Latent variable models in ecology and their inference via EM algorithm its extensions

SD, PG, SR

December 11, 2025

Contents

Acknowledgement. Relecteurs/trices: étudiant Sophie, Sarah Ouadah, Valentin Robert (ISUP/ISDS), ...

Notations

- Y = observed variable
- y = observed data, a realization of Y.
- Z = latent variables
- X = covariates
- θ = model parameters. When necessary, we'll specify $\theta = (\theta_{\text{obs}}, \theta_{\text{lat}})$ where:
  - θ_{obs} = subset of model parameters specific to the observations conditionnaly to the latent states;
  - θ_{lat} = subset of model parameters specific to the states.
- $\theta^{(h+1)} = f(\theta^{(h)})$: update
- $Q\big(\theta \mid \theta^{(h)}\big)$ = EM objective function
- p_θ = distribution under the model parameterized by θ
- $p_\theta(z \mid Y = y)$ is the conditional distribution of the latent variable given the observation $Y = y$ for a given parameter θ
- $\mathbb{P}_\theta(\mathcal{A})$ = Probability of some event $\mathcal{A}$ under the model parameterized by θ
- $\mathbb{E}_{\theta'}[\phi(y, Z, \theta) \mid Y = y] = \int_{z \in \mathcal{Z}} \phi(y, z, \theta) p_{\theta'}(z \mid Y = y)\, \mathrm{d}z$
- $\mathbb{V}_\theta[X]$ Variance (-covariance, in the multivariate case) of the random variable X under the parameter θ.
- $\mathbb{C}\text{ov}_\theta(X, Y)$ Covariance (it is a matrix in the multivariate case) of the random variables X and Y under parameter θ.
- $\mathsf{ELBO}(q, \theta, y)$ = varational lower bound
- ψ = variational parameters
- $Z_i \sim \pi$, $Z_{ik} = \mathbb{I}\{Z_i = k\}$.
- $x^\intercal$: x transposed
- $\mathbf{S}_+^d$: Set of real symmetric positive semidefinite matrices in dimension d (in which belong covariance matrices)
- $x_{1:n}$ is a notation for any collection of (random or not) variables $(x_1, \ldots, x_n)$
- $\mathbb{I}_{\mathcal{B}}(x)$ the indicatrix of the set $\mathcal{B}$, i.e. the function that equals one if $x \in \mathcal{B}$ and 0 otherwise.
- $\mathbb{I}_k(x)$ the function that equals one if $x = k$ and 0 otherwise
- $\mathbf{1}_d$ Vector of dimension d filled with ones.
- $\mathbf{1}_{d_1 d_2}$ Matrix of dimensions $d_1 \times d_2$ filled with ones.
- $\|v\|$ (Euclidean) norm of the vector v.
- $|\Sigma|$ Determinant of the matrix Σ
- $X \sim \mathcal{N}_d(\mu, \Sigma)$ = X follows a multivariate Gaussian distribution in dimension d with expectation μ and variance Σ.
- $\phi\big(x; \mu, \sigma^2\big)$ = Probability density function of a Gaussian distribution with mean μ and variance σ^2.
- $Z_i \overset{\text{ind}}{\sim} p_{\theta_i}(\cdot)$ Random variables $(Z_i)_{i=\ldots}$ are indenpendant and distributed within the same parametric family (which will often be precised in context.
- $Z_i \overset{\text{iid}}{\sim} p_\theta(\cdot)$ Random variables $(Z_i)_{i=\ldots}$ are indenpendant and identically distributed with distribution p_θ

Chapter 1

Introduction

1.1 The book

This book brings together the lecture notes from a series of courses that we have given or are still giving at the Master's level in several universities: Université Paris-Saclay, Sorbonne Université, Université Bretagne-Sud.

It aims to present the entire approach of statistical ecology, i.e., starting with a set of data collected to answer an ecological question, proposing a statistical model, designing an inference method to estimate its parameters, and then interpret the results to answer the initial question. A major focus is done on mathematical formalization, both of the model and the inference method, in the purpose of performing quantitative ecology.

Starting from data and ecological questions, this book illustrates the perpetual compromise of statistical modeling between the "realistic" character of a model, which pushes for complexity, and the need for inference, which pushes for simplicity. This illustration is made by exploring the specific but very broad category of latent variable models, and their inference using a common keystone, the Expectation-Maximization algorithm.

By fully formalizing the mathematical model adopted to answer the ecological question, we aim at identifying its specificities that make the statistical inference simple or complex. Consequently, for each model considered, all the statistical machinery is deployed and all the calculations are detailed, to enable the reader to do the same for the model she or he will have to deal with in her or his own research.

The book is organized into chapters according to the increasing difficulty of the inference. These chapters contain methodological sections that present generic tools, unrelated to any particular application, and data analysis sections that present a specific model, motivated by a particular ecological problem. Each of these sections is introduced by the presentation of an ecological data set, for which a statistical model is proposed. We try to emphasize the common structure and the differences between the successive models that we present. We then present the statistical method to perform inference, always revolving around the Expectation Maximization algorithm, designed for maximum likelihood inference. All the concepts and useful outputs of the model are illustrated on datasets or questions that we, as researchers in statistics, encountered when we worked with ecologists.

This book is therefore aimed on the one hand at Masters and PhD students with a mathematical background and interested in statistical ecology, and on the other hand at ecologists wishing to deepen their statistical training. We hope that the former will find interest in the link between well known models in statistical learning (clustering methods, hidden Markov models...) and challenging applications in ecology, and that they will be able to make bridges between methods that are often too much compartmented. We hope that ecologists will find a clear and unambiguous formulation for models that they might have used for a long time, and that this formulation would reveal some useful subtleties about the when and why of their usage.

It is always useful to precise what this book is not about. This book is not about an extensive overview of all statistical methods in ecology. For instance, we do not cover the basic (but fundamental) tools of statistical ecology which are the linear regression, the analysis-of-variance model or the generalized linear model, even if we widely cover some extensions of these. We then assume that the reader has a minimal knowledge of these models and there assumptions. This book is not meant to be purely operational, as we do not focus on software implementation and do not provide an extensive list or description of packages that performs inference for the discussed models. Finally, by making some choices about specific models, this book does not covers all latent variable models. For instance, we do not discuss the modeling of data whose dependences are induced by a continuous spatial domain, which is a subject in its own right and for which a huge number of specific models, theories and methods have been proposed [see, for example Cressie, 2015, for a fairly comprehensive overview].

With this same logic, as we decided to emphasize the broad applicability of the EM algorithm, we do not

focus on Bayesian statistics, even if several issues commonly encountered in Bayesian are extensively discussed, such as the inference of conditional distribution and, in less extent, methods for sampling in these distributions.

1.2 Latent variable models in ecology

Latent variable models are statistical models that relates the random observable variables, (denoted Y throughout the book), to a set of unobserved random variables (denoted 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 state of the system at each time-step. They may also have no physical reality and be used to enrich a simpler model to fit the data, as in mixture models where the latent variables are used to classify the observations into groups that are unknown a priori, or in mixed model where they allow to introduce some dependence between the observations.

Latent variable models have been widely used in ecology for decades: a review of some recent applications can be found in Peyrard and Gimenez [2022]. Their popularity is partly due to their ease of use, when conceiving the model. Latent variable models can be often be written as hierarchical models, represented as a graph encoding a cascade of effect going down from the unknown parameters (denoted θ in the book) to the observed variables Y, passing through one or several latent variables Z. We will use such representations, called graphical models, throughout the book.

Still, from a statistical viewpoint, the presence of hidden variables in the model formulation makes the inference more difficult. Briefly speaking, because the latent variables are unobserved, all their possible values have to be considered, which requires an integration with respect to Z, which is difficult and even infeasible in many cases. For example, the likelihood of the observations Y is an integral over the latent variables Z, and its maximisation can not be achieved in a naive manner because the calculation of the likelihood itself is complex (or infeasible).

As we shall see in this book, many classical methods in statistical learning can be seen from the viewpoint of latent variable models such as clustering methods, principal component analysis or autoencoders.

1.3 The content

In this book, we present a selection of latent variable models classically used in ecology and their inference using the Expectation-Maximization (EM) algorithm.

The EM algorithm was proposed by Dempster et al. [1977] to carry maximum likelihood inference for models involving latent variables. As the EM algorithm has been used successfully in a wide variety of contexts, making it a central tool in statistics, we have chosen to focus this book on inference strategies based on this algorithm.

The main problem when inferring models with latent variables is that only some of the variables involved in the model are actually observed, while the other variables (the latent variables) remain unknown. The rational of the EM algorithm is to recover the missing information, by examining the conditional distribution of the latent variables Z given the observed variables Y. More often than not, this step (which constitutes the 'E' step of the algorithm) is the critical one, as it can lead to elementary, complex or infeasible calculations. We have chosen to organize the book according to the difficulty of the step.

The book is organized as follows. In Chapter 2, we define the generic form of latent variable models and present the EM algorithm in detail. Models for which the E step can be performed with close form formulas are presented in chapter 3. Chapter 4 deals with models that require a greater effort: there is no close-form solution for E step, but it can still be performed exactly using, for example, recursion formulas. The last two chapters are devoted to models for which the solution of E step cannot be evaluated at all, so that an approximation to it must be considered: Chapter 5 presents deterministic approximations, while Chapter **??** deals with stochastic approximations. Reminders on some useful notions, as well as some technical proofs are devoted to the Appendices.

The reading of the book does not need be linear: not all sections of a chapter need to be read before going to the next. Still, we advise to read the whole Chapter 2, and the methodological Sections 3.1, 4.1, 5.1 to have a clear picture.

Chapter 2

Latent variable models and the EM algorithm

Contents

This first chapter formalized the pivotal tools that will be encountered all along these books, which are the latent variable models, the Expectation-Maximization algorithm and some of its theoretical properties.

The first section of this chapter (Section 2.1) is devoted to the formal definition of a latent variable model, illustrated through a simple motivating example: detecting sub-populations among Palmer penguins based on bill length. The Expectation-Maximization (EM) algorithm, a widely used method for parameter estimation in latent variable models, is then introduced in Section 2.2. Finally, the last two sections of the chapter address the asymptotic variance of the resulting estimators (Section 2.3) and present model selection criteria (Section 2.4).

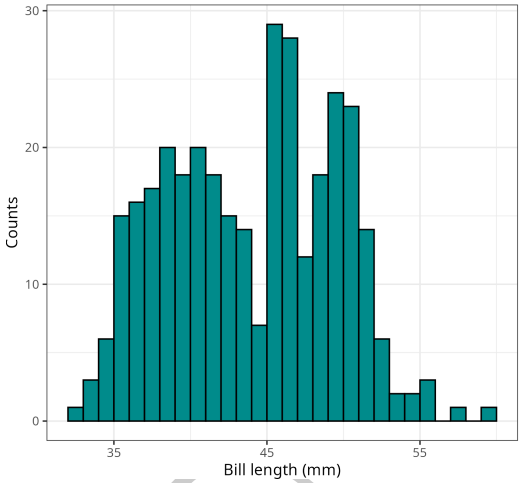
2.1 Latent variable models

2.1.1 From Palmer penguins to mixture models

Dataset 2.1 (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*)[a].

[a]Picture by Samuel Blanc, via Wikimedia Commons, licence CC BY-SA 3.0.



(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 bill length in Palmer penguins. A quick look at the histogram of bill length (Figure 2.1, right) suggests that a simple Gaussian distribution is not adequate. Indeed, the distribution exhibits multiple local modes (peaks), indicating that it may not originate from a single homogeneous population. Since the distribution of a trait within a homogeneous population is typically assumed to follow a Gaussian distribution, this observed multimodality may reflect the presence of heterogeneous subpopulations, each with distinct bill length characteristics.[1] This feature of the data naturally motivates the use of a (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 each individual belongs[2].

Let $\boldsymbol{y} = \{y_i\}_{1\leq i\leq n}$ denote the n observed bill lengths. Considering the two modes observed in the distribution in Figure 2.1, we assume these are realizations of n independent random variables $\{Y_i\}_{1\leq i\leq n}$ that follow a two-component Gaussian mixture distribution. In the case of a two-component mixture model, each observation is assumed to originate from either population 1 or population 2. Assuming that the bill lengths within each population follow a Gaussian distribution, we model the data as follows:

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

To make the model complete, we must specify how the individuals are spread among the two populations. To this aim, for each observation i, we introduce a binary random variable Z_i, which indicates whether observation i belongs to population 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 $\boldsymbol{Y} = \{Y_i\}_{1\leq i\leq n}$ and that we do not know their population of origin $\{Z_i\}_{1\leq i\leq n}$, 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.

[1]The dataset actually includes information on both species and sex. However, for the purpose of this introductory example, we intentionally ignore these variables.

[2]In the rest of this section, we use the word "population" to stick to a natural interpretation for this dataset, but from a statistical point of view, the best word would certainly be cluster, as it is the result of a clustering task.

Model 2.1 (Two-component Gaussian mixture)**.**

$$Z_i \overset{iid}{\sim} \mathrm{Cat}\left(\omega = (\omega_1, \omega_2)\right), \qquad 1 \leq i \leq n\,,$$
$$Y_i \mid \{Z_i = z_i\}\,, \overset{ind}{\sim} \begin{cases} \mathcal{N}\left(\mu_1, \sigma_1^2\right) & \text{if } Z_i = 1 \\ \mathcal{N}\left(\mu_2, \sigma_2^2\right) & \text{if } Z_i = 2 \end{cases} \qquad 1 \leq i \leq n\,,$$

where $\mathrm{Cat}\left(\omega\right)$ *stands for the categorical distribution with probability vector* ω*.*

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

$$Y_i \mid \{Z_i = z_i\} \overset{\text{ind}}{\sim} \mathcal{N}\left(\mu_{z_i}, \sigma_{z_i}^2\right).$$

where the indices of the mean and variance are given by the latent variable value Z_i.

The proposed model depends on six 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:

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

2.1.2 General definition of the latent variable models and vocabulary

Model 2.2 (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 $\boldsymbol{Z}$ *is not observed.*

The second line of this model (2.2) links the data $\boldsymbol{Y}$ to the latent variables $\boldsymbol{Z}$ and is parametrized by θ_{obs}. The first line (2.1) corresponds to the latent variable part of the model and relies on the parameter θ_{lat}. In some cases, the latent variable $\boldsymbol{Z}$ is discrete (e.g. a categorical variable), in others $\boldsymbol{Z}$ is a continuous variable (e.g. a Gaussian variable).

Back to the Palmer example. In the two-component Gaussian mixture, $p_{\theta_{\text{lat}}}(\cdot)$ is the product of n the categorical distribution 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 na Gaussian distribution 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:

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

2.1.3 Marginal and complete (log)-likelihoods

In this book, we will primarily focus on inference methods based on the likelihood of the observations $p_\theta(\boldsymbol{y})$, where $\boldsymbol{y}$ represents the observed realization of the random vector $\boldsymbol{Y}$. Consequently, we will carefully define both the likelihood of the observations (also referred to as the marginal likelihood) and the complete likelihood.

2.1.3.1 Definitions

For a given realization of latent variables $\boldsymbol{z} = \{Z_i\}_{1 \leq i \leq n}$, following the definition of the latent variable Model 2.2, 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:

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

where $\mathcal{Z}$ is the support of the latent distribution $p_{\theta_{\text{lat}}}(\boldsymbol{z})$. $p_\theta(\boldsymbol{y})$ 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.

Thus, the marginal likelihood is expressed as an integral over the latent variables. In the case where the latent variables are discrete, the integral becomes a discrete summation:

$$p_\theta(\boldsymbol{y}) = \sum_{\boldsymbol{z}\in\mathcal{Z}} p_{\theta_{\text{obs}}}(\boldsymbol{y}\mid \boldsymbol{Z}=\boldsymbol{z}) p_{\theta_{\text{lat}}}(\boldsymbol{z}).$$

As we will see throughout the book, in some cases, this integral of summation can be computed explicitly. In other cases, no explicit expression for the marginal likelihood exists, or its evaluation is computationally burdensome.

