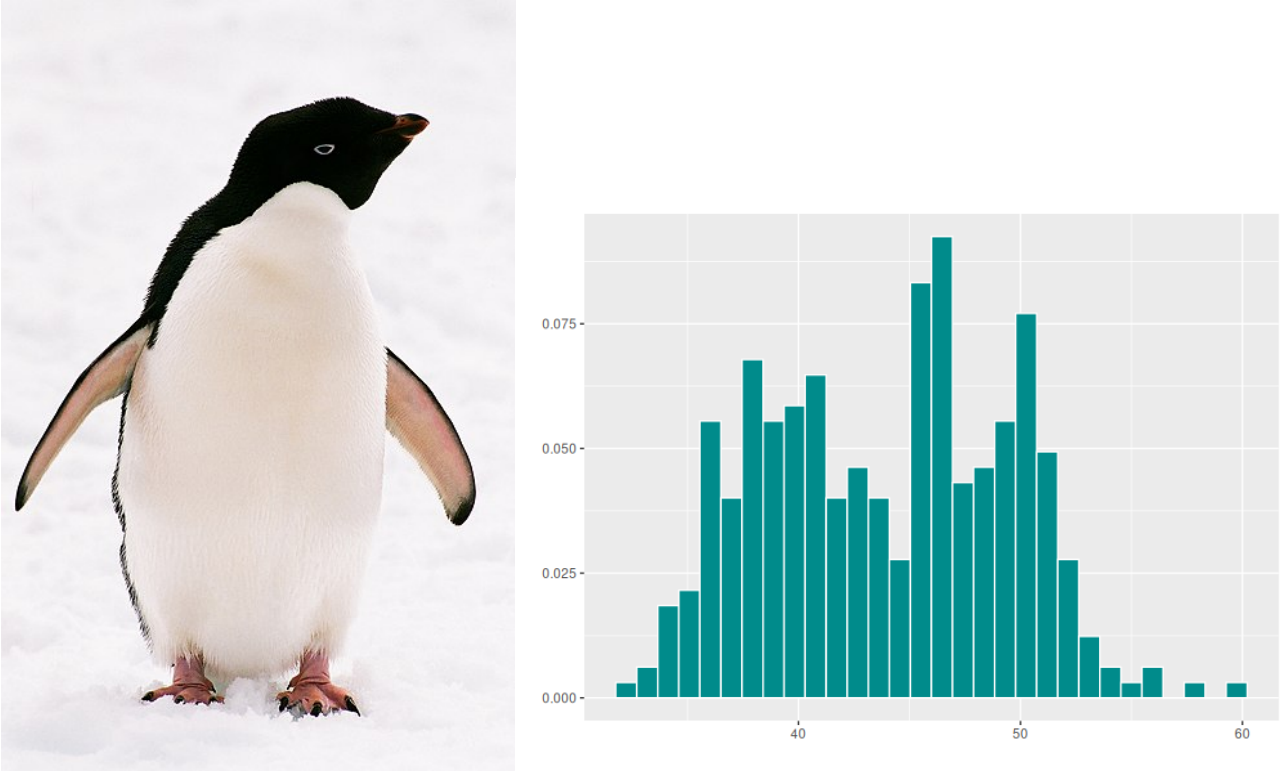
Assume that we observe $y_1, \ldots, y_n$ a set of n observations. The aim of probabilistic modelling is to propose a model $\mathcal{M}_\theta$ defined by a probability distribution function p_θ that both depend on some parameter θ (ideally of limited size) and is adapted to the data. By doing so, we assume that the observations are the realisation of the model. As an example, linear models assume that the observations are the realisation of a Gaussian distribution with expectation expressed as a linear function of covariates. Generalized linear models allows to consider non-Gaussian distributions such as Poisson, Bernoulli or binomial distributions. Latent variable models arise when one needs to use additional non observed random variables to write a model adapted to the data.

The first section of this chapter is dedicated to the formal definition of a latent variable model and is motivated by a first simple example, namely the detection of sub-populations among Palmer penguins, based on the length of their bill. The Expectation-Maximisation algorithm, which is a popular tool for the parameter estimation in latent variable models, will then be presented in Section 2.2. The asymptotic variance of the resulting estimates and criteria for model selection will be presented in the last two sections of this chapter.

2.1 Latent variable models

2.1.1 From Palmer penguins to mixture models

Dataset 1 (Palmer penguins)**.** *As a first illustrative example, let us consider the Palmer penguins dataset [Horst et al., 2020]. The dataset contains the measurement of bill length for* 342 *penguins collected from three islands in the Palmer Archipelago, Antarctica.*



(a) Adelie penguin (*Pygoscelis adeliae*)

(b) Distribution of the bill length

Figure 2.1: Palmer penguins dataset [Horst et al., 2020]. Distribution of the bill length for the complete studied population of penguins.

Our objective is to model the distribution of the bill length of Palmer penguins. Having a quick look at the histogram of the bill length (Figure 2.1, right), we can quickly convince ourselves that a simple Gaussian distribution is not adapted. Indeed, we observe several local modes (maximum) in the distribution. As it is expected that the distribution of a trait among an homogeneous population follows a Gaussian distribution, this multimodality could reflect the fact that sampled penguins come from heterogeneous populations, and that the bill length differ amongst these populations[1] This characteristic of the data naturally leads us to the (Gaussian) mixture model.

Two-component Gaussian mixture.

A mixture model is a probabilistic model assuming the presence of populations within the whole observations, without knowing to which population an individual belongs.

Let $(y_1, \dots, y_n)$ be the observations (bill lengths in our case). We assume that they are realisations of independent random variables $(Y_1, \dots, Y_n)$.
In the case of a two component mixture, each observation is supposed to belong either to population/cluster 1 or population/cluster 2. Assuming that, within each cluster, the bill lengths have a Gaussian distribution, we then state that

- if observation i belongs to cluster 1, then $Y_i \sim \mathcal{N}\big(\mu_1, \sigma_1^2\big)$ and
- if observation i belongs to cluster 2, then $Y_i \sim \mathcal{N}\big(\mu_2, \sigma_2^2\big)$.

[1] Actually, the dataset records the species and the sex of each penguin. For the purpose of this introduction, we willingly omit this information.

To make the model complete, we must specify how the individuals are spread among the two populations/clusters. To this aim, for each observation i, we introduce a binary random variable Z_i, which encodes whether observation i belongs to cluster 1 or 2 and we set that

$$\mathbb{P}(Z_i = 1) = \omega_1 \quad , \quad \mathbb{P}(Z_i = 2) = \omega_2 \,,$$

where ω_1 is the probability to belong to the first population, and $\omega_2 = 1 - \omega_1$ is the probability to belong to the second one. Considering that we only observe a realisation of the variables $\boldsymbol{Y} = (Y_1, \ldots, Y_n)$ and that we do not know their population of origin $Z_1, \ldots, Z_n$, these variables are called latent variables. In a more compact way, $\boldsymbol{Z} = (Z_i)_{1 \leq i \leq n}$ is said to be *the* latent variable of the model.

We may now define the two-component Gaussian mixture model as follows.

Model 1 (Two-component Gaussian mixture)**.**

$$\begin{aligned} &\{Z_1, \ldots Z_n\} \textit{i.i.d}: && Z_i \sim \textit{Cat}(\omega = (\omega_1, \omega_2)) \\ &\{Y_1, \ldots Y_n\} \textit{ independent } | \{Z_1, \ldots Z_n\}: && Y_i \sim \begin{cases} \mathcal{N}(\mu_1, \sigma_1^2) & \textit{if } Z_i = 1 \\ \mathcal{N}(\mu_2, \sigma_2^2) & \textit{if } Z_i = 2 \end{cases} \end{aligned}$$

where $Cat(\omega)$ stands for the categorical distribution with probability vector ω.

Note that the two conditional Gaussian distributions can be reformulated as a single one as:

$$Y_i \mid Z_i \sim \mathcal{N}\big(\mu_{Z_i}, \sigma_{Z_i}^2\big) \,.$$

where the indexes of the mean and variance are given by the latent variable value Z_i. The proposed model depends on 6 unknown quantities (parameters), namely $\omega_1, \omega_2, \mu_1, \mu_2, \sigma_1^2, \sigma_2^2$. In a concise notation, we will denote this set in a single vector θ. In this book, it will be convenient to distinguish the parameters corresponding to the distribution of the latent variable, that we will denote θ_{lat} and from those attached to the distribution of the observations (knowing the latent variables), which we will denote θ_{obs}. Here, we have:

$$\theta_{\text{lat}} = (\omega_1, \omega_2), \ \theta_{\text{lat}} = (\mu_1, \mu_2, \sigma_1^2, \sigma_2^2) \,. \quad \text{and} \quad \theta = (\theta_{\text{lat}}, \theta_{\text{obs}})$$

2.1.2 General definition of the latent variable models and vocabulary

Model 2 (Generic latent variable model)**.** *Let y be the realisation of $\boldsymbol{Y}$. $\boldsymbol{Y}$ is said to follow a latent variable model if we can write its distribution as follows:*

$$\boldsymbol{Z} \sim p_{\theta_{lat}}(\cdot), \tag{2.1}$$
$$\boldsymbol{Y} \mid \boldsymbol{Z} \sim p_{\theta_{obs}}(\cdot \mid \boldsymbol{Z}). \tag{2.2}$$

where the latent variable Z is not observed.

The first line of this model (2.2) links the data $\boldsymbol{Y}$ to the latent variables $\boldsymbol{Z}$ and is parametrized by θ_{obs}. The second line (2.1) corresponds to the latent variable part of the model and relies on the parameter θ_{lat}.

Back to the Palmer example. In the two-component Gaussian mixture, $p_{\theta_{\text{lat}}}(\cdot)$ is the product of n categorical distributions with support $\{1, 2\}$ and parameter $\theta_{\text{lat}} = \omega = (\omega_1, \omega_2)$ and $p_{\theta_{\text{obs}}}(\cdot \mid Z)$ is a product of n Gaussian distributions depending of parameters $\theta_{\text{obs}} = (\mu_1, \mu_2, \sigma_1^2, \sigma_2^2)$.

Notations. $\theta \in \Theta$ refers to the whole set of unknown parameters: $\theta = (\theta_{\text{obs}}, \theta_{\text{lat}})$. Thereafter, to avoid overloading the notations, we will often use the generic θ as follows:

$$\begin{aligned} \boldsymbol{Z} &\sim p_\theta(\cdot), \\ \boldsymbol{Y} \mid \boldsymbol{Z} &\sim p_\theta(\cdot \mid \boldsymbol{Z}) \,. \end{aligned}$$

2.1.3 Marginal and complete (log)-likelihoods

In this book, we will mostly consider (frequentist and Bayesian) inference methods based on the likelihood of the observations $p_\theta(\boldsymbol{y})$, where $\boldsymbol{y}$ stands for the observed realization of the random vector $\boldsymbol{Y}$.

2.1.3.1 Definitions

For a given realization of latent variables $\boldsymbol{z} = (z_1, \ldots, z_n)$, following the definition of the latent variable model, we are able to write the expression of the joint likelihood of $(\boldsymbol{y}, \boldsymbol{z})$:

$$p_\theta(\boldsymbol{y}, \boldsymbol{z}) = p_{\theta_{\text{obs}}}(\boldsymbol{y} \mid \boldsymbol{Z} = \boldsymbol{z}) p_{\theta_{\text{lat}}}(\boldsymbol{z}) \,. \tag{2.3}$$

This quantity is the so-called complete likelihood where "complete" refers to the fact that the observations $\boldsymbol{y}$ are enhanced by the latent variables $\boldsymbol{Z}$.

The latent variable $\boldsymbol{Z}$ being non observed, the likelihood of $\boldsymbol{y}$ in fact results from the integration of the complete likelihood over all the possibles values taken by the latent variable. Formally, the likelihood writes as:

$$p_\theta(\boldsymbol{y}) = \int_{\boldsymbol{z} \in \mathcal{Z}} p_\theta(\boldsymbol{y}, \boldsymbol{z}) \, \mathrm{d}\boldsymbol{z} = \int_{\boldsymbol{z} \in \mathcal{Z}} p_{\theta_{\text{obs}}}(\boldsymbol{y} \mid \boldsymbol{Z} = \boldsymbol{z}) p_{\theta_{\text{lat}}}(\boldsymbol{z}) d\boldsymbol{z}. \tag{2.4}$$

$p_\theta(\boldsymbol{y})$ is sometimes referred to as the marginal likelihood or incomplete data likelihood (as opposed to the complete likelihood). In this book we will prefer the marginal likelihood denomination.

2.1.3.2 Complete and marginal likelihood for the Gaussian mixture model.

For the sake of clarity, we now detail the calculation of the complete and incomplete likelihoods under Model 1, with $\boldsymbol{Z} = (Z_1, \ldots, Z_n)$ and $\boldsymbol{y} = (y_1, \ldots, y_n)$. For each observation $i \in \{1, \ldots n\}$ and each cluster $k \in \{1, 2\}$, we introduce the binary variable

$$Z_{ik} = \mathbf{1}_{\{k\}}(Z_i) = \begin{cases} 1 & \text{if} \quad Z_i = k \\ 0 & \text{otherwise} \end{cases} \tag{2.5}$$

which will prove to be very useful for calculations. Observe that, because each observation i has to belong to one of the two clusters, we have $Z_{i1} + Z_{i2} = 1$.

Proposition 1. *For the mixture of two Gaussian distribution, the expression of the complete and marginal log-likelihoods are given by the following expressions:*

$$\log p_\theta(\boldsymbol{y}, \boldsymbol{Z}) = \sum_{i=1}^{n} Z_{i1} \log \phi\left(y_i; \mu_1, \sigma_1^2\right) + Z_{i2} \log \phi\left(y_i; \mu_2, \sigma_2^2\right) + Z_{i1} \log \omega_1 + Z_{i2} \log \omega_2 \tag{2.6}$$

and

$$\log p_\theta(\boldsymbol{y}) = \sum_{i=1}^{n} \log \left[\omega_1 \, \phi\left(y_i; \mu_1, \sigma_1^2\right) + \omega_2 \phi\left(y_i; \mu_2, \sigma_2^2\right)\right] \tag{2.7}$$

where $\phi\left(x; \mu, \sigma^2\right)$ is the density of a Gaussian distribution of mean μ and variance σ^2.

Proof of Proposition 1

- **About** $p_{\theta_{\mathbf{obs}}}(\boldsymbol{y} \mid \boldsymbol{Z})$. By independence of the observations:

$$\begin{aligned} p_{\theta_{\text{obs}}}(\boldsymbol{y} \mid \boldsymbol{Z}) &= p_{\theta_{\text{obs}}}(y_1, \ldots, y_n \mid Z_1, \ldots Z_n) = \prod_{i=1}^{n} p_{\theta_{\text{obs}}}(y_i \mid Z_i) \quad \text{and} \\ \log p_{\theta_{\text{obs}}}(\boldsymbol{y} \mid \boldsymbol{Z}) &= \sum_{i=1}^{n} \log p_{\theta_{\text{obs}}}(y_i \mid Z_i). \end{aligned}$$

  Besides, we have

$$p_{\theta_{\text{obs}}}(y_i \mid Z_i = 1) = \phi\left(y_i; \mu_1, \sigma_1^2\right) \quad \text{and} \quad p_{\theta_{\text{obs}}}(y_i \mid Z_i = 2) = \phi\left(y_i; \mu_2, \sigma_2^2\right) \tag{2.8}$$

  where $\phi\left(x; \mu, \sigma^2\right)$ is the density of a Gaussian distribution of mean μ and variance σ^2. Using the binary variables Z_{ik}, these two equations can be combined into a single one as:

$$\begin{aligned} p_{\theta_{\text{obs}}}(y_i \mid Z_i) &= \left(\phi\left(y_i; \mu_1, \sigma_1^2\right)\right)^{Z_{i1}} \left(\phi\left(y_i; \mu_2, \sigma_2^2\right)\right)^{Z_{i2}} \\ \log p_{\theta_{\text{obs}}}(y_i \mid Z_i) &= Z_{i1} \log \phi\left(y_i; \mu_1, \sigma_1^2\right) + Z_{i2} \log \phi\left(y_i; \mu_2, \sigma_2^2\right). \end{aligned}$$

- **About $p_{\theta_{\text{lat}}}(\boldsymbol{Z})$.** By independence of the latent variables, we have

$$p_{\theta_{\text{lat}}}(\boldsymbol{Z}) = \prod_{i=1}^{n} p_{\theta_{\text{lat}}}(Z_i) \quad \text{and} \quad \log p_{\theta_{\text{lat}}}(Z) = \sum_{i=1}^{n} \log p_{\theta_{\text{lat}}}(Z_i)$$

with

$$\mathbb{P}_{\theta_{\text{lat}}}(Z_i = 1) = \omega_1 \quad \text{and} \quad \mathbb{P}_{\theta_{\text{lat}}}(Z_i = 2) = \omega_2. \tag{2.9}$$

As before, these equations can be reformulated into one unique equation:

$$\begin{aligned} p_\theta(Z_i) &= \omega_1^{Z_{i1}} \omega_2^{Z_{i2}} \\ \log p_\theta(Z_i) &= Z_{i1} \log \omega_1 + Z_{i2} \log \omega_2 \end{aligned}$$

- Combining the previous quantities, **the complete log-likelihood** $\log p_\theta(\boldsymbol{y}, \boldsymbol{Z})$ is equal to the following sum:

$$\begin{aligned} \log p_\theta(\boldsymbol{y}, \boldsymbol{Z}) &= \log p_{\theta_{\text{obs}}}(\boldsymbol{y} \mid \boldsymbol{Z}) + \log p_{\theta_{\text{lat}}}(\boldsymbol{Z}) \\ &= \sum_{i=1}^{n} Z_{i1} \log \phi\left(y_i; \mu_1, \sigma_1^2\right) + Z_{i2} \log \phi\left(y_i; \mu_2, \sigma_2^2\right) \\ &\quad + \sum_{i=1}^{n} Z_{i1} \log \omega_1 + Z_{i2} \log \omega_2 \end{aligned}$$

- **The marginal likelihood** $p_\theta(\boldsymbol{y})$ is obtained by integrating the latent variables:

$$\begin{aligned} p_\theta(\boldsymbol{y}) &= \prod_{i=1}^{n} p_\theta(y_i), \\ \text{where} \quad p_\theta(y_i) &= p_{\theta_{\text{obs}}}(y_i \mid Z_i = 1)\mathbb{P}_{\theta_{\text{lat}}}(Z_i = 1) + p_{\theta_{\text{obs}}}(y_i \mid Z_i = 2)\mathbb{P}_{\theta_{\text{lat}}}(Z_i = 2) \\ &= \omega_1 \, \phi\left(y_i; \mu_1, \sigma_1^2\right) + \omega_2 \phi\left(y_i; \mu_2, \sigma_2^2\right) \end{aligned}$$

Finally, we get :

$$p_\theta(\boldsymbol{y}) = \prod_{i=1}^{n} \left[\omega_1 \, \phi\left(y_i; \mu_1, \sigma_1^2\right) + \omega_2 \phi\left(y_i; \mu_2, \sigma_2^2\right)\right]$$

2.1.4 Maximum likelihood estimation

If one wants to maximise the observed likelihood with respect to θ to obtain the maximum likelihood estimation:

$$\widehat{\theta} = \arg\max_{\theta \in \Theta} \log p_\theta(\boldsymbol{y}),$$

a quick look at expression (2.7) leads us to the conclusion that the solution is not explicit (because of the logarithm of a sum of two terms). Although we could consider here any numerical optimization method, in more general cases (that we will cover throughout this book), the integral form of the likelihood of the observations in Equation (2.4) quickly becomes intractable, and the direct numerical optimisation with respect to θ becomes complicated. As an alternative to direct optimization, the EM algorithm provides an efficient tool to reach the value $\widehat{\theta}$ maximizing the likelihood by taking advantage of the latent variable structure.

2.2 The Expectation Maximisation (EM) algorithm

The EM algorithm was proposed by Dempster, Laird and Rubin in 1977 and is now one of the most cited paper of the statistical literature. Before presenting its principle, let us introduce the following notations.

- $p_\theta(\boldsymbol{z} \mid \boldsymbol{Y} = \boldsymbol{y})$ is the density of the conditional distribution of the latent variable $\boldsymbol{z}$ given the observation $\boldsymbol{Y} = \boldsymbol{y}$ for a given parameter θ. We recall that by the Bayes Formula, we have:

$$p_\theta(\boldsymbol{z} \mid \boldsymbol{Y} = \boldsymbol{y}) = \frac{p_{\theta_{\text{obs}}}(\boldsymbol{y} \mid \boldsymbol{Z} = \boldsymbol{z}) p_{\theta_{\text{lat}}}(\boldsymbol{z})}{p_\theta(\boldsymbol{y})}$$

- For any L^1-function Ψ and any couple of parameters (θ, θ'), we set:

$$\mathbb{E}_{\theta'}\left[\Psi(\boldsymbol{y}, \boldsymbol{Z}, \theta) \mid \boldsymbol{Y} = \boldsymbol{y}\right] := \int_{\boldsymbol{z} \in \mathcal{Z}} \Psi(\boldsymbol{y}, \boldsymbol{z}, \theta) p_{\theta'}(\boldsymbol{z} \mid \boldsymbol{Y} = \boldsymbol{y}) \, \mathrm{d}\boldsymbol{z}.$$

2.2.1 A central decomposition

The EM algorithm is based on the following decomposition of the incomplete data log-likelihood function $\log p_\theta(\boldsymbol{y})$.

Proposition 2 (Decomposition of the incomplete data log-likelihood)**.** *For any θ and θ'*

$$\begin{array}{rcccc} \log p_\theta(\boldsymbol{y}) & = & \mathbb{E}_{\theta'}\left[\log p_\theta(\boldsymbol{y},\boldsymbol{Z}) \mid \boldsymbol{Y}=\boldsymbol{y}\right] & - & \mathbb{E}_{\theta'}\left[\log p_\theta(\boldsymbol{Z} \mid \boldsymbol{Y}=\boldsymbol{y}) \mid \boldsymbol{Y}=\boldsymbol{y}\right]. \\ & = & Q(\theta \mid \theta') & + & H(\theta \mid \theta') \end{array} \tag{2.10}$$

Remark.

1. The decomposition of Proposition 2 is convenient for our estimation purpose because it makes a connexion between the incomplete data log-likelihood $\log p_\theta(\boldsymbol{y})$ (often intractable) and the complete data likelihood $\log p_\theta(\boldsymbol{y},\boldsymbol{z})$ (generally more manageable).
2. Note that if $\theta' = \theta$, then the second term $H(\theta \mid \theta)$ is the entropy[2] of the latent variables $\boldsymbol{Z}$ given the observed $\boldsymbol{Y}=\boldsymbol{y}$ as follows:

$$\begin{aligned} H(\theta \mid \theta) &= -\mathbb{E}_\theta[\log p_\theta(\boldsymbol{Z} \mid \boldsymbol{Y}=\boldsymbol{y}) \mid \boldsymbol{Y}=\boldsymbol{y}] \\ &= -\int_{\boldsymbol{z}\in\mathcal{Z}} \log p_\theta(\boldsymbol{z} \mid \boldsymbol{Y}=\boldsymbol{y}) p_\theta(\boldsymbol{z} \mid \boldsymbol{Y}=\boldsymbol{y})\,\mathrm{d}\boldsymbol{z} \\ &= \mathsf{Ent}[p_\theta(\boldsymbol{Z} \mid \boldsymbol{Y}=\boldsymbol{y})] \end{aligned}$$

   which leads to

$$\log p_\theta(\boldsymbol{y}) = \mathbb{E}_\theta\left[\log p_\theta(\boldsymbol{y},\boldsymbol{Z}) \mid \boldsymbol{Y}=\boldsymbol{y}\right] + \mathsf{Ent}[p_\theta(\boldsymbol{Z} \mid \boldsymbol{Y}=\boldsymbol{y})]. \tag{2.11}$$

Proof of Proposition 2

The Bayes formula states that:

$$p_\theta(\boldsymbol{Z} \mid \boldsymbol{Y}=\boldsymbol{y}) = \frac{p_\theta(\boldsymbol{y},\boldsymbol{Z})}{p_\theta(\boldsymbol{y})}.$$

So, taking the log of this expression, we obtain:

$$\log p_\theta(\boldsymbol{Z} \mid \boldsymbol{Y}=\boldsymbol{y}) = \log p_\theta(\boldsymbol{y},\boldsymbol{Z}) - \log p_\theta(\boldsymbol{y}).$$

Integrating the latent variable $\boldsymbol{Z}$ with respect to $p_{\theta'}(\boldsymbol{Z} \mid \boldsymbol{Y}=\boldsymbol{y})$ on both side leads to:

$$\begin{aligned} H(\theta \mid \theta') &= -\mathbb{E}_{\theta'}\left[\log p_\theta(\boldsymbol{Z} \mid \boldsymbol{Y}=\boldsymbol{y}) \mid \boldsymbol{Y}=\boldsymbol{y}\right] \\ &= -\mathbb{E}_{\theta'}\left[\log p_\theta(\boldsymbol{y},\boldsymbol{Z}) \mid \boldsymbol{Y}=\boldsymbol{y}\right] + \mathbb{E}_{\theta'}\left[\log p_\theta(\boldsymbol{y}) \mid \boldsymbol{Y}=\boldsymbol{y}\right] \\ &= -Q(\theta \mid \theta') + \log p_\theta(\boldsymbol{y}) \end{aligned}$$

where we go from line 2 to 3 reminding that $\log p_\theta(\boldsymbol{y})$ is a constant with respect to $\boldsymbol{Z}$ so $\mathbb{E}_{\theta'}\left[\log p_\theta(\boldsymbol{y}) \mid \boldsymbol{Y}=\boldsymbol{y}\right] = \log p_\theta(y)$. We obtain the equality of Proposition 2.

2.2.2 Principle of the EM algorithm

Now, we are looking for the maximum likelihood estimation:

$$\widehat{\theta} = \arg\max_\theta \log p_\theta(\boldsymbol{y}).$$

The algorithm EM is defined as follows.

Algorithm 1 (Expectation Maximization)**.**

- ***Initialization*** *Choose a an initial value $\theta^{(0)}$.*
- ***Iteration*** *For $h \geq 0$, let $\theta^{(h)}$ be the current value of the parameter. Repeat until convergence:*

[2] Remind that the entropy of any probability distribution q is $\mathsf{Ent}(q) = -\mathbb{E}_q[\log q(Z)]$

– ***Expectation step*** *(E-step): Compute*

$$Q(\theta \mid \theta^{(h)}) = \mathbb{E}_{\theta^{(h)}}[\log p_\theta(\boldsymbol{y}, \boldsymbol{Z}) \mid \boldsymbol{Y} = \boldsymbol{y}];$$

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

$$\theta^{(h+1)} = \arg\max_{\theta \in \Theta} Q(\theta \mid \theta^{(h)})$$

One can prove that the sequence $(\theta^{(h)})_{h \leq 0}$ increases the log-likelihood $\log p_{\theta^{(h)}}(\boldsymbol{y})$ at each iteration.

Proposition 3 (Dempster et al. [1977])**.** *The sequence $(\theta^{(h)})_{h \geq 0}$ defined by the EM Algorithm 1 is such that:*

$$\log p_{\theta^{(h+1)}}(\boldsymbol{y}) \geq \log p_{\theta^{(h)}}(\boldsymbol{y}), \quad \forall h \geq 0.$$

The proof of Proposition 3 relies on the Jensen inequality which we now remind.

Lemma 1 (Jensen inequality)**.** *Let X be an integrable real-valued random variable and let f be a convex function. Then:*

$$f(\mathbb{E}[X]) \leq \mathbb{E}[f(X)].$$

As a consequence, if f is a concave function, we have

$$f(\mathbb{E}[X]) \geq \mathbb{E}[f(X)].$$

Proof of Proposition 3

$\theta^{(h+1)}$ being the value maximizing $\theta \to Q(\theta \mid \theta^{(h)})$, we have

$$Q(\theta^{(h)} \mid \theta^{(h)}) \leq Q(\theta^{(h+1)} \mid \theta^{(h)}).$$

So, by definition of Q:

$$\mathbb{E}_{\theta^{(h)}}[\log p_{\theta^{(h)}}(\boldsymbol{y}, \boldsymbol{Z}) \mid \boldsymbol{Y} = \boldsymbol{y}] \leq \mathbb{E}_{\theta^{(h)}}[\log p_{\theta^{(h+1)}}(\boldsymbol{y}, \boldsymbol{Z}) \mid \boldsymbol{Y} = \boldsymbol{y}]$$

As a consequence,

$$\begin{aligned} \mathbb{E}_{\theta^{(h)}}[\log p_{\theta^{(h+1)}}(\boldsymbol{y}, \boldsymbol{Z}) \mid \boldsymbol{Y} = \boldsymbol{y}] - \mathbb{E}_{\theta^{(h)}}[\log p_{\theta^{(h)}}(\boldsymbol{y}, \boldsymbol{Z}) \mid \boldsymbol{Y} = \boldsymbol{y}] &\geq 0 \\ \mathbb{E}_{\theta^{(h)}}\left[\log \frac{p_{\theta^{(h+1)}}(\boldsymbol{y}, \boldsymbol{Z})}{p_{\theta^{(h)}}(\boldsymbol{y}, \boldsymbol{Z})} \mid \boldsymbol{Y} = \boldsymbol{y}\right] &\geq 0 \end{aligned} \tag{2.12}$$

Now, applying Jensen's inequality with $f = \log$, which is concave, and $X = p_{\theta^{(h+1)}}(\boldsymbol{y}, \boldsymbol{Z}) / p_{\theta^{(h)}}(\boldsymbol{y}, \boldsymbol{Z})$, we obtain:

$$\log\left(\mathbb{E}_{\theta^{(h)}}\left[\frac{p_{\theta^{(h+1)}}(\boldsymbol{y}, \boldsymbol{Z})}{p_{\theta^{(h)}}(\boldsymbol{y}, \boldsymbol{Z})} \mid \boldsymbol{Y} = \boldsymbol{y}\right]\right) \geq \mathbb{E}_{\theta^{(h)}}\left[\log \frac{p_{\theta^{(h+1)}}(\boldsymbol{y}, \boldsymbol{Z})}{p_{\theta^{(h)}}(\boldsymbol{y}, \boldsymbol{Z})} \mid \boldsymbol{Y} = \boldsymbol{y}\right] \geq 0.$$

We reformulate this left term:

$$\begin{aligned} \log\left(\mathbb{E}_{\theta^{(h)}}\left[\frac{p_{\theta^{(h+1)}}(\boldsymbol{y}, \boldsymbol{Z})}{p_{\theta^{(h)}}(\boldsymbol{y}, \boldsymbol{Z})} \mid \boldsymbol{Y} = \boldsymbol{y}\right]\right) &= \log \int \frac{p_{\theta^{(h+1)}}(\boldsymbol{y}, \boldsymbol{z})}{p_{\theta^{(h)}}(\boldsymbol{y}, \boldsymbol{z})} p_{\theta^{(h)}}(\boldsymbol{z} \mid \boldsymbol{Y} = \boldsymbol{y}) \, \mathrm{d}\boldsymbol{z} \\ &= \log \int \frac{p_{\theta^{(h+1)}}(\boldsymbol{y}, \boldsymbol{z})}{p_{\theta^{(h)}}(\boldsymbol{y}, \boldsymbol{z})} \frac{p_{\theta^{(h)}}(\boldsymbol{y}, \boldsymbol{z})}{p_{\theta^{(h)}}(\boldsymbol{y})} \, \mathrm{d}\boldsymbol{z} \\ &= \log\left(\frac{1}{p_{\theta^{(h)}}(\boldsymbol{y})} \int p_{\theta^{(h+1)}}(\boldsymbol{y}, \boldsymbol{z}) \, \mathrm{d}\boldsymbol{z}\right) \\ &= \log\left(\frac{p_{\theta^{(h+1)}}(\boldsymbol{y})}{p_{\theta^{(h)}}(\boldsymbol{y})}\right). \end{aligned}$$

From this we have that:

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

which concludes the demonstration.