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 2.1, with $\boldsymbol{Z} = \{Z_i\}_{1\le i\le n}$ and $\boldsymbol{y} = \{y_i\}_{1\le i\le n}$. For each observation $i \in \{1,\dots n\}$ and each population $k \in \{1,2\}$, we introduce the binary variable

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

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

Proposition 2.1. *For the mixture of two Gaussian distribution, the expression of the complete and marginal log-likelihoods are given by the following expressions:*

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

and

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

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

Proof of Proposition 2.1

Conditional distribution $p_{\theta_{\text{obs}}}(\boldsymbol{y}\mid\boldsymbol{Z})$. By independence of the observations:

$$p_{\theta_{\text{obs}}}(\boldsymbol{y}\mid\boldsymbol{Z}) = p_{\theta_{\text{obs}}}(y_1,\dots,y_n\mid Z_1,\dots Z_n) = \prod_{i=1}^{n} p_{\theta_{\text{obs}}}(y_i\mid Z_i)$$

$$\text{so} \qquad \log p_{\theta_{\text{obs}}}(\boldsymbol{y}\mid\boldsymbol{Z}) = \sum_{i=1}^{n}\log p_{\theta_{\text{obs}}}(y_i\mid Z_i).$$

Besides, we have

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

where $\phi\big(x;\mu,\sigma^2\big)$ 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:

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

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

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

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:

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

Complete log-likelihood $\log p_\theta(\boldsymbol{y}, \boldsymbol{Z})$. Combining the previous quantities, we get:

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

Marginal likelihood $p_\theta(\boldsymbol{y})$. Because of independence, we have:

$$p_\theta(\boldsymbol{y}) = \prod_{i=1}^{n} p_\theta(y_i)\,,$$

and, applying formula (2.4) for each y_i, we get

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

which leads to

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

2.1.4 Maximum likelihood estimation

The maximum likelihood estimate (MLE) is defined as:

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

Proposition 2.1 provides an explicit expression of the marginal likelihood for the two component Gaussian mixture. However, a closer examination of expression (2.7) reveals that, due to the presence of a logarithm applied to a sum of two terms, setting the partial derivatives of $\log p_\theta(\boldsymbol{y})$ to zero does not yield a closed-form solution for $\widehat{\theta}$. While one could apply general-purpose numerical optimization methods in this context, in more complex scenarios (which we will encounter throughout this book), the integral form of the likelihood in Equation (2.4) often becomes intractable. As a result, direct numerical optimization with respect to θ becomes computationally challenging. As an alternative to direct optimization, the EM algorithm offers an efficient approach to obtain the value $\widehat{\theta}$ that maximizes the likelihood, by leveraging the latent variable structure of the model.

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 denote by $\mathbb{E}_{\theta'}[\Psi(\boldsymbol{y}, \boldsymbol{Z}, \theta) \mid \boldsymbol{Y} = \boldsymbol{y}]$ the expectation of $\Psi(\boldsymbol{y}, \boldsymbol{Z}, \theta)$ where the latent variables $\boldsymbol{Z}$ are integrated out under the conditional distribution $p_{\theta'}(\boldsymbol{z} \mid \boldsymbol{Y} = \boldsymbol{y})$:

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

2.2.1 A central decomposition

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

Proposition 2.2 (Decomposition of the incomplete data log-likelihood)**.** *For any θ and θ'*

$$\log p_\theta(\boldsymbol{y}) = \underbrace{\mathbb{E}_{\theta'}[\log p_\theta(\boldsymbol{y}, \boldsymbol{Z}) \mid \boldsymbol{Y} = \boldsymbol{y}]}_{Q(\theta \mid \theta')} \underbrace{- \mathbb{E}_{\theta'}[\log p_\theta(\boldsymbol{Z} \mid \boldsymbol{Y} = \boldsymbol{y}) \mid \boldsymbol{Y} = \boldsymbol{y}]}_{H(\theta \mid \theta')} \tag{2.10}$$

Proof of Proposition 2.2

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:

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

and so:

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

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

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

by definition of $Q(\theta \mid \theta')$ and $H(\theta \mid \theta')$. Moreover, the left term $\log p_\theta(\boldsymbol{y})$ is a constant with respect to $\boldsymbol{Z}$, leading to

$$\mathbb{E}_{\theta'}[\log p_\theta(\boldsymbol{y}) \mid \boldsymbol{Y} = \boldsymbol{y}] = \log p_\theta(y).$$

So, $\log p_\theta(\boldsymbol{y}) = Q(\theta \mid \theta') + H(\theta \mid \theta')$, which concludes the proof.

Remarks.

- *The decomposition of Proposition 2.2 is convenient for our estimation purpose because it makes a connexion between the marginal log-likelihood $\log p_\theta(\boldsymbol{y})$ (often intractable) and the complete data log-likelihood $\log p_\theta(\boldsymbol{y}, \boldsymbol{z})$ (usually more manageable).*
- *Note that if $\theta' = \theta$, then the second term $H(\theta \mid \theta)$ is the entropy[3] of the latent variables $\boldsymbol{Z}$ given the observed $\boldsymbol{Y} = \boldsymbol{y}$:*

$$H(\theta \mid \theta) = -\mathbb{E}_\theta[\log p_\theta(\boldsymbol{Z} \mid \boldsymbol{Y} = \boldsymbol{y}) \mid \boldsymbol{Y} = \boldsymbol{y}] = \mathsf{Ent}[\boldsymbol{Z} \mid \boldsymbol{Y} = \boldsymbol{y}].$$

As a consequence, the decomposition given in Proposition 2.3 leads to

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

[3]Remind that the entropy of any random variable Y with probability distribution q, $\mathsf{Ent}[Y] = -\mathbb{E}_q[\log q(Y)]$

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 EM algorithm is defined as follows.

Algorithm 2.1 (Expectation Maximization)**.**

- ***Initialization.*** *Choose an initial value* $\theta^{(0)}$*.*
- ***Iteration.*** *For* $h \geq 0$*, let* $\theta^{(h)}$ *be the current value of the parameter. Repeat until convergence:*
  - ***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 2.3 (Dempster et al. [1977])**.** *The sequence* $(\theta^{(h)})_{h \geq 0}$ *defined by the EM Algorithm 2.1 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 2.3 relies on the Jensen inequality which we now remind.

Lemma 2.1 (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 2.3

Because $\theta^{(h+1)}$ is the maximizer of $Q(\theta \mid \theta^{(h)})$ with respect to θ, we have

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

So, by definition of Q:

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

As a consequence,

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

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:

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

We reformulate the 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 proof.

Remarks.

E step: The E-step of Algorithm 2.1 instructs us to "compute" the function $Q(\theta \mid \theta^{(h)})$. In practice, this involves calculating all the quantities necessary to evaluate this function, which is typically expressed as a sum or an integral. Since $Q(\theta \mid \theta^{(h)})$ is an expectation with respect to the distribution $p_{\theta^{(h)}}(\boldsymbol{z} \mid \boldsymbol{Y}=\boldsymbol{y})$, the core task of the E-step is to characterize this conditional distribution. This key idea will be illustrated through various examples presented throughout this book.

M step: The proof of Proposition 2.3 requires only that $Q(\theta^{(h+1)} \mid \theta^{(h)}) \geq Q(\theta^{(h)} \mid \theta^{(h)})$ and does not assume that $\theta^{(h+1)}$ is the exact maximizer of $Q(\theta \mid \theta^{(h)})$. Therefore, the M step only needs to produce an update $\theta^{(h+1)}$ that increases the objective function $Q(\theta \mid \theta^{(h)})$ in order for Proposition 2.3 to hold. This variant of the EM algorithm, known as the Generalized EM algorithm in Dempster et al. [1977], is particularly useful when the M step involves an iterative optimization procedure (e.g., gradient ascent). In such cases, performing only a few iterations may be sufficient to increase $Q(\theta \mid \theta^{(h)})$, without requiring full convergence.

2.2.3 A first application of the EM algorithm

As a first illustration, we detail the E step and M step for two-component Gaussian mixture model (Model 2.1).

Algorithm 2.2 (EM for a two-component Gaussian mixture model)**.** *Starting from $\theta^{(0)}$, repeat until convergence:*

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

$$\begin{aligned}
\tau_{i1}^{(h)} &= \frac{\omega_1^{(h)} \phi\left(y_i; \mu_1^{(h)}, \sigma_1^{2(h)}\right)}{\omega_1^{(h)} \phi\left(y_i; \mu_1^{(h)}, \sigma_1^{2(h)}\right) + \omega_2^{(h)} \phi\left(y_i; \mu_2^{(h)}, \sigma_2^{2(h)}\right)}, \\
\tau_{i2}^{(h)} &= 1 - \tau_{i1}^{(h)}
\end{aligned}$$

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*

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

In practice, the algorithm is stopped when the parameters stabilize i.e. $\left\|\theta^{(h+1)} - \theta^{(h)}\right\| < \epsilon$ 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 populations/clusters 1 and 2, given the observation $Y_i = y_i$, also referred as the individual conditional probabilities[4]. 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 population 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.

Proof of Algorithm 2.2

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

Objective function $Q(\theta \mid \theta^{(h)})$. The expression of $\log p_\theta(\boldsymbol{y}, \boldsymbol{Z})$ is given in Equation (2.6). To evaluate $Q(\theta \mid \theta^{(h)})$ we have to integrate the latent variables:

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

where the only random terms in the expectation are the Z_{i1}'s and the Z_{i2}'s. Thus,

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

where

$$\begin{aligned} \tau_{i1}^{(h)} &= \mathbb{E}_{\theta^{(h)}}[Z_{i1} \mid Y_i = y_i] = \mathbb{P}_{\theta^{(h)}}(Z_i = 1 \mid Y_i = y_i) \,, \\ \text{and} \quad \tau_{i2}^{(h)} &= \mathbb{E}_{\theta^{(h)}}[Z_{i2} \mid Y_i = y_i] = \mathbb{P}_{\theta^{(h)}}(Z_i = 2 \mid Y_i = y_i) = 1 - \tau_{i1}^{(h)} \,. \end{aligned}$$

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

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

which leads to

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

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

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

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

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

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

$$\frac{\partial}{\partial \omega_1} Q(\theta \mid \theta^{(h)}) = \sum_{i=1}^{n} \tau_{i1}^{(h)} \frac{1}{\omega_1} - \tau_{i2}^{(h)} \frac{1}{1-\omega_1} = 0 \qquad \Leftrightarrow \qquad \frac{\sum_{i=1}^n \tau_{i1}^{(h)}}{\omega_1} = \frac{\sum_{i=1}^n \tau_{i2}^{(h)}}{1-\omega_1}$$

$$\omega_1^{(h+1)} = \frac{\sum_{i=1}^n \tau_{i1}^{(h)}}{\sum_{i=1}^n \underbrace{\tau_{i1}^{(h)} + \tau_{i2}^{(h)}}_{=1}} = \frac{\sum_{i=1}^n \tau_{i1}^{(h)}}{n}$$

$\omega_1^{(h+1)}$ 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:

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

So

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

Using the expression (2.13), we obtain

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

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:

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

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 2.1, 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 marginal likelihood function. Observe that both global and local maxima 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 (Algorithm 2.2) 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: Log-likelihood ($\log p_{\widehat{\theta}}(y)$) reached by the EM starting from the various values of $\theta^{(0)}$. One can notice that the third initialization lead to a local maximum (see also the Figure 2.2).

To further investigate the behavior of the EM algorithm, we consider a special case where the parameters $(\omega_1, \sigma_1^2, \sigma_2^2)$ are known and fixed to their true values, and only (μ_1, μ_2) are to be estimated. For this experiment, we simulated data $\boldsymbol{y}$ using the parameters $\omega_1 = 0.36$, $\sigma_1^2 = \sigma_2^2 = 9$, and $(\mu_1, \mu_2) = (35, 45)$. Figure 2.3 shows the trajectories of the likelihood through EM iterations starting from five different initial points. Three distinct convergence points can be observed: one corresponds to the global maximum of the likelihood (marked by the green point), which is close to the true parameter values (red cross); another corresponds to a local maximum (bottom right); and the third corresponds to a saddle point of the likelihood surface (center of the figure). This toy example illustrates the practical importance of initializing the EM algorithm from multiple or carefully chosen starting points, in order to explore different regions of the parameter space and improve the chances of reaching the global maximum.

2.3 Evaluating the asymptotic variance of the maximum likelihood estimator

The Fisher information plays a central role in statistical inference, as it quantifies the amount of information that an observable random variable carries about an unknown parameter. It provides a lower bound on the variance

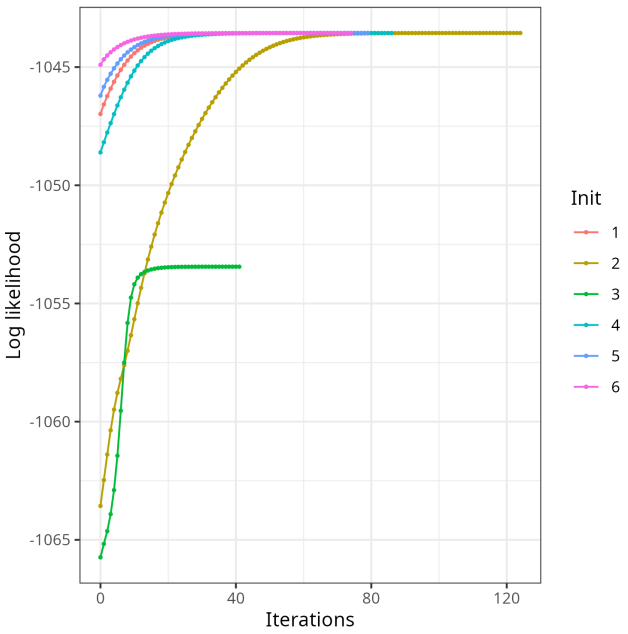


Figure 2.2: Evolution of the log-likelihood along the iterations for the different initial points provided in Table 2.1. One can see that i) the number of iterations to reach convergence strongly depends on the starting point and ii) the third initialization leads to a local maximum.

of unbiased estimators through the Cramér–Rao inequality and is essential for constructing confidence intervals and hypothesis tests. However, in models involving latent variables—such as mixture models or hidden Markov models—the standard definition of Fisher information must be adapted, since the observed data likelihood involves integration over unobserved components. This makes the computation more challenging and often requires specialized techniques, such as using the observed information matrix or leveraging the EM algorithm framework. In this section, we recall the basic concepts of Fisher information and present the main principles for computing or estimating it in the context of models with latent variables.

2.3.1 Fisher information and asymptotic normality of the MLE

The statistical theory of maximum likelihood estimation provides a foundation for constructing asymptotic confidence intervals for estimated parameters. In this section, we recall the relevant asymptotic results in the context of independent observations, which are not necessarily identically distributed due to the presence of observed covariates—such as in regression frameworks like linear or generalized linear models.

Observations Assume that we observe $\boldsymbol{y} = \{y_i\}_{1\leq i\leq n}$, realizations of independent random variables $\{Y_i\}_{1\leq i\leq n}$, whose distributions i) belong to a common parametric family ii) differ in some characteristics[5] because of different covariates, denoted $\boldsymbol{x} = \{x_i\}_{1\leq i\leq n}$ (each measure can be a vector). No particular assumptions are made on the covariates, but we assume

$$Y_i \overset{\text{ind}}{\sim} p_{\theta^*}(\cdot; x_i)$$

where θ^* is the true parameter.

Fisher information Thanks to the independence, the log-likelihood as a function of θ is[6]

$$\log p_\theta(\boldsymbol{y}) = \sum_{i=1}^{n} p_{\theta^*}(y_i; x_i)\,.$$

We denote $\widehat{\theta}_n$ the maximum likelihood estimator, i.e. the random variable:

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

[5]Typically, the expectation of the distribution is different for Y_1 and Y_2 because the measured covariates x_1 and x_2 are different

[6]Note that we omit the dependence on $\boldsymbol{x}$ on the left hand side, as we do not make any modelling assumption on $\boldsymbol{x}$. We keep it on the right hand side for this section, but dependence on covariates will often be implicit in the rest of this book.

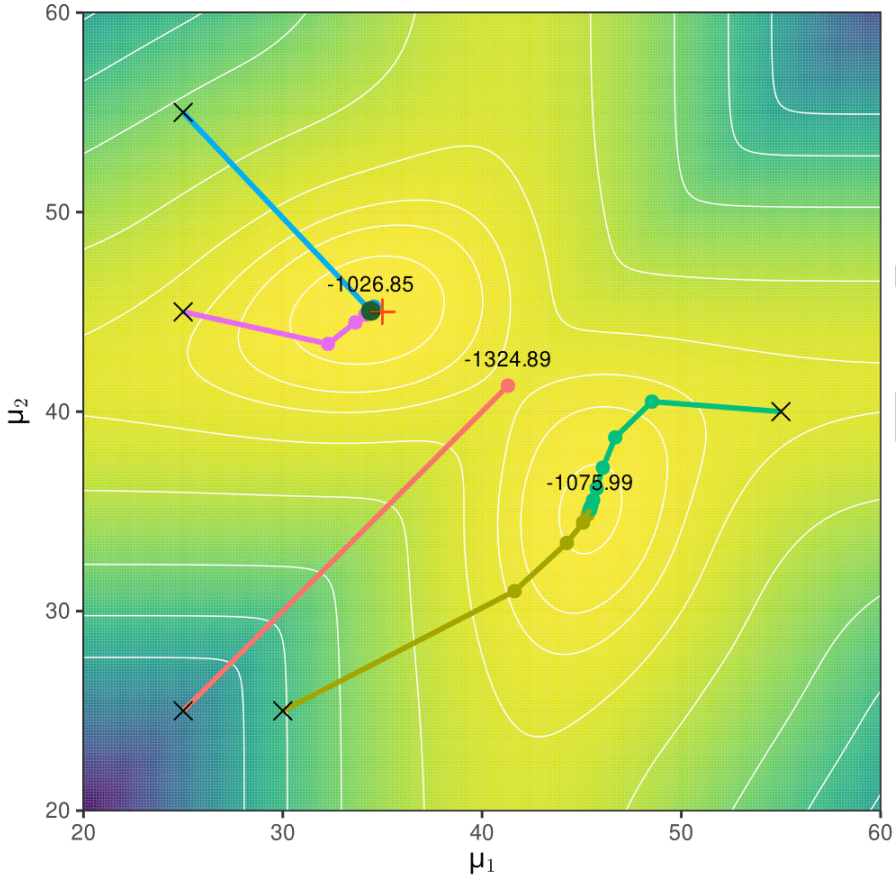


Figure 2.3: Trajectories of the log-likelihood through EM iterations from five different starting points, in a toy example where only (μ_1, μ_2) are unknown.The background color represents the value of the log-likelihood $\log p_\theta(y)$ (blue indicates low values, yellow indicates high values), with white lines showing level curves. The five black crosses denote the initial points for the EM algorithm, and the colored dots trace the likelihood values at each EM iteration. The green dot marks the global maximum, while the red cross indicates the true parameter values. Among the runs, two EM trajectories converge to the global maximum, one to a saddle point, and two to a local maximum. The numeric labels show the log-likelihood values at the three final convergence points.