Remark. (About the E step)

The E-step of Algorithm 1 states to "compute" the function $Q(\theta \mid \theta^{(h)})$. In practice, this means calculating all the quantities required to evaluate this function, which is typically expressed as either a sum or an integral. Since this function represents an expectation with respect to the distribution of $\boldsymbol{Z} \mid \boldsymbol{Y} = \boldsymbol{y}$, the central aspect of the E-step is the characterization of this distribution, as illustrated in the examples throughout this book.

2.2.3 A first application of the EM algorithm

As a first illustration, we detail the E-step and M-step for the 2 Gaussian mixture model (Model 1).

Algorithm 2 (EM for a 2 Gaussian mixture model)**.** *Starting from $\theta^{(0)}$, repeat until convergence:*

E-step. *For all $i = 1, \ldots, n$, compute:*

$$\tau_{i1}^{(h)} = \frac{\omega_1^{(h)}\phi\left(y_i; \mu_1^{(h)}, \sigma_1^{2(h)}\right)}{\omega_1^{(h)}\phi\left(y_i; \mu_1^{(h)}, \sigma_1^{2(h)}\right) + \omega_2^{(h)}\phi\left(y_i; \mu_2^{(h)}, \sigma_2^{2(h)}\right)},$$
$$\tau_{i2}^{(h)} = \frac{\omega_2^{(h)}\phi\left(y_i; \mu_2^{(h)}, \sigma_2^{2(h)}\right)}{\widehat{\omega}_1^{(h)}\phi\left(y_i; \mu_1^{(h)}, \sigma_1^{2(h)}\right) + \omega_2^{(h)}\phi\left(y_i; \mu_2^{(h)}, \sigma_2^{2(h)}\right)}.$$

M-step. *Update the estimate of θ. Define $N_1^{(h)} = \sum_{i=1}^n \tau_{i1}^{(h)}$ and $N_2^{(h)} = \sum_{i=1}^n \tau_{i2}^{(h)}$. Then*

$$\omega_1^{(h+1)} = \frac{N_1^{(h)}}{n}, \qquad \omega_2^{(h+1)} = \frac{N_2^{(h)}}{n},$$
$$\mu_1^{(h+1)} = \frac{\sum_{i=1}^n \tau_{i1}^{(h)} y_i}{N_1^{(h)}}, \qquad \mu_2^{(h+1)} = \frac{\sum_{i=1}^n \tau_{i2}^{(h)} y_i}{N_2^{(h)}},$$
$$\sigma_1^{2(h+1)} = \frac{\sum_{i=1}^n \tau_{i1}^{(h)} (y_i - \mu_1^{(h+1)})^2}{N_1^{(h)}}, \qquad \sigma_2^{2(h+1)} = \frac{\sum_{i=1}^n \tau_{i2}^{(h)} (y_i - \mu_1^{(h+1)})^2}{N_2^{(h)}}.$$

In practice, the algorithm is stopped when the parameters stabilize i.e. $\|\theta^{(h+1)} - \theta^{(h)}\| < \epsilon$ with $\epsilon = 10^{-6}$ for instance.

The $\tau_{i1}^{(h)}, \tau_{i2}^{(h)}$ are the probabilities, under parameter $\theta^{(h)}$ for each individual i to be in clusters 1 and 2, given the observation $Y_i = y_i$, also referred as the individual conditional probabilities[3]. The estimate at step h of $\mathbb{P}(Z = 1)$, namely $\omega_1^{(h)}$ is then obtained by averaging the individual conditional probabilities to be in class 1. In the same spirit $\mu_1^{(h+1)}$ and $\sigma_1^{2(h+1)}$ are the empirical mean and variance where the observations are weighted by the individual conditional probabilities to be in class 1.

To write the EM algorithm we first have to write the quantity $Q(\theta \mid \theta^{(h)})$. Then we will be able to explicit the E and M steps.

Proof of Algorithm 2

About $Q(\theta \mid \theta^{(h)})$;

We gave the expression of $\log p_\theta(\boldsymbol{y}, \boldsymbol{Z})$ in Equation (2.6). To evaluate $Q(\theta \mid \theta^{(h)})$ we have to integrate

[3]Or, in a more formal way: $\tau_{ik}^{(h)} = \mathbb{P}_{\theta^{(h)}}(Z_i = k | Y_i = y_i)$

the latent variables:

$$\begin{aligned}
Q(\theta \mid \theta^{(h)}) &= \mathbb{E}_{\theta^{(h)}}\left[\log p_\theta(\boldsymbol{y}, \boldsymbol{Z}) \mid \boldsymbol{Y} = \boldsymbol{y}\right] \\
&= \mathbb{E}_{\theta^{(h)}}\left[\sum_{i=1}^{n} Z_{i1} \log \phi\left(y_i; \mu_1, \sigma_1^2\right) + Z_{i2} \log \phi\left(y_i; \mu_2, \sigma_2^2\right) + \sum_{i=1}^{n} Z_{i1} \log \omega_1 + Z_{i2} \log \omega_2 \mid \boldsymbol{Y} = \boldsymbol{y}\right] \\
&= \sum_{i=1}^{n} \mathbb{E}_{\theta^{(h)}}\left[Z_{i1} \mid Y_i = y_i\right] \log \phi\left(y_i; \mu_1, \sigma_1^2\right) + \mathbb{E}_{\theta^{(h)}}\left[Z_{i2} \mid Y_i = y_i\right] \log \phi\left(y_i; \mu_2, \sigma_2^2\right) \\
&\quad + \sum_{i=1}^{n} \mathbb{E}_{\theta^{(h)}}\left[Z_{i1} \mid Y_i = y_i\right] \log \omega_1 + \mathbb{E}_{\theta^{(h)}}\left[Z_{i2} \mid Y_i = y_i\right] \log \omega_2 \\
&= \sum_{i=1}^{n} \left[\tau_{i1}^{(h)} \log \phi\left(y_i; \mu_1, \sigma_1^2\right) + \tau_{i2}^{(h)} \log \phi\left(y_i; \mu_2, \sigma_2^2\right)\right] + \sum_{i=1}^{n} \tau_{i1}^{(h)} \log \omega_1 + \tau_{i2}^{(h)} \log \omega_2
\end{aligned}$$

where

$$\begin{aligned}
\tau_{i1}^{(h)} &= \mathbb{E}_{\theta^{(h)}}[Z_{i1} \mid Y_i = y_i] = \mathbb{P}_{\theta^{(h)}}(Z_i = 1 \mid Y_i = \boldsymbol{y}_i) \\
\text{and} \quad \tau_{i2}^{(h)} &= \mathbb{E}_{\theta^{(h)}}[Z_{i2} \mid Y_i = y_i] = \mathbb{P}_{\theta^{(h)}}(Z_i = 1 \mid Y_i = \boldsymbol{y}_i).
\end{aligned}$$

E-step This step requires to compute the $\tau_{i1}^{(h)}$ and $\tau_{i2}^{(h)}$ defined above. Now, by the Bayes formula, we get, for $z_i \in \{1, 2\}$:

$$p_{\theta^{(h)}}(z_i \mid Y_i = \boldsymbol{y}_i) = \frac{p_{\theta^{(h)}}(y_i \mid z_i) p_{\theta^{(h)}}(Z_i)}{p_{\theta^{(h)}}(y_i)},$$

which leads to

$$\begin{aligned}
\tau_{i1}^{(h)} = \mathbb{P}_{\theta^{(h)}}(Z_i = 1 \mid Y_i = \boldsymbol{y}_i) &= \frac{\omega_1^{(h)} \phi\left(y_i; \mu_1^{(h)}, \sigma_1^{2(h)}\right)}{\omega_1^{(h)} \phi\left(y_i; \mu_1^{(h)}, \sigma_1^{2(h)}\right) + \omega_2^{(h)} \phi\left(y_i; \mu_2^{(h)}, \sigma_2^{2(h)}\right)}, \\
\tau_{i2}^{(h)} = \mathbb{P}_{\theta^{(h)}}(Z_i = 2 \mid Y_i = \boldsymbol{y}_i) &= \frac{\omega_2^{(h)} \phi\left(y_i; \mu_2^{(h)}, \sigma_2^{2(h)}\right)}{\omega_1^{(h)} \phi\left(y_i; \mu_1^{(h)}, \sigma_1^{2(h)}\right) + \omega_2^{(h)} \phi\left(y_i; \mu_2^{(h)}, \sigma_2^{2(h)}\right)}
\end{aligned}$$

using the assumptions of the model on numerator and Equation (2.7) on the denominator. Finally, we obtain:

$$Q(\theta \mid \theta^{(h)}) = \sum_{i=1}^{n} \tau_{i1}^{(h)} \left(\log \omega_1 + \log \phi\left(y_i; \mu_1, \sigma_1^2\right)\right) + \tau_{i2}^{(h)} \left(\log \omega_2 + \log \phi\left(y_i; \mu_2, \sigma_2^2\right)\right). \tag{2.13}$$

M-step We are now able to maximise this last equation (2.13) with respect to θ by setting its partial derivatives to zero.

- We first compute the derivative with respect to ω_1 (remember that $\omega_2 = 1 - \omega_1$):

$$\begin{aligned}
\frac{\partial}{\partial \omega_1} Q(\theta \mid \theta^{(h)}) = \sum_{i=1}^{n} \tau_{i1}^{(h)} \frac{1}{\omega_1} - \tau_{i2}^{(h)} \frac{1}{1 - \omega_1} = 0 \qquad &\Leftrightarrow \qquad \frac{\sum_{i=1}^{n} \tau_{i1}^{(h)}}{\omega_1} = \frac{\sum_{i=1}^{n} \tau_{i2}^{(h)}}{1 - \omega_1} \\
\omega_1^{(h+1)} = \frac{\sum_{i=1}^{n} \tau_{i1}^{(h)}}{\sum_{i=1}^{n} \underbrace{\tau_{i1}^{(h)} + \tau_{i2}^{(h)}}_{=1}} &= \frac{\sum_{i=1}^{n} \tau_{i1}^{(h)}}{n}
\end{aligned}$$

  $\omega_1^{(h+1)}$ is obtained by averaging the individual conditional probabilities to be in cluster 1. The symmetric formula holds for $\omega_2^{(h+1)}$:

$$\omega_2^{(h+1)} = \frac{\sum_{i=1}^{n} \tau_{i2}^{(h)}}{n},$$

  so we may check that $\omega_2^{(h+1)} + \omega_2^{(h+1)} = 1$.

- We now consider the optimization with respect $(\mu_1, \mu_2, \sigma_1^2, \sigma_2^2)$. We recall that:

$$\log \phi\left(y_i; \mu, \sigma^2\right) = -\frac{1}{2}\left[\log(2\pi) + \log(\sigma^2) + \frac{(y_i - \mu)^2}{\sigma^2}\right]$$

So

$$\frac{\partial}{\partial \mu} \log \phi\left(y_i; \mu, \sigma^2\right) = \frac{y_i - \mu}{\sigma^2} \qquad \text{and} \qquad \frac{\partial}{\partial \sigma^2} \log \phi\left(y_i; \mu, \sigma^2\right) = -\frac{1}{2}\left[\frac{1}{\sigma^2} - \frac{(y_i - \mu)^2}{(\sigma^2)^2}\right].$$

Using the expression (2.13), we obtain

$$\begin{aligned}
& \tfrac{\partial}{\partial \mu_1} Q(\theta \mid \theta^{(h)}) && = 0 \\
\Leftrightarrow \quad & \sum_{i=1}^{n}\left[\tau_{i1}^{(h)} \frac{\partial}{\partial \mu_1} \log \phi\left(y_i; \mu_1, \sigma_1^2\right)\right] && = 0 \\
\Leftrightarrow \quad & \sum_{i=1}^{n}\left[\tau_{i1}^{(h)} \frac{y_i - \mu_1}{\sigma_1^2}\right] && = 0 \\
\Leftrightarrow \quad & \sum_{i=1}^{n} \tau_{i1}^{(h)} y_i && = \sum_{i=1}^{n} \tau_{i1}^{(h)} \mu_1 \\
\Leftrightarrow \quad & \mu_1^{(h+1)} && = \frac{\sum_{i=1}^{n} \tau_{i1}^{(h)} y_i}{\sum_{i=1}^{n} \tau_{i1}^{(h)}}.
\end{aligned}$$

The same formula holds for $\mu_2^{(h+1)}$:

$$\mu_2^{(h+1)} = \frac{\sum_{i=1}^{n} \tau_{i2}^{(h)} y_i}{\sum_{i=1}^{n} \tau_{i2}^{(h)}}.$$

Finally, the derivation with respect to σ_1^2 leads to:

$$\begin{aligned}
& \tfrac{\partial}{\partial \sigma_1^2} Q(\theta \mid \theta^{(h)}) && = 0 \\
\Leftrightarrow \quad & \sum_{i=1}^{n}\left[\tau_{i1}^{(h)} \frac{\partial}{\partial \sigma_1^2} \log \phi\left(y_i; \mu_1, \sigma_1^2\right)\right] && = 0 \\
\Leftrightarrow \quad & -\tfrac{1}{2}\sum_{i=1}^{n} \tau_{i1}^{(h)}\left[\frac{1}{\sigma_1^2} - \frac{(y_i - \mu_1)^2}{(\sigma_1^2)^2}\right] && = 0 \\
\Leftrightarrow \quad & \sum_{i=1}^{n} \tau_{i1}^{(h)} && = \sum_{i=1}^{n} \tau_{i1}^{(h)} \frac{(y_i - \mu_1)^2}{\sigma_1^2} \\
\Leftrightarrow \quad & \sigma_1^{2(h+1)} && = \frac{\sum_{i=1}^{n} \tau_{i1}^{(h)} (y_i - \mu_1^{(h+1)})^2}{\sum_{i=1}^{n} \tau_{i1}^{(h)}}.
\end{aligned}$$

The symmetric formula holds for $\sigma_2^{2(h+1)}$:

$$\sigma_2^{2(h+1)} = \frac{\sum_{i=1}^{n} \tau_{i2}^{(h)} (y_i - \mu_2^{(h+1)})^2}{\sum_{i=1}^{n} \tau_{i2}^{(h)}}.$$

2.2.4 Convergence

General properties There is no general guarantee about the convergence of the EM algorithm towards the MLE $\widehat{\theta}$. The only property we demonstrated before is that the sequence $(\log p_{\theta^{(h)}}(\boldsymbol{y}))_{h \geq 0}$ is non decreasing. The convergence properties of the EM algorithm are discussed in detail by Wu [1983] and McLachlan and Krishnan [2008]. In particular, the convergence towards the unique maximum likelihood is established if the likelihood is unimodal and under differentiability conditions with respect to θ. In the case of the Gaussian mixture Model 1, because the labels of the clusters can be exchanged, the likelihood $p_\theta(\boldsymbol{y})$ is the same for $\theta = (\omega_1, \omega_2, \mu_1, \mu_2, \sigma_1^2, \sigma_2^2)$ and $\theta' = (\omega_2, \omega_1, \mu_2, \mu_1, \sigma_2^2, \sigma_1^2)$, so the likelihood function is not unimodal. If the complete log-likelihood belongs to the exponential family (as defined in the Appendix section A.2.1) with compact parameter space, all the limit points of any EM sequence (i.e. starting from any initial point) are stationary points of the likelihood function. In some cases, it is shown that it is possible for the algorithm to converge to local minima or saddle points. This algorithm is very sensible to the initialisation point as will be illustrated here after. In practice, it will be initialized on many starting values or with carefully chosen ones.

Illustration of the convergence of the EM algorithm on the 2 Gaussian mixture The previously presented EM has been implemented to estimate the parameters for the dataset plotted in Figure 2.1. Table 2.1 provides 5 initial points $\theta^{(0)}$ and the likelihood reached by the EM starting from theses points. The initial points 1,2,4,5, and 6 lead to the same value of the likelihood ($p_{\widehat{\theta}}(y) = -1043.56$) and to the same value of parameter $\widehat{\theta}$ (not shown here). However, the third initial point does not allow the EM algorithm to reach the global maximum. In Figure 2.2 we represent the log-likelihood along the iterations of the algorithm starting from the initial points $\theta^{(0)}$ reported in Table 2.1.

	$\theta^{(0)}$					
	$\mu_1^{(0)}$	$\mu_2^{(0)}$	$\sigma_1^{2(0)}$	$\sigma_2^{2(0)}$	$\omega_1^{(0)}$	$\log p_{\widehat{\theta}}(y)$
Init 1	40.00	50.00	5.00	5.00	0.50	-1043.56
Init 2	20.00	50.00	5.00	5.00	0.50	-1043.56
Init 3	35.00	70.00	5.00	5.00	0.60	-1053.44
Init 4	50.00	40.00	10.00	10.00	0.40	-1043.56
Init 5	40.00	50.00	1.00	1.00	0.50	-1043.56
Init 6	39.07	48.49	3.00	3.00	0.50	-1043.56

Table 2.1: **EM for 2 Gaussian mixture**. Log-likelihood $\log p_{\widehat{\theta}}(y)$ reached by the EM starting from the various values of $\theta^{(0)}$.

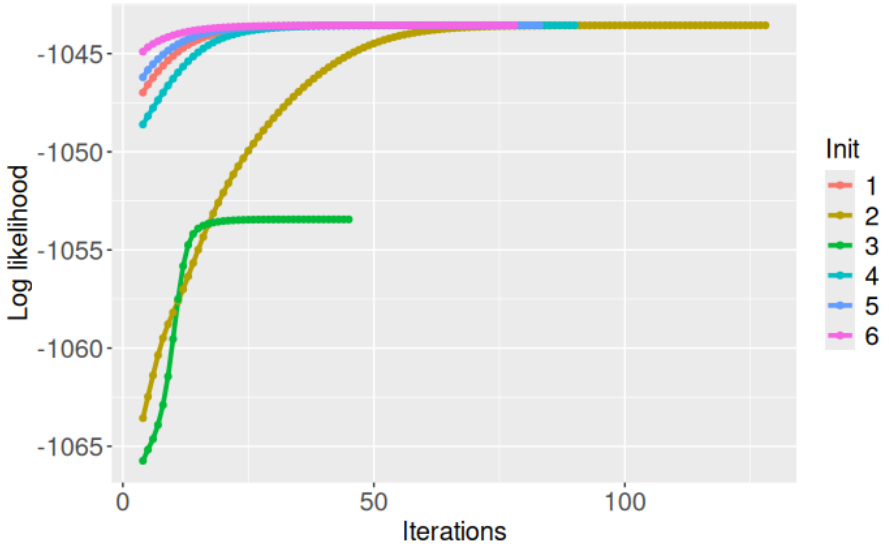


Figure 2.2: **EM for the estimation of a Gaussian mixture on pinguins dataset**. Evolution of the log-likelihood along the iterations for several in initial points provided in Table 2.1.

In order to illustrate a bit more the behaviour of the EM algorithm, we consider the special case where the parameters $(\omega_1, \sigma_1^2, \sigma_2^2)$ are known and fixed to their true value and we only want to estimate (μ_1, μ_2). For simplicity, for this particular experiment, we simulated data $\boldsymbol{y}$ with parameters $\omega_1 = 0.36$, $\sigma_2^2 = \sigma_1^2 = 9$ and $(\mu_1, \mu_2) = (35, 45)$

When only estimating μ_1 and μ_2, we can represent the 2 variables-function $(\mu_1, \mu_2) \mapsto \log p_\theta(y)$ with a heat map as represented in Figure 2.3. The true parameter $(\mu_1^\star, \mu_2^\star)$ is represented by a black point at coordinates $(35, 45)$. It is visibly close to the global maximum of the likelihood. In this figure we also observe a local maximum at coordinates about $(48, 35)$ and a saddle point at around $(40, 40)$. On the same plot, we plotted the trajectories of various EM algorihtm starting respectively at the values provided in the legend. The initial values 4 and 5 lead the EM to the global maximum (-1197.48). Starting from the initial points 2 and 3, the EM converges to a local maximum, while starting at the first initial value, the algorithm reaches a saddle point.

As a consequence, in practice the EM algorithm must be initialized on several points and/or well chosen points, trying to explore various regions of the parameter space.

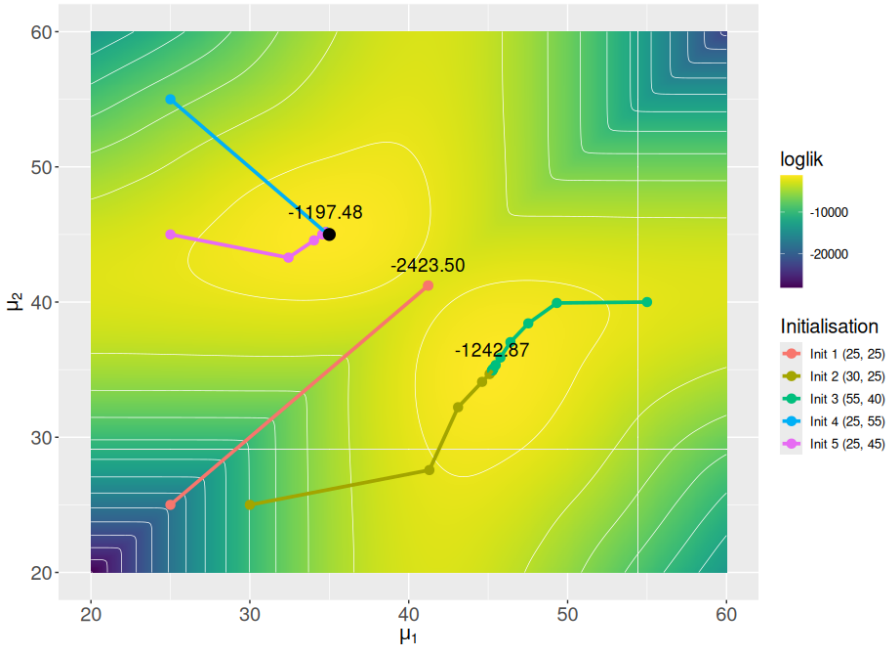


Figure 2.3: **EM for 2 Gaussian mixture**. Heatmap for the function $(\mu_1, \mu_2) \mapsto \log p_\theta(y)$ with the other parameters fixed to their true value. The 5 lines show several trajectories starting from various initial values. The black point is the true parameters used to simulate the data.

2.3 Evaluation of the asymptotic variance of the MLE

Asymptotic normality When providing an estimate for a parameter, it is critical to also give a confidence interval. The statistical theory of maximum likelihood estimator states that an asymptotic confidence interval can be deduced from the Fisher information. More precisely, assume that the observations y_i are the realization of independent and identically distributed variables

$$Y_i \overset{\text{i.i.d}}{\sim} p_{\theta^*}(\cdot)$$

where θ^* is the true parameter. Then, under regularity assumptions on p_{θ^*}, the maximum likelihood estimator $\widehat{\theta}_n$ is asymptotically Gaussian:

$$\sqrt{n}(\widehat{\theta}_n - \theta^*) \xrightarrow[n\to\infty]{\mathcal{L}} \mathcal{N}(0, I(\theta^*)^{-1})$$

where $I(\theta)$ is the Fisher information matrix defined hereafter. From this asymptotic result, we able to say that: $\sqrt{n}(\widehat{\theta}_n - \theta^*) \approx_{n\to\infty} \mathcal{N}(0, I(\theta^*)^{-1})$.. However, θ^* is unknown but it can be estimated by $\widehat{\theta}_n$. $\widehat{\theta}_n$ begin consistent, we have:

$$\sqrt{n}(\widehat{\theta}_n - \theta^*) \approx_{n\to\infty} \mathcal{N}(0, I(\widehat{\theta}_n)^{-1}) \tag{2.14}$$

We now have to explicit the Fisher information $I(\theta)$ for any value of θ.

About the Fisher information $I(\theta)$ is defined as:

$$I(\theta) = \mathbb{E}_\theta\left[S_\theta(Y) S_\theta(Y)^\intercal\right] \quad \text{where} \quad S_\theta(Y) = \nabla_\theta \log p_\theta(Y) = \frac{\nabla_\theta p_\theta(Y)}{p_\theta(Y)}\,. \tag{2.15}$$

where the expectation $\mathbb{E}_\theta[\cdot]$ is with respect to $Y \sim p_\theta$. $S_\theta(y)$ is called the score function. Under certain regularity conditions, the Fisher information matrix may also be written as

$$I(\theta) = -\mathbb{E}_\theta\left[\text{Hess}_\theta \log p_\theta(Y)\right] = -\mathbb{E}_\theta\left[\mathbf{J}_\theta S_\theta(Y)\right]$$

where Hess_θ and $\mathbf{J}_\theta$ are respectively the Hessian and Jacobian operator[4] .

[4]Remember that if $\psi : \Theta \mapsto \mathbb{R}$, the gradient of ψ, denoted $\nabla_\theta \psi(\theta)$, is the vector of the partial derivatives of ψ with respect to each component of θ. If $\Phi : \Theta \mapsto \mathbb{R}^k$, $\mathbf{J}_\theta \Phi(\theta)$ is the Jacobian i.e. the partial derivatives of each component of Φ with respect to each component of θ. $\text{Hess}_\theta\, \psi(\theta)$ is the Hessian matrix i.e. the matrix of the second derivatives of ϕ : $\text{Hess}_\theta\, \psi(\theta) = \mathbf{J}_\theta \nabla_\theta \psi(\theta)$.

Calculation the Fisher information for latent variable models When dealing with latent variable models, the integration against $p_\theta(\cdot)$ is complicated since, as said before this quantity is already an integral against the latent variables:

$$\mathbb{E}_\theta[\psi(Y)] = \int_y \psi(y) p_\theta(y)\,\mathrm{d}y = \int_{y,Z} \psi(y) p_\theta(y,Z)\,\mathrm{d}y\,\mathrm{d}z\,.$$

As a consequence, in practice, this expectation is often replaced by an empirical mean over all the observations, i.e. $I(\theta)$ is approximated by $\widehat{I}(\theta)$:

$$\widehat{I}(\theta) = -\frac{1}{n}\sum_{i=1}^{n} \mathrm{Hess}_\theta \log p_\theta(y_i) = -\frac{1}{n}\sum_{i=1}^{n} \mathbf{J}_\theta S_\theta(y_i)$$

If the observations of independent, this quantity can be reformulated as:

$$\begin{aligned}\widehat{I}(\theta) &= -\frac{1}{n}\,\mathrm{Hess}_\theta\left[\sum_{i=1}^{n}\log p_\theta(y_i)\right] = -\frac{1}{n}\mathbf{J}_\theta\left[\sum_{i=1}^{n} S_\theta(y_i)\right] \\ &= -\frac{1}{n}\,\mathrm{Hess}_\theta\left[\log p_\theta(\boldsymbol{y})\right] = -\frac{1}{n}\mathbf{J}_\theta S_\theta(\boldsymbol{y}) \end{aligned} \tag{2.16}$$

where $\log p_\theta(\boldsymbol{y})$ is the likelihood of the vector of observations $\boldsymbol{y} = (y_1, \dots, y_n)$.

Still, in latent variable models, the function $\log p_\theta(\boldsymbol{y})$ is defined as an integral over the latent variable, and so may be complicated to evaluate, and so are its derivatives. The Louis's formulae [Louis, 1982] provides a convenient way to compute $\mathbf{J}_\theta S_\theta(\boldsymbol{y})$ in the case of latent variable models which only uses by-products of the EM algorithm.