Assuming that this function admits derivatives with respect to all components of θ, and these derivatives are continuous, we define the score function as the gradient (*i.e.* the vector of all partial derivatives) of the log-likelihood:

$$S_\theta(\boldsymbol{y}) = \nabla_\theta \log p_\theta(\boldsymbol{y})\,. \tag{2.14}$$

Note that we can write $S_\theta(\boldsymbol{y}) = \sum_{i=1}^n S_\theta(y_i)$ where $S_\theta(y_i)$ is an overloaded notation for $\nabla_\theta \log p_\theta(y_i, x_i)$. Assume now that each component of the score function admits a finite order two moment with respect to $p_\theta(\boldsymbol{y})$, we define the Fisher information matrix as:

$$I_n(\theta) = \mathbb{E}_\theta\big[S_\theta(\boldsymbol{Y})S_\theta(\boldsymbol{Y})^\top\big] = \int S_\theta(\boldsymbol{y})S_\theta(\boldsymbol{y})^\top p_\theta(\boldsymbol{y})\,\mathrm{d}\boldsymbol{y} = \sum_{i=1}^n \mathbb{E}_\theta\big[S_\theta(Y_i)S_\theta(Y_i)^\top\big]\,. \tag{2.15}$$

Under certain regularity conditions that we will assume for the rest of this section, the Fisher information matrix may also be written as

$$I_n(\theta) = -\mathbb{E}_\theta[\mathbf{H}_\theta\left(\log p_\theta(\boldsymbol{Y})\right)] = -\mathbb{E}_\theta[\mathbf{J}_\theta\left(S_\theta(\boldsymbol{Y})\right)]\,, \tag{2.16}$$

where $\mathbf{H}_\theta$ and $\mathbf{J}_\theta$ are respectively the Hessian and Jacobian operator [7].

Now, assume that $\frac{1}{n}I_n(\theta)$ converges[8] to some limit $\bar{I}(\theta)$. Then we have the following central limit theorem for the maximum likelihood estimator $\widehat{\theta}_n$

$$\sqrt{n}(\widehat{\theta}_n - \theta^*) \xrightarrow[n\to\infty]{\mathcal{L}} \mathcal{N}(0, \bar{I}(\theta^*)^{-1})\,. \tag{2.17}$$

[7] 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 $\Phi : \Theta \mapsto \mathbb{R}^k$, $\mathbf{J}_\theta\left(\Phi(\theta)\right)$ is the Jacobian *i.e.* the partial derivatives of each component of Φ with respect to each component of θ. $\mathbf{H}_\theta\,\psi(\theta)$ is the Hessian matrix i.e. the matrix of the second derivatives of $\phi : \mathbf{H}_\theta\left(\psi(\theta)\right) = \mathbf{J}_\theta\left(\nabla_\theta\psi(\theta)\right)$.

[8] In the regression context, this would require some assumptions on how $\boldsymbol{x}$ is obtained. Typically, the $\boldsymbol{x}_i$ may be assumed to be realisations of iid random variables with finite variance.

Observe that this may rewritten as

$$\sqrt{n}\,\bar{I}(\theta^*)^{1/2}\,(\widehat{\theta}_n - \theta^*) \xrightarrow[n\to\infty]{\mathcal{L}} \mathcal{N}(0, \mathbf{I}_n)\,.$$

Now, provided that $n^{-1}I_n(\widehat{\theta}_n)$ is a consistent estimator of $\bar{I}(\theta^*)$, Slutsky's lemma yields

$$I_n(\widehat{\theta}_n)^{1/2}\,(\widehat{\theta}_n - \theta^*) \xrightarrow[n\to\infty]{\mathcal{L}} \mathcal{N}(0, \mathbf{I}_n)\,, \tag{2.18}$$

which is used to compute confidence intervals for elements of θ.

Note that if $Y_1, \ldots, Y_n$ are identically distributed, we retrieve the classical result as $I_n(\theta) = nI(\theta)$ where

$$I_\theta = \mathbb{E}_\theta\big[S_\theta(Y_1)S_\theta(Y_1)^\top\big] \tag{2.19}$$

is the Fisher information matrix common to all samples. In theory, the generic approximation (2.18) can be used for any parameters in any statistical model. In practice, however, computing the Fisher information matrix involves evaluating an integral that is often intractable, depending on the (marginal) distribution of the observations.

In those cases, the expectation has to be approximated, for instance by its empirical version using the observations. For instance when $\boldsymbol{Y}$ consists in a n sample of independent and identically distributed random variables $Y_1, \ldots, Y_n$, such that in that case, $I(\widehat{\theta}_n)$ of (2.19) can be estimated by the empirical mean:

$$\widehat{I}(\widehat{\theta}_n) = \frac{1}{n}\sum_{i=1}^{n} S_{\widehat{\theta}_n}(y_i)S_{\widehat{\theta}_n}^\top(y_i)\,.$$

Thanks to alternative formulation (2.16) of the Fisher information, alternative estimators can be designed (see Delattre and Kuhn [2023] for a comparison of the approaches).

2.3.2 Fisher information in latent variable models

For the latent variable models, computing the Fisher information matrix appears to be a challenging task as it depends on the marginal density of observations, which is not directly specified. Indeed, as the likelihood writes as an integral (see Equation (2.4)), it is not clear how to obtain, in general, in this context the score function and the hessian of the log-likelihood. However, the Louis's formulae [Louis, 1982] provides a convenient way to compute these quantities, only using by-products of the EM algorithm.

Proposition 2.4 (Louis [1982])**.** *Let $S_\theta(\boldsymbol{y}) = \nabla_\theta \log p_\theta(\boldsymbol{y})$ be the score function. Provided that differentiation and integration can be exchanged and that all given integrals are finite, then we have:*

$$S_\theta(\boldsymbol{y}) = \int_{\boldsymbol{z}\in\mathcal{Z}} \nabla_\theta \log p_\theta(\boldsymbol{y}, \boldsymbol{Z}) p_\theta\left(\boldsymbol{z} \mid \boldsymbol{Y} = \boldsymbol{y}\right) = \mathbb{E}_\theta[S_\theta(\boldsymbol{y}, \boldsymbol{Z}) \mid \boldsymbol{Y} = \boldsymbol{y}]\,, \tag{2.20}$$

where we introduced the notation $S_\theta(\boldsymbol{y}, \boldsymbol{Z}) = \nabla_\theta \log p_\theta(\boldsymbol{y}, \boldsymbol{Z})$, and the expectation is then taken with respect to the distribution of $\boldsymbol{Z} \mid \{\boldsymbol{Y} = \boldsymbol{y}\}$. Moreover:

$$\mathbf{H}_\theta\left(\log p_\theta(\boldsymbol{y})\right) = \mathbf{J}_\theta\left(S_\theta(\boldsymbol{y})\right) = \mathbb{E}_\theta[\mathbf{J}_\theta\left(S_\theta(\boldsymbol{y}, \boldsymbol{Z})\right) \mid \boldsymbol{Y} = \boldsymbol{y}] + \mathbb{V}_\theta[S_\theta(\boldsymbol{y}, \boldsymbol{Z}) \mid \boldsymbol{Y} = \boldsymbol{y}]\,. \tag{2.21}$$

The first result (2.20) is referred as the Louis' trick. Note that the formulation of Proposition 2.4 presents the main advantage that it relies on the complete likelihood and can, most of the times, be easily computed. In the next chapter we provide an example of such calculation for the ZIP model (Section 3.2.6) and for the Gaussian linear mixed model (Section 3.4.5). Before proving this result, let's show how it is used in practice.

Providing confidence intervals in practice Thanks to Equation (2.21), the hessian of the log-likelihood can be computed as long as the expectations of the right hand side can be computed (using the conditional distribution of $\boldsymbol{Z} \mid \{\boldsymbol{Y} = \boldsymbol{y}\}$). Then, thanks to Equation (2.16), the Fisher information matrix can be computed at the MLE by:

$$I_n\left(\widehat{\theta}_n\right) = -\mathbb{E}_{\widehat{\theta}_n}\left[\mathbf{H}_\theta\left(\log p_{\widehat{\theta}_n}(\boldsymbol{Y})\right)\right],$$

where the expectation is taken with the respect to the marginal distribution of $\boldsymbol{Y}$. As said earlier, if this expectation cannot be computed, it will be approximated by its empirical version. In all cases, the variance of $\widehat{\theta}_n$ will be estimated by:

$$\widehat{\mathbb{V}}[\widehat{\theta}_n] = \left(I_n\left(\widehat{\theta}_n\right)\right)^{-1}\,.$$

Then, for $\alpha \in]0;1[$, we can use the Gaussian approximation (2.18) to obtain an asymptotic confidence interval of level $1-\alpha$ for the j-th element of θ^*, denoted θ_j^*:

$$\mathsf{IC}\left(\theta_j^*\right) = \left[\widehat{\theta}_{n,j} \pm q_{1-\frac{\alpha}{2}}\sqrt{V_{jj}}\right],$$

where $q_{1-\frac{\alpha}{2}}$ is the quantile of order $1-\frac{\alpha}{2}$ of a standard Gaussian distribution, and V_{jj} is the j-th element of the diagonal of $\widehat{\mathbb{V}}[\widehat{\theta}_n]$.

Proof of Proposition 2.4

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

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

thus proving the first part of Proposition 2.4. Now, let us compute $\mathbf{J}_\theta S_\theta(y)$. Because the Hessian matrix of $\log f$ is

$$\mathbf{H}(\log f) = \frac{\mathbf{H}(f)}{f} - \left(\frac{\nabla f}{f}\right)\left(\frac{\nabla f}{f}\right)^\top, \tag{2.22}$$

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

$$\begin{aligned}
\mathbf{J}_\theta\left(S_\theta(\boldsymbol{y})\right) &= \mathbf{H}_\theta\left(\log p_\theta(\boldsymbol{y})\right) \\
&= \frac{\mathbf{H}_\theta\left(p_\theta(\boldsymbol{y})\right)}{p_\theta(\boldsymbol{y})} - \left[\frac{\nabla_\theta p_\theta(\boldsymbol{y})}{p_\theta(\boldsymbol{y})}\right]\left[\frac{\nabla_\theta p_\theta(\boldsymbol{y})}{p_\theta(\boldsymbol{y})}\right]^\top \\
&= \frac{\mathbf{H}_\theta\left(p_\theta(\boldsymbol{y})\right)}{p_\theta(\boldsymbol{y})} - S_\theta(\boldsymbol{y})S_\theta(\boldsymbol{y})^\top \quad \text{from Equation (2.14)} \\
&= \frac{\int_{\boldsymbol{z}\in\mathcal{Z}} \mathbf{H}_\theta\left(p_\theta(\boldsymbol{y},\boldsymbol{z})\right)\mathrm{d}\boldsymbol{z}}{p_\theta(\boldsymbol{y})} - \mathbb{E}_\theta[S_\theta(\boldsymbol{y},\boldsymbol{Z})\mid\boldsymbol{Y}=\boldsymbol{y}]\mathbb{E}_\theta[S_\theta(\boldsymbol{y},\boldsymbol{Z})\mid\boldsymbol{Y}=\boldsymbol{y}]^\top
\end{aligned} \tag{2.23}$$

where the last line uses the Louis' trick. We now concentrate on the first term of Equation (2.23). The same trick used to demonstrate the Louis' trick can be combined with (2.22) to get

$$\begin{aligned}
\frac{\int_{\boldsymbol{z}\in\mathcal{Z}} \mathbf{H}_\theta\left(p_\theta(\boldsymbol{y},\boldsymbol{Z})\right)\mathrm{d}\boldsymbol{z}}{p_\theta(\boldsymbol{y})} &= \int_{\boldsymbol{z}\in\mathcal{Z}} \frac{\mathbf{H}_\theta\left(p_\theta(\boldsymbol{y},\boldsymbol{z})\right)}{p_\theta(\boldsymbol{y},\boldsymbol{z})} \underbrace{\frac{p_\theta(\boldsymbol{y},\boldsymbol{z})}{p_\theta(\boldsymbol{y})}}_{=p_\theta(\boldsymbol{z}\mid\boldsymbol{Y}=\boldsymbol{y})}\mathrm{d}\boldsymbol{z} \\
&= \int_{\boldsymbol{z}\in\mathcal{Z}}\left[\frac{\mathbf{H}_\theta\left(p_\theta(\boldsymbol{y},\boldsymbol{z})\right)}{p_\theta(\boldsymbol{y},\boldsymbol{z})} - \frac{\nabla_\theta p_\theta(\boldsymbol{y},\boldsymbol{z})}{p_\theta(\boldsymbol{y},\boldsymbol{z})}\left(\frac{\nabla_\theta p_\theta(\boldsymbol{y},\boldsymbol{z})}{p_\theta(\boldsymbol{y},\boldsymbol{z})}\right)^\top\right] p_\theta(\boldsymbol{z}\mid\boldsymbol{Y}=\boldsymbol{y})\,\mathrm{d}\boldsymbol{z} \\
&\quad + \int_{\boldsymbol{z}\in\mathcal{Z}}\left[\frac{\nabla_\theta p_\theta(\boldsymbol{y},\boldsymbol{z})}{p_\theta(\boldsymbol{y},\boldsymbol{z})}\left(\frac{\nabla_\theta p_\theta(\boldsymbol{y},\boldsymbol{z})}{p_\theta(\boldsymbol{y},\boldsymbol{z})}\right)^\top\right] p_\theta(\boldsymbol{z}\mid\boldsymbol{Y}=\boldsymbol{y})\,\mathrm{d}\boldsymbol{z} \\
&= \int_{\boldsymbol{z}\in\mathcal{Z}}\left[\mathbf{H}_\theta\left(\log p_\theta(\boldsymbol{y},\boldsymbol{z})\right) + S_\theta(\boldsymbol{y},\boldsymbol{z})S_\theta(\boldsymbol{y},\boldsymbol{z})^\top\right] p_\theta(\boldsymbol{z}\mid\boldsymbol{Y}=\boldsymbol{y})\,\mathrm{d}\boldsymbol{z} \\
&= \mathbb{E}_\theta[\mathbf{J}_\theta\left(S_\theta(\boldsymbol{y},\boldsymbol{Z})\right)\mid\boldsymbol{Y}=\boldsymbol{y}] + \mathbb{E}_\theta\left[S_\theta(\boldsymbol{y},\boldsymbol{Z})S_\theta(\boldsymbol{y},\boldsymbol{Z})^\top\mid\boldsymbol{Y}=\boldsymbol{y}\right]
\end{aligned} \tag{2.24}$$

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 $\mathbf{J}_\theta\left(S_\theta(\boldsymbol{y})\right)$ can be reformulated as in the last equation of the proposition. Combining this last equation to (2.24) completes the proof.

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 which covariates should be kept in the model and which can be removed. The same issue can arise in latent variable models: a typical question for mixture models is to determine how many components/clusters are needed to describe the observed data.

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

All the criteria presented here assume that the observations $y_1, \ldots, y_n$ are the realisations of independent and identically distributed Y_i. In presence of covariates, the Y_i's remain independent, but are not identically distributed. As in Section 2.3, mild assumptions can be made on the covariates to derive similar criteria in this. In case of non-independent observations, the criteria need to be amended [see e.g. Delattre et al., 2014].

2.4.1 Akaike's Information Criterion (AIC)

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

$$\mathrm{AIC}(m) \coloneqq \log p_{\widehat{\theta}_m}(\boldsymbol{y}) - D_m. \tag{2.25}$$

When considering a collection of models $\mathcal{M} = \{m_1, m_2, \ldots\}$, 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.

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* (2.25).

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

1. *the model M is drawn from $\mathcal{M} = \{m_1, m_2, \ldots\}$, 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 $p(\boldsymbol{Y}, \theta, M) = p(M)p(\theta \mid M)p(\boldsymbol{Y} \mid \theta, M)$.

In the Bayesian literature, the distribution $p(M)$ is named the prior distribution (or simply 'the prior') over the models and $p(\theta \mid M = m)$ the prior of the model parameters (conditional to the model). In the framework of Model 2.3, 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

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

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 Bayesian information criterion (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 A.2 from Appendix A.4 yields

$$\log\left(\int p(\boldsymbol{y}, \theta_m, m)\, \mathrm{d}\theta_m\right) = \log p(\boldsymbol{y} \mid \widehat{\theta}_m, m) - D_m \frac{\log n}{2} + O_n(1), \tag{2.26}$$

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 (2.26) in the case of independent and identically distributed observations is given in Appendix A.4. 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].

2.4.3 Integrated Completed Likelihood (ICL)

Latent variable models In presence of a latent variable $\boldsymbol{Z}$, the model selection problem can be stated in the framework of the Bayesian Model 2.4.

Model 2.4 (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 $p(\boldsymbol{Y}, \boldsymbol{Z}, \theta, M) = p(M)p(\theta \mid M)p(\boldsymbol{Z} \mid \theta, M)p(\boldsymbol{Y} \mid \boldsymbol{Z}, \theta, M)$.

Under Model 2.4, the integral from Equation (2.26) becomes

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

involving an additional integration over the latent variables $\boldsymbol{Z}$ to compute criteria such as BIC.

In some specific cases, such as mixture models (Section 3.1) or multivariate Poisson log-normal models (Section 5.3), we can compute explicitly $p(y_i, \theta, m)$ (where the Z_i are explicitly marginalized) 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 (when the latent variables can not be integrated out), 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 (for instance in Section 5.2 of Chapter 5).

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 A.2 (Appendix A.4) to get

$$\log\left(\int p(\boldsymbol{y}, \boldsymbol{z}, \theta, m)\, \mathrm{d}\theta\right) = \log p(\boldsymbol{y}, \boldsymbol{z} \mid \widehat{\theta}_m, m) - D_m \frac{\log n}{2} + O_n(1). \tag{2.27}$$

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 criterion. 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 (2.27):

$$\mathrm{ICL}_1(m) \coloneqq \log p(\boldsymbol{y}, \widehat{\boldsymbol{z}} \mid \widehat{\theta}_m, m) - D_m \frac{\log n}{2}. \tag{2.28}$$

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 $\mathbb{E}_{\widehat{\theta}_m}\big[\log p(\boldsymbol{y}, \boldsymbol{Z} \mid \theta = \widehat{\theta}_m, M = m) \mid \boldsymbol{Y} = \boldsymbol{y}\big]$.

$$\mathrm{ICL}_2(m) = \mathbb{E}_{\widehat{\theta}_m}\big[\log p(\boldsymbol{y}, \boldsymbol{Z} \mid \theta = \widehat{\theta}_m, M = m) \mid \boldsymbol{Y} = \boldsymbol{y}\big] - D_m \frac{\log n}{2}\,. \tag{2.29}$$

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

Comments

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

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

  The BIC penalty $\mathrm{pen}(m) = D_m \log(n)/2$ only refers to the complexity of the model m, 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 $\mathsf{Ent}\big[\boldsymbol{Z} \mid \boldsymbol{Y} = \boldsymbol{y}, M = m, \theta = \widehat{\theta}_m\big]$ becomes $\mathsf{Ent}_{\widehat{\theta}_m}[\boldsymbol{Z} \mid \boldsymbol{Y} = \boldsymbol{y}]$. The BIC and ICL criteria then adopt their most common form, which is similar to that of the AIC, namely:

$$\begin{aligned}\mathrm{AIC}(m) &= \log p_{\widehat{\theta}_m}(\boldsymbol{y}) - D_m, \\ \mathrm{BIC}(m) &= \log p_{\widehat{\theta}_m}(\boldsymbol{y}) - D_m \frac{\log n}{2}, \\ \mathrm{ICL}_2(m) &= \log p_{\widehat{\theta}_m}(\boldsymbol{y}) - D_m \frac{\log n}{2} - \mathsf{Ent}_{\widehat{\theta}_m}[\boldsymbol{Z} \mid \boldsymbol{Y} = \boldsymbol{y}]\,.\end{aligned} \tag{2.30}$$

Remarks.

- *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.*
- *Likewise several other model selection criteria, both are penalized criterion model selection criterion in the sense that are defined as the difference between a measure of fit (e.g. the maximized log-likelihood) for model* m *(*$\log p_{\widehat{\theta}_m}(\boldsymbol{y})$*) and a model-specific penalty.*
- *Lastly, the AIC, BIC and ICL criteria require the evaluation of the log-likelihood. In Chapter 5, we will see how they can be adapted when the likelihood can not be evaluated.*

This first chapter provided a general introduction to latent variable models and the EM algorithm. To illustrate these concepts, we explored a simple latent variable model: the Gaussian mixture model with two components. The next chapter will introduce more advanced models where the EM algorithm can be applied effectively.

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 mixture model, the zero-inflated Poisson model and a mixture model for genetic data. These three models are mixture models.
- In Sections 3.4 and 3.5 we present respectively the linear mixed-effects model and the probabilistic principal component analysis where the latent variables are continuous and more precisely Gaussian.

Each section follows a consistent structure. First, we introduce a motivating dataset from the fields of ecology or evolution. We then propose a probabilistic model and provide detailed derivations of its complete and marginal likelihoods. The E step and M step of the EM algorithm are thoroughly described, and model selection issues are discussed when relevant. Finally, we apply the inference procedure to the dataset previously introduced.

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 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 3.1 (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 a variety of techniques, including hierarchical clustering and k-means clustering (see Everitt et al. [2011] and references therein). In this chapter, we focus on the most widely used model-based approach: the finite Gaussian mixture model.

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

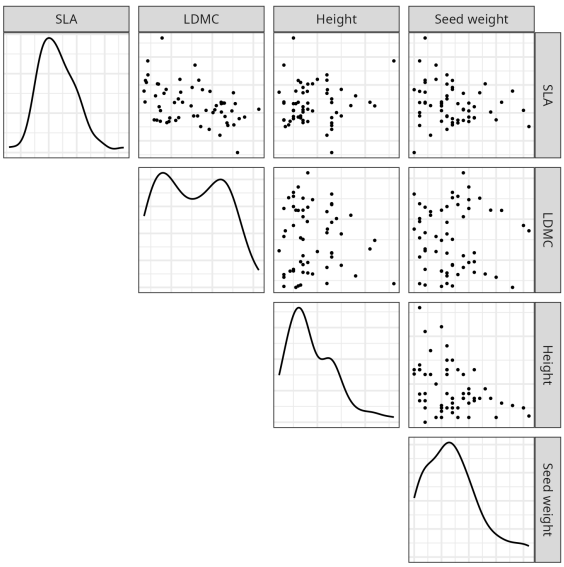


Figure 3.1: Pair plot of vegetal species characteristics. The diagonal shows the empirical density of each variable. Scales are omitted as they are not relevant for illustrating our point. See Table 3.1 for signification of variables.

In this probabilistic framework, the observed data are assumed to be realizations of random variables whose distribution depends on a latent variable. This latent variable takes values in a finite set of size K, where K represents the number of clusters (or communities) present in the data.

In this context, when the number of clusters K is known or fixed, the goals are: i) to recover the cluster assignments from the observed data, and ii) to characterize the distribution of the observed data within each cluster. A third, equally important objective is to estimate the number of clusters K. We postpone this discussion to the end of the section and begin by assuming that K is known.

Notations Let's assume we have observed data $\boldsymbol{y} = \{y_i\}_{1 \le i \le n}$ 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 $1 \le i \le n$,

- $Z_i \in \{1, \dots, K\}$, 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 distribution 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 $d_y \times d_y$ invertible symmetric 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 characterize the distribution of observations within cluster in this Gaussian setting.

In this context $\theta = \{\theta_{\mathrm{lat}}, \theta_{\mathrm{obs}}\}$ is the set of unknown parameters, to be estimated using observations. The Gaussian mixture model is summarized as follow.

Model 3.1 (Gaussian mixture model)**.**

$$Z_i \overset{iid}{\sim} \mathrm{Cat}\left((\omega_1, \ldots, \omega_K)\right), \qquad 1 \le i \le n\,,$$
$$Y_i \mid \{Z_i = z_i\} \overset{ind}{\sim} \mathcal{N}\big(\mu_{z_i}, \sigma_{z_i}^2\big) \qquad 1 \le i \le n\,.$$

Graphical model. A graphical model is a type of probabilistic model in which a graph encodes the conditional independence structure among random variables. In such models, 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 directed graphical model (DAG) associated with the joint distribution $p_\theta(Y, Z)$ for Model 3.1, where the absence of link between components indicates the independence assumption.

By inspecting the directed acyclic graph (DAG) or by applying the rules of conditional probability, we see that,

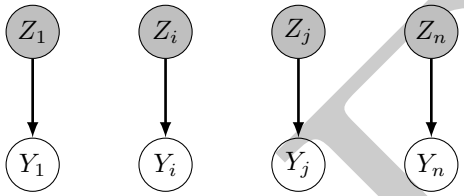


Figure 3.2: Graphical representation of the mixture model (Model 3.1).

conditionally on the observations $\boldsymbol{Y} = \boldsymbol{y}$, the latent variable Z_i depends only on the corresponding observed variable $Y_i = y_i$. That is,

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

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

Definition 3.1 (Mixture distribution)**.** *Let $K \ge 2$ be an integer. A random variable Y on $\mathbb{R}^d$ is said to follow a K-mixture distribution if there exist K probability density functions $p_1(y), \ldots, p_K(y)$ and weights $\omega_1, \ldots, \omega_k$ 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 kth mixture component.

Proposition 3.1. *Under Model 3.1, 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 3.1

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

$1 \leq i \leq n$, 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 3.2. *The marginal log-likelihood of the Gaussian mixture model (Model 3.1) 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 and $x^\top$ is the transposed of x. Its complete log-likelihood is:

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

Equation (3.3) theoretically allows for maximum likelihood estimation via a numerical optimization algorithm. However, this approach can suffer from significant numerical instability, primarily due to the presence of a logarithm applied to a sum of exponential terms, which prevents straightforward simplification. In this context, the Expectation-Maximization (EM) algorithm offers a powerful and numerically stable alternative, avoiding the need for direct approximation of the observed likelihood.

Proof of Proposition 3.2

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

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

Then, using Proposition 3.1, the log-likelihood of Model 3.1 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

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

can be compacted into a unique expression as:

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

The complete log-likelihood in this case writes:

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

3.1.4 EM algorithm

The EM for the Gaussian mixture model writes as:

Algorithm 3.1 (EM for a Gaussian mixture model). *Starting from $\theta^{(0)}$, repeat until convergence:*

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

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

M step. *For all $k = 1, \dots, K$, 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. $\left\|\theta^{(h+1)} - \theta^{(h)}\right\| < \epsilon$ 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 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:*

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

- *$N_k^{(h)}$ 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 since it represents the expected proportion of observations assigned to cluster k under the parameter $\theta^{(h)}$. Similarly, the cluster expectation (μ_k) is*

estimated as the empirical weighted mean of the observations, where the weights correspond to the posterior probabilities of each observation to belong to cluster k under $\theta^{(h)}$. The same principle applies to the estimation of the cluster variance Σ_k, which is computed as a weighted empirical covariance.

Proof of Algorithm 3.1

Objective function $Q(\theta \mid \theta^{(h)})$. 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:

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

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 $Q\big(\theta \mid \theta^{(h)}\big)$ with respect to θ. Note that $Q\big(\theta \mid \theta^{(h)}\big)$ can be separated into $Q\big(\boldsymbol{\omega} \mid \theta^{(h)}\big)$ and $Q\big(\boldsymbol{\mu}, \boldsymbol{\Sigma} \mid \theta^{(h)}\big)$:

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

- Maximization with respect to $\boldsymbol{\omega}$

  We want to maximize $Q\big(\boldsymbol{\omega} \mid \theta^{(h)}\big)$ 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

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

We easily see that $\nabla_{\boldsymbol{\omega},\lambda} Q\big(\boldsymbol{\omega}, \lambda \mid \theta^{(h)}\big) = 0$ 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

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

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:

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

Then, as Σ_k is positive definite:

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

- Maximization with respect to Σ_k Let's consider the derivative[2]with respect to Σ_k

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

Then, it is direct to see that:

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

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

3.1.5 About the clustering

If clustering is central to the statistical analysis, it may be necessary to assign each data point to a single cluster—this is known as *hard clustering*. Such an assignment can be obtained using the MAP (maximum a posteriori) criterion:

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

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

And so (thanks to the product form):

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

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 3.2 (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*

$$\ell_i^{(h)} = \arg\min_j d(y_i, \mu_j^{(h)}) \qquad \text{Index of the closest mean}$$

$$\tau_{i,k}^{(h)} = \mathbb{I}_{\ell_i^{(h)}}(k), \qquad k \in \{1,\dots,K\}$$

- *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 assignment, 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.

Clustering uncertainty In contrast to hard clustering methods, the mixture model combined with the EM algorithm provides a probabilistic assignment of each observation to clusters. Specifically, the quantities $\widehat{\tau}_{ik} = \mathbb{P}_{\widehat{\theta}}(Z_i = k \mid Y_i = y_i)$ represent the posterior probabilities of observation i belonging to cluster k, a procedure commonly referred to as soft clustering.

The uncertainty associated with the clustering can be quantified using the entropy of the conditional distribution of the latent variables given the observations:

$$\mathsf{Ent}_{\widehat{\theta}}[\boldsymbol{Z} \mid \boldsymbol{Y} = \boldsymbol{y}] = -\sum_{i=1}^n \sum_{k=1}^K \widehat{\tau}_{ik} \log \widehat{\tau}_{ik} \tag{3.11}$$

This entropy measures the fuzziness of the clustering assignment and ranges from 0 to $n \log(K)$ for n independent observations taking K values. If each vector $(\widehat{\tau}_{i1}, \dots, \widehat{\tau}_{iK})$ is close to a vector such as $(0,\dots,0,1,0,\dots,0)$, then each observation is confidently assigned to a unique cluster, and the entropy will be close to 0. Conversely, higher entropy values indicate more uncertainty in cluster assignments. The maximum entropy, achieved when $\widehat{\tau}_{ik} = \frac{1}{K}$ for all i and k, corresponds to complete uncertainty in clustering—an extreme and typically unrealistic scenario in practice.

As a conclusion, on the one hand, the probabilistic mixture model for clustering relies on specific distributional assumptions about the data, which should be carefully assessed in practice. On the other hand the probabilistic mixture model for clustering has the advantage of providing a principled measure of uncertainty in cluster assignments, in contrast to methods such as K-means which yield hard partitions.

3.1.6 Choosing the number of components

The previous section detailed how to estimate model parameters when the number of components is known. However, a crucial remaining question is: how can we estimate the number of components K? As discussed in Section 2.4, the log-likelihood alone is not a suitable criterion for this purpose. This is because a model with $K-1$ components is nested within a model with K components, and the likelihood inevitably increases as components are added. To address this, the number of components K is usually selected using penalized criteria such as AIC, BIC or ICL, defined in Equations (2.30), 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 the case of

Gaussian mixtures, 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 :

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

It is important to note that these selection criteria are not designed with the same goal. For instance, the ICL criterion includes an additional penalty term accounting for the uncertainty in classification, measured by the conditional entropy defined in Equation (3.11). As a result, ICL tends to favor models with fewer, more clearly separated clusters compared to BIC (see the following section and Figure 3.3 for an illustration).

3.1.7 Illustration on the Bohemia meadow dataset

We now go back to be Bohemia meadow Dataset 3.1. We estimate the parameters of the multivariate Gaussian mixture model.

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 evolution of the log likelihood along the iterations for the multiple starting points: note that all the curves of the likelihood along the iterations are non-decreasing, as stated in Proposition 2.3. Figure 3.3 also shows the differences in the final log-likelihood, therefore illustrating the numerous 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 $K \in \{1, \dots, 6\}$, and choosing the best final point among the 200 trials[3]. For the 6 models, we compute the three model selection criteria 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 3.4) 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.

3.1.8 Extensions

In this section, we have detailed the computations involved in the multivariate Gaussian mixture model, including the complete and marginal likelihoods as well as the EM algorithm updates or the model selection criterion. However, it is worth emphasizing that the same principles apply to a wide range of other models. These include, for instance, mixtures of linear regressions, Poisson mixtures, mixtures of log-normal distributions, or more generally to any mixtures of distributions belonging to the exponential family (see Appendix Section A.2). While the specific forms of the likelihood and parameter updates may differ, the general EM framework and inference strategy remain conceptually similar across these models.