Proposition 4 (Louis [1982])**.** *Provided that differentiation and integration can be exchanged and that all given integrals are finite, the following equalities hold:*

$$S_\theta(\boldsymbol{y}) = \mathbb{E}_\theta\left[S_\theta(\boldsymbol{y},\boldsymbol{Z}) \mid \boldsymbol{Y} = \boldsymbol{y}\right] \quad \textit{where} \quad S_\theta(\boldsymbol{y},\boldsymbol{Z}) = \nabla_\theta \log p_\theta(\boldsymbol{y},\boldsymbol{Z}) \tag{2.17}$$

where the expectation is over the latent variables $Z \sim p_\theta(\cdot \mid \boldsymbol{Y} = \boldsymbol{y})$ and

$$\begin{aligned}\mathbf{J}_\theta S_\theta(\boldsymbol{y}) &= \mathbb{E}_\theta[\mathbf{J}_\theta S_\theta(y,Z) \mid \boldsymbol{Y} = \boldsymbol{y}] + \mathbb{E}_\theta\left[S_\theta(\boldsymbol{y},\boldsymbol{Z}) S_\theta(\boldsymbol{y},\boldsymbol{Z})^\top \mid \boldsymbol{Y} = \boldsymbol{y}\right] - \mathbb{E}_\theta[S_\theta(\boldsymbol{y},\boldsymbol{Z}) \mid \boldsymbol{Y} = \boldsymbol{y}]\,\mathbb{E}_\theta[S_\theta(\boldsymbol{y},\boldsymbol{Z}) \mid \boldsymbol{Y} = \boldsymbol{y}]^\top \\ &= \mathbb{E}_\theta[\mathbf{J}_\theta S_\theta(y,Z) \mid \boldsymbol{Y} = \boldsymbol{y}] + \mathbb{V}_\theta[S_\theta(\boldsymbol{y},\boldsymbol{Z}) \mid \boldsymbol{Y} = \boldsymbol{y}].\end{aligned}$$

The first result (2.17) is referred as the Louis' trick. Note that the formulation of Proposition 4 presents the main advantage that it relies on the complete likelihood and can, most of the times, be easily computed. We provide in the next chapter an example of such calculation for the ZIP model (see page 37). The proof of this proposition is provided at the end of this Section. We first comment how such a result can be used in practice.

Providing confidence intervals in practice From results in Equations (2.14) and (2.16) we get that

$$\mathbb{V}[\widehat{\theta}_n] \approx_{n\to\infty} \frac{1}{n}\left[\widehat{I}(\widehat{\theta}_n)\right]^{-1} \quad \text{with} \quad \widehat{I}(\widehat{\theta}_n) = -\frac{1}{n}\mathbf{J}_\theta S_{\widehat{\theta}_n}(\boldsymbol{y})$$

which can be reformulated using Proposition 4 with the integrated complete likelihood. So in practice, we have to compute the complete log-likelihood, its first and second derivatives and integrate the latent variables in theses quantities. Once this matrix is calculated, it has to be inverted to get the variance-covariance matrix.

- The term is null when $\theta = \widehat{\theta} = \arg\max_\theta \log p_\theta(\boldsymbol{y})$. Indeed

  $$\mathbb{E}_\theta[S_\theta(\boldsymbol{y},\boldsymbol{Z})|Y = y] = S_\theta(\boldsymbol{y}) = \frac{\nabla_\theta p_\theta(\boldsymbol{y})}{p_\theta(\boldsymbol{y})},$$

  which is equal to 0 for $\theta = \widehat{\theta}$ since $\nabla_\theta p_\theta(\boldsymbol{y})\big|_{\theta=\widehat{\theta}} = 0$

Proof of Proposition 4

Recalling that $\log p_\theta(\boldsymbol{y}) = \log\left[\int_{\boldsymbol{z}\in\mathcal{Z}} p_\theta(\boldsymbol{y},\boldsymbol{Z})\,\mathrm{d}\boldsymbol{z}\right]$, we have

$$\begin{aligned}
S_\theta(\boldsymbol{y}) &= \frac{\nabla_\theta p_\theta(\boldsymbol{y})}{p_\theta(\boldsymbol{y})} = \frac{\nabla_\theta \int_{\boldsymbol{z}\in\mathcal{Z}}(p_\theta(\boldsymbol{y},\boldsymbol{z}))\,\mathrm{d}\boldsymbol{z}}{p_\theta(\boldsymbol{y})} \\
&= \frac{\int_{\boldsymbol{z}\in\mathcal{Z}} \nabla_\theta p_\theta(\boldsymbol{y},\boldsymbol{z})\,\mathrm{d}\boldsymbol{z}}{p_\theta(\boldsymbol{y})} \qquad \text{provided we can intervert } \textstyle\int \text{ and } \nabla \\
&= \int_{\boldsymbol{z}\in\mathcal{Z}} \frac{\nabla_\theta p_\theta(\boldsymbol{y},\boldsymbol{z})}{p_\theta(\boldsymbol{y},\boldsymbol{z})}\frac{p_\theta(\boldsymbol{y},\boldsymbol{z})}{p_\theta(\boldsymbol{y})}\,\mathrm{d}\boldsymbol{z} \\
&= \int_{\boldsymbol{z}\in\mathcal{Z}} \frac{\nabla_\theta p_\theta(\boldsymbol{y},\boldsymbol{z})}{p_\theta(\boldsymbol{y},\boldsymbol{z})} p_\theta(\boldsymbol{Z}\mid\boldsymbol{Y}=\boldsymbol{y})\,\mathrm{d}\boldsymbol{z} \\
&= \int_{\boldsymbol{z}\in\mathcal{Z}} \nabla_\theta \log p_\theta(\boldsymbol{y},\boldsymbol{z}) p_\theta(\boldsymbol{z}\mid\boldsymbol{Y}=\boldsymbol{y})\,\mathrm{d}\boldsymbol{z} \\
&= \mathbb{E}_\theta\left[S_\theta(\boldsymbol{y},\boldsymbol{Z})\mid\boldsymbol{Y}=\boldsymbol{y}\right] \qquad \text{where} \qquad S_\theta(\boldsymbol{y},\boldsymbol{Z}) = \nabla_\theta \log p_\theta(\boldsymbol{y},\boldsymbol{Z})
\end{aligned}$$

thus proving the first part of Proposition 4. Now, let us compute $\mathbf{J}_\theta S_\theta(y)$. Because the Hessian matrix of $\log f$ is

$$\mathrm{Hess}(\log f) = \frac{\mathrm{Hess}\, f}{f} - \left(\frac{\nabla f}{f}\right)\left(\frac{\nabla f}{f}\right)^\top, \tag{2.18}$$

then the Hessian of $\log p_\theta(\boldsymbol{y})$ is

$$\begin{aligned}
\mathbf{J}_\theta S_\theta(\boldsymbol{y}) &= \mathrm{Hess}_\theta[\log p_\theta(\boldsymbol{y})] \\
&= \frac{\mathrm{Hess}_\theta\, p_\theta(\boldsymbol{y})}{p_\theta(\boldsymbol{y})} - \left[\frac{\nabla_\theta p_\theta(\boldsymbol{y})}{p_\theta(\boldsymbol{y})}\right]\left[\frac{\nabla_\theta p_\theta(\boldsymbol{y})}{p_\theta(\boldsymbol{y})}\right]^\top \\
&= \frac{\mathrm{Hess}_\theta\, p_\theta(\boldsymbol{y})}{p_\theta(\boldsymbol{y})} - S_\theta(\boldsymbol{y})S_\theta(\boldsymbol{y})^\top \quad \text{from Equation (2.15)} \\
&= \frac{\int_{\boldsymbol{z}\in\mathcal{Z}} \mathrm{Hess}_\theta\, p_\theta(\boldsymbol{y},\boldsymbol{z})\,\mathrm{d}\boldsymbol{z}}{p_\theta(\boldsymbol{y})} - \mathbb{E}_\theta[S_\theta(\boldsymbol{y},\boldsymbol{Z})\mid\boldsymbol{Y}=\boldsymbol{y}]\mathbb{E}_\theta[S_\theta(\boldsymbol{y},\boldsymbol{Z})\mid\boldsymbol{Y}=\boldsymbol{y}]^\top
\end{aligned} \tag{2.19}$$

where the last line uses the Louis' trick. We now concentrate on the first term of Equation (2.19). The same trick used to demonstrate the Louis' trick can be combined with (2.18) to get

$$\begin{aligned}
\frac{\int_{\boldsymbol{z}\in\mathcal{Z}} \mathrm{Hess}_\theta\, p_\theta(\boldsymbol{y},\boldsymbol{Z})\,\mathrm{d}\boldsymbol{z}}{p_\theta(\boldsymbol{y})} &= \int_{\boldsymbol{z}\in\mathcal{Z}} \frac{\mathrm{Hess}_\theta\, p_\theta(\boldsymbol{y},\boldsymbol{z})}{p_\theta(\boldsymbol{y},\boldsymbol{z})} \underbrace{\frac{p_\theta(\boldsymbol{y},\boldsymbol{z})}{p_\theta(\boldsymbol{y})}}_{=p_\theta(z\mid\boldsymbol{Y}=\boldsymbol{y})}\,\mathrm{d}\boldsymbol{z} \\
&= \int_{\boldsymbol{z}\in\mathcal{Z}} \left[\frac{\mathrm{Hess}_\theta\, p_\theta(\boldsymbol{y},\boldsymbol{z})}{p_\theta(\boldsymbol{y},\boldsymbol{z})} - \frac{\nabla_\theta p_\theta(\boldsymbol{y},\boldsymbol{z})}{p_\theta(\boldsymbol{y},\boldsymbol{z})}\left(\frac{\nabla_\theta p_\theta(\boldsymbol{y},\boldsymbol{z})}{p_\theta(\boldsymbol{y},\boldsymbol{z})}\right)^\top\right] p_\theta(z\mid\boldsymbol{Y}=\boldsymbol{y})\,\mathrm{d}z \\
&\quad + \int_{\boldsymbol{z}\in\mathcal{Z}} \left[\frac{\nabla_\theta p_\theta(\boldsymbol{y},\boldsymbol{z})}{p_\theta(\boldsymbol{y},\boldsymbol{z})}\left(\frac{\nabla_\theta p_\theta(\boldsymbol{y},\boldsymbol{z})}{p_\theta(\boldsymbol{y},\boldsymbol{z})}\right)^\top\right] p_\theta(\boldsymbol{z}\mid\boldsymbol{Y}=\boldsymbol{y})\,\mathrm{d}\boldsymbol{z} \\
&= \int_{\boldsymbol{z}\in\mathcal{Z}} \left[\mathrm{Hess}_\theta \log p_\theta(\boldsymbol{y},\boldsymbol{z}) + S_\theta(\boldsymbol{y},\boldsymbol{z})S_\theta(\boldsymbol{y},\boldsymbol{z})^\top\right] p_\theta(\boldsymbol{z}\mid\boldsymbol{Y}=\boldsymbol{y})\,\mathrm{d}\boldsymbol{z} \\
&= \mathbb{E}_\theta[\mathbf{J}_\theta S_\theta(\boldsymbol{y},\boldsymbol{Z})\mid\boldsymbol{Y}=\boldsymbol{y}] + \mathbb{E}_\theta\left[S_\theta(\boldsymbol{y},\boldsymbol{Z})S_\theta(\boldsymbol{y},\boldsymbol{Z})^\top\mid\boldsymbol{Y}=\boldsymbol{y}\right]
\end{aligned} \tag{2.20}$$

which completes the proof, combining Equations (2.19) and (2.20). Finally, noting that

$$\mathbb{E}_\theta\left[S_\theta(\boldsymbol{y},\boldsymbol{Z})S_\theta(\boldsymbol{y},\boldsymbol{Z})^\top\mid\boldsymbol{Y}=\boldsymbol{y}\right] - \mathbb{E}_\theta[S_\theta(\boldsymbol{y},\boldsymbol{Z})\mid\boldsymbol{Y}=\boldsymbol{y}]\,\mathbb{E}_\theta[S_\theta(\boldsymbol{y},\boldsymbol{Z})\mid\boldsymbol{Y}=\boldsymbol{y}]^\top = \mathbb{V}_\theta[S_\theta(\boldsymbol{y},\boldsymbol{Z})\mid\boldsymbol{Y}=\boldsymbol{y}]$$

then $\mathbf{J}_\theta S_\theta(\boldsymbol{y})$ can be reformulated as in the last equation of the proposition.

2.4 Model selection for latent variable models

In many practical situations, different models can be considered to describe the data at hand and/or answer the question of interest. For example, when several covariates are available, the question is to decide which covariates should be kept in the model, and which can be removed. The question can also arise in latent variable models: the typical question for mixture models is to decide how many components/clusters are needed to describe the

observed data.

It is quite intuitive that the maximized likelihood $p_{\widehat{\theta}}(\boldsymbol{y})$ is not a relevant criterion to choose among models as, by construction, models involving a larger set of parameters will achieve a higher maximized likelihood. Model selection must therefore make a balance between the fit to the data –which can be measured by $p_{\widehat{\theta}}(\boldsymbol{y})$– and some measure of the complexity of the model. In this section, we make a quick review of the most used criteria.

All the criteria presented here assume that the observations $y_1, \dots, y_n$ are the realisations of independent and identically distributed Y_i. In case where this assumption does not hold anymore, then the criteria have to be adapted.

2.4.1 Akaike's Information Criterion (AIC) (SR).

One of the most popular model selection criterion is due to Akaike [1973] and consists in a penalized version of the log-likelihood, the penalty being proportional to the number of parameters. More specifically, considering a model m involving D_m independent parameters (gathered in θ_m), the AIC is defined as

$$\mathrm{AIC}(m) \coloneqq \log p_{\widehat{\theta}_m}(\boldsymbol{y}) - D_m. \tag{2.21}$$

When considering a collection of models $\mathcal{M} = \{m_1, m_2, \dots\}$, the best model according to Akaike's criterion is the one with highest AIC. The reader may refer to the original paper [Akaike, 1973] or to Lebarbier and Mary-Huard [2006] for the precise derivation of this criterion and for the justification of the penalty. Note that in the case of latent variable models, if $\log p_{\widehat{\theta}_m}(\boldsymbol{y})$ can not be computed easily, the use of AIC will be compromised.

Remark. Note that, in many publications (including the original one), the criterion is expressed as $2D_m - 2\log p_{\widehat{\theta}_m}(\boldsymbol{y})$, so the selected model corresponds to the lowest value. For all the criteria presented here, we adopt the common penalized log-likelihood representation of Equation (2.21).

2.4.2 Bayesian Information Criterion (BIC)

Bayesian viewpoint. The model selection problem can be easily stated in a Bayesian framework, as proposed by Schwarz [1978]. The model M itself is considered as a random variable taken from a finite set of models $\mathcal{M}$. The full model is then built in the following way.

Model 3 (Bayesian setting for model selection)**.** *The complete model involves three steps:*

1. *the model M is drawn from $\mathcal{M} = \{m_1, m_2, \dots\}$, with distribution $p(M)$,*
2. *the parameter set θ is drawn with conditional distribution $p(\theta \mid M)$ and*
3. *the data set $\boldsymbol{Y}$ is drawn conditionally on the parameter, with distribution $p(\boldsymbol{Y} \mid \theta, M)$.*

The full joint distribution is then $p(\boldsymbol{Y}, \theta, M) = p(M)p(\theta \mid M)p(\boldsymbol{Y} \mid \theta_M, M)$.

In the Bayesian literature, the distribution $p(M)$ is named the prior distribution (or simply 'the prior') over the models and $p(\theta \mid M)$ the prior of the model parameters (conditionnal to the model). In the framework of Model 3, the model selection problem can be translated into the determination of the conditional probability (also called 'posterior' probability) of a model $m \in \mathcal{M}$ given the observed data set $\boldsymbol{y}$, that is

$$p(M = m \mid \boldsymbol{Y} = \boldsymbol{y}) = \frac{1}{p(\boldsymbol{y})} \int p(\boldsymbol{y}, \theta, m)\, \mathrm{d}\theta_m\,,$$

where the normalizing constant $p(\boldsymbol{y}) = \sum_{m \in \mathcal{M}} \int p(\boldsymbol{y}, \theta, m)\, \mathrm{d}\theta$ does not depend on m. As a consequence, the models can be compared based on the numerator $\int p(\boldsymbol{y}, \theta, m)\, \mathrm{d}\theta$ alone.

The BIC criterion. The BIC is a first order approximation of the logarithm of the integral $\int p(\boldsymbol{y}, \theta, m)\, \mathrm{d}\theta$. More specifically, denoting by $\widehat{\theta}_m = \arg\max_\theta p(\boldsymbol{y} \mid \theta, M = m)$ the maximum likelihood estimate of θ under model m, the Laplace approximation given in Lemma 3 from Appendix A.4.1 yields

$$\log\left(\int p(\boldsymbol{y}, \theta_m, m)\, \mathrm{d}\theta_m\right) = \log p(\boldsymbol{y} \mid \widehat{\theta}_m, m) - D_m \frac{\log n}{2} + O_n(1), \tag{2.22}$$

the dominant term of which defines the BIC:

$$\mathrm{BIC}(m) := \log p(\boldsymbol{y} \mid \widehat{\theta}_m, m) - D_m \frac{\log n}{2}.$$

A sketch of proof of Equation (2.22) in the case of independent and identically distributed observations is given in Appendix A.4.1. A precise derivation of BIC (and a comparison with the Akaike Information Criterion, AIC, recalled at the end of this section) can be found in Lebarbier and Mary-Huard [2006]. Still, this criterion can be used if we are able to compute $\log p(\boldsymbol{y} \mid \widehat{\theta}_m, m)$, which is not always the case in latent variable models, as we will see in the next chapters.

2.4.3 Integrated Completed Likelihood (ICL)

Models with latent variables. In presence of a latent variable $\boldsymbol{Z}$, the model selection problem can be stated in the framework of the Bayesian Model 4.

Model 4 (Bayesian setting for model selection with latent variables)**.** *The complete model involves three steps:*

1. *the model M is drawn from $\mathcal{M} = \{m_1, m_2, \dots\}$, with ('prior') distribution $p(M)$,*
2. *the parameter set θ is drawn with conditional ('prior') distribution $p(\theta \mid M)$,*
3. *the set of latent variables $\boldsymbol{Z}$ is drawn conditionally on the parameter, with distribution $p(\boldsymbol{Z} \mid \theta, M)$.*
4. *the data set $\boldsymbol{Y}$ is drawn conditionally on the parameter and the latent, with distribution $p(\boldsymbol{Y} \mid \boldsymbol{Z}, \theta, M)$.*

The full joint distribution is then $p(\boldsymbol{Y}, \boldsymbol{Z}, \theta, M) = p(M)p(\theta \mid M)p(\boldsymbol{Z} \mid \theta, M)p(\boldsymbol{Y} \mid \boldsymbol{Z}, \theta, M)$.

In the framework of Model 4, the integral from Equation (2.22) becomes

$$\iint p(\boldsymbol{y}, \boldsymbol{z}, \theta, m)\, \mathrm{d}\boldsymbol{z}\, \mathrm{d}\theta = \iint p(\boldsymbol{y} \mid \boldsymbol{z}, \theta, m) p(\boldsymbol{z}, \mid \theta, m) p(\theta \mid m)\, \mathrm{d}\boldsymbol{z}\, \mathrm{d}\theta$$

so we have to deal with the additional integration over the latent variables $\boldsymbol{Z}$.

In some specific cases, such as mixture models (Section **??**) or multivariate Poisson log-normal models (Section **??**), we may resort to the distribution of each observation y_i, conditional on θ and m, but marginalized over the corresponding Z_i:

$$p(y_i, \theta, m) = \int p(y_i, z_i, \theta, m)\, \mathrm{d}z_i,$$

which brings us back to the setting of preceding paragraph, so the Laplace approximation still holds and the BIC is defined in the same way. In the general case, the Laplace approximation of the integral $\iint p(\boldsymbol{y}, \boldsymbol{z}, \theta, m)\, \mathrm{d}\boldsymbol{z}\, \mathrm{d}\theta$ is more intricate. Some such examples will be studied in Sections **??**, **??** and **??**.

Laplace approximation of the complete likelihood. To circumvent the additional difficulty induced by the additional integration over the latent variable $\boldsymbol{Z}$, one may directly consider the so-called (log-)integrated complete likelihood:

$$\log \int p(\boldsymbol{y}, \boldsymbol{z}, \theta, m)\, \mathrm{d}\theta$$

where, as before, complete means that $\boldsymbol{z}$ is supposed to be known in some way. Biernacki et al. [2000] apply the Laplace approximation from Lemma 3 (Appendix A.4.1) to get

$$\log\left(\int p(\boldsymbol{y}, \boldsymbol{z}, \theta, m)\, \mathrm{d}\theta\right) = \log p(\boldsymbol{y}, \boldsymbol{z} \mid \widehat{\theta}_m, m) - D_m \frac{\log n}{2} + O_n(1). \tag{2.23}$$

Since the latent variables $\boldsymbol{z}$ are not observed, they have to be "estimated" or "integrated", giving rise to two versions of the Integrated Completed Likelihood criteria (ICL).

ICL critrerion. In the context of mixture models, Biernacki et al. [2000] propose to simply set $\boldsymbol{Z}$ to its posterior mode

$$\widehat{\boldsymbol{z}} = \arg\max_{z} p(\boldsymbol{y}, \boldsymbol{z} \mid \theta = \widehat{\theta}_m, M = m)$$

and define the ICL criterion as the dominant term of Equation (2.23):

$$\mathrm{ICL}_1(m) \coloneqq \log p(\boldsymbol{y}, \widehat{\boldsymbol{z}} \mid \widehat{\theta}_m, m) - D_m \frac{\log n}{2}. \tag{2.24}$$

McLachlan and Peel [2000] propose an alternative version of ICL, where $\log p(\boldsymbol{y}, \boldsymbol{z} \mid \theta = \widehat{\theta}_m, M = m)$ is averaged over $\boldsymbol{z}$ with respect to its conditional distribution $p(\boldsymbol{Z} \mid \boldsymbol{Y} = \boldsymbol{y}, \theta = \widehat{\theta}_m, M = m)$, that is to replace $\log p(\boldsymbol{y}, \boldsymbol{z} \mid \theta = \widehat{\theta}_m, M = m)$ with $\mathbb{E}_{\widehat{\theta}_m}[\log p(\boldsymbol{y}, \boldsymbol{Z} \mid \theta = \widehat{\theta}_m, M = m) \mid \boldsymbol{Y} = \boldsymbol{y}]$.

$$\mathrm{ICL}_2(m) = \mathbb{E}_{\widehat{\theta}_m}[\log p(\boldsymbol{y}, \boldsymbol{Z} \mid \theta = \widehat{\theta}_m, M = m) \mid \boldsymbol{Y} = \boldsymbol{y}] - D_m \frac{\log n}{2}. \tag{2.25}$$

In the sequel, we will most often prefer the ICL_2 version of the ICL criterion to the ICL_1.

Comments

- Note that the ICL's are very convenient in latent variable models since the fit term is a by-product of the EM algorithm.
- Note that, interestingly, thanks to the decomposition (2.11), the resulting ICL_2 criterion can be reformulated as:

$$\begin{aligned}\mathrm{ICL}_2(m) &= \log p(\boldsymbol{y} \mid \theta = \widehat{\theta}_m, M = m) - \mathsf{Ent}[p(\boldsymbol{Z} \mid \boldsymbol{Y} = \boldsymbol{y}, \theta = \widehat{\theta}_m, M = m] - D_m \log(n)/2 \\ &= BIC(m) - \mathsf{Ent}[p(\boldsymbol{Z} \mid \boldsymbol{Y} = \boldsymbol{y}, \theta = \widehat{\theta}_m, M = m)]\end{aligned}$$

  The BIC penalty only refers to the complexity of the model m ($\mathrm{pen}(m) = D_m \log(n)/2$), whereas the ICL criterion also penalizes the conditional entropy of the latent variable $\boldsymbol{Z}$, that is for the uncertainty about $\boldsymbol{Z}$ given the observed data $\boldsymbol{y}$. Depending on the problem at hand, either BIC or ICL can be preferred.

2.4.4 Summary

Although derived in a Bayesian framework, both BIC and ICL criteria are widely used for model selection in a frequentist setting. The notations are then slightly different as the likelihood $p(\boldsymbol{y} \mid \theta = \widehat{\theta}_m, M = m)$ is then denoted $p_{\widehat{\theta}_m}(\boldsymbol{y})$ and the entropy $\mathsf{Ent}[p(\boldsymbol{Z} \mid \boldsymbol{Y} = \boldsymbol{y}, \theta = \widehat{\theta}_m, M = m)]$ becomes $\mathsf{Ent}[p_{\widehat{\theta}_m}(\boldsymbol{Z} \mid \boldsymbol{Y} = \boldsymbol{y})]$. The BIC and ICL criteria then adopt their most common form, which is similar to that of the AIC, namely:

$$\begin{aligned}AIC(m) &= \log p_{\widehat{\theta}_m}(\boldsymbol{y}) - D_m, \\ BIC(m) &= \log p_{\widehat{\theta}_m}(\boldsymbol{y}) - D_m \frac{\log n}{2}, \\ ICL_2(m) &= \log p_{\widehat{\theta}_m}(\boldsymbol{y}) - D_m \frac{\log n}{2} - \mathsf{Ent}[p_{\widehat{\theta}_m}(\boldsymbol{Z} \mid \boldsymbol{Y} = \boldsymbol{y})].\end{aligned} \tag{2.26}$$

Note that because the AIC penalty is $\log(n)/2$ times smaller than the BIC penalty, AIC will tend to select more complex models that BIC, especially when the number of observations is large.

Observe that, likewise several other model selection criteria, both are penalized criterion model selection criterion in the sense that are defined as the difference between the maximized log-likelihood under model m ($\log p_{\widehat{\theta}_m}(\boldsymbol{y})$) and a model-specific penalty.

Chapter 3

Explicit E step

Contents

In this chapter, we present a set of models for which it is possible to apply the EM algorithm directly. The models we chose to present fall into two main categories depending on the nature of the latent variables.

- Sections 3.1, 3.2 and 3.3 are dedicated to models where the latent variables take a finite number of values, namely the multivariate Gaussian mixutre model, the Zero-inflated Poisson model and a mixture model for genetic data.
- In Section **??** we present the linear mixed-effects models where the latent variables are continuous and more precisely Gaussian.

Each section is structured as follows. First, we introduce a motivating dataset derived from the fields of ecology or evolution. Next, we propose a probabilistic model and provide a detailed calculation of its complete and marginal likelihoods. The E-step and M-step are then thoroughly explained, and any potential model selection issues are addressed. Finally, we perform statistical inference on the previously presented dataset.

3.1 Multivariate Gaussian mixture model for species clustering

3.1.1 Data and question

To identify patterns and structures within ecological data, a common statistical technique is clustering. Clustering aims to group similar observations together to summarize potentially complex datasets into simpler ones. From

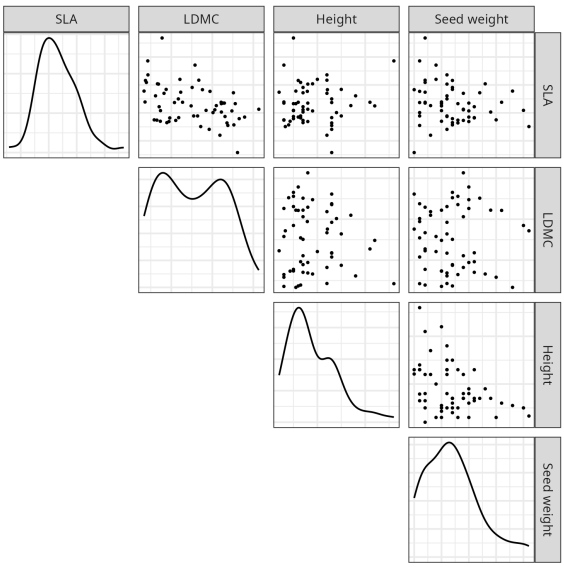


Figure 3.1: Pair plot of vegetal species characteristics. The diagonal shows the empirical density of each variable. Scales are omitted as they are not relevant for illustraing our point. See Table 3.1 for signification of variables.

a statistical point of view, we expect that i) observations within a group are "alike," and ii) two different groups are well separated. From an ecological perspective, another constraint is that the created clusters remain meaningful. For instance, in community ecology, clustering groups species or individuals with similar ecological characteristics (phenotypical or functional traits, for instance) into communities, which then become the object of study. Another example would be the clustering of habitats, which is useful for ecosystem management, as well as providing simplified inputs for species distribution models.

Dataset 2 (Meadow vegetal species traits)**.** *Lepš et al. [2011] consider a study of vegetation composition in meadows of Bohemia, Czech Republic. Four specific traits, namely the specific leaf area (SLA), the leaf dry matter content (LDMC), the reproductive plant height and the seed weight are measured over 58 species* [1]*. Our objective is the identify groups of species that share similar traits. An extract of the table corresponding to this dataset is Table 3.1. They are plotted in Figure 3.1.*

Species	SLA	LDMC	Height	Seed weight
Sp1	28.30	275.90	60.50	0.00
Sp2	27.70	294.30	39.50	0.10
Sp3	23.80	227.40	105.00	0.20
Sp4	30.40	166.20	45.00	1.30
Sp5	28.40	187.40	75.00	0.70
Sp6	23.30	322.00	27.50	0.50

Table 3.1: Traits of vegetal species of Bohemia meadow respectively the specific leaf area, the leaf dry matter content, the reproductive plant height and the seed weight. Data is an extract of Lepš et al. [2011], available in the `R` (R Core Team [2022]) package `traitor` (Götzenberger [2015]).

3.1.2 Gaussian mixture model

Clustering can be performed using multiple techniques such as hierarchical clustering or k-means clustering (see Everitt et al. [2011] and the references therein). In this chapter we focus on the most popular model based approach, the finite Gaussian mixture. In this probabilistic setting, we assume that the observed data are realizations of random variables, whose distribution depend on a latent variable. This latent variable belongs to a finite set of size K, where K is the number of clusters in our data (the number of communities).

In this context, when the number of clusters K is known/fixed, the goal is to i) retrieve the clusters from observed data and ii) characterize the distribution of observed data within each cluster. A third objective is then to estimate K. We let this important objective for the end of the section, and first assume that K is known.

[1] We considered the data of the package `traitor` package ([Götzenberger, 2015]) from which we removed two outliers.