3.2 Zero-inflated Poisson for species distribution

We now turn to another important example of latent variable models: the Zero-Inflated Poisson (ZIP) model, which is particularly useful for count data with an excess of zero observations.

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

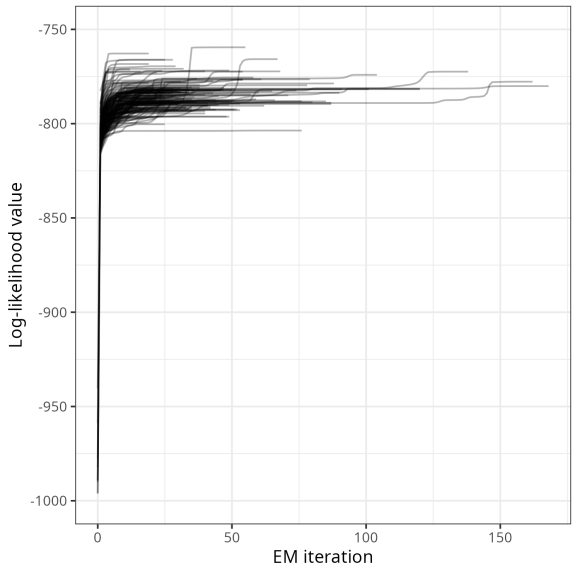


Figure 3.3: Evolution of the log-likelihood when performing EM from 200 different starting points on data set of Table 3.1, with $K = 3$. The algorithm stops when the increase of the log-likelihood is lower than 10^{-8} (hence the different number of iterations for each curve).

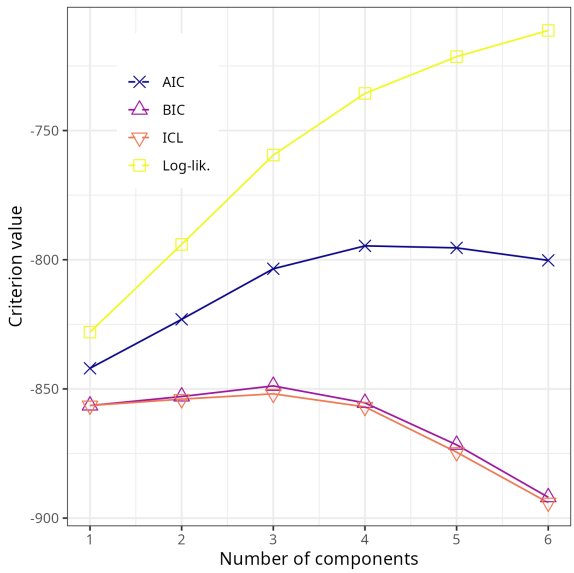


Figure 3.4: Log-likelihood and penalized likelihood criteria for different value of K (Example 3.1).

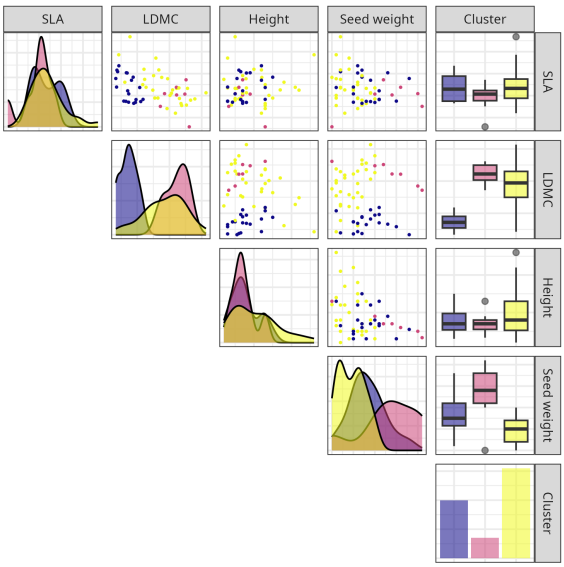


Figure 3.5: Best clustering (in terms of log-likelihood) for $K = 3$.

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.

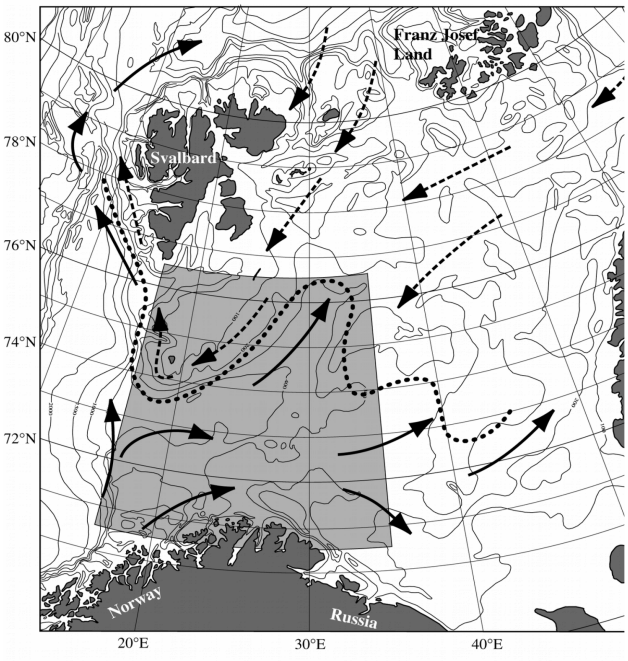


Figure 3.6: Map of the Barents sea, where data from Example 3.2 were collected.

Dataset 3.2 (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

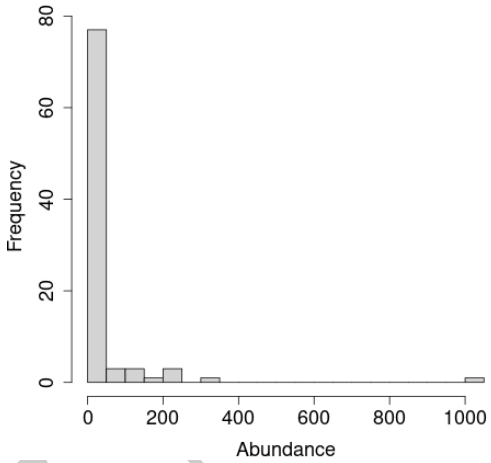


Figure 3.7: Cod abundance in the Barents sea. Left: head of the data table. Right: Histogram of the observed abundances.

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:

$$\forall 1 \leq i \leq n,\ Y_i \overset{\text{iid}}{\sim} \mathcal{P}(\lambda_i), \qquad \log(\lambda_i) = x_i^\intercal \beta\,. \tag{3.12}$$

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

$$\log(\lambda_i) = o_i + x_i^\intercal \beta. \tag{3.13}$$

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 (3.12) 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 $\widetilde{Y}_i = \mathbb{I}_{Y_i>0}$:

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

and to use a logistic regression model:

$$\forall 1 \leq i \leq n,\ \widetilde{Y}_i \overset{\text{iid}}{\sim} \mathcal{B}\text{ern}(\pi_i), \qquad \log\left(\frac{\pi_i}{1-\pi_i}\right) = x_i^\intercal \alpha. \tag{3.14}$$

This simple approach binarizing the observations results in a loss of information, as it ignores the actual count values beyond the binary outcome.

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 3.2 (Zero-inflated Poisson regression model)**.** *For $1 \leq i \leq n$:*

$$Z_i \overset{iid}{\sim} \mathcal{B}(\pi_i),$$
$$Y_i \mid \{Z_i = z_i\} \overset{ind}{\sim} \begin{cases} \delta_{\{0\}} & \text{if } z_i = 0\,, \\ \mathcal{P}(\lambda_i) & \text{if } z_i = 1\,, \end{cases},$$

where $\delta_{\{0\}}$ stands for the Dirac mass in zero[a], and the probability of presence and intensity of presence are linked to covariates by:

$$\operatorname{logit}(\pi_i) := \log\left(\frac{\pi_i}{1-\pi_i}\right) = x_i^\intercal \alpha\,,$$
$$\log(\lambda_i) = x_i^\intercal \beta\,.$$

[a] Formally, $Y \sim \delta_{\{0\}} \Leftrightarrow \mathbb{P}(Y = 0) = 1$

Remarks.

- *Again, an offset term o_i can be added to the Poisson regression to account for heterogeneous sampling efforts.*
- *Note that Model (3.2) is sometimes parametrized 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 3.1) 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 (3.2) 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 de-conditioning on Z_i and turns out to be a zero-inflated distribution.

Definition 3.2. *The random variable Y over $\mathbb{N}$ has a zero-inflated distribution $ZIP(\pi, \lambda)$ iff*

$$\mathbb{P}(Y = 0) = (1 - \pi) + \pi e^{-\lambda}, \qquad \text{and, for } y \geq 1, \quad \mathbb{P}(Y = y) = \pi e^{-\lambda} \frac{\lambda^y}{y!}. \tag{3.15}$$

Formula (3.15) which can be reformulated into a unique formula:

$$\mathbb{P}_{\text{ZIP}}(Y = y) = (1 - \pi)\mathbb{I}_{\{0\}}(y) + \pi e^{-\lambda} \frac{\lambda^y}{y!}. \tag{3.16}$$

Proposition 3.3. *Under Model* (3.2)*, the marginal distribution of the observed abundance Y_i is a zero-inflated Poisson $ZIP(\pi_i, \lambda_i)$.*

Proof of Proposition 3.3

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 $y_i \geq 1$, 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 (3.2) is the same as this of the mixture model, given in Figure 3.2. According to this model the couples $\{(Y_i, Z_i)\}_{1\leq i\leq n}$ are all independent.

3.2.3 Marginal and complete log-likelihoods

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

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

Denoting $\boldsymbol{Z} = \{Z_i\}_{1\leq i\leq n}$, the complete likelihood of Model (3.2) is

$$\begin{aligned}\log p_\theta(\boldsymbol{y}, \boldsymbol{Z}) &= \log p_\theta(\boldsymbol{Z}) + \log p_\theta(\boldsymbol{y} \mid \boldsymbol{Z}) \\ &= \sum_{i=1}^{n} Z_i \log \pi_i + (1-Z_i)\log(1-\pi_i) + \sum_{i=1}^{n} Z_i\left(-\lambda_i + y_i \log\lambda_i - \log(y_i!)\right). \end{aligned} \tag{3.17}$$

3.2.4 EM algorithm for the ZIP model

Algorithm 3.3 (EM for the ZIP model)**.** *Starting from $\theta^{(0)}$, repeat until convergence:*

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

$$\tau_i^{(h)} = \mathbb{P}_\theta(Z_i = 1 \mid Y_i = y_i) = \frac{(1-\pi_i^{(h)})\mathbb{I}_{y_i>0} + \pi_i^{(h)} e^{-\lambda_i^{(h)}}}{(1-\pi_i^{(h)}) + \pi_i^{(h)} e^{-\lambda_i^{(h)}}}. \tag{3.18}$$

M step. *Update the estimate of θ as*

$$\alpha^{(h+1)} = \arg\max_{\alpha} \sum_{i=1}^{n} \tau_i^{(h)} \log\pi_i + (1-\tau_i^{(h)})\log(1-\pi_i) \qquad \textit{with} \qquad \operatorname{logit}(\pi_i) = x_i^\top \alpha \tag{3.19}$$

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

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 3.3

Objective function $Q(\theta \mid \theta^{(h)})$. From the expression of the complete log-likelihood provided in Equation (3.17), the integration of the latent variables Z_i leads to the following formula for $Q(\theta \mid \theta^{(h)})$:

$$Q(\theta \mid \theta^{(h)}) = \sum_{i=1}^{n} \tau_i^{(h)} \log \pi_i + (1 - \tau_i^{(h)}) \log(1 - \pi_i) + \sum_{i=1}^{n} \tau_i^{(h)} \left(-\lambda_i + y_i \log \lambda_i - \log(y_i!)\right), \tag{3.20}$$

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

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 $\tau_i^{(h)} = \mathbb{E}_{\theta^{(h)}}[Z_i \mid \boldsymbol{Y} = \boldsymbol{y}]$. 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 (3.2), we easily see that if the observed count is not zero ($y_i > 0$), then the species is surely present, so

$$\mathbb{P}_{\theta^{(h)}}(Z_i = 1 \mid Y_i > 0) = 1. \tag{3.21}$$

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

$$\begin{aligned} \mathbb{P}_{\theta^{(h)}}(Z_i = 1 \mid Y_i = 0) &= \frac{\mathbb{P}_{\theta^{(h)}}(Y_i = 0, Z_i = 1)}{\mathbb{P}_{\theta^{(h)}}(Y_i = 0)} = \frac{\mathbb{P}_{\theta^{(h)}}(Y_i = 0 \mid Z_i = 1)\mathbb{P}_{\theta^{(h)}}(Z_i = 1)}{\mathbb{P}_{\theta^{(h)}}(Y_i = 0)} \\ &= \frac{\pi_i^{(h)} \mathrm{e}^{-\lambda_i^{(h)}}}{(1 - \pi_i^{(h)}) + \pi_i^{(h)} \mathrm{e}^{-\lambda_i^{(h)}}} \quad \text{(using Equation (3.16) with } y_i = 0) \end{aligned} \tag{3.22}$$

Now, combining Equations (3.21) and (3.22), we obtain a global formula:

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

M step. We may now update the parameter θ by maximizing the objective function of the EM algorithm provided in Equation (3.20). 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 Analysis of the Cod abundance in the Barent sea

We now compare the ZIP regression (3.2) with the logistic regression (3.14) and Poisson regression (3.12) on the cod abundances introduced in Dataset 3.2. To ease the interpretation and the comparison of the regression coefficients, the four covariates were centered and their variances were set to one. Models (3.14) and (3.12) can be fitted with the R `glm` R function, and model (3.2) 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 (3.2) 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 centred, 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 (α)					Abundance (β)				
	Inter.	Lat.	Long.	Depth	Temp.	Inter.	Lat.	Long.	Depth	Temp.
Logistic (3.14)	-1.275	-0.251	0.301	-0.387	1.994	–	–	–	–	–
Poisson (3.12)	–	–	–	–	–	0.010	-0.721	-0.043	0.917	2.479
ZIP (3.2)	-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.

Presence probability. Figure 3.8 (left) gives the estimated probability $\widehat{\pi}_i^{ZIP}$ of presence in each station for Example 3.2, 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 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 (3.2), the sites can classified in term of actual presence or absence of the 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 (3.18). 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.

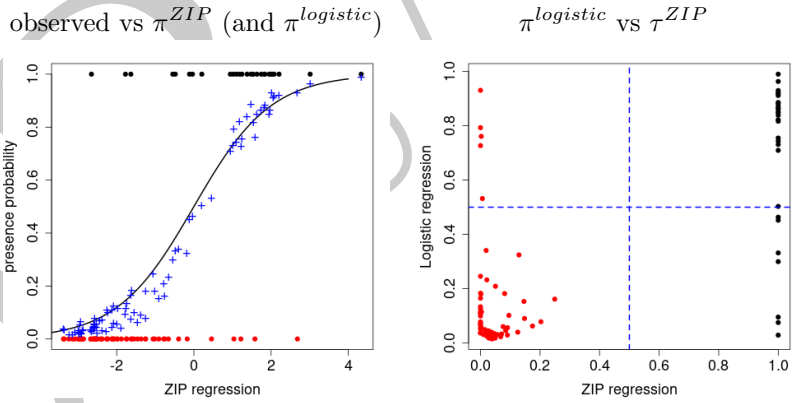


Figure 3.8: Cod abundance in the Barents sea. Left: probability of presence according to the ZIP model ($\widehat{\pi}_i^{ZIP}$: black curve), dots = observed presence (black dots (•): $Y_i > 0$, red dots (•): $Y_i = 0$), blue crosses (+): probability of presence according to logistic regression $\pi_i^{logistic}$.
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 (middle): 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 account 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.