Notations Let's assume we have observed data $(y_1, y_2, \dots, y_n)$ in $\mathbb{R}^{d_y}$. In our example, $n = 74$ and $d_y = 6$. We suppose that these are realizations from **independent** random variables $Y_1, \dots, Y_n$, defined as the marginals of independent joint random variables $(Y_1, Z_1), \dots (Y_n, Z_n)$ such that, for $1 \leq i \leq n$,

- $Z_i \in \{1, \dots, K\}$, and, for each $k \in \{1, \dots, K\}$:

$$\mathbb{P}(Z_i = k) = \omega_k \tag{3.1}$$

  where ω_k are unknown probabilities, therefore satisfying, for each k, $\omega_k > 0$ and $\sum_{k=1}^{K} \omega_k = 1$. We denote $\boldsymbol{\omega} = (\omega_1, \dots, \omega_K)$: this parameter gives the proportions of the clusters in the entire population. In what follows, we use the categorical distirbution to write equation (3.1)

$$Z_i \sim \mathrm{Cat}(\boldsymbol{\omega}).$$

  Using the general notations of Chapter 2, $\boldsymbol{\omega}$ parametrizes the distribution of the latent variables and so

$$\theta_{\mathrm{lat}} = \{\boldsymbol{\omega}\}.$$

- $Y_i \mid \{Z_i = k\} \sim \mathcal{N}(\mu_k, \Sigma_k)$, where μ_k and Σ_k are respectively the expectation and the covariance matrix of the observations of cluster k. μ_k is a vector of dimension d_y ($\mu_k \in \mathbb{R}^{d_y}$) while Σ_k is $d_y \times d_y$ inversible symetric matrix ($\Sigma_k \in \mathbf{S}_+^{d_y}$, positive definite). We denote

$$\boldsymbol{\mu} = (\mu_1, \dots, \mu_K) \quad , \quad \boldsymbol{\Sigma} = (\Sigma_1, \dots, \Sigma_K) \quad \text{and} \quad \theta_{\mathrm{obs}} = \{\boldsymbol{\mu}, \boldsymbol{\Sigma}\}.$$

  These parameters characterizes the distribution of observations within cluster is this Gaussian setting.

In this context $\theta = \{\theta_{\mathrm{lat}}, \theta_{\mathrm{obs}}\}$ is the set of unkown parameters, to estimate using observations. The Gaussian mixture model is summarized in the following box.

Model 5 (Gaussian mixture model)**.**

$$\begin{cases} \{(Y_i, Z_i)\}_{1 \leq i \leq n} \ \textit{independent} \\ Z_i \sim \mathit{Cat}(\boldsymbol{\omega}) \\ Y_i \mid \{Z_i = z_i\} \sim \mathcal{N}_{d_y}(\mu_{z_i}, \Sigma_{z_i}) \end{cases}.$$

Graphical model. A graphical model is a type of probabilistic model where a graph encodes the conditional independence structure: nodes represent random variables, and edges signify the conditional independence relations between these variables. By examining the graph, we can understand how the joint distribution breaks down into a product of smaller components, each involving only a subset of the variables. More general properties on graphical models are provided in the Appendix section A.3. Figure 3.2 displays the oriented graphical model associated with the joint distribution $p_\theta(Y, Z)$ for Model 5, where the absence of link between components indicates the independence assumption.

Reading the DAG or by calculus of conditional probabilities, we obtain that conditionally to the observations

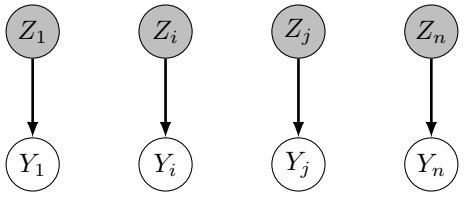


Figure 3.2: Graphical representation of the mixture model (Model 5).

$\boldsymbol{Y} = \boldsymbol{y}$, the latent variable i Z_i only depends on $Y_i = y_i$.

$$\mathbb{P}(Z_i = k \mid \boldsymbol{Y} = \boldsymbol{y}) = \mathbb{P}(Z_i = k \mid Y_i = y_i)$$

Alternative formulation using a mixture density The mixture model (or mixture distribution) can be alternatively defined by its density, refered as the mixture density.

Definition 1 (Mixture distribution)**.** *Let K be an integer greater or equal to 2. A random variable Y on $\mathbb{R}^d$ is said to have a K-mixture distribution if there exist K probability density functions $p_1(y), \dots, p_K(y)$ and weights $\omega_1, \dots, \omega_k$ satisfying*

- *for each k, $\omega_k > 0$,*
- *$\sum_{k=1}^{K} \omega_k = 1$,*

such that the probability density function of x, denoted $p(x)$ satisfies:

$$p(y) = \sum_{k=1}^{K} \omega_k p_k(y). \tag{3.2}$$

In this case, $y \mapsto p(y)$ is said to be a mixture density, and p_k is called the k-th mixture component.

Proposition 5. *Under Model 5, the marginal distribution of the observation Y_i is a mixture distribution where each mixture component is the p.d.f. of a Gaussian random variable.*

Proof of Proposition 5

Let $\mathcal{B}$ be a non-zero measure subset of $\mathbb{R}^{d_y}$, and $p_k(\cdot)$ be the p.d.f. of a $\mathcal{N}(\mu_k, \Sigma_k)$ random variable. For $1 \le i \le n$, we have:

$$\begin{aligned}\mathbb{P}(Y_i \in \mathcal{B}) &= \sum_{k=1}^{K} \mathbb{P}(Y_i \in \mathcal{B}, Z_i = k) = \sum_{k=1}^{K} \mathbb{P}(Z_i = k)\mathbb{P}(Y_i \in \mathcal{B} \mid Z_i = k) \\ &= \sum_{k=1}^{K} \omega_k \int_{\mathcal{B}} p_k(y)\,\mathrm{d}y = \int_{\mathcal{B}} \sum_{k=1}^{K} \omega_k p_k(y)\,\mathrm{d}y\,.\end{aligned}$$

3.1.3 Complete and marginal log-likelihoods

Proposition 6. *The marginal log-likelihood of the Gaussian mixture model (Model 5) is given by*

$$\log p_\theta(\boldsymbol{y}) = \sum_{i=1}^{n} \log\left(\frac{1}{(2\pi)^{\frac{d_y}{2}}} \sum_{k=1}^{K} \omega_k \left|\Sigma_k\right|^{-\frac{1}{2}} e^{-\frac{1}{2}(y_i-\mu_k)^\top \Sigma_k^{-1}(y_i-\mu_k)} \right). \tag{3.3}$$

where $|\Sigma|$ is the determinant of the matrix Σ. Its complete log-likelihood is:

$$\log p_\theta(\boldsymbol{y}, \boldsymbol{Z}) = \sum_{i=1}^{n} \sum_{k=1}^{K} Z_{ik} \left(\log \omega_k - \frac{d_y}{2} \log 2\pi - \frac{1}{2} \log |\Sigma_k| - \frac{1}{2} (y_i - \mu_k)^\top \Sigma_k^{-1} (y_i - \mu_k) \right) \tag{3.4}$$

Equation (3.3) enables, in theory, to perform maximum likelihood estimation using a numerical optimization algorithm. However, one can sense that this would be subject to high numerical instability due to the sum of exponential terms that cannot be simplified in the logarithm terms. The use of the EM algorithm here is a powerful alternative to avoid any numerical approximation and instability.

Proof of Proposition 6

Marginal log-likelihood Remind that the density of a multivariate (of dimension d_y) Gaussian distribution is:

$$\phi(y; \mu, \Sigma) = \frac{1}{(2\pi)^{\frac{d_y}{2}}} \frac{1}{\sqrt{|\Sigma|}} \exp\left\{ -\frac{1}{2} (y-\mu)^\top \Sigma^{-1} (y_i - \mu) \right\}.$$

Then, using Proposition 5, the log-likelihood of Model 5 is given by

$$\log p_\theta(\boldsymbol{y}) = \log p_\theta(y_{1:n}) = \sum_{i=1}^{n} \log p_\theta(y_i) = \sum_{i=1}^{n} \log\left(\sum_{k=1}^{K} \omega_k p_{k,\theta_{\text{obs}}}(y_i) \right).$$

Injecting the expression of the density of a multivariate Gaussian variable, we obtain;

$$\log p_\theta(\boldsymbol{y}) = \sum_{i=1}^{n} \log\left(\frac{1}{(2\pi)^{\frac{d_y}{2}}} \sum_{k=1}^{K} \omega_k \left|\Sigma_k\right|^{-\frac{1}{2}} \mathrm{e}^{-\frac{1}{2}(y_i-\mu_k)^\top \Sigma_k^{-1}(y_i-\mu_k)} \right).$$

Complete log likelihood It is convenient to express, for every $1 \leq i \leq n$, $p_\theta(y_i, Z_i)$ as a product by noticing that

$$p_\theta(y_i, Z_i) = p_{\theta_{\text{lat}}}(Z_i) p_{\theta_{\text{obs}}}(y_i|Z_i) = \begin{cases} \omega_1 p_{1,\theta_{\text{obs}}}(y_i) & \text{if } Z_i = 1 \\ \vdots & \\ \omega_K p_{K,\theta_{\text{obs}}}(y_i) & \text{if } Z_i = K \end{cases}$$

can be compacted into a unique expression as:

$$p_\theta(y_i, Z_i) = \prod_{k=1}^{K} \left(\omega_k p_{k,\theta_{\text{obs}}}(y_i)\right)^{Z_{ik}} \qquad \text{where} \qquad Z_{ik} = \mathbf{1}_{\{k\}}(Z_i).$$

The complete log-likelihood in this case writes:

$$\begin{aligned}
\log p_\theta(\boldsymbol{y}, \boldsymbol{Z}) &= \sum_{i=1}^{n} \log p_\theta(y_i, Z_i) \\
&= \sum_{i=1}^{n} \log\left[\prod_{k=1}^{K} \left(\omega_k p_{k,\theta_{\text{obs}}}(y_i)\right)^{Z_{ik}} \right] = \sum_{i=1}^{n} \prod_{k=1}^{K} Z_{ik} \log\left(\omega_k p_{k,\theta_{\text{obs}}}(y_i)\right) \\
&= \sum_{i=1}^{n} \prod_{k=1}^{K} Z_{ik} \left[\log \omega_k + \log\left(p_{k,\theta_{\text{obs}}}(y_i)\right)\right] \\
&= \sum_{i=1}^{n} \sum_{k=1}^{K} Z_{ik} \left(\log \omega_k - \frac{d_y}{2} \log 2\pi - \frac{1}{2} \log\left|\Sigma_k\right| - \frac{1}{2} \left(y_i - \mu_k\right)^\top \Sigma_k^{-1} \left(y_i - \mu_k\right) \right).
\end{aligned}$$

3.1.4 EM algorithm

The EM for the Gaussian mixture model writes as:

Algorithm 3 (EM for a 2 Gaussian mixture model)**.** *Starting from $\theta^{(0)}$, repeat until convergence:*

E-step. *For all $i = 1, \ldots, n$, and all $k = 1, \ldots, K$ compute:*

$$\tau_{ik}^{(h)} = \mathbb{P}_{\theta^{(h)}}\left(Z_i = k \mid \boldsymbol{Y} = \boldsymbol{y}\right) = \frac{\omega_k^{(h)} \left|\Sigma_k^{(h)}\right|^{-\frac{1}{2}} e^{-\frac{1}{2}\left(y_i-\mu_k^{(h)}\right)^\top (\Sigma_k^{(h)})^{-1}\left(y_i-\mu_k^{(h)}\right)}}{\sum_{\ell=1}^{K} \omega_\ell^{(h)} \left|\Sigma_\ell^{(h)}\right|^{-\frac{1}{2}} e^{-\frac{1}{2}\left(y_i-\mu_\ell^{(h)}\right)^\top (\Sigma_\ell^{(h)})^{-1}\left(y_i-\mu_\ell^{(h)}\right)}}.$$

M-step. *For all $k = 1, \ldots, K$, set:*

$$N_k^{(h)} = \sum_{i=1}^{n} \tau_{ik}^{(h)}, \tag{3.5}$$

and update the estimate of θ as

$$\omega_k^{(h+1)} = \frac{1}{n} N_k^{(h)}, \tag{3.6}$$

$$\mu_k^{(h+1)} = \frac{\sum_{i=1}^{n} \tau_{i,k}^{(h)} y_i}{N_k^{(h)}} \tag{3.7}$$

$$\Sigma_k^{(h+1)} = \frac{1}{N_k^{(h)}} \sum_{i=1}^{n} \tau_{i,k}^{(h)} \left(y_i - \mu_k^{(h+1)}\right)\left(y_i - \mu_k^{(h+1)}\right)^\top. \tag{3.8}$$

In practice, the algorithm is stopped when the parameters stabilize i.e. $\|\theta^{(h+1)} - \theta^{(h)}\| < \epsilon$ with $\epsilon = 10^{-6}$ for instance.

Remarks.

- Note that the quantity in Equation (3.6) is of great interest as it gives, for the current estimate $\theta^{(h)}$, the probability for the observation i to be in the cluster k. A natural estimator for the cluster of y_i is then the maximum a posteriori (MAP) i.e. the most probable cluster given the observation:

$$\hat{z}_i^{(h)} = \underset{k=1,\dots,K}{\arg\max}\, \tau_{ik}^{(h)} \tag{3.9}$$

- $N_k^{(h)}$ is the expected number of observations belonging to cluster k under estimate $\theta^{(h)}$. Note that

$$\sum_{k=1}^{K} N_k^{(h)} = \sum_{k=1}^{K}\sum_{i=1}^{n} \tau_{ik}^{(h)} = \sum_{i=1}^{n} \underbrace{\sum_{k=1}^{K} \tau_{ik}^{(h)}}_{=1} = n. \tag{3.10}$$

- ω_k is estimated by $\omega_k^{(h+1)}$ which is a natural estimator, as it represents the expected proportion of cluster k under the parameter $\theta^{(h)}$. Similarly, the expectation in cluster k μ_k is estimated as the empirical weighted mean of observations, where weights are their probability of being in cluster k under $\theta^{(h)}$. The same holds for the variance Σ_k.

Proof of Algorithm 3

About $Q\big(\theta|\theta^{(h)}\big)$ Suppose we have a current value $\theta^{(h)} = \left\{\boldsymbol{\omega}^{(h)}, \boldsymbol{\mu}^{(h)}, \boldsymbol{\Sigma}^{(h)}\right\}$, then, we use the expression of the complete log-likelihood (3.4) to get:

$$\begin{aligned} Q\big(\theta|\theta^{(h)}\big) &= \mathbb{E}_{\theta^{(h)}}[\log p_\theta(\boldsymbol{y}, \boldsymbol{Z}) \mid \boldsymbol{Y} = \boldsymbol{y}] \\ &= \sum_{i=1}^{n}\sum_{k=1}^{K} \mathbb{E}_{\theta^{(h)}}[Z_{ik} \mid \boldsymbol{Y} = \boldsymbol{y}] \left(\log \omega_k - \frac{d_y}{2}\log 2\pi - \frac{1}{2}\log|\Sigma_k| - \frac{1}{2}\left(y_i - \mu_k\right)^\top \Sigma_k^{-1}\left(y_i - \mu_k\right)\right). \end{aligned}$$

E step Then, we have to evaluate:

$$\tau_{ik}^{(h)} = \mathbb{E}_{\theta^{(h)}}[Z_{ik} \mid \boldsymbol{Y} = \boldsymbol{y}] = \mathbb{P}_{\theta^{(h)}}(Z_i = k \mid \boldsymbol{Y} = \boldsymbol{y}).$$

As seen in the previous chapter for a mixture of two univariate Gaussians, we use the conditional independence to write:

$$\tau_{ik}^{(h)} = \mathbb{P}_{\theta^{(h)}}(Z_i = k \mid \boldsymbol{Y} = \boldsymbol{y}) = \mathbb{P}_{\theta^{(h)}}(Z_i = k \mid Y_i = y_i).$$

Now, by the Bayes formula, we obtain:

$$\begin{aligned} \tau_{ik}^{(h)} &= \frac{\mathbb{P}_{\theta^{(h)}}(Z_i = k) p_{k,\theta^{(h)}}(y_i)}{p_{\theta^{(h)}}(y_i)} \\ &= \frac{\omega_k^{(h)} \left|\Sigma_k^{(h)}\right|^{-\frac{1}{2}} \mathrm{e}^{-\frac{1}{2}\left(y_i - \mu_k^{(h)}\right)^\top (\Sigma_k^{(h)})^{-1}\left(y_i - \mu_k^{(h)}\right)}}{\sum_{\ell=1}^{K} \omega_\ell^{(h)} \left|\Sigma_\ell^{(h)}\right|^{-\frac{1}{2}} \mathrm{e}^{-\frac{1}{2}\left(y_i - \mu_\ell^{(h)}\right)^\top (\Sigma_\ell^{(h)})^{-1}\left(y_i - \mu_\ell^{(h)}\right)}}. \end{aligned}$$

M step We denote: $N_k^{(h)} = \sum_{i=1}^{n} \tau_{ik}^{(h)}$, the expected number of observations in cluster k under estimate $\theta^{(h)}$. The M step consists in maximizing $Q\big(\theta|\theta^{(h)}\big)$ with respect to θ. Note that $Q\big(\theta|\theta^{(h)}\big)$ can be separated into $Q\big(\boldsymbol{\omega}|\theta^{(h)}\big)$ and $Q\big(\boldsymbol{\mu}, \boldsymbol{\Sigma}|\theta^{(h)}\big)$:

$$
\begin{aligned}
Q\big(\theta|\theta^{(h)}\big) &= \sum_{i=1}^{n}\sum_{k=1}^{K}\tau_{ik}^{(h)}\left(\log\omega_k - \frac{d_y}{2}\log 2\pi - \frac{1}{2}\log|\Sigma_k| - \frac{1}{2}\left(y_i-\mu_k\right)^\top\Sigma_k^{-1}\left(y_i-\mu_k\right)\right)\\
&= \sum_{k=1}^{K}\underbrace{\sum_{i=1}^{n}\tau_{ik}^{(h)}}_{N_k^{(h)}\,(3.5)}\left(\log\omega_k - \frac{d_y}{2}\log 2\pi - \frac{1}{2}\log|\Sigma_k| - \frac{1}{2}\left(y_i-\mu_k\right)^\top\Sigma_k^{-1}\left(y_i-\mu_k\right)\right)\\
&= \underbrace{\sum_{k=1}^{K}N_k^{(h)}\log\omega_k}_{Q(\boldsymbol{\omega}|\theta^{(h)})} + \underbrace{\sum_{k=1}^{K}N_k^{(h)}\left(-\frac{1}{2}\log|\Sigma_k| - \frac{1}{2}\left(y_i-\mu_k\right)^\top\Sigma_k^{-1}\left(y_i-\mu_k\right)\right)}_{Q(\boldsymbol{\mu},\boldsymbol{\Sigma}|\theta^{(h)})}.
\end{aligned}
$$

Maximization with respect to $\boldsymbol{\omega}$ We want to maximize $Q\big(\boldsymbol{\omega}|\theta^{(h)}\big)$ with respect to $\boldsymbol{\omega}$, subject to the constraint that $\sum_{k=1}^{K}\omega_k^{(h)} = 1$. For a Lagrange multiplier λ, we therefore want to find zeros of the gradient of the function

$$
Q\big(\boldsymbol{\omega},\lambda|\theta^{(h)}\big) = \sum_{k=1}^{K}N_k^{(h)}\log\omega_k - \lambda\left(\sum_{k=1}^{K}\omega_k - 1\right).
$$

We easily see that $\nabla_{\boldsymbol{\omega},\lambda}Q\big(\boldsymbol{\omega},\lambda|\theta^{(h)}\big) = 0$ is equivalent to:

$$
\begin{cases}
N_1^{(h)} &= \lambda\omega_1\\
&\vdots\\
N_k^{(h)} &= \lambda\omega_K\\
\sum_{k=1}^{K}\omega_k &= 1
\end{cases}
$$

Summing the K first rows, we have that

$$
\lambda = \frac{\sum_{k=1}^{K}N_k^{(h)}}{\sum_{k=1}^{K}\omega_k} = \frac{n}{1} = n,
$$

where the denominator equals to 1 is due to the constraint, and the numerator equal to n, as proved in Equation (3.10). It follows that the update is given by

$$
\omega_k^{(h+1)} = \frac{1}{n}N_k^{(h)}.
$$

Maximization with respect to μ_k Let's consider the gradient with respect to μ_k:

$$
\begin{aligned}
\nabla_{\mu_k}Q\big(\theta|\theta^{(h)}\big) &= -\frac{1}{2}\sum_{i=1}^{n}\tau_{i,k}^{(h)}\nabla_{\mu_k}\left(y_i-\mu_k\right)^\top\Sigma_k^{-1}\left(y_i-\mu_k\right)\\
&= -\Sigma_k^{-1}\left(\sum_{i=1}^{n}\tau_{i,k}^{(h)}\left(y_i-\mu_k\right)\right).
\end{aligned}
$$

Then, as Σ_k is positive definite:

$$
\begin{aligned}
&\nabla_{\mu_k}Q\big(\theta^{(h+1)}|\theta^{(h)}\big) &&= 0\\
\Rightarrow\quad &\textstyle\sum_{i=1}^{n}\tau_{i,k}^{(h)}\left(y_i-\mu_k^{(h+1)}\right) &&= 0\\
\Leftrightarrow\quad &\mu_k^{(h+1)} &&= \frac{\sum_{i=1}^{n}\tau_{i,k}^{(h)}y_i}{N_k^{(h)}}.
\end{aligned}
\tag{3.11}
$$

Maximization with respect to Σ_k Let's consider the derivative[a] with respect to Σ_k

$$\begin{aligned}\nabla_{\Sigma_k} Q\big(\theta|\theta^{(h)}\big) &= -\frac{1}{2}\sum_{i=1}^{n} \tau_{i,k}^{(h)} \nabla_{\Sigma_k}\left(\ln|\Sigma_k| + (y_i-\mu_k)^\top \Sigma_k^{-1}(y_i-\mu_k)\right)\\ &= -\frac{1}{2} N_k^{(h)} \nabla_{\Sigma_k}|\Sigma_k| - \frac{1}{2}\sum_{i=1}^{n} \tau_{i,k}^{(h)} \nabla_{\Sigma_k}(y_i-\mu_k)^\top \Sigma_k^{-1}(y_i-\mu_k)\\ &= -\frac{1}{2} N_k^{(h)} \Sigma_k^{-1} - \frac{1}{2}\Sigma_k^{-1}\left(\sum_{i=1}^{n} \tau_{i,k}^{(h)}(y_i-\mu_k)(y_i-\mu_k)^\top\right)\Sigma_k^{-1}\end{aligned}$$

Then, it is direct to see that:

$$\begin{aligned}&\nabla_{\Sigma_k} Q\big(\theta^{(h+1)}|\theta^{(h)}\big) = 0\\ \Rightarrow\quad &\Sigma_k^{(h+1)} = \tfrac{1}{N_k^{(h)}}\textstyle\sum_{i=1}^{n} \tau_{i,k}^{(h)}\left(y_i-\mu_k^{(h+1)}\right)\left(y_i-\mu_k^{(h+1)}\right)^\top,\end{aligned} \tag{3.12}$$

which is the empirical covariance of observations weighted by the probabilities of being in cluster k under $\theta^{(h)}$.

[a] We here use convenient results on derivatives with respect to matrices, that can be find in Petersen and Pedersen [2008]

3.1.5 About the clustering

If the clustering is at the core of the statistical analysis, one may need to affect each data to a unique cluster and so to perform a hard clustering. This can be supplied by the MAP (maximum a posteriori):

$$\widehat{\boldsymbol{z}} = (\widehat{z}_1,\dots,\widehat{z}_n) = \underset{(z_1,\dots,z_n)\in\{1,\dots K\}^n}{\arg\max} \prod_{i=1}^{n} \mathbb{P}_{\theta^*}(z_i|Y_i=y_i)$$

And so (thanks to the product form):

$$\widehat{z}_i = \underset{z_i\in\{1,\dots K\}}{\arg\max}\ \mathbb{P}(z_i|Y_i=y_i) = \underset{k\in\{1,\dots K\}}{\arg\max}\ \widehat{\tau}_{ik}\,.$$

Link with the K-means algorithm However, the reader may be familiar with the K-means algorithm for clustering. A natural question is the link between this approach and the Gaussian mixture clustering through the EM algorithm. Let's briefly recall the K-means algorithm in $\mathbb{R}^{d_y}$.

Algorithm 4 (K-means)**.** *Starting from K means $\mu_1^{(0)},\dots\mu_K^{(0)}$, and given a certain distance $d(\cdot,\cdot)$, for every $h\geqslant 0$, alternate the two following steps:*

- *Assignment: For every $1\leqslant i\leqslant n$, compute*

$$\ell_i^{(h)} = \underset{j}{\arg\min}\, d(y_i,\mu_j^{(h)}) \qquad \text{Index of the closest mean}$$
$$\tau_{i,k}^{(h)} = \mathbf{1}_{\ell_i^{(h)}}(k), \qquad k\in\{1,\dots,K\}$$

- *Update: for $k\in\{1,\dots,K\}$*

$$\mu_k^{(h+1)} = \frac{\sum_{i=1}^{n}\tau_{i,k}^{(h)} y_i}{\sum_{i=1}^{n}\tau_{i,k}^{(h)}}\,.$$

Written this way, it is clear that the EM algorithm (in the case of Gaussian mixture) is a generalization of the K-means, where the E step is an assignment using probabilities for $\tau_{i,k}^{(h)}$ for each observation, instead of a 0-1 assignement, and the M step for the mean is completely analogous to the update step, having the exact same formula. However, the k-means does not provide uncertainty measure on the clustering

Classification uncertainty On the contrary, adopting the mixture model and EM approach, the final quantities $\widehat{\tau}_{ik} = \mathbb{P}_{\widehat{\theta}}(z_i|Y_i=y_i)$ provide a probability for each observation to be in each cluster k, also called soft

clustering. The entropy of the conditional distribution of the clustering

$$\mathsf{Ent}_{\widehat{\theta}}[\boldsymbol{Z} \mid \boldsymbol{Y} = \boldsymbol{y}] = -\sum_{i=1}^{n}\sum_{k=1}^{K} \widehat{\tau}_{ik} \log \widehat{\tau}_{ik} \tag{3.13}$$

quantifies its fuzziness. $\mathsf{Ent}[p_{\widehat{\theta}}(\boldsymbol{Z} \mid \boldsymbol{Y} = \boldsymbol{y})]$ varies from 0 to $\log(K)$ in this case. If all the vectors $(\widehat{\tau}_{i1}, \ldots, \widehat{\tau}_{iK})$ are nearly equal to a vector of the type $(0, \ldots, 0, 1, 0 \ldots, 0)$, then all the observations are affected to a clear cluster and the entropy will be ≈ 0. On the contrary, if the soft clustering is much less picked, the entropy will be higher. The entropy will be maximum if $\widehat{\tau}_{ik} = \frac{1}{K}$, which corresponds to a total incertitude on the clustering (which never appends in practice).

As a conclusion, this probabilistic mixture model for clustering, on the one hand makes assumption on the distribution of the data, which should be checked, and on the other hand provides uncertainty in clustering, contrarily the the K-means procedure.

3.1.6 Choosing the number of components

The previous section showed how to perform estimation when the number of components is known. However, a crucial question remains how to estimate K? As seen in Section 2.4, the log-likelihood itself does not provide a relevent criterion because a model with $K-1$ components is nested in a model with K components, so the likelihood automatically increases when adding components in the model.
The number of components K is usually selected using penalized criteria such as AIC, BIC or ICL, defined in Equations (2.26), which all rely on both the maximized log-likelihood and the number of independent parameters, which we denote by D_K for a model with K components. In our Gaussian mixture settings, the D_K term consists in $K-1$ proportions, $K \times d_y$ parameters for means and $K \times d_y(d_y+1)/2$ free parameters for covariance matrices :

$$D_K = K \times \left(1 + d_y + \frac{d_y(d_y+1)}{2}\right) - 1\,.$$

We remind that the different criteria do not have the exact same aim. For example the ICL includes an additional penalty term accounting for the uncertainty of the classification, as measured by the conditional entropy given in (3.13). As a consequence, the ICL criterion will often yield in a smaller number of well separated components as the BIC criterion see Figure 3.3).

3.1.7 Illlustrations

Influence of the starting point First, we illustrate an important feature of the EM algorithm which is the influence of the starting parameter $\theta^{(0)}$. For $K = 3$, we chose randomly 200 different starting points and run the subsequent algorithms. The log-likelihood is monitored through the algorithm, and the algorithm stops when the increase in the log-likelihood becomes lower than 10^{-8}. Figure 3.3 shows the differences in the final log-likelihood, therefore illustrating the numbers of local maxima. It is worth noting that this is not a problem *per se*, as one can run (in parallel) the algorithm from multiple starting points, and choose the best based on the equation (3.3). This however illustrates the difficulty to find a global maxima of the likelihood in complex settings.

Choosing the number of components. The previous procedure was performed for $K \in \{1, \ldots, 6\}$, and choosing the bet final point among the 200 trials[2]. For the 6 models, we compute the 3 model selection criterion discussed above, together with the negative log-likelihood and show the results on Figure 3.4. We can notice that the AIC finds a 4 components model to be the best, while the two other criteria lead to $K = 3$, which is kept as the final model in the following.

Clustering results. For $K = 3$, the best parameter (in the sense of Figure 3.3) is chosen. Estimator (3.9) is then computed to cluster observations. The results are shown on Figure 3.5. The three clusters red green blue gather respectively 20, 7, and 31 observations. The first one gathers observations having a low lead dry matter content while the second one gathers plants with high seed weight and high LDMC. It is worth noting that this cluster might contain an outlier (having a small SLA). Gaussian mixture models are indeed popular models for anomaly detection (see Chandola et al. [2009], section 7). The third cluster is a class having no clear specificity. Such cluster gathering together disparate observations often occur in clustering.

[2]The case $K = 1$ does not require any EM algorithm, has it boils down to the simple estimation of a mean a covariance matrix.

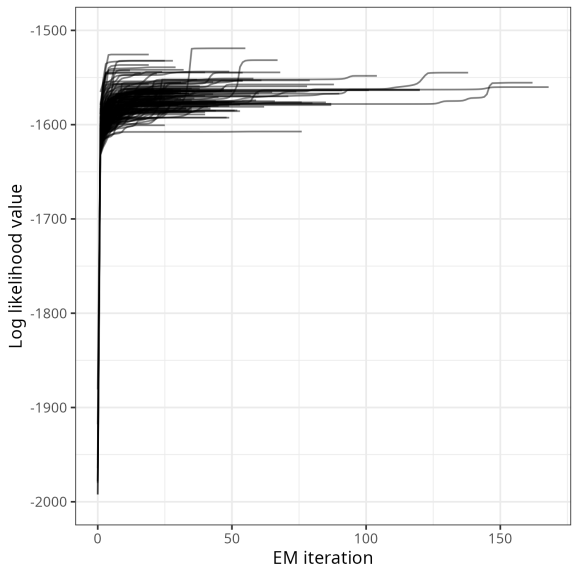


Figure 3.3: Evolution of the log-likelihood when performing EM from 200 different starting points on data set of Table 3.1. The algorithm stops when the increase of the log-likelihood is lower than 10^{-8} (hence the different number of iterations for each curve).