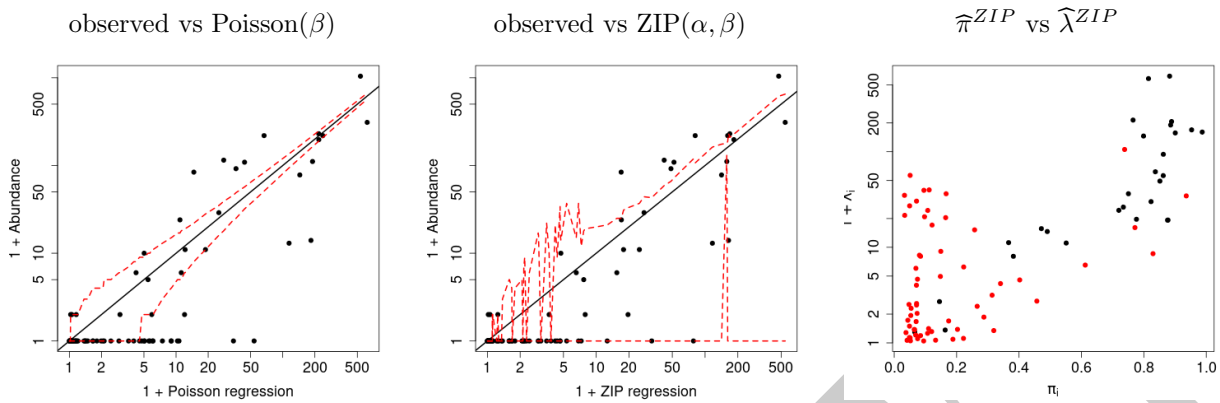


Figure 3.9: Cod abundance in the Barents sea. Left: observed abundances vs predicted abundances (in log-scale) with the Poisson regression (3.12), dotted red lines = 95% interval for the Poisson distribution (1 is added to abundances to allow log-scale). Center: observed abundances vs predicted abundances $\widehat{\lambda}_i$ (in log-scale) with the ZIP regression (3.2), 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 (3.2). Red dots = sites where the species was not observed.

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

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

Models (3.12) and (3.2) 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 3.2 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 $1 \le i \le n$:

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

Proposition 3.4. *For the ZIP model without covariates, let us define:*

$$P = \sum_{i=1}^{n} \mathbb{I}_{y_i>0}, \quad Y_+ = \sum_{i=1}^{n} y_i$$

and

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

Then we have:

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

Proof of Proposition 3.4

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 (3.17) becomes

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

where

$$Z_+ = \sum_{i=1}^n Z_i, \qquad Y_+ = \sum_{i=1}^n Z_i y_i = \sum_{i=1}^n y_i \qquad \text{and} \qquad L = \sum_{i=1}^n Z_i \log(y_i!) = \sum_{i=1}^n \log(y_i!).$$

The expressions 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

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

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

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

The Hessian is then obtained by calculating the second derivatives:

$$\partial^2_{\pi^2} \log p_\theta(\boldsymbol{y}, \boldsymbol{Z}) = -\frac{Z_+}{\pi^2} - \frac{n - Z_+}{(1-\pi)^2}, \qquad \partial^2_{\pi\lambda} \log p_\theta(\boldsymbol{y}, \boldsymbol{Z}) = 0, \qquad \partial_{\lambda^2} \log p_\theta(\boldsymbol{y}, \boldsymbol{Z}) = -\frac{Y_+}{\lambda^2}. \tag{3.23}$$

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

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

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

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

we have from Equation (3.18) that $\tau(y_i) = (1-\eta)\mathbb{I}_{y_i > 0} + \eta$, so

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

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

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

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 (3.18), so So

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

that is :

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

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

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

$$\mathbb{E}[\partial_{\lambda^2}\log p_\theta(\boldsymbol{y},\boldsymbol{Z}) \mid \boldsymbol{Y} = \boldsymbol{y}] = -\frac{y_+}{\lambda^2}\,,$$

and

$$\mathbb{V}[\partial_\pi \log p_\theta(\boldsymbol{y},\boldsymbol{Z}) \mid \boldsymbol{Y} = \boldsymbol{y}] = \gamma^2\mathbb{V}[Z_+ \mid \boldsymbol{Y} = \boldsymbol{y}]\,.$$

Example. When considering the data from Example 3.2, using the ZIP model (3.2) 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:

$$\mathsf{IC}\,(\pi) = [0.2181, 0.4111], \qquad \mathsf{IC}\,(\lambda) = [100.5, 108.0]\,.$$

3.2.7 Conclusion

The ZIP model fits naturally within the framework of latent variable models, where the excess zeros are explained by an unobserved binary indicator; as such, it benefits from the EM algorithm for parameter estimation, following the same principles detailed in previous sections. Its extension to multiple species will be discussed at the end of Section 5.3 of Chapter 5.

3.3 Genetic structure of a population: mixture model

We now turn to a mixture model specifically tailored to population genetics data, where the goal is to identify underlying genetic populations (or clusters) based on multilocus genotype information.

3.3.1 Data and question

Understanding genetic diversity within a population is a central question in population genetics. A natural approach to modeling this diversity is to assume the existence of ancestral populations, from which the genomes of individuals in the current population are derived.

Dataset 3.3 (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 $1 \leq i \leq n$, 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

		Markers							
	i	$j=1$	$j=2$	$j=3$	$j=4$	$j=5$	$j=6$	$j=7$	Location
$Y_1 = \{Y_{1j}\}_{1\leq j\leq 7}$	1	1	1	6	1	2	2	5	Ngangao
	1	3	2	6	1	2	4	5	
$Y_2 = \{Y_{2j}\}_{1\leq j\leq 7}$	2	3	2	1	1	2	2	5	Ngangao
	2	3	3	6	1	2	10	8	
$Y_3 = \{Y_{3j}\}_{1\leq j\leq 7}$	3	3	3	6	1	3	2	1	Ngangao
	3	7	5	6	2	NA	2	4	

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

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 3.3 (Mixture model for the genetic structure of a population)**.** *Denoting by Z_i the population of origin of individual i and $Y_i = \{Y_{ij}\}_{1\leq j\leq p}$ the whole genotype of individual i, the model states that*

$$Z_i \overset{iid}{\sim} \text{Cat}\,(\omega = (\omega_1, \ldots, \omega_K)), \qquad 1 \leq i \leq n\,,$$
$$Y_{ij} \mid \{Z_i = z_i\} \overset{ind}{\sim} \text{Cat}\,(\boldsymbol{\phi_{z_i j}}) \qquad 1 \leq i \leq n, 1 \leq j \leq n\,.$$

where for $1 \leq k \leq K$, $\boldsymbol{\phi_{kj}}$ is the set of probabilities for each pair of alleles $\{a, b\} \in \{1, \ldots m_j\}^2$, such that:

$$\phi_{kj}(\{a, b\}) = \begin{cases} 2\gamma_{kja}\gamma_{kjb} & \text{if } a \neq b \quad (\textit{heterozygous}), \\ \gamma_{kja}^2 & \text{if } a = b \quad (\textit{homozygous}). \end{cases} \tag{3.24}$$

The parameters of the model are

- *$\omega = [\omega_k]_{1\leq k\leq K}$ the vector of probabilities for an individual to come from a given population,*
- *γ_{kja} the allelic frequency of allele $a \in \{1, \ldots m_j\}$ at locus $j \in \{1, \ldots p\}$ in population $k \in \{1, \ldots K\}$,*
- *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 3.3 relies on three key assumptions:

1. The entire genome of a given individual originates from a single ancestral population, meaning that Z_i is 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 (3.24) .

Assumption 3 determines the form of the emission probability $\phi_{kj}(\{a,b\})$ as defined in Equation (3.24). This assumption is both biologically reasonable for ancient or well-established populations and statistically advantageous, as it significantly reduces the number of emission parameters γ to be estimated.

Assumption 2 is valid primarily when the loci are sufficiently distant along the genome, thereby limiting the influence of linkage disequilibrium. This is the case in Example 3.3, which involves only seven widely spaced genetic markers. A more flexible framework that accommodates potential dependencies between neighboring loci is introduced in Section 4.1.7.

Finally, Assumption 1 is quite restrictive, as it excludes the possibility of admixture—i.e., individuals having ancestry from multiple populations—a phenomenon that is biologically plausible and often observed. This limitation is addressed by the generalization presented later in Section 4.1.7, which relaxes this assumption.

Graphical model. The graphical model of Model 3.3 is given in Figure 3.10. Its structure mainly results from the fact that the couples $\{(Z_i, Y_i)\}_{1\leq i\leq n}$ are iid.

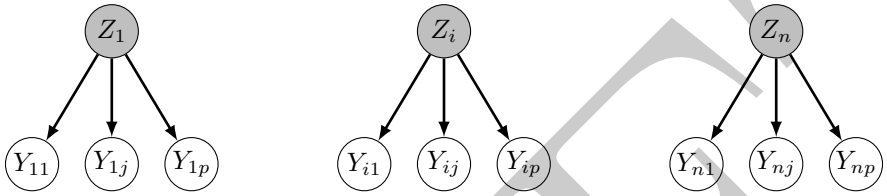


Figure 3.10: Graphical representation of the mixture Model 3.3 for the genetic structure of a population.

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

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

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

$$\log p_\theta(\boldsymbol{y}) = \sum_{i=1}^{n} \log\left[\sum_{k=1}^{K} \omega_k \phi_k(y_i)\right] = \sum_{i=1}^{n} \log\left[\sum_{k=1}^{K} \omega_k \left(\prod_{j=1}^{p} \phi_{kj}(y_{ij})\right)\right] \tag{3.25}$$

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 3.4 (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:*

$$\tau_{ik}^{(h)} = \frac{\omega_k \phi_k(y_i)}{\sum_{\ell=1}^{K} \omega_\ell \phi_\ell(y_i)} = \frac{\omega_k \prod_{j=1}^{p} \phi_{kj}(y_{ij})}{\sum_{\ell=1}^{K} \omega_\ell \prod_{j=1}^{p} \phi_{\ell j}(y_{ij})}. \tag{3.26}$$

M step. *Update the estimate of θ as*

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

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

Proof of Algorithm 3.4

Objective function $Q(\theta \mid \theta^{(h)})$. Using the formula of the complete likelihood provided in Equation (3.25). The conditional expectation of the complete likelihood is

$$Q(\theta \mid \theta^{(h)}) = \sum_{i=1}^{n} \sum_{k=1}^{K} \tau_{ik}^{(h)} \left(\log \omega_k + \sum_{j=1}^{p} \log \phi_{kj}(y_{ij}) \right) \tag{3.27}$$

where

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

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 (3.27), 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:

$$\phi_{kj}(y_{ij}) = \phi_{kj}(\{y_{ij}^1, y_{ij}^2\}) = \prod_{a=1}^{m_j} \gamma_{kja}^{y_{ija}^1} \prod_{b=1}^{m_j} \gamma_{kjb}^{y_{ijb}^2} \prod_{a=1}^{m_j} 2^{y_{ija}^1 - y_{ija}^1 y_{ija}^2}$$

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

To get the update formulas for the allelic frequencies, we set to zero the derivative of (3.27), 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

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

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

$$1 = \sum_{a=1}^{m_j} \frac{1}{\lambda_{kj}} \sum_{i=1}^{n} \tau_{ik}^{(h)} (y_{ija}^1 + y_{ija}^2) \quad \Leftrightarrow \quad \lambda_{kj} = \sum_{i=1}^{n} \tau_{ik}^{(h)} \left(\underbrace{\sum_{a=1}^{m_j} y_{ija}^1}_{=1} + \underbrace{\sum_{a=1}^{m_j} y_{ija}^2}_{=1} \right) = 2 \sum_{i=1}^{n} \tau_{ik}^{(h)}$$

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

3.3.5 Selection of the number of founder populations

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 3.3 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 Analysis of the Taita Thrush dataset

Model 3.3 can be used to analyse the genetic structure of the Taita Thrush population introduced in Dataset 3.3. Figure 3.11 gives the log-likelihood, the BIC and the ICL criterion for $K = 1, \dots 5$ 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 to a low conditional entropy $\mathsf{Ent}[Z \mid Y = y]$, which indicates a low uncertainty in the classification: τ_{ik} are all either close to 1 or to 0.

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 (orange) of marker 2 is predominant in populations 2 and 3, but rare in population 1 and that the first allele (green) of marker 3 is quite frequent in population 2, whereas it is quite rare in populations 1 and 3. Figure 3.13 gives the estimated probabilities τ_{ik} that the bird i originates from population k given its genotype y_i for all birds, sorted by their capture location. We observe a strong consistency between the estimated population of origin and the capture location, as the estimated probabilities are almost the same (and either 0 or 1) for every areas. One can see an exception for two birds from Mbololo, which have a high probability to originate from the same population as Ngangao. Finally, we also see that the birds from Yale seem to originate from the same founding population as those from Mbololo.

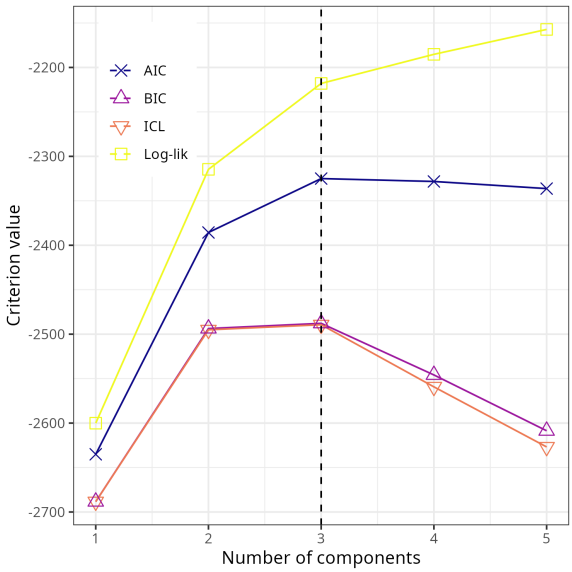


Figure 3.11: Log-likelihood and penalized criteria for Model 3.3 for the Taita thrush data set (Dataset 3.3) as a function of the number of populations K.

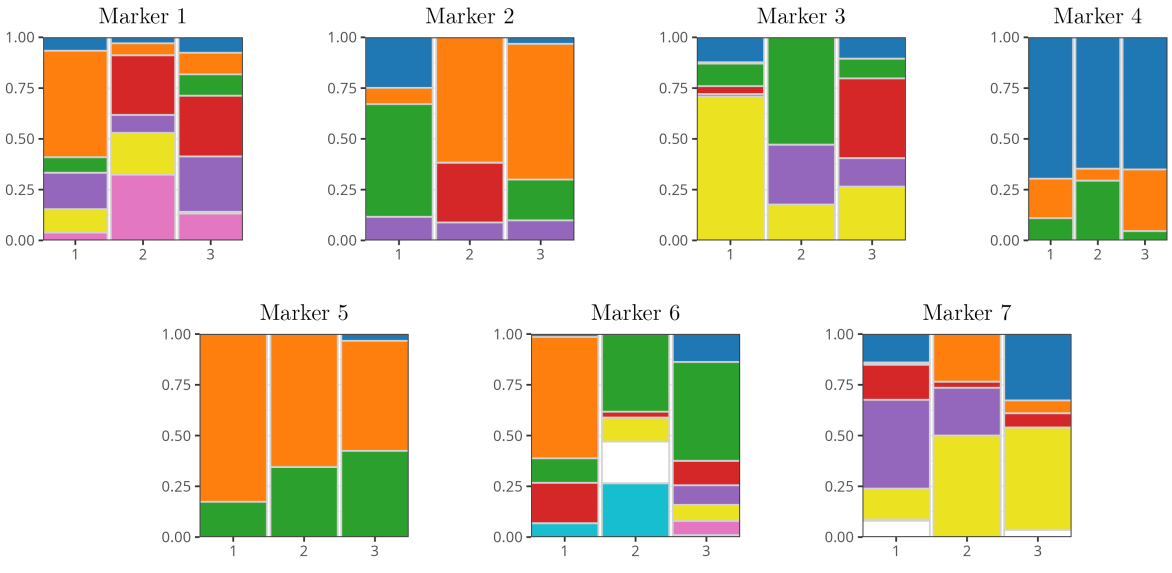


Figure 3.12: Allelic frequencies γ_{kja} for the $p = 7$ markers ($1 \leq j \leq p$) in each of the $\widehat{K} = 3$ populations ($1 \leq k \leq \widehat{K}$) of origin of Taita thrushes (Example 3.3). Colors refer the alleles $1 \leq a \leq m_j$ of each marker j.