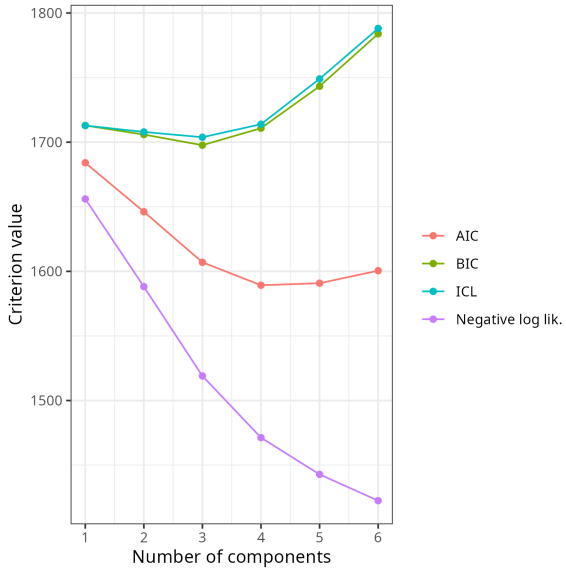


Figure 3.4: Evolution of the negative log-likelihood and the penalized likelihood criteria on the meadow vegetation data set).

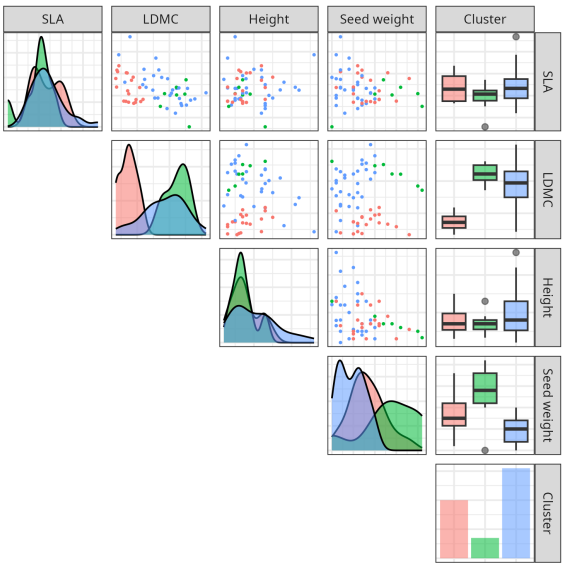


Figure 3.5: Best clustering (in terms of log-likelihood) for $K = 3$).

3.2 Zero-inflated Poisson for species distribution

3.2.1 Data and question

Species distribution models (SDM) aim at understanding how environmental conditions affect the abundance of a given species in a given site. The data are typically collected in the following way: n sites are visited and in each site i ($1 \leq i \leq n$) a d-dimensional vector x_i of environmental descriptors is recorded, as well as the number y_i of individuals of the species observed in the site.

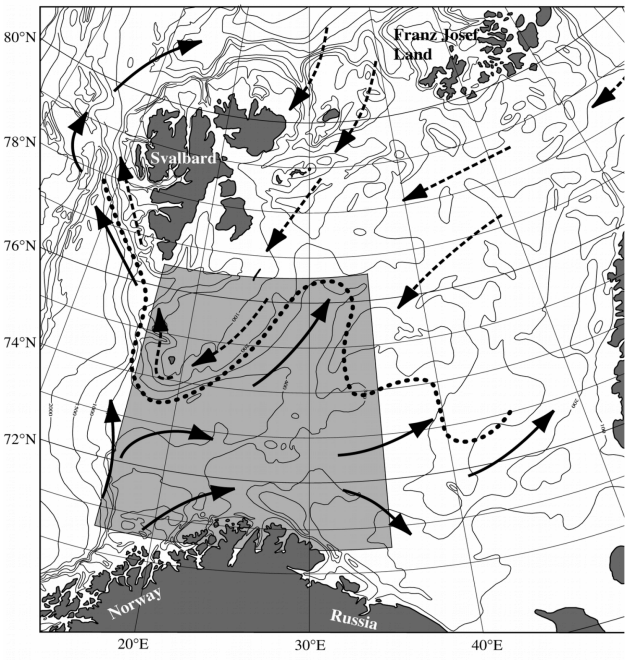


Figure 3.6: Map of the Barents sea, where data from Example 3 were collected.

Dataset 3 (Cod in the Barents sea)**.** *Fossheim et al. [2006] measured the abundance of cod (Gadus morhua) measured in $n = 89$ stations of the Barents sea. In each station, fishes were captured according to the same protocole, the latitude and longitude of each site were measured together with two environmental covariates: depth and temperature of the water. The data are available from the* `PLNmodels` *R package [Chiquet et al.,*

2021]. Figure 3.6 gives a map of the Barents sea where the data were collected. Figure 3.7 gives the first few lines of the dataset and the histogram of the observed abundances, which display a large variance and a high number of observations equal to 0: the species is actually not observed (i.e. $y_i = 0$) in $n_0 = 61$ stations.

	Latit.	Longit.	Depth	Temp.	Abundance
1	71.10	22.43	349	3.95	309
2	71.32	23.68	382	3.75	1041
3	71.60	24.90	294	3.45	218
4	71.27	25.88	304	3.65	77
5	71.52	28.12	384	3.35	13
6	71.48	29.10	344	3.65	196

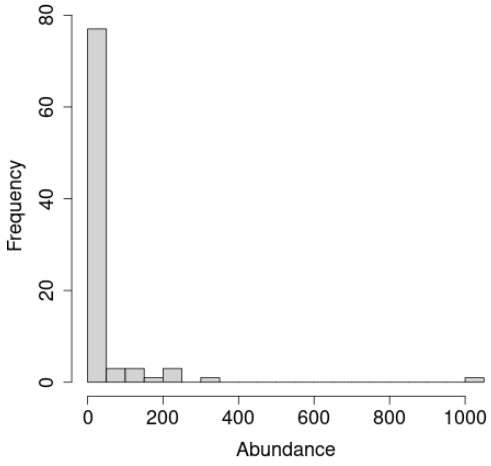


Figure 3.7: Cod abundance in the Barents sea. Left: head of the data table. Right: Histogram of the observed abundances.

Classical Poisson or logistic regression approaches. The Poisson regression model (which is special instance of generalized linear models, see Appendix A.2) provides a natural and well established framework for such count data. This model states that the sites are all independent and that the mean number of observed individuals in site i depends linearly on the covariates, through the log link function:

$$\{Y_i\}_{1\le i\le n} \text{ independent}, \qquad Y_i \sim \mathcal{P}(\lambda_i), \qquad \log(\lambda_i) = x_i^\top \beta. \tag{3.14}$$

The model can be adapted to account for heterogeneous sampling efforts (e.g. different observation times) by adding an known site-specific offset term o_i to the regression model:

$$\log(\lambda_i) = o_i + x_i^\top \beta. \tag{3.15}$$

The unknown parameter θ is only the vector of regression coefficients β, its estimation (by maximizing the likelihood) and interpretation are straightforward.
Still, this model suffers an important limitation because, if the species is actually absent from the site, the parameter λ_i of the Poisson regression model should be zero (whatever the sampling effort), but Model (3.14) is not defined in this case.

Alternatively, one may aim at understanding the drivers of the simple presence of the species in each site. One way is to consider the binary variable $\widetilde{Y}_i = \mathbf{1}_{Y_i>0}$:

$$\widetilde{Y}_i = \begin{cases} 1 & \text{the species has been observed in site } i \\ 0 & \text{otherwise,} \end{cases}$$

and to use a logistic regression model:

$$\{\widetilde{Y}_i\}_{1\le i\le n} \text{ independent}, \qquad \widetilde{Y}_i \sim \mathcal{B}\text{ern}(\pi_i), \qquad \log\left(\frac{\pi_i}{1-\pi_i}\right) = x_i^\top \alpha. \tag{3.16}$$

3.2.2 The ZIP model

A way to circumvent both limitations is to include in the model a variable Z_i, that indicates whether the species is actually present or not:

$$Z_i = \begin{cases} 0 & \text{if the species is actually absent (and not only unobserved) in site } i\,, \\ 1 & \text{if the species is actually present (but possibly not observed) in site } i\,. \end{cases}$$

The variable Z_i is obviously latent, because not observed. The distribution observed abundance Y_i can then be defined conditionally on Z_i, yielding the zero-inflated Poisson (ZIP) model, which states that

- the sites are independent,
- the binary variable Z_i depends on the environment through a logistic regression model,
- if the species is absent ($Z_i = 0$), then the observed abundance Y_i can only be zero, whereas if it is present, the observed abundance depends on the covariates through a Poisson regression model.

Model 6 (Zero-inflated Poisson regression model)**.**

$$\{(Y_i, Z_i)\}_{1\le i\le n} \text{ independent}, \qquad Z_i \sim \mathcal{B}(\pi_i), \qquad \operatorname{logit}(\pi_i) := \log\left(\frac{\pi_i}{1-\pi_i}\right) = x_i^\intercal \alpha,$$
$$Y_i \mid Z_i = 0 \sim \delta_0,$$
$$Y_i \mid Z_i = 1 \sim \mathcal{P}(\lambda_i), \qquad \log(\lambda_i) = x_i^\intercal \beta$$

where δ_0 stands for the Dirac mass in zero: $Y \sim \delta_0 \Leftrightarrow \mathbb{P}(Y = 0) = 1$.

Remarks.

- Again, an offset term o_i can be added to the Poisson regression to account for heterogeneous sampling efforts.
- Note that Model (2) is sometimes parameterized in terms of absence probability, which amounts at replacing the presence probability π_i with the absence probability $1 - \pi_i$ and the vector of regression coefficients α with $-\alpha$.
- The ZIP model is in fact a particular mixture model as defined in the previous section (Section 1) where the first component of the mixture is a Dirac distribution at 0, and the second component is a Poisson distribution with parameter λ. The weight is π

The parameters of Model (6) are

$$\theta_{\text{obs}} = \beta, \qquad \theta_{\text{lat}} = \alpha, \qquad \theta = (\alpha, \beta)$$

where α and β are both vectors of regression coefficients: α encodes the effects of the environmental covariates on the presence probability π_i while β encodes the effects of the same covariates on the mean observed abundance λ_i of the species, provided it is present in the site.

The marginal distribution of the observed abundance Y_i can be obtained by deconditioning on Z_i and turns out to be a zero-inflated distribution.

Definition 2. *The random variable Y over $\mathbb{N}$ has a zero-inflated distribution $ZIP(\pi, \lambda)$ iff*

$$\mathbb{P}(Y = 0) = (1 - \pi) + \pi e^{-\lambda}, \qquad \text{and, for } y \ge 1, \quad \mathbb{P}(Y = y) = \pi e^{-\lambda}\frac{\lambda^y}{y!}. \tag{3.17}$$

Formula (3.17) which can be reformulated into a unique formula:

$$\mathbb{P}_{\text{ZIP}}(Y = y) = (1 - \pi)\mathbf{1}_{\{0\}}(y) + \pi e^{-\lambda}\frac{\lambda^y}{y!}. \tag{3.18}$$

Proposition 7. *Under Model (6), the marginal distribution of the observed abundance Y_i is a zero-inflated Poisson $ZIP(\pi_i, \lambda_i)$.*

Proof of Proposition 7

The observed abundance Y_i is zero either if the species is absent (with probability $1 - \pi_i$) or if it is present (with probability π_i), but unseen (which occurs with probability $e^{-\lambda_i}$). Then, for the observed abundance to be $y_i \ge 1$, we need to the species to be present (with probability π_i) and the count to be y_i (with probability $e^{-\lambda_i}\lambda_i^{y_i}/y_i!$).

Graphical model. The graphical model associated with the joint distribution $p_\theta(\boldsymbol{Y}, \boldsymbol{Z})$ for Model (6) is the same as this of the mixture model, given in Figure 3.2. According to this model the couples $\{(Y_i, Z_i)\}_{1\le i\le n}$ are all independent.

3.2.3 Marginal and complete log-likelihoods

We denote $\boldsymbol{y} = \{y_i\}_{1\le i\le n}$. Because the sites are independent, the marginal log-likelihood is

$$\log p_\theta(\boldsymbol{y}) = \sum_{i=1}^{n} \log\left((1-\pi_i)\mathbf{1}_{\{0\}}(y_i) + \pi_i e^{-\lambda_i}\lambda_i^{y_i}/y_i!\right).$$

Denoting $\boldsymbol{Z} = \{Z_i\}_{1\le i\le n}$, the complete likelihood of Model (6) is

$$\begin{aligned}\log p_\theta(\boldsymbol{y}, \boldsymbol{Z}) &= \log p_\theta(\boldsymbol{Z}) + \log p_\theta(\boldsymbol{y} \mid \boldsymbol{Z}) \\ &= \sum_{i=1}^{n} Z_i \log\pi_i + (1-Z_i)\log(1-\pi_i) + \sum_{i=1}^{n} Z_i\left(-\lambda_i + y_i\log\lambda_i - \log(y_i!)\right). \end{aligned} \tag{3.19}$$

3.2.4 EM algorithm for the ZIP model

Algorithm 5 (EM for the ZIP model)**.** *Starting from $\theta^{(0)}$, repeat until convergence:*

E-step. *For all $i = 1, \dots, n$, compute:*

$$\tau_i^{(h)} = \mathbb{P}_\theta(Z_i = 1 \mid Y_i = y_i) = \frac{(1-\pi_i^{(h)})\mathbf{1}_{y_i>0} + \pi_i^{(h)} e^{-\lambda_i^{(h)}}}{(1-\pi_i^{(h)}) + \pi_i^{(h)} e^{-\lambda_i^{(h)}}}. \tag{3.20}$$

M-step. *Update the estimate of θ as*

$$\alpha^{(h+1)} = \arg\max_\alpha \sum_{i=1}^{n} \tau_i^{(h)}\log\pi_i + (1-\tau_i^{(h)})\log(1-\pi_i) \qquad \textit{with} \qquad \operatorname{logit}(\pi_i) = x_i^\top\alpha \tag{3.21}$$

$$\beta^{(h+1)} = \arg\max_\beta \sum_{i=1}^{N} \tau_i^{(h)}\left(-\lambda_i + y_i\log\lambda_i - \log(y_i!)\right) \qquad \textit{with} \qquad \lambda_i = \exp(x_i^\top\beta).$$

Remark. In this model, the update of the parameters at the M-step is not explicit. However, having a look at the quantities they have to maximise, we observe that they have the same form as the log-likelihood of a classical logistic regression ($\widetilde{Y}_i$ being replaced with $\tau_i^{(h)}$) for α and of a Poisson regression (with weights $\tau_i^{(h)}$) for β. The optimization with respect to α and β can be achieved numerically with standard libraries dedicated to generalized linear models.

Proof of Algorithm 5

About $Q(\theta \mid \theta^{(h)})$. From the expression of the complete log-likelihood provided in Equation (3.19), the integration of the latent variables Z_i leads to the following formula for $Q(\theta \mid \theta^{(h)})$:

$$Q(\theta \mid \theta^{(h)}) = \sum_{i=1}^{n} \tau_i^{(h)}\log\pi_i + (1-\tau_i^{(h)})\log(1-\pi_i) + \sum_{i=1}^{n} \tau_i^{(h)}\left(-\lambda_i + y_i\log\lambda_i - \log(y_i!)\right), \tag{3.22}$$

where $\tau_i^{(h)} = \mathbb{E}_{\theta^{(h)}}[Z_i \mid \boldsymbol{Y} = \boldsymbol{y} = \boldsymbol{y}]$.

E-step. To evaluate $Q(\theta \mid \theta^{(h)})$, we only need to evaluate the conditional expectation of each Z_i given the data $\boldsymbol{y}$, that is to evaluate $\tau_i^{(h)} = \mathbb{E}_{\theta^{(h)}}[Z_i \mid \boldsymbol{Y} = \boldsymbol{y} = \boldsymbol{y}]$. This can be done in closed form.
First, observe that, because the couples (Y_i, Z_i) are all independent from each other, the conditional distribution of Z_i given $\boldsymbol{Y}$ is the same as its conditional distribution given the corresponding Y_i only:

$\tau_i^{(h)} = \mathbb{E}_{\theta^{(h)}}[Z_i \mid Y_i = y_i]$. Furthermore, because the Z_i are 0/1, we know that

$$\tau_i^{(h)} = \mathbb{E}_{\theta^{(h)}}[Z_i \mid Y_i = y_i] = \mathbb{P}_{\theta^{(h)}}(Z_i = 1 \mid Y_i = y_i).$$

From Model (6), we easily see that if the observed count is not zero ($y_i > 0$), then the species is surely present, so

$$\mathbb{P}_{\theta^{(h)}}(Z_i = 1 \mid Y_i > 0) = 1. \tag{3.23}$$

If the observed count is $y_i = 0$, we can apply the Bayes formula:

$$\begin{aligned}\mathbb{P}_{\theta^{(h)}}(Z_i = 1 \mid Y_i = 0) &= \frac{\mathbb{P}_{\theta^{(h)}}(Y_i = 0, Z_i = 1)}{\mathbb{P}_{\theta^{(h)}}(Y_i = 0)} = \frac{\mathbb{P}_{\theta^{(h)}}(Y_i = 0 \mid Z_i = 1)\mathbb{P}_{\theta^{(h)}}(Z_i = 1)}{\mathbb{P}_{\theta^{(h)}}(Y_i = 0)} \\ &= \frac{\pi_i^{(h)} e^{-\lambda_i^{(h)}}}{(1-\pi_i^{(h)}) + \pi_i^{(h)} e^{-\lambda_i^{(h)}}} \quad \text{(using Equation (3.18) with } y_i = 0) \end{aligned} \tag{3.24}$$

Now, combining Equations (3.23) and (3.24), we obtain a global formula:

$$\tau_i^{(h)} := \tau(y_i) = \mathbb{E}_{\theta^{(h)}}[Z_i \mid Y_i = y_i] = \mathbf{1}_{y_i>0} + \frac{\pi_i^{(h)} e^{-\lambda_i^{(h)}}}{(1-\pi_i^{(h)}) + \pi_i^{(h)} e^{-\lambda_i^{(h)}}} \mathbf{1}_{y_i=0} = \frac{(1-\pi_i^{(h)})\mathbf{1}_{y_i>0} + \pi_i^{(h)} e^{-\lambda_i^{(h)}}}{(1-\pi_i^{(h)}) + \pi_i^{(h)} e^{-\lambda_i^{(h)}}}.$$

M step. We may now update the parameter θ by maximizing the objective function of the EM algorithm provided in Equation (3.22). Reminding that

$$\pi_i = \exp(x_i^\top \alpha)/(1+\exp(x_i^\top \alpha)) \quad \text{and} \quad \lambda_i = \exp(x_i^\top \beta),$$

we observe that $Q(\theta \mid \theta^{(h)})$ can be decomposed into a sum of two terms depending respectively in α and β:

$$\begin{aligned} A^{(h)}(\alpha) &= \sum_{i=1}^n \tau_i^{(h)} \log \pi_i + (1-\tau_i^{(h)}) \log(1-\pi_i) \\ B^{(h)}(\beta) &= \sum_{i=1}^n \tau_i^{(h)} \left(-\lambda_i + y_i \log \lambda_i - \log(y_i!)\right) \end{aligned}$$

which can be optimized separately:

$$\arg\max_\alpha Q(\theta \mid \theta^{(h)}) = \arg\max_\alpha A^{(h)}(\alpha) \qquad \text{and} \qquad \arg\max_\beta Q(\theta \mid \theta^{(h)}) = \arg\max_\beta B^{(h)}(\beta).$$

3.2.5 Illustration

We now compare the ZIP regression (6) with the logistic regression (3.16) and Poisson regression (3.14) on the cod abundances introduced in Dataset 3. To ease the interpretation and the comparison of the regression coefficients, the four covariates were centered and their variances were set to one. Models (3.16) and (3.14) can be fitted with the R `glm` R function, and model (6) with the `zeroinfl` function of the `pscl` R package.

Parameter estimates. Table 3.2 gives the MLE of the regression coefficients for the Poisson regression , the logistic regression and the ZIP regression (6) models. We observe that the regression coefficients for both the presence probability (α) and the abundance (β) are different when dealing with both aspect separately (i.e logistic regression or Poisson regression) or jointly (ZIP regression). As the covariates have been centered, one may focus on the intercepts, which control the presence probability and the abundance, respectively, in a 'mean' site. The ZIP regression yields a higher mean presence probability than the logistic, because it accounts for the fact that the species can be present, when it is actually not observed. As for the Poisson part (which deals with the mean abundance), the Poisson regression yields a smaller mean abundance, as it needs to accommodate for the numerous zeros in the data set, whereas the abundance part of the ZIP regression only deals with case where the species is actually present.

Presence probability. Figure 3.8 (left) gives the estimated probability $\widehat{\pi}_i^{ZIP}$ of presence in each station for Example 3, as a function of the linear predictor $x_i^\top \widehat{\alpha}^{ZIP}$. The blue crosses indicate the probability of presence according to the logistic regression $\pi_i^{logistic}$: we see that the two models yield similar probabilities. Still, the

	Presence (α)					Abundance (β)				
	Inter.	Lat.	Long.	Depth	Temp.	Inter.	Lat.	Long.	Depth	Temp.
Logistic (3.16)	-1.275	-0.251	0.301	-0.387	1.994	–	–	–	–	–
Poisson (3.14)	–	–	–	–	–	0.010	-0.721	-0.043	0.917	2.479
ZIP (6)	-0.95	-0.287	0.374	-0.578	1.59	1.543	-0.371	-0.265	0.864	1.858

Table 3.2: Cod abundance in the Barents sea.

binary part of the ZIP model (encoded in π_i^{ZIP}) does not contain all information, regarding the prediction of the actual presence of the species in a given site: the abundance part must also be accounted for.

Indeed, under the ZIP model (6), the sites can classified in term of actual presence or absence of the species of species, using the same rule as this used to classify observations into components under a mixture model, as seen in Section 3.1. This ZIP classification is based on the estimate of the conditional probability τ_i, given in Equation (3.20). The right panel of Figure 3.8 compares the conditional probability τ_i^{ZIP} resulting from the ZIP model, with the presence probability $\pi_i^{logistic}$. We observed the classification based on τ_i^{ZIP} is much more contrasted than this based on $\pi_i^{logistic}$, which predicts very low probabilities of presence in sites where the species has actually been observed. This difference is greatly do the the fact that the logistic regression relies on degraded data, that is the $\widetilde{Y}_i$, instead of the observed counts Y_i.

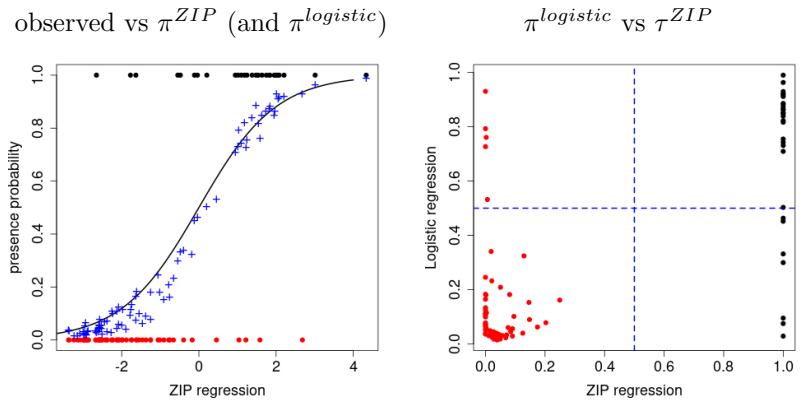


Figure 3.8: Cod abundance in the Barents sea. Left: probability of presence according to the ZIP model ($\widehat{\pi}_i^{ZIP}$: black curve), dots = observed presence (black dots (•): $Y_i > 0$, red dots (•): $Y_i = 0$), blue crosses (+): probability of presence according to logistic regression $\pi_i^{logistic}$.
Right: prediction of presence according to the ZIP model (x axis) vs prediction of presence according to the logistic regression (y axis). Blue dotted lines = 50% thresholds.

Abundance prediction. Figure 3.9 displays the fit of the Poisson regression model (left) and of the ZIP model (center): obviously, the variability of the data does not fit the expected variability under the simple Poisson assumption. The prediction intervals of the ZIP model better accounts for the additional variability due to the excess of zeros, but are much larger.
The predictions provided by the ZIP regression model must be carefully analysed as they combine estimates of both the presence probability π_i of the species in site i, and of its expected abundance λ_i *conditional on its presence*. Because the regression parameters are different for the two parameters, a high expected abundance λ_i may coincide with a low presence probability π_i, as shown in the right panel of Figure 3.9. This explains the apparently erratic behavior of the prediction interval displayed in the center panel of Figure 3.9.

Model comparison. We may compare the ZIP model with the Poisson regression, as they both deal with the observed counts Y_i, (whereas the logistic regression deals with the $\widetilde{Y}_i$). Their respective log-likelihood are

$$\log p_{\widehat{\alpha}}^{Poisson}(y) = -1142.8 \qquad \text{(with } p = 5 \text{ independent parameters)},$$
$$\log p_{\widehat{\alpha},\widehat{\beta}}^{ZIP}(y) = -892.2 \qquad \text{(with } 2p = 10 \text{ independent parameters)}.$$

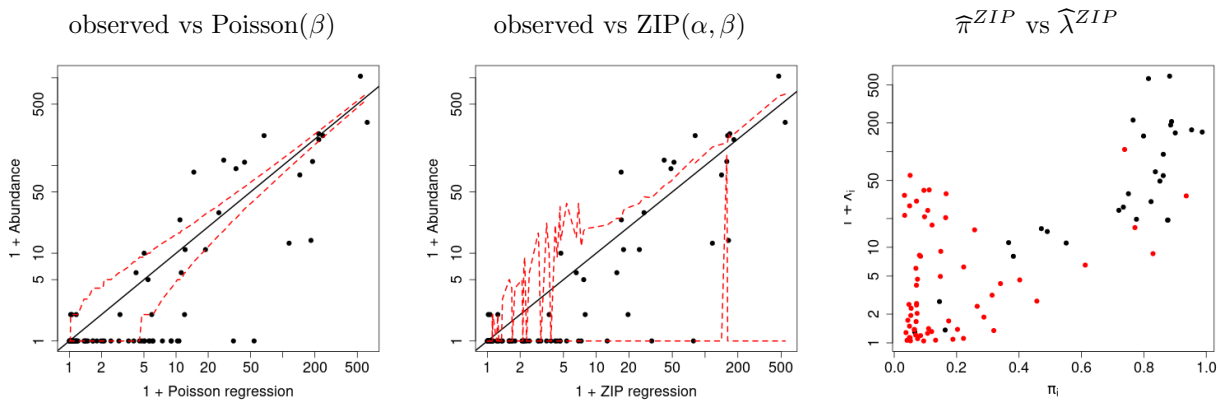


Figure 3.9: Cod abundance in the Barents sea. Left: observed abundances vs predicted abundances (in log-scale) with the Poisson regression (3.14), dotted red lines = 95% interval for the Poisson distribution (1 is added to abundances to allow log-scale). Center: observed abundances vs predicted abundances $\widehat{\lambda}_i$ (in log-scale) with the ZIP regression (6), dotted red lines = 95% interval for the ZIP distribution. Right: presence probability $\widehat{\pi}_i$ and expected abundance $\widehat{\lambda}_i$ estimated with the ZIP regression (6). Red dots = sites where the species was not observed.

Models (3.14) and (6) can be compared with AIC or BIC:

$$AIC(Poisson) = -1148, \qquad AIC(ZIP) = -902.2$$
$$BIC(Poisson) = -1154, \qquad BIC(ZIP) = -914.6.$$

Both criteria concur to conclude to a much better fit of the ZIP regression model.

3.2.6 Using the Louis' formula to get the asymptotic variance

To conclude this section, we use the zero-inflated Poisson Model 6 to illustrate the use of the Louis's formula [Louis, 1982] introduced in Section 2.3 to estimate the asymptotic variance of the MLE $\widehat{\theta}$. To make the calculations lighter, we consider a model with no covariates, that is where, for all $1 \leq i \leq n$:

$$\pi_i = \pi, \quad \lambda_i = \lambda, \quad \text{and } \theta = (\pi, \lambda)$$

Proposition 8. *For the ZIP model without covariates, let us define:*

$$\boldsymbol{P} = \sum_{i=1}^{n} \mathbf{1}_{y_i>0}, \quad \boldsymbol{Y}_+ = \sum_{i=1}^{n} y_i$$

and

$$\gamma = \frac{1}{\pi(1-\pi)}, \quad \eta = \frac{\pi e^{-\lambda}}{(1-\pi)+\pi e^{-\lambda}}, \qquad V = (1-\eta)\eta(n-\boldsymbol{P}).$$

Then we have:

$$\mathbf{J}_\theta S_\theta(\boldsymbol{y}) = \begin{pmatrix} -\frac{\boldsymbol{P}+(n-\boldsymbol{P})\eta}{\pi^2} - \frac{(n-\boldsymbol{P})(1-\eta)}{(1-\pi)^2} & 0 \\ 0 & -\frac{\boldsymbol{Y}_+}{\lambda^2} \end{pmatrix} + V \begin{pmatrix} \gamma^2 & -\gamma \\ -\gamma & 1 \end{pmatrix}$$

Proof of Proposition 8

The proof is composed of three steps: first we calculate the derivatives of the complete likelihood, then we compute their conditional expectation given the observed data, and, finally, we gather all these results into the estimated Fisher information matrix.

Complete log-likelihood In absence of covariate, the complete likelihood (3.19) becomes

$$\begin{aligned}\log p_\theta(\boldsymbol{y}, \boldsymbol{Z}) &= \sum_{i=1}^n Z_i \log\pi + (1-Z_i)\log(1-\pi) + \sum_{i=1}^n Z_i\left(-\lambda + y_i\log\lambda - \log(y_i!)\right)\\ &= Z_+\log\pi + (n-Z_+)\log(1-\pi) - \boldsymbol{Z}_+\lambda + y_+\log\lambda - L\end{aligned}$$

where

$$\boldsymbol{Z}_+ = \sum_{i=1}^n Z_i, \qquad \boldsymbol{Y}_+ = \sum_{i=1}^n Z_i y_i = \sum_{i=1}^n y_i \qquad \text{and} \qquad L = \sum_{i=1}^n Z_i\log(y_i!) = \sum_{i=1}^n \log(y_i!).$$

The expressions $\boldsymbol{Y}_+$ and L derive from the fact that $Z_i \in \{0,1\}$ and when $Z_i = 0$, $y_i = 0$ so $y_i = y_i Z_i$. The same holds for $\log(y_i!)$.

Derivatives of the complete log-likelihood. From this expression of the complete likelihood, we get the derivatives

$$\partial_\pi \log p_\theta(\boldsymbol{y}, \boldsymbol{Z}) = \frac{\boldsymbol{Z}_+}{\pi} - \frac{n-Z_+}{1-\pi} = \frac{1}{\pi(1-\pi)}\boldsymbol{Z}_+ - \frac{n}{1-\pi} = \gamma\boldsymbol{Z}_+ - \frac{n}{1-\pi}$$

where $\gamma = 1/\pi(1-\pi)$. Besides,

$$\partial_\lambda \log p_\theta(\boldsymbol{y}, \boldsymbol{Z}) = -\boldsymbol{Z}_+ + \frac{y_+}{\lambda},$$

The Hessian is then obtained by calculating the second derivatives:

$$\partial^2_{\pi^2} \log p_\theta(\boldsymbol{y}, \boldsymbol{Z}) = -\frac{\boldsymbol{Z}_+}{\pi^2} - \frac{n-\boldsymbol{Z}_+}{(1-\pi)^2}, \qquad \partial^2_{\pi\lambda} \log p_\theta(\boldsymbol{y}, \boldsymbol{Z}) = 0, \qquad \partial_{\lambda^2} \log p_\theta(\boldsymbol{y}, \boldsymbol{Z}) = -\frac{\boldsymbol{Y}_+}{\lambda^2}. \tag{3.25}$$

Integration of the latent variables. Louis' formulas then require to evaluate the conditional expectation of the Hessian matrix and the conditional variance of the gradient vector. A quick look at their formula enables us to conclude that we need to compute the conditional expectation and variance of $Z_+ = \sum_{i=1}^n Z_i$, that is

$$\mathbb{E}[\boldsymbol{Z}_+ \mid \boldsymbol{Y} = \boldsymbol{y}] = \sum_{i=1}^n \mathbb{E}[Z_i \mid \boldsymbol{y}] = \sum_{i=1}^n \mathbb{E}[Z_i \mid Y_i = y_i] = \sum_{i=1}^n \tau(y_i)$$

Denoting by η the probability for the species to be present given that $Y_i = 0$

$$\eta = \frac{\pi e^{-\lambda}}{(1-\pi) + \pi e^{-\lambda}},$$

we have from Equation (3.20) that $\tau(y_i) = (1-\eta)\mathbf{1}_{y_i>0} + \eta$, so

$$\mathbb{E}[\boldsymbol{Z}_+ \mid \boldsymbol{Y} = \boldsymbol{y}] = \sum_{i=1}^n (1-\eta)\mathbf{1}_{y_i>0} + \eta = (1-\eta)\boldsymbol{P} + n\eta = \boldsymbol{P} + (n-\boldsymbol{P})\eta,$$

where $\boldsymbol{P}$ stands for the number of sites where the species is observed: $\boldsymbol{P} = \sum_{i=1}^n \mathbf{1}_{y_i>0}$.
Let us now consider the variance: by conditional independance of the $Z_i \mid \boldsymbol{Y} = \boldsymbol{y}$, we have that

$$\mathbb{V}[\boldsymbol{Z}_+ \mid \boldsymbol{Y} = \boldsymbol{y}] = \sum_{i=1}^n \mathbb{V}[Z_i \mid Y_i = y_i].$$

Besides, we have demonstrated that $Z_i \mid \boldsymbol{Y} = \boldsymbol{y}$ is distributed as a Bernoulli with parameter $\tau_i = \tau(y_i)$ provided in Equation (3.20), so So

$$\begin{aligned}\mathbb{V}[Z_i \mid \boldsymbol{Y} = \boldsymbol{y}] = \tau_i(1-\tau_i) &= \left((1-\eta)\mathbf{1}_{y_i>0} + \eta\right)\left(1 - (1-\eta)\mathbf{1}_{y_i>0} - \eta\right)\\ &= \left((1-\eta)\mathbf{1}_{y_i>0} + \eta\right)(1-\eta)(1-\mathbf{1}_{y_i>0})\\ &= (1-\eta)[(1-\eta)\underbrace{\mathbf{1}_{y_i>0}(1-\mathbf{1}_{y_i>0})}_{=0} + \eta(1-\mathbf{1}_{y_i>0})]\\ &= (1-\eta)\eta(1-\mathbf{1}_{y_i>0}),\end{aligned}$$

that is :

$$\mathbb{V}[\boldsymbol{Z}_+ \mid \boldsymbol{Y} = \boldsymbol{y}] = \sum_{i=1}^{n} (1-\eta)\eta(1 - \mathbf{1}_{y_i>0}) = (1-\eta)\eta(n-P).$$

Expression of $\widehat{I}(\theta)$. By injecting the expectation and variance of Z_+ into the Hessian and the gradient, we can now derive the required conditional moments, that is

$$\mathbb{E}[\partial^2_{\pi^2} \log p_\theta(\boldsymbol{y}, \boldsymbol{Z}) \mid \boldsymbol{Y} = \boldsymbol{y}] = -\frac{P + (n-P)\eta}{\pi^2} - \frac{(n-P)(1-\eta)}{(1-\pi)^2}, \quad \mathbb{E}[\partial^2_{\pi\lambda} \log p_\theta(\boldsymbol{y}, \boldsymbol{Z}) \mid \boldsymbol{Y} = \boldsymbol{y}] = 0,$$
$$\mathbb{E}[\partial_{\lambda^2} \log p_\theta(\boldsymbol{y}, \boldsymbol{Z}) \mid \boldsymbol{Y} = \boldsymbol{y}] = -\frac{y_+}{\lambda^2}$$

and

$$\begin{aligned}
\mathbb{V}[\partial_\pi \log p_\theta(\boldsymbol{y}, \boldsymbol{Z}) \mid \boldsymbol{Y} = \boldsymbol{y}] &= \gamma^2 \mathbb{V}[Z_+ \mid \boldsymbol{Y} = \boldsymbol{y}], \\
\mathbb{V}[\partial_\lambda \log p_\theta(\boldsymbol{y}, \boldsymbol{Z}) \mid \boldsymbol{Y} = \boldsymbol{y}] &= \mathbb{V}[Z_+ \mid \boldsymbol{Y} = \boldsymbol{y}], \\
\mathbb{C}\mathrm{ov}[\partial_\pi \log p_\theta(\boldsymbol{y}, \boldsymbol{Z}), \partial_\lambda \log p_\theta(\boldsymbol{y}, \boldsymbol{Z}) \mid \boldsymbol{Y} = \boldsymbol{y}] &= -\gamma \mathbb{V}[Z_+ \mid \boldsymbol{Y} = \boldsymbol{y}],
\end{aligned}$$

Example. When considering the data from Example 3, using the ZIP model (6) with no covariate, the parameter estimates are $\widehat{\alpha} = 0.7787$ and $\widehat{\beta} = 4.647$, which correspond to

$$\widehat{\pi} = 0.3146, \qquad \widehat{\lambda} = 104.2.$$

The respective (estimated) asymptotic standard deviations of the estimators are 0.04922 for $\widehat{\pi}$ and 1.930 for $\widehat{\lambda}$ and, in the present case, the asymptotic covariance between the two estimators is negligible ($< 10^{-10}$). The resulting 95% confidence intervals are then:

$$CI(\pi) = [0.2181, 0.4111], \qquad CI(\lambda) = [100.5, 108.0].$$

3.3 Genetic structure of a population: mixture model

3.3.1 Data and question

Understanding the genetic diversity within a population is one of the most prominent questions in population genetics. A natural way to model this diversity is to assume the existence of ancestral populations from which the genome of each individual in the studied population is derived.

Dataset 4 (Taita thrush)**.** *We consider the data collected by Galbusera et al. [2000] and further analysed by Pritchard et al. [2000]*[a]*. It consists in $p = 7$ markers recorded for $n = 155$ birds (thrushes). The markers are microsatellites (i.e. repetitions of di- or tri-nucleotides), with respectively $m_1 = 8$, $m_2 = 5$, $m_3 = 6$, $m_4 = 3$, $m_5 = 4$, $m_6 = 10$ and $m_7 = 8$ alleles*[b]*. Because the birds are diploid, two alleles were recorded for each individual $1 \le i \le n$, as presented in Table 3.3. The genotype of individual $i = 2$ for marker $j = 3$ is hence given by the un-ordered couple $y_{ij} = \{1, 6\}$. Observe that some data are missing.*

The birds were captured in 4 different locations in the south-west of Kenya: Chawia (17 individuals), Ngangao (54), Mbololo (80), and Yale (4).

[a] The data are available from `web.stanford.edu/group/pritchardlab/software/structure-data_v.2.3.1.html`.
[b] For the sake of clarity, the 22 alleles of the first marker were clustered into 8 categories, further considered as alleles

3.3.2 A mixture model for genetic structure

Mixture model. A simple model assumes that each individual originates from one of K founder populations, each characterized by specific allele frequencies. This model, with K populations of origin, can be described as follows

Model 7 (Mixture model for the genetic structure of a population)**.** *Denoting by Z_i the population of*

	i	Markers $j=1$	$j=2$	$j=3$	$j=4$	$j=5$	$j=6$	$j=7$	Location
Y_{1j}	1	1	1	6	1	2	2	5	Ngangao
	1	3	2	6	1	2	4	5	Ngangao
Y_{2j}	2	3	2	1	1	2	2	5	Ngangao
	2	3	3	6	1	2	10	8	Ngangao
	3	3	3	6	1	3	2	1	Ngangao
	3	7	5	6	2	NA	2	4	Ngangao

Table 3.3: Genotypes of the first three birds from Taita thrush data set (Dataset 4) for the $p = 7$ markers. The genotype of each individual is recorded on two lines of the table.

origin of individual i and $Y_i = (Y_{i1}, \dots Y_{ip})$ the whole genotype of individual i, the model states that

$$\begin{aligned} \{Z_i\}_{1\le i\le n} \text{ iid}: &\qquad Z_i \sim Cat(\omega)\,, \\ \{Y_i\}_{1\le i\le n} \text{ independent} \,|\, \{Z_i\}_{1\le i\le n}, & \\ \{Y_{ij}\}_{1\le j\le p} \text{ independent} \,|\, Z_i: &\qquad \mathbb{P}\{Y_{ij} = \{a,b\} \mid Z_i = k\} = \phi_{kj}(\{a,b\}) \end{aligned}$$

where

$$\phi_{kj}(\{a,b\}) = \begin{cases} 2\gamma_{kja}\gamma_{kjb} & \text{if } a \neq b \quad (\textit{heterozygous}), \\ \gamma_{kja}^2 & \text{if } a = b \quad (\textit{homozygous}). \end{cases} \tag{3.26}$$

The parameters of the model are

- $\omega = [\omega_k]_{1\le k\le K}$ *the vector of probabilities for an individual to come from a given population,*
- γ_{kja} *the allelic frequency of allele $a \in \{1, \dots m_j\}$ at locus $j \in \{1, \dots p\}$ in population $k \in \{1, \dots K\}$,*
- γ *the set of all allelic frequencies at each locus in each population,*

that is

$$\theta = (\omega, \gamma).$$

The latent variables are the memberships Z_i and the observed variables are the genotypes Y_i.

Assumptions. Model 7 is based on three main underlying assumptions.

1. The whole genome of a given individual originates from a single founding population (Z_i unique for each individual i).
2. The genotype of a given individual at different loci (markers) are independent, conditional on its population of origin.
3. Each population adheres to the Hardy-Weinberg principle, which assumes random mating among parents, resulting in the genotype probabilities specified in Equation (3.26) .

Assumption 3 gives the shape of the emission probability $\phi_{kj}(\{a,b\})$ defined in (3.26): it is both reasonable for old populations and statistically useful as it reduces the number of emission parameters γ to be estimated. Assumption 2 only make sense if the loci are very distant along the genome, which is the case in Example 4 where only seven markers are recorded. The generalization introduced in Section **??** will account for possible linkage desequilibria (correlations) between neighbor loci. Assumption 1 is very strong as it does not account for crossing between populations, which are very likely to have occurred. The two generalizations introduced in Sections **??** and **??** will get rid of this assumption.

Graphical model. The graphical model of Model 7 is given in Figure 3.10. Its structure mainly results from the fact that the couples $\{(Z_i, Y_i)\}_{1\le i\le n}$ are iid.

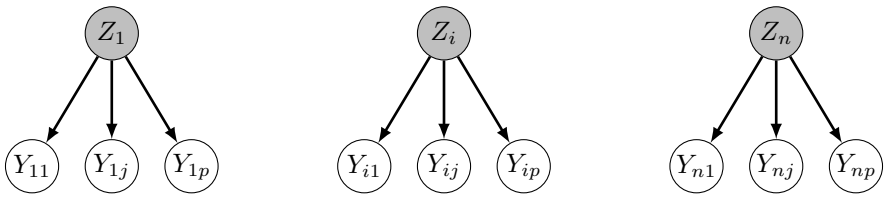


Figure 3.10: Graphical representation of the mixture Model 7 for the genetic structure of a population.

3.3.3 Complete and marginal likelihoods

Let ϕ_k denote the emission distribution of the whole genotype of a given individual i belonging to population k. Because the genotypes at the different locus are conditionally independent, we have that

$$\phi_k(y_i) \coloneqq \mathbb{P}\{Y_i = y_i \mid Z_i = k\} = \prod_{j=1}^{p} \phi_{kj}(y_{ij}).$$

Because the individuals are independent, the log-likelihood of the observed data is then

$$\log_\theta(\boldsymbol{y}) = \sum_{i=1}^{n} \log\left[\sum_{k=1}^{K} \omega_k \phi_k(y_i)\right] = \sum_{i=1}^{n} \log\left[\sum_{k=1}^{K} \omega_k \left(\prod_{j=1}^{p} \phi_{kj}(y_{ij})\right)\right] \tag{3.27}$$

where $\boldsymbol{y}$ stands for the set of all observed genotypes and Z for the set of all individuals' membership.

Defining again the binary variable Z_{ik} ($i \in \{1, \dots n\}$, $k \in \{1, \dots K\}$)), which is 1 if $Z_i = k$ and 0 otherwise, we get a similar form as in Section 3.1:

$$\log_\theta(\boldsymbol{y}, \boldsymbol{Z}) = \sum_{i=1}^{n}\sum_{k=1}^{K} Z_{ik} \left(\log \omega_k + \log \phi_k(y_i)\right) = \sum_{i=1}^{n}\sum_{k=1}^{K} Z_{ik} \left(\log \omega_k + \sum_{j=1}^{p} \log \phi_{kj}(y_{ij})\right).$$

3.3.4 EM for the population genetic mixture model

Algorithm 6 (EM for the population genetic mixture model)**.** *Starting from $\theta^{(0)}$, repeat until convergence:*

E-step. *For all $i = 1, \dots, n$, and all $k = 1, \dots, K$, compute:*

$$\tau_{ik}^{(h)} = \frac{\omega_k \phi_k(y_i)}{\sum_{\ell=1}^{K} \omega_\ell \phi_\ell(y_i)} = \frac{\omega_k \prod_{j=1}^{p} \phi_{kj}(y_{ij})}{\sum_{\ell=1}^{K} \omega_\ell \prod_{j=1}^{p} \phi_{\ell j}(y_{ij})}. \tag{3.28}$$

M-step. *Update the estimate of θ as*

$$\omega^{(h+1)} = \frac{1}{n} \sum_{i=1}^{n} \tau_{ik}^{(h)}.$$

$$\gamma_{kja}^{(h+1)} = \sum_{i=1}^{n} \tau_{ik}^{(h)} (y_{ija}^1 + y_{ija}^2) \Big/ \sum_{i=1}^{n} \tau_{ik}^{(h)}.$$

Proof of Algorithm 6

Integrated complete likelihood Using the formula of the complete likelihood provided in Equation (3.27). The conditional expectation of the complete likelihood is

$$Q(\theta \mid \theta^{(h)}) = \sum_{i=1}^{n}\sum_{k=1}^{K} \tau_{ik}^{(h)} \left(\log \omega_k + \sum_{j=1}^{p} \log \phi_{kj}(y_{ij})\right) \tag{3.29}$$

where

$$\tau_{ik}^{(h)} \coloneqq \mathbb{E}_{\theta^{(h)}}[Z_{ik} \mid \boldsymbol{Y} = \boldsymbol{y}] = \mathbb{E}_{\theta^{(h)}}[Z_{ik} \mid Y_i = y_i]$$

because the couples (Z_i, Y_i) are all independent.

E step. As in Section 3.1, τ_{ik} derives from Bayes' formula:

$$\tau_{ik} = \frac{\omega_k \phi_k(y_i)}{\sum_{\ell=1}^K \omega_\ell \phi_\ell(y_i)} = \frac{\omega_k \prod_{j=1}^p \phi_{kj}(y_{ij})}{\sum_{\ell=1}^K \omega_\ell \prod_{j=1}^p \phi_{\ell j}(y_{ij})}.$$

At step h of the EM algorithm, the conditional probability $\tau_{ik}^{(h)}$ is calculated using the current estimate $\theta^h = (\omega^{(h)}, \gamma^{(h)})$, plugging the estimates $\gamma_{kja}^{(h)}$ in the respective emission distributions ϕ_{kj}.

M step. Setting to zero the derivative of $Q(\theta \mid \theta^{(h)})$ with respect to the ω_k's yields the same update formula for the probabilities ω_k as in Section 3.1: at step h, we get

$$\omega^{(h+1)} = \frac{1}{n} \sum_{i=1}^n \tau_{ik}^{(h)}.$$

To explicitly show the allelic frequencies γ_{kja} in the objective function (3.29), we introduce the binary variable y_{ija}^1,

$$y_{ija}^1 = \begin{cases} 1 & \text{if the first allele of individual } i \text{ at locus } j \text{ is } a \\ 0 & \text{otherwise} \end{cases}$$

We define y_{ijb}^2 in the same way for the second allele. Then, the formula provided for $\log \phi_{kj}(\{y_{ij}^1, y_{ij}^2\})$ can be contracted into a unique formula as:

$$\begin{aligned} \log \phi_{kj}(y_{ij}) &= \sum_{a,b=1}^{m_j} y_{ija}^1 \log \gamma_{kja} + y_{ijb}^2 \log \gamma_{kjb} + y_{ija}^1 y_{ijb}^2 \log 2 \\ &= \sum_{a=1}^{m_j} (y_{ija}^1 + y_{ija}^2) \log \gamma_{kja} + \log 2 \sum_{a,b=1}^{m_j} y_{ija}^1 y_{ijb}^2. \end{aligned}$$

To get the update formulas for the allelic frequencies, we set to zero the derivative of (3.29), accounting for the constraint that the allelic frequencies sum to 1 for each locus j in each population k: $\sum_{a=1}^{m_j} \gamma_{kja} = 1$. Applying the Lagrange multipliers methods, we have

$$\partial_{\gamma_{kja}} \left[Q(\theta \mid \theta^{(h)}) + \lambda_{kj} \left(\sum_{a=1}^{m_j} \gamma_{kja} - 1 \right) \right] = \frac{1}{\gamma_{kja}} \sum_{i=1}^n \tau_{ik}^{(h)} (y_{ija}^1 + y_{ija}^2) - \lambda_{kj},$$

which is zero for

$$\gamma_{kja}^{(h)} \propto \sum_{i=1}^n \tau_{ik}^{(h)} (y_{ija}^1 + y_{ija}^2),$$

which can be seen as the total number of copies of allele a observed at locus j in population k, weighted by the probability for each individual i to belong to this population. Applying the constraint $\sum_{a=1}^{m_j} \gamma_{kja} = 1$ yields

$$\gamma_{kja}^{(h+1)} = \sum_{i=1}^n \tau_{ik}^{(h)} (y_{ija}^1 + y_{ija}^2) \Big/ \sum_{i=1}^n \tau_{ik}^{(h)}.$$

3.3.5 Model selection

The number of founding populations K is usually unknown and needs to be estimated. The BIC criterion introduced in Section 2.4 can be used for this purpose. Because of the sum constraints, the number of independent probability parameters ω_k is $K-1$ and the number of independent allelic frequencies for marker j in population k is $m_j - 1$. Hence, the total number of parameters of Model 7 with K populations is

$$D_K = (K-1) + K(m_+ - p)$$

where $m_+ = \sum_{j=1}^p m_j$ is the total number of alleles at all loci. The BIC criterion for K populations is hence

$$BIC_K = \log p_{\widehat{\theta}_K}(y) - \frac{\log(n)}{2} [(K-1) + K(m_+ - p)],$$

where $\widehat{\theta}_K$ is the maximum likelihood estimate of θ with K populations.
The ICL criterion can defined in the same way as

$$ICL_K = \log p_{\widehat{\theta}_K}(y) - \frac{\log(n)}{2}[(K-1) + K(m_+ - p)] + \sum_{i=1}^{n}\sum_{k=1}^{K} \tau_{ik} \log \tau_{ik}.$$

3.3.6 Illustration

Model 7 can be used to analyse the genetic structure of the Taita Thrush population introduced in Example 4. Figure 3.11 gives the log-likelihood, the BIC and the ICL criterion for $K = 1, \ldots 5$ populations of origin: both the BIC and ICL criterion opt for $\widehat{K} = 3$ populations. The small difference between the BIC and ICL criteria for all K is due a low conditional entropy $\mathcal{H}(Z \mid Y = y)$, which indicates a small uncertainty in the classification: τ_{ik} are all either close to 1 or to 0.

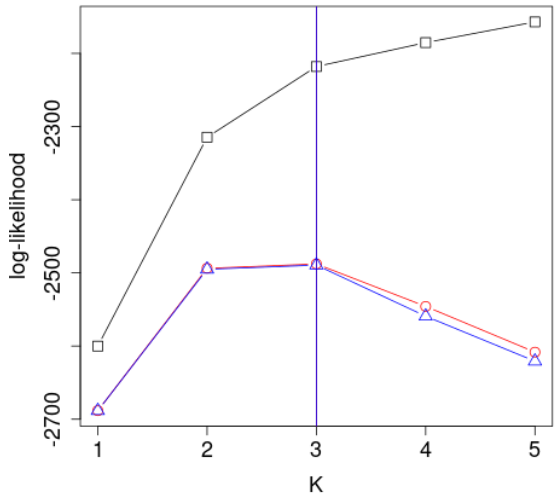


Figure 3.11: Log-likelihood for Model 7 for the Taita thrush data set (Example 4) as a function of the number of populations K. Log-likelihood $\log p_{\widehat{\theta}_K}(y)$: black squares (□), BIC criterion BIC_K: red circles (○), ICL criterion ICL_K: blue triangles (△).

Figure 3.12 gives the estimated allelic frequencies for the $p = 7$ markers in each of the $\widehat{K} = 3$ estimated populations of origin. For example, we observe that the distribution of the alleles of marker 1 are all different, that the second allele (green) of marker 2 is predominant in populations 1 and 3, but rare in population 2 and that the first allele (blue) of marker 3 is quite frequent in population 3, whereas it is quite rare in populations 1 and 2.

Figure 3.13 gives the estimated probability τ_{ik} that the bird i originates from population k given its genotype y_i. We see that this probability is away from 0 and 1 for only two individuals (from Mbolo). We observe a strong consistency between the population of origin and the capture location for the three main locations (Chawia, Ngangao and Mbolo). Only two birds from Mbolo have a high probability to originate from the same population as Ngangao. We also see that the birds from Yale seem to originate from the same founding population as these from Mbolo.

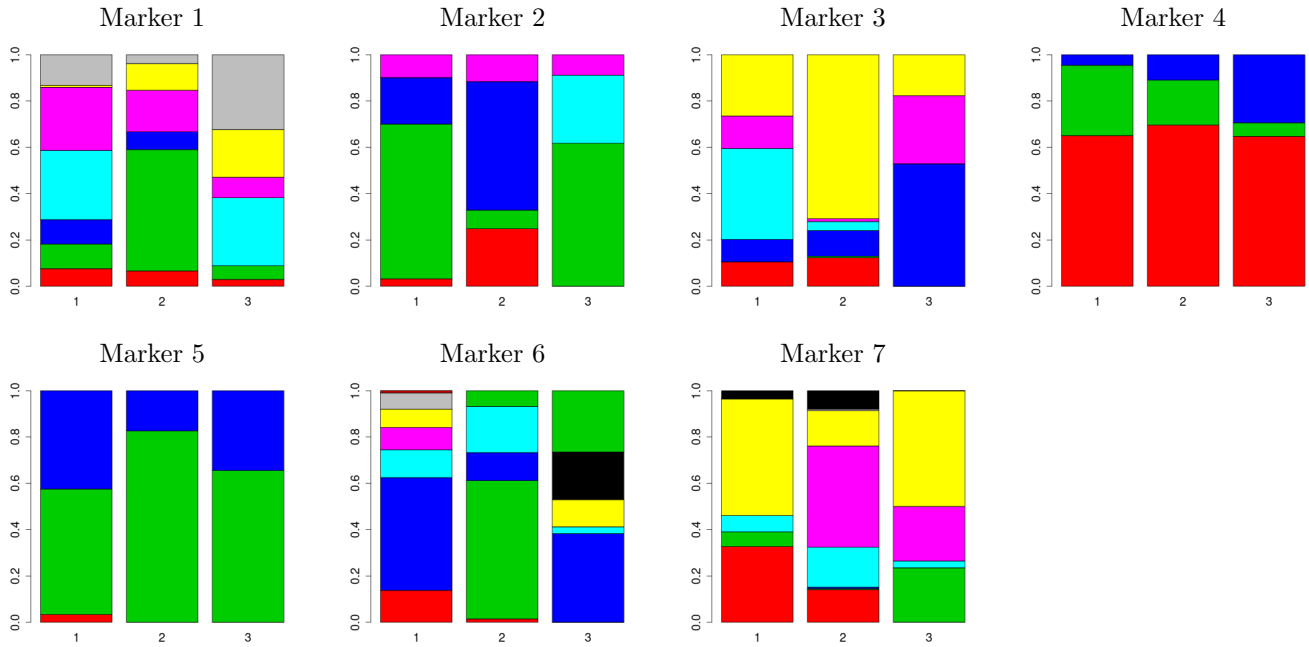


Figure 3.12: Allelic frequencies for the $p = 7$ markers in each of the $\widehat{K} = 3$ populations of origin of Taita thrushes (Example 4).

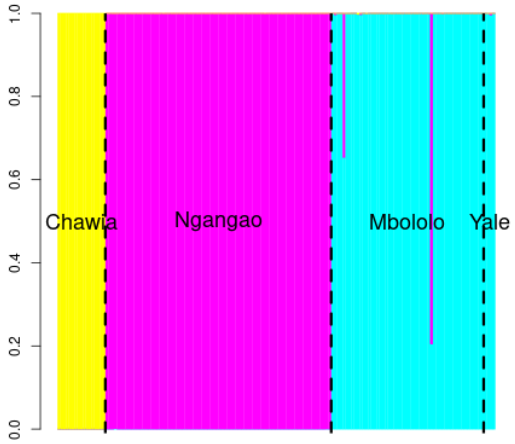


Figure 3.13: Classification of the Taita thrushes (Example 4) into the $\widehat{K} = 3$ inferred populations of origin and the location where the were captured. Names between dashed vertical lines: capture location.

Bibliography

Hirotogu Akaike. Information Theory and an Extension of the Maximum Likelihood Principle, pages 199–213. Springer New York, New York, NY, 1973.

C. Biernacki, G. Celeux, and G. Govaert. Assessing a mixture model for clustering with the integrated completed likelihood. IEEE Trans. Pattern Anal. Machine Intel., 22(7):719–25, 2000.

Varun Chandola, Arindam Banerjee, and Vipin Kumar. Anomaly detection: A survey. ACM computing surveys (CSUR), 41(3):1–58, 2009.

J. Chiquet, M. Mariadassou, and S. Robin. The Poisson-lognormal model as a versatile framework for the joint analysis of species abundances. Frontiers in Ecology and Evolution, 9:188, 2021. doi: 10.3389/fevo.2021.588292. URL `https://www.frontiersin.org/article/10.3389/fevo.2021.588292`.

Noel Cressie. Statistics for spatial data. John Wiley & Sons, 2015.

A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the em algorithm. Journal of the Royal Statistical Society. Series B (Methodological), 39(1):1–38, 1977. ISSN 00359246. URL `http://www.jstor.org/stable/2984875`.

Brian S Everitt, Sabine Landau, Morven Leese, and Daniel Stahl. Cluster analysis. John Wiley & Sons, 2011.

M. Fossheim, E. M Nilssen, and M. Aschan. Fish assemblages in the Barents Sea. Marine Biology Research, 2 (4):260–269, 2006.

P. Galbusera, L. Lens, T. Schenck, E. Waiyaki, and E. Matthysen. Genetic variability and gene flow in the globally, critically-endangered taita thrush. Conservation Genetics, 1:45–55, 2000.

Lars Götzenberger. traitor: Tools For Functional Diversity Assessment With Missing Trait Data, 2015. R package version 0.0.0.9001.

Xavier A. Harrison, Lynda Donaldson, Maria Eugenia Correa-Cano, Julian Evans, David N. Fisher, Cecily E.D. Goodwin, Beth S. Robinson, David J. Hodgson, and Richard Inger. A brief introduction to mixed effects modelling and multi-model inference in ecology. PeerJ, 2018, 2018. ISSN 21678359. doi: 10.7717/peerj.4794.

Allison Marie Horst, Alison Presmanes Hill, and Kristen B Gorman. palmerpenguins: Palmer Archipelago (Antarctica) penguin data, 2020. URL `https://allisonhorst.github.io/palmerpenguins/`. R package version 0.1.0.