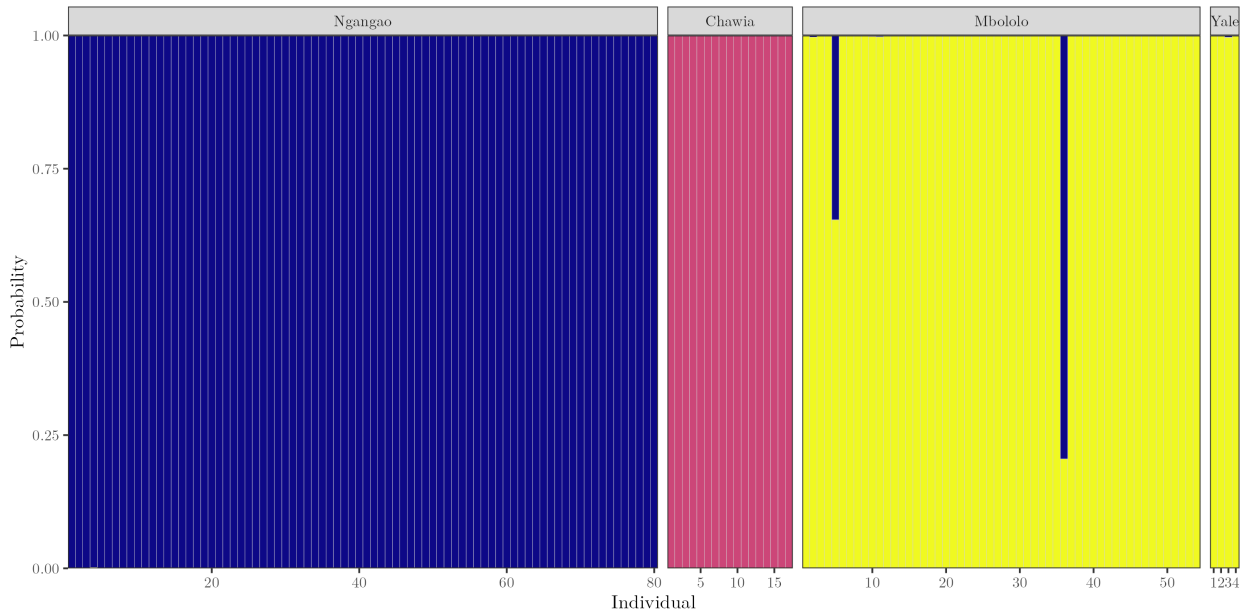


Figure 3.13: Clustering of the Taita thrushes (Example 3.3) into the $\widehat{K} = 3$ inferred populations of origin. The blue is for population 1, the the red for 2, and the yellow for 3. Data were separated depending on their 4 different capture locations.

3.4 Linear mixed model

Linear Mixed Models (LMM) and Generalized Linear mixed models (GLMM) have witnessed a popularity increase in the last decades, in every domain, including ecology [see Harrison et al., 2018, and references therein]. Mixed models extend the classical linear models by introducing a random mixed effect in order to account for dependencies between the observations. The simplest case is when one observes repeatedly a variable of interest several times on the same statistical unit.

3.4.1 Data and question

Dataset 3.4 (Ammonium concentration in Borneo soil [Sellan et al., 2021])**.** *We consider the ($\log_{10}$) concentration of ammonium (NH_4) in 180 soil samples in the Malaysian part of Borneo. The samples were obtained for 3 different soil typologies (alluvial, heath and sandstones soils) on 9 different geographical areas (hereafter named plots). Figure 3.14 displays the distribution of the concentrations among each plot within each soil.*

Among other questions (see Sellan et al. [2021]), a question was whether the variations of ammonium could be explained by soil typology. This naturally suggests the use a linear model (more specifically, analysis of variance, or ANOVA). However, the fact that some observations originate from the same geographical area (statistical unit) breaks the classical independence assumption in linear models. Indeed, one might expect that a (positive) correlation exists between measurements taken from the same plot. Linear mixed models explicitly account for this correlation structure of the data by introducing a latent variable in the modelling framework. In the specific context of mixed models, this latent variable is commonly referred to as a random effect.

3.4.2 The linear mixed model

3.4.2.1 Guiding example: A single random effect on the intercept

Notations. We consider a set of n observations coming from J ($J < n$) statistical units, typically a geographical area or a lineage, such that for each statistical unit j, there are n_j observations, and so $n = \sum_{j=1}^{J} n_j$. We denote y_{ij} the ith observation of the jth statistical unit ($1 \leq j \leq J$, $1 \leq i \leq n_j$); $y_{ij} \in \mathbb{R}$. $x_{ij} \in \mathbb{R}^p$ is the set of covariates for this observation. As classically done in statistical models, the first element of x_{ij} is 1 (corresponding to the intercept of the linear model), and qualitative variables are encoded with binary variables[4]. We denote by $\boldsymbol{y}$ the vector of observations sorted by statistical unit (*i.e*, $\boldsymbol{y} = [y_{11}, \dots y_{n_1 1}, y_{12}, \dots, y_{n_J J}]^\top$), and $\boldsymbol{X}$ the $n \times p$ design matrix whose rows are given by the x_{ij}. We denote by $\boldsymbol{U}$ the $n \times J$ design matrix that describes how the observations are spread among the statistical units. More specifically, we define $\boldsymbol{U} = [u_{ij}]_{1 \leq in \leq n, 1 \leq j \leq J}$, where u_{ij} is 1 if the observation i has been made on the statistical unit j, and zero otherwise.

Considering Example 3.4, we assume that the statistical unit (the plot, in our example) only affects the response through the addition of the constant. $\boldsymbol{y}$ is the realization of $\boldsymbol{Y}$ whose distribution is defined below.

Model 3.4. *Linear mixed model (one random effect)*

Using the above notations, we assume the following model:

$$\begin{aligned} Z_j &\overset{iid}{\sim} \mathcal{N}\big(0, \gamma^2\big), && 1 \leq j \leq J\,, \\ Y_{ij} \mid \{\boldsymbol{Z} = \boldsymbol{z}\}, &\overset{iid}{\sim} \mathcal{N}\big(x_{ij}^\top \beta + z_j, \sigma^2\big) && 1 \leq j \leq J, 1 \leq i \leq n_j\,, \end{aligned}$$

where $\beta \in \mathbb{R}^p, \sigma^2 \in \mathbb{R}_+, \gamma^2 \in \mathbb{R}_+$ are unknown parameters. This model can be written equivalently in its matrix form:

$$\boldsymbol{Z} \sim \mathcal{N}_J\big(0, \gamma^2 \mathbf{I}_J\big), \qquad \boldsymbol{Y} \mid \{\boldsymbol{Z} = \boldsymbol{z}\} \sim \mathcal{N}_n\big(\boldsymbol{X}\beta + \boldsymbol{U}\boldsymbol{z}, \sigma^2 \mathbf{I}_n\big)\,.$$

where $\mathbf{I}_J$ is the identity matrix of size J.

[4]More precisely, a level of reference is set, and the level to which an observation belong is encoded using $K-1$ values with only zeros except, at most, a 1. For instance, for soil typology of Dataset 3.4, we set "alluvial" as the reference level, "heath" as the second level and "sandstone" as the third one. Thus, an observation of type "alluvial" will be encoded as (0,0), of type "heath" as (1, 0) and of type "sandstone" as (0,1). The corresponding x_i is then either (1, 0, 0), (1, 1, 0) or (0, 1, 0).

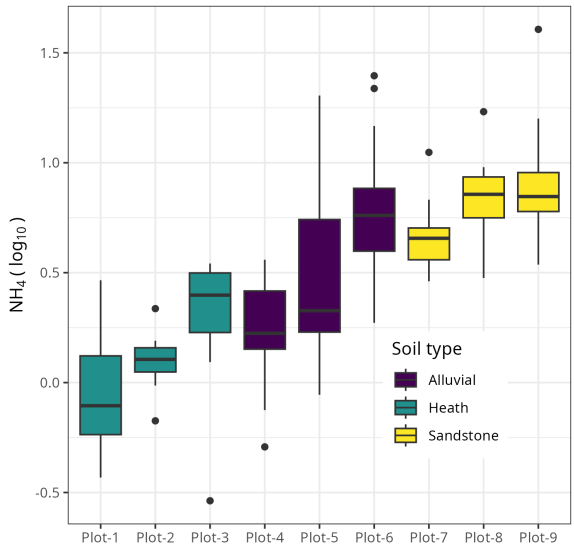


Figure 3.14: Concentration (in $\log_{10}$ scale) of NH_4 in Borneo soil samples. The different plots correspond to different geographical areas. Each boxplot is built from 20 observations.

Graphical model and dependence structure. The graphical model corresponding to Model 3.4 is depicted on Figure 3.15.

It is worth noting that this model is equivalent to modelling the covariance. Indeed, one can reformulate the model as follows: for any observation i ($1 \leq i \leq n$), let $u(i) = j$ if the ith observation belongs to statistical unit j. Then we have

$$Y_i = x_i^\top \beta + Z_{u(i)} + E_i, \;\; E_i \overset{\text{iid}}{\sim} \mathcal{N}\left(0, \sigma^2\right), Z_j \overset{\text{iid}}{\sim} \mathcal{N}\left(0, \gamma^2\right), 1 \leq j \leq J, E_i \perp Z_j \,.$$

Then, we have that:

$$\mathbb{E}[Y_i] = x_i^\top \beta, \qquad \mathbb{V}[Y_i] = \sigma^2 + \gamma^2, \qquad \mathbb{C}\text{ov}\left(Y_i, Y_{i'}\right) = \begin{cases} 0 & \text{if } u(i) \neq u(i') \,, \\ \gamma^2 & \text{if } u(i) = u(i') \,, \end{cases}$$

so

$$\mathbb{C}\text{or}(Y_i, Y_{i'}) = \begin{cases} 0 & \text{if } u(i) \neq u(i') \,, \\ \gamma^2 \big/ \left(\gamma^2 + \sigma^2\right) & \text{if } u(i) = u(i') \,. \end{cases}$$

This last equation illustrates an interesting feature of linear mixed models, which is the decomposition of the variance of the observations. The correlation is a percentage of total variance (the sum at the denominator) due to the statistical unit (whose variance is at the numerator). This ratio is a possible measure for heritability in studies where the latent variable models the lineage effect, as it measures the percentage of variability due to the lineage.

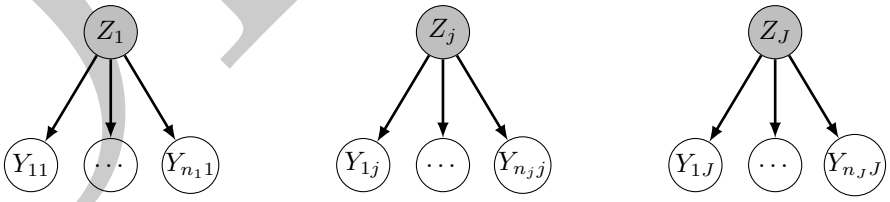


Figure 3.15: Graphical model for Model 3.4.

3.4.2.2 Towards multiple random effects

Motivating example and notations One could imagine extending Model 3.4 to different statistical unit (for instance, different plots but also different lineages), but also that the change of statistical unit modifies the relation between the observations and the covariates (in other words, it does not only modify the first element of the vector β), resulting in a random interaction term. For instance, let's imagine we observe the yield of crops. The crops are planted on J different fields, and come from K different lineages. We denote y_{ijk} the observed yield for the ith crop of the kth lineage on the jth field ($1 \leq j \leq J, 1 \leq k \leq K, 1 \leq i \leq n_{jk}$. Each crop is watered

with an amount of water $x_{ijk} \in \mathbb{R}$. The assumption is that the field affects the overall yield, and that the lineage affects the effect of watering on the yield of the crop (thus both the intercept and the slope of the linear relation). This implies to have one latent variable per field and two latent variables per lineage. Model 3.4 then becomes:

$$\begin{aligned} Z_j^1 &\overset{\text{iid}}{\sim} \mathcal{N}\big(0, \gamma_1^2\big), && 1 \le j \le J \quad \text{(Field effect)}, \\ Z_k^2 &\overset{\text{iid}}{\sim} \mathcal{N}\big(0, \gamma_2^2\big), && 1 \le k \le K \quad \text{(Lineage effect on the intercept)}, \\ Z_k^3 &\overset{\text{iid}}{\sim} \mathcal{N}\big(0, \gamma_3^2\big), && 1 \le k \le K \quad \text{(Lineage effect on the slope)} \end{aligned}$$

and

$$Y_{ijk} \mid \{\boldsymbol{Z} = \boldsymbol{z}\} \overset{\text{ind}}{\sim} \mathcal{N}\big(\beta_0 + z_j^1 + z_k^2 + (\beta_1 + z_k^3) x_{ijk}, \sigma^2\big), \qquad 1 \le j \le J, \quad 1 \le k \le K, \quad 1 \le i \le n_{jk}\,.$$

Thus, we have $d_z = J + 2K$ independent Gaussian latent variables, which play a role analogous to the β coefficients in the conditional distribution of $\boldsymbol{Y}$ given $\boldsymbol{Z}$, as they affect linearly the expectation of $\boldsymbol{Y} \mid \boldsymbol{Z}$. Suppose in general that we have d_z independent Gaussian latent variables, defined with r blocks of random variables. Each block is made of m_j i.i.d. centered Gaussian variables having the same variance γ_j^2 $(1 \le j \le r)$. In our motivating example, $r = 3$, and $m_1 = J$, $m_2 = m_3 = K$. We denote $\boldsymbol{Z}_j$ the vector of m_j i.i.d.random variables having variance γ_j^2.

This leads to the following generic formulation of the linear mixed model.

Model 3.5. *Linear mixed model (general formulation)*

$$\begin{aligned} \boldsymbol{Z} &\sim \mathcal{N}_{d_z}(0, \Gamma) \\ \boldsymbol{Y} \mid \boldsymbol{Z} &\sim \mathcal{N}_n\big(\boldsymbol{X}\beta + \boldsymbol{U}\boldsymbol{Z}, \sigma^2 \mathbf{I}_n\big)\,. \end{aligned}$$

where

Γ *is a* $d_z \times d_z$ *diagonal matrix with* r *different values* $\gamma_1^2, \dots, \gamma_r^2$, *each repeated* $m_1, \dots, m_r$ *times respectively*

$$\Gamma = diag(\gamma_1^2 \mathbf{1}_{m_1}, \dots, \gamma_r^2 \mathbf{1}_{m_r}),$$

- $\beta \in \mathbb{R}^p$ *is an unknown vector of coefficients,*
- $\boldsymbol{X}$ *(respectively* $\boldsymbol{U}$*) is a design matrix of dimension* $n \times p$ *(resp .* $n \times d_z$*) for the fixed (resp. random) effects.*

The parameters of this model are:

$$\theta = \big(\beta, \sigma^2, \gamma_1^2, \dots, \gamma_r^2\big)\,.$$

3.4.3 Complete and marginal log-likelihoods

We first observe that, under Model 3.5, the marginal distribution of the observed variables $\boldsymbol{Y}$ has an explicit form.

Proposition 3.5 (Marginal distribution of $\boldsymbol{Y}$ in linear mixed model)**.** *The marginal distribution of* $\boldsymbol{Y}$ *in Model 3.5 is given by*

$$\boldsymbol{Y} \sim \mathcal{N}_n\big(\boldsymbol{X}\beta, \sigma^2 \mathbf{I}_n + \boldsymbol{U}\Gamma\boldsymbol{U}^\top\big)\,. \tag{3.28}$$

Proof of Proposition 3.5

This is a direct application of Proposition A.6 (Appendix A.1.2) to Model 3.5.

Proposition 3.5 leads to an alternative formulation of the linear mixed model that does not rely on latent variables.

Model 3.6 (Equivalent formulation of the linear mixed model with one random effect)**.** *Using the above notations, and assuming that observations* $Y_1, \dots, Y_n$ *are ordered thanks to the statistical units (i.e., the*

first n_1 observations belong to the first statistical unit, and so on...), Model 3.5 is equivalent to:

$$\boldsymbol{Y} = \boldsymbol{X}\beta + E\,, \qquad E \sim \mathcal{N}_n(0,\Sigma)\,, \qquad \Sigma = \sigma^2\mathbf{I}_n + \boldsymbol{U}\Gamma\boldsymbol{U}^\top\,.$$

The following lemma shows the expression of covariance matrix Σ, its determinant and inverse, in the context of Model 3.4. These expressions will enable us to write the log-likelihood.