Gerard Janssen and Mulder Saskia. Zonation of macrofauna across sandy beaches and surf zones along the dutch coast. Oceanologia, 47, 06 2005.

A. Jeliazkov, D. Mijatovic, S. Chantepie, N. Andrew, R. Arlettaz, L. Barbaro, N. Barsoum, A. Bartonova, E. Belskaya, and N. Bonada. A global database for metacommunity ecology, integrating species, traits, environment and space. Scientific data, 7(1):6, 2020.

J. Josse, J. Pagès, and F. Husson. Multiple imputation in principal component analysis. Advances in Data Analysis and Classification, 5:231–246, 2011.

Steffen L. Lauritzen. Graphical Models. Oxford University Press, 1996. ISBN 0-19-852219-3.

E. Lebarbier and T. Mary-Huard. Une introduction au critère BIC : fondements théoriques et interprétation. J. Soc. Française Statis., 147(1):39–57, 2006.

Jan Lepš, Francesco de Bello, Petr Šmilauer, and Jiří Doležal. Community trait response to environment: disentangling species turnover vs intraspecific trait variability effects. Ecography, 34(5):856–863, 2011.

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

K.V. Mardia, J.T. Kent, and J.M. Bibby. Multivariate analysis. Academic press, 1979.

G. McLachlan and D. Peel. Finite Mixture Models. Wiley, 2000.

Geoffrey J. McLachlan and Thriyambakam Krishnan. The EM Algorithm and Extensions, 2E. [John Wiley & Sons, Inc]., 2 2008. ISBN 9780470191613. doi: 10.1002/9780470191613.

Kaare Brandt Petersen and Michael Syskind Pedersen. The matrix cookbook. Technical University of Denmark, 7(15):510, 2008. URL https://www.math.uwaterloo.ca/~hwolkowi/matrixcookbook.pdf.

N. Peyrard and O. Gimenez. Statistical Approaches for Hidden Variables in Ecology. Wiley, 2022. ISBN 9781119902782. URL https://books.google.fr/books?id=kG9jEAAAQBAJ.

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

R Core Team. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria, 2022. URL https://www.R-project.org/.

Donald B Rubin. Inference and missing data. Biometrika, 63(3):581–592, 1976.

Gideon Schwarz. Estimating the dimension of a model. The annals of statistics, pages 461–464, 1978.

M. E Tipping and C. M Bishop. Probabilistic principal component analysis. Journal of the Royal Statistical Society: Series B, 61(3):611–622, 1999.

S. Villéger, J. R. Miranda, D. F. Hernandez, and D. Mouillot. Low functional β-diversity despite high taxonomic β-diversity among tropical estuarine fish communities. PloS one, 7(7):e40679, 2012.

C. F. Jeff Wu. On the Convergence Properties of the EM Algorithm. The Annals of Statistics, 11(1):95 – 103, 1983. doi: 10.1214/aos/1176346060. URL https://doi.org/10.1214/aos/1176346060.

Alain F. Zuur, Elena N. Ieno, Neil Walker, Anatoly A. Saveliev, and Graham M. Smith. Mixed effects models and extensions in ecology with R. Springer New York, 2009. ISBN 978-0-387-87457-9. doi: 10.1007/978-0-387-87458-6. URL https://link.springer.com/book/10.1007/978-0-387-87458-6#bibliographic-information.

Appendix A

Appendix

A.1 Multivariate distributions

A.1.1 General properties (SR)

Proposition 9. *Let U be a random real vector with mean vector $\mathbb{E}(U) = \mu$ and variance matrix $\mathbb{V}(U) = \Sigma$ and let A be symmetric matrix, denoting $\|U\|_A^2 = U^\top AU$, we have*

$$\mathbb{E}(\|U\|_A^2) = \|\mu\|_A^2 + \operatorname{tr}(A\Sigma).$$

Proof of Proof

Because A is symmetric, we may write it as $A = BB^\top$, so taking $V = BU$ we get

$$\mathbb{E}(V) = \mathbb{E}(BU) = B\mu, \qquad \mathbb{V}(V) = \mathbb{V}(BU) = B\Sigma B^\top, \qquad \mathbb{E}(\|U\|_A^2) = \mathbb{E}(V^\top V).$$

Furthermore, by definition of the variance, we have that

$$\mathbb{E}(V^\top V) = \mathbb{E}(V)^\top \mathbb{E}(V) + \operatorname{tr}(\mathbb{V}(V)) = \mu^\top B^\top B\mu + \operatorname{tr}(B\Sigma B^\top),$$

which gives the result because $\operatorname{tr}(B\Sigma B^\top) = \operatorname{tr}(B^\top B\Sigma)$.

A.1.2 Multivariate normal distribution (SR, PG)

Proposition 10. *The entropy of the multivariate Gaussian distribution in dimension p is*

$$\mathsf{Ent}[\mathcal{N}(\mu, \Sigma)] = \frac{p}{2}(1 + \log(2\pi)) + \frac{1}{2}\log|\Sigma|.$$

Proof of Proof

The entropy of a distribution $\mathcal{F}$ with density f is $-\mathbb{E}_f[\log f(X)]$, so, taking $X \sim \mathcal{N}(\mu, \Sigma)$, we have

$$\mathsf{Ent}[\mathcal{N}(\mu, \Sigma)] = \frac{1}{2}\left(\log|\Sigma| + p\log(2\pi) + \mathbb{E}\left[\|X - \mu\|_{\Sigma^{-1}}^2\right]\right).$$

The result follows from Proposition 9 because $\mathbb{E}(X - \mu) = 0_p$ and $\operatorname{tr}(\Sigma\Sigma^{-1}) = \operatorname{tr}(I_p) = p$.

Proposition 11. *Let (U, V) be a couple of real random vectors with respective mean vectors $\mathbb{E}(U) = \mu_U$ and $\mathbb{E}(V) = \mu_V$ and respective variance and covariance matrices $\mathbb{V}(U) = \Sigma_{UU}$, $\mathbb{V}(V) = \Sigma_{VV}$ and $\mathbb{C}ov(U, V) =$*

$\Sigma_{UV} = \Sigma_{VU}^\top = \mathbb{C}ov(V,U)^\top$ *and joint multivariate Gaussian distribution :*

$$\begin{bmatrix} U \\ V \end{bmatrix} \sim \mathcal{N}\left(\mu = \begin{bmatrix} \mu_U \\ \mu_V \end{bmatrix}, \Sigma = \begin{bmatrix} \Sigma_{UU} & \Sigma_{VU} \\ \Sigma_{UV} & \Sigma_{VV} \end{bmatrix}\right),$$

then the conditional distribution of V *given* U *is Gaussian and*

$$\mathbb{E}(V \mid U) = \mu_V + \Sigma_{VU}\Sigma_{UU}^{-1}(U - \mu_U), \qquad \mathbb{V}(V \mid U) = \Sigma_{VV} - \Sigma_{VU}\Sigma_{UU}^{-1}\Sigma_{UV}.$$

Proof of Proof

See Mardia et al. [1979], Section 3.1.

Conditional distributions in the linear regression Let us consider the following model

$$Y \sim \mathcal{N}\left(X\beta, \sigma^2\mathbb{I}_n\right), \quad \beta \sim \mathcal{N}(\mu, \Omega)$$

then

$$\beta|Y \sim \mathcal{N}(m, \Sigma)$$

with

Indeed

$$\begin{bmatrix} Y \\ \beta \end{bmatrix} \sim \mathcal{N}\left(\mu = \begin{bmatrix} X\mu_0 \\ \mu_0 \end{bmatrix}, \Sigma = \begin{bmatrix} X\Omega X^\top + \sigma^2\mathbf{I}_n & X\Omega \\ \Omega X^\top & \Omega \end{bmatrix}\right),$$

So

$$\begin{aligned} \mathbb{V}[\beta|Y] &= \Omega - X\Omega(X\Omega X^\top + \sigma^2\mathbf{I}_n)^{-1}\Omega X^\top \\ \mathbb{E}[\beta|Y] &= \mu_0 + X\Omega(X\Omega X^\top + \sigma^2\mathbf{I}_n)^{-1}(Y - X\mu_0) \end{aligned}$$

If $\mu_0 = 0_p$ and $\Omega = \omega^2\mathbf{I}_p$ then:

$$\begin{aligned} \mathbb{V}[\beta|Y] &= \omega^2\mathbf{I}_p - \omega^2 X(\omega^2 XX^\top + \sigma^2\mathbf{I}_n)^{-1}\omega^2 X^\top \\ \mathbb{E}[\beta|Y] &= X\omega^2(\omega^2 XX^\top + \sigma^2\mathbf{I}_n)^{-1}Y \end{aligned}$$

A.1.2.1 Useful results about Gaussian vectors

Let's consider the following model.

Model 8. *Gaussian prior and linear Gaussian likelihood*

$$\begin{aligned} Z &\sim \mathcal{N}_{d_Z}(\mu, \Omega) \\ Y|Z &= \mathcal{N}_{d_Y}\left(AZ + b, \Sigma\right), \end{aligned}$$

where A *is a* $d_Y \times d_Z$ *matrix.*

Then, the two following Propositions gives us the law of $Z \mid Y$ and the marginal distribution of Y.

Proposition 12 (Posterior distribution of Z in Model 8)**.** *Let's consider a random couple* (Z,Y) *satisfying model 8. Let's write* $\Lambda = \Omega^{-1}$ *and* $\Gamma = \Sigma^{-1}$ *the corresponding precision matrices in this model. Then, we have:*

$$\begin{aligned} Z \mid Y &\sim \mathcal{N}_{d_Z}\left(\left(\Lambda + A^\top\Gamma A\right)^{-1}\left(A^\top\Gamma(Y - b) + \Lambda\mu\right), \left(\Lambda + A\Gamma A^\top\right)^{-1}\right) \\ \textit{Or, equivalently, } Z \mid Y &\sim \mathcal{N}_{d_Z}\left(\mu + K(Y - b - A\mu), (I_{d_Z} - KA)\,\Omega\right) \\ &\quad \textit{where } K = \Omega A^\top(\Sigma + A\Omega A^\top)^{-1} \end{aligned}$$

The proof (and the one of the following Proposition) will require these useful results to manipulate inverse of matrices:

Lemma 2 (Woodbury identity and inverse of a blockwise defined matrix)**.** *Let A and D be invertible square matrices of size $d_A \times d_A$ and $d_D \times d_D$ respectively, and B and C be rectangular matrices of size $d_A \times d_D$ and $d_D \times d_A$ respectively. We have the following results:*

1. *Woodbury identity*
$$(A + BD^{-1}C)^{-1} = \left(I_{d_A} - A^{-1}B(D + CA^{-1}B)^{-1}C\right)A^{-1}.$$
2. *Let M be an invertible $d_M \times d_M$ matrix that can be written block wise:*
$$M = \begin{pmatrix} A & B \\ C & D \end{pmatrix},$$
Then, we have that:
$$M^{-1} = \begin{pmatrix} S & -SBD^{-1} \\ -D^{-1}CS & D^{-1} + D^{-1}CSBD^{-1} \end{pmatrix},$$
where
$$S = (A - BD^{-1}C)^{-1}.$$

Proof of Proof

For both results, the proof is obtained by multiplying doing matrix multiplication and checking that it gives the identity matrix.

Proof of Proof of Proposition 12

First, let's recall a general fact about gaussian vectors. If $X \sim \mathcal{N}_d(m, V)$, and writing $P = V^{-1}$ then, it's probability density function satisfies, for $x \in \mathbb{R}^d$

$$\log p(x) = -\frac{1}{2}(x-m)^\top P(x-m) + \text{Cst} = -\frac{1}{2}x^\top Px + x^\top Pm + \text{Cst}, \tag{A.1}$$

where the Cst[a] stands for a constant that does not depend on X. Moreover, every probability density functions that has the form of Equation (A.1) is the one of a Gaussian vector with mean m and variance $V = P^{-1}$. The technique of the proof, well known by bayesian users as *completing the square*, consists in identifying such form for the law of $Z|Y$ and Y.

Let's then consider the p.d.f. of $Z|Y$. In the following, we only keep the terms depending on z.

$$\begin{aligned}
\log p(z|y) &= \log p(z) + \log p(y|z) + \text{Cst} \\
&= -\frac{1}{2}z^\top \Lambda z + z^\top \Lambda \mu - \frac{1}{2}z^\top A^\top \Gamma A z + z^\top A^\top \Gamma (y-b) + \text{Cst} \\
&= -\frac{1}{2}z^\top (\Lambda + A^\top \Gamma A) + z^\top (A^\top \Gamma (y-b) + \Lambda \mu) + \text{Cst}.
\end{aligned}$$

We (almost) recognize the form of Equation (A.1) where $P = \Lambda + A^\top \Gamma A$ (which is the wanted result). However, in (A.1), P also appears in the linear term, to multiply m. We then make the precision matrix appears:

$$\begin{aligned}
\log p(z|y) &= -\frac{1}{2}z^\top (\Lambda + A^\top \Gamma A) + z^\top \left(A^\top \Gamma (y-b) + \Lambda \mu\right) + \text{Cst} \\
&= -\frac{1}{2}z^\top (\Lambda + A^\top \Gamma A) + z^\top (\Lambda + A^\top \Gamma A)(\Lambda + A^\top \Gamma A)^{-1} A^\top \Gamma (y-b) + \text{Cst}.
\end{aligned}$$

Which prove the result, as we have a p.d.f. of the form (A.1). Then, $Z|Y$ follows a multivariate gaussian distribution with variance $(\Lambda + A^\top \Gamma A)^{-1}$ and mean $(\Lambda + A^\top \Gamma A)^{-1} A^\top \Gamma (y-b)$. Now, by Woodbury identity of Lemma 2, we have that (retransforming precision matrices into variances):

$$\left(\Lambda + A^\top \Gamma A\right)^{-1} = \left(I_{d_Z} - \Omega A(\Sigma + A\Omega A^\top)^{-1} A\right)\Omega,$$

And therefore:

$$\begin{aligned}\left(\Lambda + A^\top \Gamma A\right)^{-1}\left(A^\top \Gamma (Y-b) + \Lambda\mu\right) &= \left(I_{d_Z} - \Omega A(\Sigma + A\Omega A^\top)^{-1}A\right)\Omega\left(A^\top\Sigma^{-1}(Y-b) + \Omega{-}1\mu\right)\\ &= \mu - KA\mu + \Omega A^\top\Sigma^{-1}(Y-b) - KA\Omega A^\top\Sigma^{-1}(Y-b)\\ &= \mu - KA\mu + K(\Sigma + A\Omega A^\top)\Sigma^{-1}(Y-b) - KA\Omega A^\top\Sigma^{-1}(Y-b\\ &= \mu + K(Y - b - A\mu)\end{aligned}$$

[a] Note that this generic term potentially refers to a different constant at each step of the computation.

Proposition 13 (Marginal distribution of Y in Model 8)**.** *Let's consider a random couple (Z,Y) satisfying Model 8. Then the marginal distribution of Y is given by: Then, we have:*

$$Y \quad\sim\quad \mathcal{N}_{d_Y}\left(A\mu + b, \Sigma + A\Omega A^\top\right).$$

Proof of Proof

We start by writing the joint p.d.f. of (Z,Y) (without omitting the terms involving y this time).

$$\begin{aligned}\log p(z,y) &= \log p(z) + \log p(y|z) + \text{Cst}\\ &= -\frac{1}{2}z^\top\Lambda z + z^\top\Lambda\mu - \frac{1}{2}z^\top A^\top\Gamma A z - \frac{1}{2}y^\top\Gamma y + z^\top A^\top\Gamma y + y^\top\Gamma b - z^\top A^\top\Gamma b + \text{Cst}\\ &= -\frac{1}{2}\begin{pmatrix}z^\top & y^\top\end{pmatrix}\begin{pmatrix}\Lambda + A^\top\Gamma A & -A^\top\Gamma\\ -\Gamma A & \Gamma\end{pmatrix}\begin{pmatrix}z\\ y\end{pmatrix} + \begin{pmatrix}z^\top & y^\top\end{pmatrix}\begin{pmatrix}\Lambda\mu - A^\top\Gamma b\\ \Gamma b\end{pmatrix}\end{aligned}$$

Now, let's stop here. Denoting $x = (z,y)^\top$, we again (almost) recognize the form of Equation (A.1), where we have the precision matrix

$$P = \begin{pmatrix}\Lambda + A^\top\Gamma A & -A^\top\Gamma\\ -\Gamma A & \Gamma\end{pmatrix}.$$

To higlight the mean vector, we then have to make P appear in the linear term. We have that

$$\log p(z,y) = -\frac{1}{2}\begin{pmatrix}z^\top & y^\top\end{pmatrix}P\begin{pmatrix}z\\ y\end{pmatrix} + \begin{pmatrix}z^\top & y^\top\end{pmatrix}PP^{-1}\begin{pmatrix}\Lambda\mu - A^\top\Gamma b\\ \Gamma b\end{pmatrix}$$

Then, we have that the mean is given by $P^{-1}\begin{pmatrix}\Lambda\mu - A^\top\Gamma b\\ \Gamma b\end{pmatrix}$. Now, Lemma 2 gives us a formula to compute P^{-1}. We easily check that the corresponding S matrix is given by $\Lambda^{-1} = \Omega$, and thus, that the variance-covariance matrix of (Z,Y) and the expectation are given by[a]:

$$\begin{aligned}P^{-1} &= \begin{pmatrix}\Omega & \Omega A^\top\\ A\Omega & \Sigma + A\Omega A^\top\end{pmatrix}\\ m = P^{-1}\begin{pmatrix}\Lambda\mu - A^\top\Gamma b\\ \Gamma b\end{pmatrix} &= \begin{pmatrix}\mu\\ A\mu + b\end{pmatrix}.\end{aligned}$$

Thus, having the expectation and variance-covariance of the Gaussian vector (Z,Y), the marginal in Y is also a Gaussian vector with variance-covariance $\Sigma + A\Omega A^\top$ and expectation $A\mu + b$.

[a] Recalling that $\Gamma^{-1} = \Sigma$

The last proof actually gave us the following result:

Proposition 14 (Joint distribution of (Y,Z) in Model 8)**.** *Let's consider a random couple (Z,Y) satisfying*

Model 8. Then the joint distribution of (X, Y) is given by:

$$\begin{pmatrix} Z \\ Y \end{pmatrix} \sim \mathcal{N}_{d_z+d_y}\left(\begin{pmatrix} \mu \\ A\mu + b \end{pmatrix}, \begin{pmatrix} \Omega & \Omega A^\top \\ A\Omega & \Sigma + A\Omega A^\top \end{pmatrix}\right).$$

A.1.3 Other multivariate distributions (SR)

A.1.3.1 Dirichlet distribution

Definition 3 (Dirichlet distribution)**.** *The Dirichlet distribution with parameter $\alpha \in \mathbb{R}^{+d}$, denoted by $\mathcal{D}(\alpha)$, is defined over the simplex of $\mathbb{R}^d$: $\{x \in (\mathbb{R}^+)^d : \sum_{k=1}^d x_k = 1\}$. Its probability mass function is*

$$p_\alpha(x) = \frac{1}{D(\alpha)} \prod_{k=1}^d x_k^{\alpha_k - 1}$$

where $D(\alpha) = \left(\prod_{k=1}^d \Gamma(\alpha_k)\right) \Big/ \Gamma(\alpha_+)$ and $\alpha_+ = \sum_{k=1}^d \alpha_k$.

Proposition 15. *Let $X \sim \mathcal{D}(\alpha)$, for $k \in \{1, \dots K\}$, it holds that*

$$\mathbb{E}(X_k) = \frac{\alpha_k}{\alpha_+}, \qquad \mathbb{V}(X_k) = \frac{\alpha_k(\alpha_+ - \alpha_k)}{\alpha_+^2(\alpha_+ + 1)}, \qquad \mathbb{E}(\log X_k) = \psi(\alpha_k) - \psi(\alpha_+),$$

where ψ stands for the di-gamma function, that is, the derivative of the log-gamma function: $\psi(x) = \partial_x \log(\Gamma(x))$, and, for $1 \le k < \ell \le d$,

$$\mathbb{C}ov(X_k, X_\ell) = \frac{\alpha_k \alpha_\ell}{\alpha_+^2(\alpha_+ + 1)}.$$

Furthermore, its entropy is

$$\mathsf{Ent}(X) = \log D(\alpha) + (\alpha_+ - K)\log\psi(\alpha_+) - \sum_{k=1}^K (\alpha_k - 1)\log\psi(\alpha_k).$$

The Dirichlet distribution id the conjugate of the multinomial distribution: let $W \sim \mathcal{D}(\alpha)$ and $X \mid W \sim \mathcal{M}(n, W)$:

$$p(x \mid W) = \binom{n}{x_1 \dots x_d} \prod_{k=1}^d W_k^{x_k},$$

then the conditional distribution of W given $X = x$ is Dirichlet with parameter $\widetilde{\alpha} = \alpha + x$:

$$W \mid X = x \sim \mathcal{D}(\alpha + x).$$

In this setting, the marginal distribution of X is a Dirichlet-multinomial $\mathcal{DM}(n, \alpha)$.

Definition 4 (Dirichlet-multinomial distribution)**.** *The Dirichlet-multinomial distribution with parameters $n \in \mathbb{N}^*$ and $\alpha \in (\mathbb{R}^+)^d$, denoted $\mathcal{DM}(n, \alpha)$, is defined for $x \in \mathbb{N}^d$ such that $\sum_{k=1}^d x_i = n$. Its probability mass function is*

$$p(x) = \binom{n}{x_1 \dots x_d} \frac{D(\widetilde{\alpha})}{D(\alpha)} = \frac{\Gamma(\alpha_+)\Gamma(n+1)}{\Gamma(\widetilde{\alpha}_+)} \prod_{k=1}^d \frac{\Gamma(\alpha_k)\Gamma(x_k+1)}{\Gamma(\widetilde{\alpha}_k)}$$

where $\widetilde{\alpha}_k = \alpha_k + x_k$, $\widetilde{\alpha} = [\widetilde{\alpha}_k]_{1 \le k \le K}$ and $\widetilde{\alpha}_+ = \sum_{k=1}^d \widetilde{\alpha}_k$

A.2 Exponential family and generalized linear models

A.2.1 The natural exponential family (SR, SD)

The natural exponential family is a family of probability distributions that includes such common distributions as the normal distribution, Bernoulli's distribution, the binomial distribution, Poisson's distribution, Gamma distribution and others. What these distributions have in common is that they are written in exponential form. unified presentation of results.

Definition 5. *The distribution $f(;\gamma)$ belongs to exponential family with parameter $\gamma \in \mathbb{R}^d$ if*

$$f(y;\gamma) = \exp[\gamma^\top t(y) - a(y) - b(\gamma)] \tag{A.2}$$

where $t(y) \in \mathbb{R}^d$ is the vector of the sufficient statistics , $\gamma \mapsto b(\gamma) \in \mathbb{R}$ is a differentiable function.

The parameter γ is linked to the sufficient statistics as follows:

Proposition 16.

$$\mathbb{E}_\gamma[t(Y)] = \nabla b(\gamma), \qquad \mathbb{V}_\gamma(t(Y)] = \operatorname{Hess} b(\gamma)$$

Proof of Proof

Using the fact that $\int_{\mathcal{Y}} f_Y(y;\gamma)dy = 1$, we derive that expression with respect to γ (interverting the derivative and the $\int$ by regularity, with no demonstration).

$$\begin{aligned} 0 = \int \nabla_\gamma f(y;\gamma)\,\mathrm{d}y &= \int \nabla_\gamma \exp[\gamma^\top t(y) - a(y) - b(\gamma)]\,\mathrm{d}y \\ &= \int [t(y) - \nabla_\gamma b(\gamma)] \exp[\gamma^\top t(y) - a(y) - b(\gamma)]\,\mathrm{d}y \\ &= \int [t(y) - \nabla_\gamma b(\gamma)]\, f(y;\gamma)\,\mathrm{d}y \end{aligned}$$

that is

$$0 = [\mathbb{E}[t(Y)] - b'(\gamma)] \underbrace{\int f_Y(y;\gamma,\phi)\,\mathrm{d}y}_{=1}.$$

Differentiating one more time, we get

$$\begin{aligned} 0 &= \int \mathbf{J}_\gamma \left[(t(y) - \nabla_\gamma b(\gamma))\, f(y;\gamma) \right] \mathrm{d}y \\ &= \int -\operatorname{Hess} b(\gamma) f(y;\gamma)\,\mathrm{d}y + \int [t(y) - \nabla_\gamma b(\gamma)]^\top [t(y) - \nabla_\gamma b(\gamma)]\, f(y;\gamma)\,\mathrm{d}y \\ &= -\operatorname{Hess} b(\gamma) + \int [t(y) - \mathbb{E}(t(Y))]^\top [t(y) - \mathbb{E}(t(Y))] f(y;\gamma)\,\mathrm{d}y \\ &= -\operatorname{Hess} b(\gamma) + \mathbb{V}((Y)) \end{aligned}$$

Proposition 17. *For an iid sample $(Y_1, \dots Y_n)$ of $f(\cdot;\gamma)$, the MLE of γ, $\widehat{\gamma}$ 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 of Proof

The log-likelihood of an iid sample is

$$\log p_\gamma(Y_1, \dots Y_n) = \sum_{i=1}^n \left[\gamma^\top t(Y_i) - a(Y_i) - b(\gamma) \right] \tag{A.3}$$

and its derivative wrt γ is

$$\partial_\gamma \log p_\gamma(Y_1, \dots Y_n) = \sum_{i=1}^n t(Y_i) - n b'(\gamma).$$

Setting it to zero gives the result.

Proposition 18. *The log-likelihood* (A.3) *is concave wrt γ.*

Proof of Proof

On the one hand, the Hessian matrix of the log-likelihood (A.3) is $-n \operatorname{Hess} b(\gamma)$. On the other hand, Proposition 16 states that $\operatorname{Hess} b(\gamma)$ is a variance matrix, so it is positive definite. As a consequence, the Hessian matrix of the log-likelihood is negative definite and the log-likelihood is concave.

A.2.1.1 Canonical parameter

If, in addition, $y \mapsto t(y)$ is the identity then γ is called the canonical parameter (or natural parameter) and is related to the mean through

$$\mathbb{E}[Y] = \nabla b(\gamma) \tag{A.4}$$

A.2.2 Generalized linear models (SR)

The Generalized Linear Model (GLM) is an extension of the linear model that allows us to deal with observations whose probability distribution belongs to an extended family of distributions.

More precisely, let $y = (y_1, \ldots, y_n)$ be the vector of observations, X is the matrix of explanatory variables. The linear model is written as

$$Y \sim \mathcal{N}(X\beta, \sigma^2 I_n).$$

Definition 6. *The generalized linear is given by a probability distribution for Y_i and a function g called the link function such that*

$$g(\mathbb{E}[Yi]) = x_i\beta.$$

This establishes a non-linear relationship between the expectation of the variable to be explained and the explanatory variables and allows us to consider observations of a varied nature, such as presence/absence data presence/absence data, success rates for treatments, species count data, or even lifetimes or other positive asymmetric variables.

Examples

- If $y_i \in \{0, 1\}$ then naturally we'll use Bernoulli distribution to model y_i:

  $$Y_i \sim \mathcal{B}ern(p_i) \quad \text{with} \quad \mathbb{E}[Y_i] = p_i \in [0, 1]$$

  Let $g : [0, 1] \mapsto \mathbb{R}$, be bijective we pose

  $$g(\mathbb{E}[Y_i]) = x_i\beta \Leftrightarrow \mathbb{E}[Y_i] = g^{-1}(x_i\beta)$$

  We can choose for $g^{-1}(u) = \frac{e^u}{1+e^u}$ i.e. $g(p) = \operatorname{logit}(p) = \log \frac{p}{1-p}$ $(p \in [0, 1])$. Any other function $g : [0, 1] \mapsto \mathbb{R}$ can be used, for example we can choose for g the distribution function of a reduced centered normal distribution (g = probit).

- If $y_i \in \mathbb{N}$ then we may want to use Poisson's distribution to model y_i: $Y_i \sim \mathcal{P}(\mu_i)$. In this case, $sp[Y_i] = \mu_i \in \mathbb{R}^+$. We're looking for $g : \mathbb{R}^+ \mapsto \mathbb{R}$, bijective to set

  $$g(\mathbb{E}[Y_i]) = x_i\beta \Leftrightarrow \mathbb{E}[Y_i] = g^{-1}(x_i\beta).$$

  For example, we can choose $g^{-1}(u) = e^u$ i.e. $g(\mu) = \log \mu$ $(\mu \in \mathbb{R}^+)$.

Choose the link function Any bijection from the space of $\mathbb{E}[Y]$ into $\mathbb{R}$ can be chosen as a link function. However, very often we choose as link function the function that transforms the expectation $E[Y]$ into the natural parameter i.e.

$$g = (b')^{-1}$$

g defined in this way is called the canonical link function.

Therefore, since $g(\mathbb{E}[Y_i]) = x_i\beta$ and furthermore, by Equation A.4, we have : $\mathbb{E}[Y_i] = g^{-1}(x_i\beta) = b'(\gamma_i)$, we obtain

$$\gamma_i = (b')^{-1} g^{-1}(x_i\beta) = x_i\beta$$

for this particular choice.

Examples

- For the exponential distribution,

$$b(\gamma) = -\log(-\gamma), \quad \text{so} \quad b'(\gamma) = \frac{-1}{\gamma}, \quad \text{so} \quad g(\mu) = -\frac{1}{\mu}.$$

- For the Bernoulli distribution,

$$f_Y(y) = \exp\{y\log\pi + (1-y)\log(1-\pi)\} = \exp\left\{y\log\left(\frac{\pi}{1-\pi}\right) + \log(1-\pi)\right\}$$

So $\gamma = \log\left(\frac{\pi}{1-\pi}\right)$ i.e. $\pi = \frac{e^\gamma}{1+e^\gamma}$ et $b(\gamma) = -\log(1-\pi) = -\log\left(\frac{1}{1+e^\gamma}\right) = \log(1+e^\gamma)$. As a consequence

$$b'(\gamma) = \frac{e^\gamma}{1+e^\gamma} \quad \text{and} \quad g(\mu) = (b')^{-1}(\mu) = \log\left(\frac{\mu}{1-\mu}\right)$$

we obtain the logit function.

The choice of link function is an additional freedom in the modelling process. The specific choice of the canonical (or natural) link function is motivated by theoretical considerations. In fact, it ensures the convergence of the estimation algorithm algorithm (Newton-Raphson algorithm) towards the maximum likelihood. In practice, if there is no reason to choose a specific link function, the default choice is to choose the canonical link function. The Table refMLGusuels presents some of the best-known generalized linear models. For each choice of $Y|X = x$ distribution, there is a canonical link function $g(\cdot)$ which gives the regression its name.

Choice of the of $Y\|x$	Bernoulli Binomial	Poisson	Gamma	Gaussian
Link function	$g(\mu) = \text{logit}(\mu)$	$g(\mu) = \log(\mu)$	$g(\mu) = -\frac{1}{\mu}$	$g(\mu) = \mu$
Link name	logit	log	Inverse	Identity

Table A.1: MLG usuels. $\mathbb{E}_\beta[Y] := \mu$

Other non-canonical link functions are used in practice.

- The link probit: $g(\mu) = \Phi^{-1}(\mu)$ where $\Phi(\cdot)$ is the distribution function of a reduced central Gaussian.
- The link log-log: $g(\mu) = \log(-\log(1-\mu))$ with $\mu \in]0,1[$.

A.3 Graphical models (SD)

In this section we only provide a few notions on DAG. For more formalism and theory, please refer to the book by Lauritzen Lauritzen [1996].

A graphical model is a type of probabilistic model where a graph encodes the conditional independence structure. In this model, nodes represent random variables, and edges signify the conditional independence relations between these variables. By examining the graph, we can understand how the joint distribution breaks down into a product of smaller components, each involving only a subset of the variables.

A.3.1 Directed acyclic graph (DAG)

A graph is a collection of nodes (in this case representing the random variables) $\{V_1, \dots, V_N\}$ and edges which are couples of nodes. In a directed graph, the edges have directions (and so are represented as arrows). Figure A.1 is an exemple of directed graph.

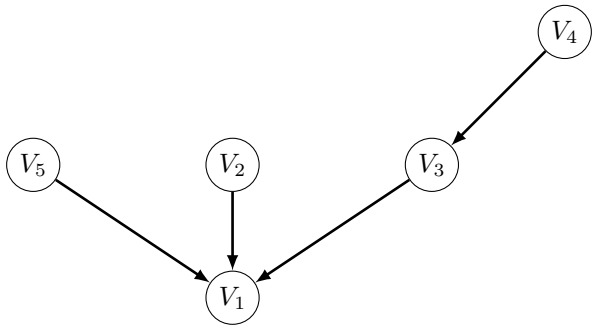
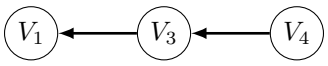


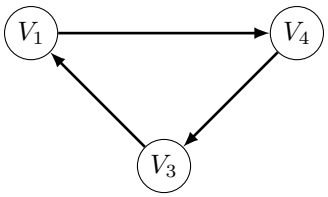
Figure A.1: An example of directed acyclic graph with vertices $V = \{V_1, \dots, V_5\}$ and edges $E = \{(V_5, V_1), (V_2, V_1), (V_3, V_1), (V_4, V_3)\}$

If there is an arrow from X to Y then X is said to be a parent for Y. We denote $pa(X, G)$ all the parents of X in graph G. In Figure A.1, the parents of node V_1 are $pa(V_1, G) = \{V_5, V_2, V_3\}$ and $pa(V_3, G) = \{V_4\}$. A directed path between two variables is a set of arrows in the same direction linking one node to the other as a chain:



X is an ancestor for Y is there is a path from X to Y or equivalently, Y is a descendant of X.

A cycle is a directed path that starts and ends at the same node.



A directed graph is said to be acyclic (DAG) if it has no cycle. Graphical models are only defined for DAG.

Remark. Note that directed graph associated to a probabilistic distribution is also refered as Bayesian network which is quite confusing since the terminology does not refer to the inference method (a Bayesian network may be infered by a frequentist approach).

A.3.2 DAG and probability

Let us consider a set of variables $V = \{V_i\}_{1:N}$ and a DAG $G(V, E)$. The distribution P on V is said to be factorized with respect to G si est

$$\mathbb{P}(V_{1:N}) = \prod_{i=1}^{N} \mathbb{P}(V_i | pa(V_i, G))$$

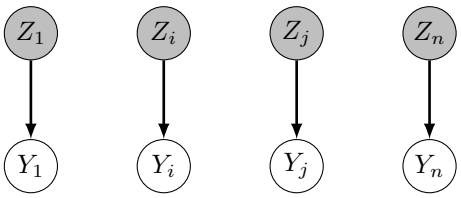
where $pa(V_i, G)$ refers to the parents of V_i in graph G.

Examples of DAG and corresponding joint probability

- The joint distribution corresponding to the DAG of Figure A.1

$$\mathbb{P}(V_1, V_2, V_3, V_4, V_5) = \mathbb{P}(V_1|V_2, V_3, V_5)\mathbb{P}(V_2)\mathbb{P}(V_5)\mathbb{P}(V_3|V_4)\mathbb{P}(V_4)$$

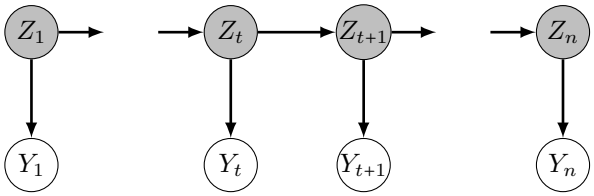
- DAG for the mixture model (Section 3.1) or for the zero-inflated Poisson model (Section 3.2) :



resulting into:

$$\mathbb{P}(Y_{1:n}, X_{1:n}) = \prod_{i=1}^{n} \mathbb{P}(Y_i|Z_i)\mathbb{P}(Z_i)$$

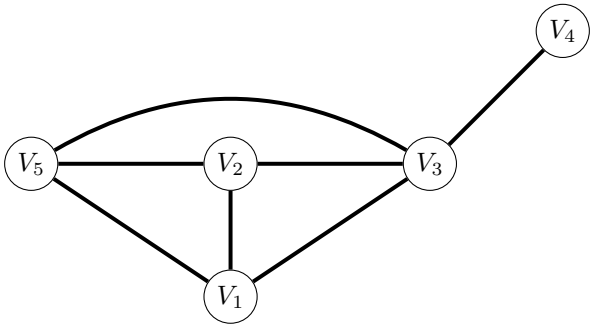
- DAG for Hidden Markow model



resulting into:

$$\mathbb{P}(Y_{1:n}, X_{1:n}) = \prod_{t=1}^{n} \mathbb{P}(Y_t|Z_t) \prod_{t=2}^{n} \mathbb{P}(Z_t|Z_{t-1})\mathbb{P}(Z_1)$$

Moralization of a DAG The moral version of a graph G is obtained by marrying the parents (setting an edge) and then by removing the directions on the edges. The moral version of the DAG of Figure A.1 is the following one:



An undirected path between two nodes is a set of edges (ignoring the directions) linking one node to the other.

Independancy properties

Proposition 19. *Let I, J and K be 3 subsets of V.*

1. *In the moral graph deduced from G, if all the paths from I to J pass through K then*

$$(V_i)_{i\in I} \perp\!\!\!\perp (V_j)_{j\in J} | (V_k)_{k\in K}.$$

As a consequence:

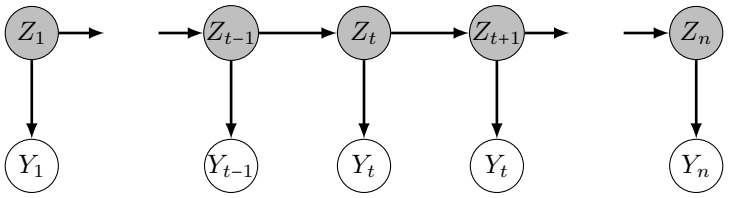
$$\mathbb{P}(V_I|V_J, V_K) = \mathbb{P}(V_I|V_K)$$

2. In a DAG G, conditionnally to its parents, a variable is independant from its non-descendants.

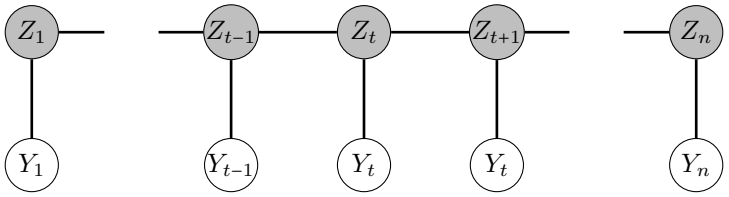
Illustration of Proposition 19 on the DAG of Figure A.1

1. Let us set $I = \{5, 2, 1\}, J = \{4\}, K = \{3\}$. All paths from I to J go through K so:
$$\mathbb{P}(V_5, V_2, V_1, V_4|V_3) = \mathbb{P}(V_5, V_2, V_1|V_3)\mathbb{P}(V_4|V_3)$$
2. $\mathbb{P}(V_1|V_2, V_3, V_4, V_5) = \mathbb{P}(V_1|V_2, V_3, V_5)$

A.3.3 Independance properties for HMM



Moral graph for the HMM



Proposition 20. *For the HMM whose DAG is provided before, the following independance properties hold:*

1. $\mathbb{P}(Z_{t+1}|Y_{1:t}, Z_{1:t}) = \mathbb{P}(Z_{t+1}|Z_t)$
2. $\mathbb{P}(Z_{t+1}|Z_{1:t}) = \mathbb{P}(Z_{t+1}|Z_t)$
3. $\mathbb{P}(Y_{t+1}|Y_{1:t}, Z_{1:t+1}) = \mathbb{P}(Y_{t+1}|Z_{t+1})$

Proof of Proof of Proposition 20

1. All paths from $I = Y_{1:t}$ to $J = Z_{t+1}$ go through $K = Z_{1:t}$ so Z_{t+1} is independent from $Y_{1:t}$ conditionally on $Z_{1:t}$ and we get:
$$\mathbb{P}(Z_{t+1}|Y_{1:t}, Z_t) = \mathbb{P}(Z_{t+1}|Z_t)$$
2. All paths from $Z_{1:t-1}$ to Z_{t+1} go through Z_t, meaning that Z_{t+1} is independent from $Z_{1:t-1}$ conditionally on Z_t (i.e. (Z_t) is a Markov chain);
3. All paths from $Y_{1:t}$ to Y_{t+1} go through Z_{t+1} meaning that Y_{t+1} is independent from $Y_{1:t}$ conditionally on Z_{t+1}
$$\mathbb{P}(Y_{t+1}|Y_{1:t}, Z_{t+1}) = \mathbb{P}(Y_{t+1}|Z_{t+1})$$

Proposition 21. *Conditionally on the observed data $Y = Y_{1:n}$ (Z_t) is still a Markov chain. In addition,*
$$\mathbb{P}(Z_{t+1}|Z_{1:t}, Y_{1:n}) = \mathbb{P}(Z_{t+1}|Z_t, Y_{t+1:n})$$

We propose 2 versions of the proof of this proposition, one relying on the DAG and the other one without the DAG.

Proof of Proof 1 of Proposition 21

1. We have to prove that conditionnally on the observations, (Z_t) is still a Markov Chain, i.e.

$$\mathbb{P}(Z_{t+1}|Z_{1:t}, Y_{1:n}) = \mathbb{P}(Z_{t+1}|Z_t, Y_{1:n}).$$

   Using that $\mathbb{P}(Z_{t+1}|Z_{1:t}, Y_{1:n}) = \mathbb{P}(\underbrace{Z_{t+1}}_{I} \mid \underbrace{Z_{1:t-1}}_{J}, \underbrace{Z_t, Y_{1:n}}_{K})$, let us set

$$I = Z_{t+1}, \quad J = Z_{1:t-1}, \quad K = \{Z_t, Y_{1:n}\}.$$

   All paths from I to J go through K.

2. We now need to prove that $\mathbb{P}(Z_{t+1}|Z_{1:t}, Y_{1:n}) = \mathbb{P}(Z_{t+1}|Z_t, Y_{t+1:n})$. We have:

$$\mathbb{P}(Z_{t+1}|Z_{1:t}, Y_{1:n}) = \mathbb{P}(\underbrace{Z_{t+1}}_{I} \mid \underbrace{Z_{1:t-1}, Y_{1:t}}_{J}, \underbrace{Z_t, Y_{t+1:n}}_{K}).$$

   All paths from I to J go through K and we can conclude.

Proof of Proof 2 of Proposition 21

Without the use of the DAG methodology, the proof takes more time. Here is a proposal.

$$\begin{aligned}
p(Z_{t+1}|Z_{1:t}, Y_{1:n})1 &= p(Z_{t+1}|Z_{1:t}, Y_{1:t}, Y_{t+1:n}) = \frac{p(Z_{t+1}, Z_{1:t}, Y_{1:t}, Y_{t+1:n})}{p(Z_{1:t}, Y_{1:t}, Y_{t+1:n})} \\
&= \frac{p(Y_{t+1:n}|\cancel{Y_{1:t}}, Z_{t+1}, \cancel{Z_{1:t}})p(Y_{1:t}, Z_{t+1}, Z_{1:t})}{p(Y_{t+1:n}|Z_{1:t}, Y_{1:t})p(Z_{1:t}, Y_{1:t})} \\
&= \frac{p(Y_{t+1:n}|Z_{t+1})p(Y_{1:t}|Z_{t+1}, Z_{1:t})p(Z_{t+1}|Z_{1:t}^t)\cancel{p(Z_{1:t})}}{p(Y_{t+1:n}|Z_{1:t}, Y_{1:t})p(Y_{1:t}|Z_{1:t})\cancel{p(Z_{1:t})}}
\end{aligned}$$

But $p(Y_{1:t}|Z_{t+1}, Z_{1:t}) = p(Y_{1:t}|Z_{1:t})$ So

$$p(Z_{t+1}|Z_{1:t}, Y_{1:n}) = \frac{p(Y_{t+1:n}|Z_{t+1})\cancel{p(Y_{1:t}|Z_{t+1}, Z_{1:t})}p(Z_{t+1}|Z_t)}{p(Y_{t+1:n}|Z_{1:t}, Y_{1:t})\cancel{p(Y_{1:t}|Z_{1:t})}}$$

Finally:

$$\begin{aligned}
p(Z_{t+1}|Z_{1:t}, Y_{1:n}) &= \frac{p(Y_{t+1:n}|Z_{t+1})\cancel{p(Y_{1:t}|Z_{t+1}, Z_{1:t})}p(Z_{t+1}|Z_t)}{p(Y_{t+1:n}|Z_{1:t}, Y_{1:t})\cancel{p(Y_{1:t}|Z_{1:t})}} \\
&= \frac{p(Y_{t+1:n}|Z_{t+1})p(Z_{t+1}|Z_t)}{p(Y_{t+1:n}|Z_t)} \\
&= p(Z_{t+1}|Z_t, Y_{t+1:n})
\end{aligned}$$

A.4 Model selection (SR)

A.4.1 Bayesian Information Criterion (BIC) (SR)

The definition of the BIC introduce in Section 2.4 relies on a Laplace approximation of the integral of an exponential function.

Lemma 3 (Laplace approximation). *Consider $L : \mathbb{R}^D \mapsto \mathbb{R}$ with a unique maximum at u^*, with full rank Hessian matrix $L''(u^*)$, it holds that*

$$\int_{\mathbb{R}^D} e^{nL(u)}\, du = e^{nL(u^*)} \left(\frac{(2\pi)^D}{n^D |-L''(u^*)|} \right)^{1/2} (1 + o_n(1)).$$

Proof of Proof

First remind that, if A is a $D \times D$ positive matrix ($A \succ 0$), we have that

$$\int_{\mathbb{R}^D} \exp\left(-\frac{1}{2} \|u\|_A^2 \right) \mathrm{d}u = |A|^{-1/2} (2\pi)^{D/2} \tag{A.5}$$

(we may recognize the normalizing constant of the multivariate Gaussian distribution). Then consider the second order Taylor expansion of nL about u^*: because its first derivative $L'(u^*)$ is zero, we get

$$nL(u) = nL(u^*) - \frac{1}{2} \|u - u^*\|^2_{(-nL''(u^*))} + o_n(\|u - u^*\|^2).$$

The result follows from (A.5), taking $A = -nL''(u^*)$ (which is positive definite because, u^* being a maximum, $L''(u^*)$ is negative definite) and observing that $|-nL''(u^*)| = n^D |-L''(u^*)|$.

A.4.1.1 Derivation of the BIC.

[*PG: Il y a ici inconsistance de notation sur les $o(1)$ et $O_n(1)$ dans le texte principal*]

We only provide here an idea of the derivation of Equation (2.22) in the independent and identically distributed setting, that is: we assume that the data are $Y = (Y_1, \dots Y_n)$, where the Y_i are all independent conditional on θ and m, with distribution $p(y_i \mid \theta, m)$.

Sketch of proof of Equation (2.22) in the iid case. We define the normalized (conditional) log-likelihood of the observe dataset $y = (y_1, \dots y_n)$ as

$$L(\theta) = \frac{1}{n} \log p(y \mid \theta, m) = \frac{1}{n} \sum_{i=1}^{n} \log p(y_i \mid \theta, m)$$

which is maximal for $\theta = \widehat{\theta}_m$ and converges in probability to $\mathbb{E}[\log p(Y_1 \mid \theta, m)]$ as n tends to infinity. Now, we may write the integral of interest as

$$\begin{aligned} \int p(y, \theta, m)\, \mathrm{d}\theta &= \int \exp\left(\log p(y \mid \theta, m) + \log p(\theta \mid m) + \log p(m) \right) \mathrm{d}\theta \\ &= \int \exp\left(n \left[L(\theta) + o_n(1) \right] \right) \mathrm{d}\theta \end{aligned}$$

where the latter equality holds as long as neither $\log p(\theta \mid m)$ nor $\log p(m)$ depend on n. Then, using Lemma 3, we have that

$$\int p(y, \theta, m)\, \mathrm{d}\theta = e^{nL(\widehat{\theta}_m)} (2\pi)^{D_m/2} n^{-D_m/2} |-L''(\theta)|^{-1/2} (1 + o_n(1))$$

where D_m is the rank $L''(\theta)$, that is the number of independent parameters in θ under model m. Taking the logarithm of it gives Equation (2.22).gives

$$\begin{aligned} \log \int p(y, \theta, m)\, \mathrm{d}\theta &= nL(\widehat{\theta}_m) - D_m \frac{\log n}{2} + \frac{D_m}{2} \log(2\pi) - \frac{1}{2} \log |-L''(\theta)| + o_n(1) \\ &= nL(\widehat{\theta}_m) - D_m \frac{\log n}{2} + O_n(1) \end{aligned}$$

because neither D_m nor $|-L''(\theta)|$ depend on n.

■

Remark. Observe that the Laplace approximation holds whenever $n^{-1}\log p(y \mid \theta, m)$ converges in probability, which is true for many other models such as hidden Markov models (Section **??**), or the multivariate Poisson log-normal model (Section **??**).

A.4.2 Variationnal approximations of ICL (SR ? ou SD ?)

A.4.2.1 Variational ICL for the stochastic block model (SBM).

We consider here the binary or Poisson SBM, where the conditional distribution of each edge Y_{ij} conditional on Z_i and Z_j depends on a one-dimensional parameter. The proof (and the resulting penalties) need be adapted to other emission distributions.

Sketch of proof of Equations (**??**) *and* (**??**). Using the conditional independence of α and $\boldsymbol{\omega}$ given K, the integral can be factorised as

$$\begin{aligned}\int p(\boldsymbol{y}, \boldsymbol{z}, \theta, K)\,\mathrm{d}\theta &= \left(\iint p(\boldsymbol{y} \mid \boldsymbol{z}, \alpha, K) p(\boldsymbol{z}, \mid \boldsymbol{\omega}, K) p(\alpha \mid K) p(\boldsymbol{\omega} \mid K)\,\mathrm{d}\alpha\,\mathrm{d}\boldsymbol{\omega}\right) \\ &= \left(\int p(\boldsymbol{y} \mid \boldsymbol{z}, \alpha, K) p(\alpha \mid K)\,\mathrm{d}\alpha\right)\left(\int p(\boldsymbol{z}, \mid \boldsymbol{\omega}, K) p(\boldsymbol{\omega} \mid K)\,\mathrm{d}\boldsymbol{\omega}\right),\end{aligned}$$

that is

$$\log\left(\int p(\boldsymbol{y}, \boldsymbol{z}, \theta, K)\,\mathrm{d}\theta\right) = \log\left(\int p(\boldsymbol{y} \mid \boldsymbol{z}, \alpha, K) p(\alpha \mid K)\,\mathrm{d}\alpha\right) + \log\left(\int p(\boldsymbol{z}, \mid \boldsymbol{\omega}, K) p(\boldsymbol{\omega} \mid K)\,\mathrm{d}\boldsymbol{\omega}\right).$$

A Laplace approximation can then be applied to each term, following the same line a in Section A.4.1 to get Equation (2.22). Hence we get for the first term

$$\log\left(\int p(\boldsymbol{y} \mid \boldsymbol{z}, \alpha, K) p(\alpha \mid K)\,\mathrm{d}\alpha\right) = \log p(\boldsymbol{y} \mid \boldsymbol{z}, \alpha = \widehat{\alpha}, K) - K(K+1)\frac{\log[n(n-1)/2]}{2} + O_n(1) \tag{A.6}$$

because the $n(n-1)/2$ edges of the network Y are conditionally independent given (Z, α, K) and because there are $K(K+1)/2$ independent parameters in α in the binary or Poisson version of Model **??**. As for the second term, we have that

$$\log\left(\int p(\boldsymbol{z}, \mid \boldsymbol{\omega}, K) p(\boldsymbol{\omega} \mid K)\,\mathrm{d}\boldsymbol{\omega}\right) = \log p(\boldsymbol{z} \mid \boldsymbol{\omega} = \widehat{\boldsymbol{\omega}}, K) - (K-1)\frac{\log n}{2} + O_n(1) \tag{A.7}$$

because the n latent variables Z_i underlying the network Y are conditionally independent given $(\boldsymbol{\omega}, K)$ and because there are $(K-1)$ independent parameters in $\boldsymbol{\omega}$. Gathering Equations (A.6) and (A.7) gives (**??**) and (**??**) using frequentist notations.

■

A.4.2.2 Variational ICL for the stochastic block model (LBM).

Sketch of proof of Equations (**??**) *and* (**??**). We follow the same line as for the SBM model: using the conditional independence of α , $\boldsymbol{\omega}^{(1)}$ and $\boldsymbol{\omega}^{(2)}$ given K and L, the integral can be factorised in three terms to get

$$\begin{aligned}&\log\left(\int p(\boldsymbol{y}, \boldsymbol{z}, \boldsymbol{w}, \theta, K, L)\,\mathrm{d}\theta\right) \\ &= \log\left(\iiint p(\boldsymbol{y} \mid \boldsymbol{z}, \boldsymbol{w}, \alpha, K, L) p(\boldsymbol{z}, \mid \boldsymbol{\omega}^{(1)}, K) p(\boldsymbol{w}, \mid \boldsymbol{\omega}^{(2)}, L) p(\alpha \mid K) p(\boldsymbol{\omega}^{(1)} \mid K) p(\boldsymbol{\omega}^{(2)} \mid L)\,\mathrm{d}\alpha\,\mathrm{d}\boldsymbol{\omega}^{(1)}\,\mathrm{d}\boldsymbol{\omega}^{(2)}\right) \\ &= \log\left(\int p(\boldsymbol{y} \mid \boldsymbol{z}, \boldsymbol{w}, \alpha, K) p(\alpha \mid K)\,\mathrm{d}\alpha\right) + \log\left(\int p(\boldsymbol{z}, \mid \boldsymbol{\omega}^{(1)}, K) p(\boldsymbol{\omega}^{(1)} \mid K)\,\mathrm{d}\boldsymbol{\omega}\right) \\ &\quad + \log\left(\int p(\boldsymbol{w}, \mid \boldsymbol{\omega}^{(2)}, K) p(\boldsymbol{\omega}^{(2)} \mid K)\,\mathrm{d}\boldsymbol{\omega}\right).\end{aligned}$$

Then a Laplace approximation can be derived for each term, assuming that the number of nodes of each type go to infinity at the same rate (that is: $\lim_{n\to\infty} n/p = \text{cst}$):

$$\log\left(\int p(\boldsymbol{y} \mid \boldsymbol{z}, \boldsymbol{w}, \alpha, K) p(\alpha \mid K) \, \mathrm{d}\alpha\right) = \log_{\widehat{\theta}}(\boldsymbol{y} \mid \boldsymbol{z}, \boldsymbol{w}) - \frac{KL}{2} \log(np) + O_n(1)$$

$$\log\left(\int p(\boldsymbol{z}, \mid \boldsymbol{\omega}^{(1)}, K) p(\boldsymbol{\omega}^{(1)} \mid K) \, \mathrm{d}\boldsymbol{\omega}\right) = \log_{\widehat{\theta}}(\boldsymbol{z}) - \frac{K-1}{2} \log(n) + O_n(1)$$

$$\log\left(\int p(\boldsymbol{w}, \mid \boldsymbol{\omega}^{(2)}, K) p(\boldsymbol{\omega}^{(2)} \mid K) \, \mathrm{d}\boldsymbol{\omega}\right) = \log_{\widehat{\theta}}(\boldsymbol{w}) - \frac{L-1}{2} \log(p) + O_n(1).$$

Gathering the three equations gives Equations (**??**) and (**??**).

■

A.5 Proofs (SD, PG, SR)

Proof of Proposition ??

q is a maximizer of $\mathcal{F}(q)$ if, for any direction h, the derivative of $\mathcal{F}(q)$ in direction h is zero, i.e.

$$\forall h, \qquad \partial_t \mathcal{F}(q+th)|_{t=0} = 0.$$

Under regularity conditions, we can move the derivative into the integral so

$$\partial_t \mathcal{F}(q+th) = \int \partial_t L(x, q(x)+th(x))\,\mathrm{d}x = \int h(x) L(x, q(x)+th(x))\,\mathrm{d}x$$

which is, at $t=0$,

$$\int \left[\partial_{q(x)} L(x, q(x))\right] h(x)\,\mathrm{d}x.$$

Because this holds for any function h, we may use the fundamental lemma of calculus of variations, which states that

$$\forall h, \quad \int f(x)h(x)\,\mathrm{d}x = 0 \qquad \Rightarrow \qquad f = 0,$$

to conclude that $\partial_{q(x)} L(x, q(x)) = 0$.

Proof of Proposition ??

We have to prove each recursion formulas.

Upward recursion: The initialization is straightforward, than the recursion is the same as the recursion given in **??** to evaluate the likelihood of the observed variables $p_\theta(Y_{1:n})$, observing that, at each step

$$\alpha_j(z) \propto p_\theta(Z_j = z) p_\theta(Y_{sub(j)} \mid Z_j = z) = p_\theta(Z_j = z)\ell_j(z).$$

where the conditional distribution $\ell_j(z)$, defined in (**??**), is evaluated thanks to (**??**) and the marginal distribution $p_\theta(Z_j)$ is evaluated with (**??**) and (**??**).

Downward recursion: Again, the initialization is obvious as

$$\alpha_{2n-1}(z) = \alpha_{MRCA}(z) = p_\theta(Z_{MRCA} = z \mid Y_{1:n}) = \tau_{MRCA}(z).$$

Then the recursion is based on a decomposition similar to this used for the HMM in Proposition **??**: let us write $\tau_j(z)$ as

$$\tau_j(z) = \int \xi_j(u, z)\,\mathrm{d}u, \qquad \text{where} \quad \xi_j(u,z) = \frac{p_\theta(Z_j = z, Z_{pa(j)} = u, Y_{sub(j)}, Y_{\overline{sub(j)}})}{p_\theta(Y_{1:n})}$$

splitting $Y_{1:n}$ into $Y_{sub(j)}$ and $Y_{\overline{sub(j)}}$, where $\overline{sub(j)}$ is the set of leaves that are not downstream of node j. Now, because Z_j is independent from $Y_{\overline{sub(j)}}$ given $Z_{pa(j)}$ and conversely, we have that

$$\begin{aligned} &p_\theta(Z_j = z, Z_{pa(j)} = u, Y_{sub(j)}, Y_{\overline{sub(j)}}) \\ &= p_\theta(Y_{sub(j)}) p_\theta(Z_j = z \mid Y_{sub(j)}) p_\theta(Z_{pa(j)} = u \mid Z_j = z) p_\theta(Y_{\overline{sub(j)}} \mid Z_{pa(j)} = u). \end{aligned}$$

Multiplying and dividing by $p_\theta(Y_{sub(j)} \mid Z_{pa(j)})$, we get that

$$\frac{p_\theta(Y_{\overline{sub(j)}} \mid Z_{pa(j)} = u)\, p_\theta(Y_{sub(j)})}{p_\theta(Y_{1:n})} = \frac{p_\theta(Y_{1:n} \mid Z_{pa(j)} = u)\, p_\theta(Y_{sub(j)})}{p_\theta(Y_{sub(j)} \mid Z_{pa(j)} = u)\, p_\theta(Y_{1:n})} = \frac{\tau_{pa(j)}(u)}{p_\theta(Z_{pa(j)} = u \mid Y_{sub(j)})}$$

so we are left with

$$\xi_j(u,z) = \alpha_j(z)\tau_{pa(j)}(u)\beta_j(u,z)$$

where

$$\beta_j(u,z) = \frac{p_\theta(Z_{pa(j)} = u \mid Z_j = z)}{p_\theta(Z_{pa(j)} = u \mid Y_{sub(j)})} = \frac{p_\theta(Z_{pa(j)} = u \mid Z_j = z)}{\int p_\theta(Z_{pa(j)} = u \mid Z_j = v) p_\theta(Z_j = v \mid Y_{sub(j)})\,\mathrm{d}v}$$
$$= \frac{p_\theta(Z_j = z) p_\theta(Z_j = z \mid Z_{pa(j)} = u)}{\int p_\theta(Z_j = u) p_\theta(Z_j = v \mid Z_{pa(j)} = u) \alpha_j(v)\,\mathrm{d}v}.$$