Lemma 3.1. *Under Model 3.4, using the notations of Model 3.6, we have*

$$\Sigma = \begin{pmatrix} B_{n_1} & 0 & \dots & 0 \\ 0 & B_{n_2} & \dots & \vdots \\ \vdots & \ddots & \ddots & \vdots \\ 0 & \dots & \dots & B_{n_J} \end{pmatrix} \tag{3.29}$$

where for $1 \le j \le J$, B_{n_j} is a $n_j \times n_j$:

$$B_{n_j} = \begin{pmatrix} \gamma^2+\sigma^2 & \gamma^2 & \dots & \gamma^2 \\ \gamma^2 & \gamma^2+\sigma^2 & \dots & \gamma^2 \\ \vdots & \ddots & \ddots & \vdots \\ \gamma^2 & \dots & \dots & \gamma^2+\sigma^2 \end{pmatrix}_{n_j\times n_j}.$$

Moreover, we have

$$\Sigma^{-1} = \begin{pmatrix} B_{n_1}^{-1} & 0 & \dots & 0 \\ 0 & B_{n_2}^{-1} & \dots & \vdots \\ \vdots & \ddots & \ddots & \vdots \\ 0 & \dots & \dots & B_{n_J}^{-1} \end{pmatrix},$$

where, for $1 \le j \le J$

$$B_{n_j}^{-1} = \frac{1}{\sigma^2}\mathbf{I}_{n_j} - \frac{\gamma^2}{\sigma^2(\sigma^2+n_j\gamma^2)}\mathbf{1}_{n_j\times n_j}\,, \tag{3.30}$$

where $\mathbf{1}_{n_j\times n_j}$ is a $n_j \times n_j$ matrix filled with 1. Finally, we have :

$$|\Sigma| = (\sigma^2)^{n-J} \times \prod_{j=1}^{J}(\sigma^2+n_j\gamma^2)\,, \tag{3.31}$$

Proof of Lemma 3.1

Under Model 3.4, we have $\Gamma = \gamma^2 I_J$ and U is a $n \times J$ matrix:

$$\boldsymbol{U} = \underbrace{\begin{pmatrix} 1 & 0 & \dots & 0 \\ \vdots & \vdots & \dots & \vdots \\ 1 & 0 & \dots & 0 \\ 0 & 1 & \dots & 0 \\ \vdots & \vdots & \dots & \vdots \\ 0 & 1 & \dots & 0 \\ \vdots & \vdots & \dots & \vdots \\ 0 & \vdots & \dots & 1 \\ \vdots & \vdots & \dots & \vdots \\ 0 & \vdots & \dots & 1 \end{pmatrix}}_{J} \begin{matrix} \left.\begin{matrix}\\\\\\\end{matrix}\right\} n_1 \\ \left.\begin{matrix}\\\\\\\end{matrix}\right\} n_2 \\ \left.\begin{matrix}\\\\\\\end{matrix}\right\} n_J \end{matrix} = \begin{pmatrix} \mathbf{1}_{n_1} & 0 & \dots & 0 \\ 0 & \mathbf{1}_{n_2} & \dots & \vdots \\ \vdots & & \ddots & \vdots \\ 0 & \dots & \dots & \mathbf{1}_{n_J} \end{pmatrix}.$$

So

$$\boldsymbol{U}\Gamma\boldsymbol{U}^\top = \gamma^2 \begin{pmatrix} \mathbf{1}_{n_1\times n_1} & 0 & \dots & 0 \\ 0 & \mathbf{1}_{n_2\times n_2} & \dots & \vdots \\ \vdots & \ddots & \ddots & \vdots \\ 0 & \dots & \dots & \mathbf{1}_{n_J\times n_J} \end{pmatrix},$$

which proves Equation (3.29) (after addition of $\sigma^2\mathbf{I}_n$): $\Sigma = \boldsymbol{U}\Gamma\boldsymbol{U} + \sigma^2\mathbf{I}_n$
Then, as Σ is block diagonal, it's inverse is also block diagonal. Remarking that, for $1 \leq j \leq J$, we have

$$B_{n_j} = \sigma^2 I_{n_j} + \gamma^2 \mathbf{1}_{n_j \times n_j} ,$$

we can check easily (using that $\mathbf{1}_{n_j \times n_j} \times \mathbf{1}_{n_j \times n_j} = n_j \mathbf{1}_{n_j \times n_j}$) that $B_{n_j}^{-1}$ of Equation (3.30) is indeed the inverse of B_{n_j}. Now, we have that:

$$|\Sigma| = \prod_{j=1}^{J} \left|B_{n_j}\right| . \tag{3.32}$$

Note that $\mathbf{1}_{n_j \times n_j}$ is a matrix of rank one and that the vector $\mathbf{1}_{n_j}$ is a eigenvector with eigenvalue n_j.
Then, if we consider any orthogonal basis $(\mathbf{1}_{n_j}, v_2, \ldots, v_{n_j})$ of $\mathbb{R}^{n_j}$, all these vectors are orthogonal eigenvectors of B_{n_j} and their eigenvalues are, respectively, $(\sigma^2 + n_j\gamma^2)$ for the vector $\mathbf{1}_{n_j}$ and σ^2 for the vectors $v_2, \ldots, v_{n_j}$. So, as the determinant is the product of eigenvalues, we have

$$\left|B_{n_j}\right| = (\sigma^2)^{n_j - 1}(\sigma^2 + n_j\gamma^2)$$

which we can plug in (3.32) to get

$$|\Sigma| = \prod_{j=1}^{J} \left((\sigma^2)^{n_j-1}(\sigma^2 + n_j\gamma^2)\right) = (\sigma^2)^{n-J} \times \prod_{j=1}^{J} (\sigma^2 + n_j\gamma^2) .$$

Log-likelihood. Then, thanks to Model 3.6 and Lemma 3.1, we have an explicit expression for the log-likelihood:

$$\log p_\theta(\boldsymbol{y}) = -\frac{n}{2}\log(2\pi) - \frac{1}{2}\log|\Sigma| - \frac{1}{2}\left(\boldsymbol{y} - \boldsymbol{X}\beta\right)^\top \Sigma^{-1}\left(\boldsymbol{y} - \boldsymbol{X}\beta\right) . \tag{3.33}$$

For instance, in the guiding example (Model 3.4), we can write all terms of the log-likelihood explicitly thanks to Lemma 3.1.

Complete log-likelihood. Following the latent variable approach of this book, we focus on the latent variable formulation given by Model 3.5, and we derive the quantities required for an EM algorithm. Let's start by writing the complete log-likelihood.

$$\begin{aligned}
\log p_\theta(\boldsymbol{y}, \boldsymbol{Z}) &= \sum_{j=1}^{J} \log p_\theta(\boldsymbol{y} \mid \boldsymbol{Z}) + \log p_\theta(\boldsymbol{Z}) \\
&= -\frac{n}{2}\log\sigma^2 - \frac{1}{2\sigma^2}\|\boldsymbol{y} - \boldsymbol{X}\beta - \boldsymbol{U}\boldsymbol{Z}\|^2 - \frac{1}{2}\log|\Gamma| - \frac{1}{2}\boldsymbol{Z}^\top\Gamma^{-1}\boldsymbol{Z} + \text{cst} \\
&= -\frac{n}{2}\log\sigma^2 - \frac{1}{2\sigma^2}\left(\|\boldsymbol{y} - \boldsymbol{X}\beta\|^2 - 2\boldsymbol{Z}^\top\boldsymbol{U}^\top(\boldsymbol{y} - \boldsymbol{X}\beta) + \boldsymbol{Z}^\top\boldsymbol{U}^\top\boldsymbol{U}\boldsymbol{Z}\right) - \frac{1}{2}\log|\Gamma| - \frac{1}{2}\boldsymbol{Z}^\top\Gamma^{-1}\boldsymbol{Z} + \text{cst} \\
&= -\frac{1}{2}\left(n\log\sigma^2 + \frac{1}{\sigma^2}\left(\|\boldsymbol{y} - \boldsymbol{X}\beta\|^2 - 2\boldsymbol{Z}^\top\boldsymbol{U}^\top(\boldsymbol{y} - \boldsymbol{X}\beta) + \text{tr}\left(\boldsymbol{Z}\boldsymbol{Z}^\top\boldsymbol{U}^\top\boldsymbol{U}\right)\right) + \log|\Gamma| + \text{tr}\left(\boldsymbol{Z}\boldsymbol{Z}^\top\Gamma^{-1}\right)\right) + \text{cst}.
\end{aligned} \tag{3.34}$$

The last equation derives from the fact that $\boldsymbol{Z}^\top\boldsymbol{U}^\top\boldsymbol{U}\boldsymbol{Z} \in \mathbb{R}$ so : $\boldsymbol{Z}^\top\boldsymbol{U}^\top\boldsymbol{U}\boldsymbol{Z} = \text{tr}\left(\boldsymbol{Z}^\top\boldsymbol{U}^\top\boldsymbol{U}\boldsymbol{Z}\right)$. Moreover, $\text{tr}\,(AB) = \text{tr}\,(BA)$, so $\boldsymbol{Z}^\top\boldsymbol{U}^\top\boldsymbol{U}\boldsymbol{Z} = \text{tr}\left(\boldsymbol{Z}\boldsymbol{Z}^\top\boldsymbol{U}^\top\boldsymbol{U}\right)$.

3.4.4 EM algorithm

Using the previous formulation of the complete log-likelihood, we are able to write de objective function and we obtain the EM algorithm.

Algorithm 3.5 (EM for the linear mixed model)**.** *Starting from* $\theta^{(0)} = \left(\beta^{(0)}, \sigma^{2,(0)}, \Gamma^{(0)}\right)$ *as defined in Model 3.5, repeat until convergence:*

E step. *Compute:*

$$O^{(h)} := \mathbb{V}_{\theta^{(h)}}[\boldsymbol{Z} \mid \boldsymbol{Y} = \boldsymbol{y}] = \left(\left(\Gamma^{(h)}\right)^{-1} + \frac{1}{\sigma^{2,(h)}}\boldsymbol{U}^\top\boldsymbol{U}\right)^{-1}, \tag{3.35}$$

$$m^{(h)} := \mathbb{E}_{\theta^{(h)}}[\boldsymbol{Z} \mid \boldsymbol{Y} = \boldsymbol{y}] = \frac{1}{\sigma^{2,(h)}} O^{(h)}\boldsymbol{U}^\top\left(\boldsymbol{y} - \boldsymbol{X}\beta^{(h)}\right). \tag{3.36}$$

M step. *Update the estimate of θ as*

$$\beta^{(h+1)} = \left(\boldsymbol{X}^\top\boldsymbol{X}\right)^{-1}\boldsymbol{X}^\top\left(\boldsymbol{y} - \boldsymbol{U}m^{(h)}\right)$$
$$\sigma^{2,(h+1)} = \frac{1}{n}\left(\left\|\boldsymbol{y} - \boldsymbol{X}\beta^{(h+1)}\right\|^2 - 2\left(\boldsymbol{U}m^{(h)}\right)^\top\left(\boldsymbol{y} - \boldsymbol{X}\beta^{(h+1)}\right) + tr\left(\boldsymbol{U}^\top\boldsymbol{U}\left(O^{(h)} + m^{(h)}\left(m^{(h)}\right)^\top\right)\right)\right)$$

For $1 \le j \le r$:

$$\gamma_j^{2,(h+1)} = \frac{1}{m_j} tr\left(O_j^{(h)} + m_j^{(h)}\left(m_j^{(h)}\right)^\top\right),$$

where $m_j^{(h)}$ is the block of size m_j of $m^{(h)}$ corresponding to the posterior mean of vector $\boldsymbol{Z}_j$ and $O_j^{(h)}$ is the corresponding diagonal block of size $m_j \times m_j$ of $O^{(h)}$.

Proof of Algorithm 3.5

Objective function $Q\left(\theta \mid \theta^{(h)}\right)$: We derive the objective function of the EM algorithm from the complete log-likelihood (3.34) as

$$\begin{aligned} Q\left(\theta \mid \theta^{(h)}\right) &= \mathbb{E}_{\theta^{(h)}}[\log p_\theta(\boldsymbol{y}, \boldsymbol{Z}) \mid \boldsymbol{Y} = \boldsymbol{y}] \\ &= -\frac{1}{2}\left(n\log\sigma^2 + \frac{1}{\sigma^2}\left(\|\boldsymbol{y} - \boldsymbol{X}\beta\|^2 - 2\mathbb{E}_{\theta^{(h)}}[\boldsymbol{Z} \mid \boldsymbol{Y} = \boldsymbol{y}]^\top\boldsymbol{U}^\top(\boldsymbol{y} - \boldsymbol{X}\beta)\right)\right) \\ &\quad - \frac{1}{2}\left(\operatorname{tr}\left(\mathbb{E}_{\theta^{(h)}}\left[\boldsymbol{Z}\boldsymbol{Z}^\top \mid \boldsymbol{Y} = \boldsymbol{y}\right]\boldsymbol{U}^\top\boldsymbol{U}\right) + \log|\Gamma| + \operatorname{tr}\left(\mathbb{E}_{\theta^{(h)}}\left[\boldsymbol{Z}\boldsymbol{Z}^\top \mid \boldsymbol{Y} = \boldsymbol{y}\right]\Gamma^{-1}\right)\right) + \text{cst}, \end{aligned} \tag{3.37}$$

E step: From Equation (3.37), we see that the E step requires the evaluation of $m^{(h)} = \mathbb{E}_{\theta^{(h)}}[\boldsymbol{Z} \mid \boldsymbol{Y} = \boldsymbol{y}]$ and $O^{(h)} = \mathbb{V}_{\theta^{(h)}}[\boldsymbol{Z} \mid \boldsymbol{Y} = \boldsymbol{y}]$. Proposition A.5 states that the distribution of $\boldsymbol{Z} \mid \{\boldsymbol{Y} = \boldsymbol{y}\}$ is Gaussian with variance and expectation given by Equations (3.35) and (3.36). Proposition A.1 then gives

$$\mathbb{E}_{\theta^{(h)}}\left[\boldsymbol{Z}\boldsymbol{Z}^\top \mid \boldsymbol{Y} = \boldsymbol{y}\right] = O^{(h)} + m^{(h)}(m^{(h)})^\top.$$

M step: The update formulas for the model parameters are obtained by setting the derivatives of (3.37) to zero. Note that the expression of $\beta^{(h+1)}$ is the expression of the least square estimator of β in the linear model, where the observations are here $\boldsymbol{y} - \boldsymbol{U}m^{(h)}$.

3.4.5 Confidence intervals for the fixed effects

The linear mixed model is a case where the computation of Louis's identity is feasible. We illustrate it here on the parameter β which is the parameter of interest in practice, for example to assess the significance of the effect of a covariate.

Proposition 3.6 (Asymptotic variance of $\hat{\beta}$)**.** *Let's denote $\hat{\beta}$, $\hat{\sigma}^2$, $\hat{\gamma}^2_{1:r}$, $\widehat{O}$ and $\widehat{m}$ respectively the maximum likelihood estimators of the parameters and the EM estimators for the posterior moments. Under Model 3.5, $\hat{\beta}$ is asymptotically independent of $\hat{\sigma}^2$ and $\hat{\gamma}^2_{1:r}$, and the asymptotic variance is estimated by*

$$\widehat{\mathbb{V}}[\hat{\beta}] = \widehat{\sigma}^2\left(\boldsymbol{X}^\top\boldsymbol{X} - \frac{1}{\widehat{\sigma}^2}\boldsymbol{X}^\top\boldsymbol{U}\widehat{O}\boldsymbol{U}^\top\boldsymbol{X}\right)^{-1}.$$

A reader familiar with the linear model would recognize an expression close to the one of the variance of $\hat{\beta}$ in the classical linear model. One can therefore directly identify the modification of the variance due to the inclusion of the random effects. The proof of this Proposition is postponed to Appendix B.1.

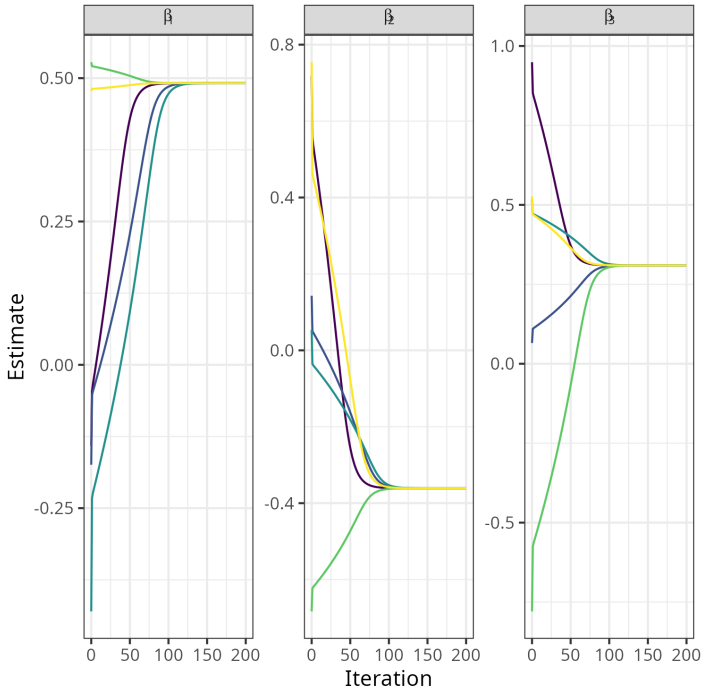


Figure 3.16: Ammonium dataset (Example 3.4). Evolution of the estimates of β along EM iteration for the linear mixed model (Model 3.4). The five colours correspond to five different starting points.

3.4.6 Illustration on the concentration of ammonium on Borneo soil

We illustrate Algorithm 3.5 on the Dataset 3.4. As indicated in Section 3.4.2.1, the fixed effect is given by a qualitative variable (the soil type). As classically done in this context, the parameter $\beta = (\beta_1, \beta_2, \beta_3)$ gathers respectively the expected value of the $\log_{10}\mathrm{NH}_4$ concentration on the reference soil (which is the alluvial). Thus, the expected $\log_{10}\mathrm{NH}_4$ value is β_1 for the alluvial soil, $\beta_1 + \beta_2$ for the heath soil, and $\beta_1 + \beta_3$ for the sandstone soil. Figure 3.16 displays the evolution of the estimates for these three parameters along the EM algorithm, starting from five different initial values. One can see that all trajectories converges toward the same point. Table 3.4 gives the estimates for the five parameters of the model. The estimated covariance matrix for $\hat{\beta}$ is

$$\widehat{\mathbb{V}}\big[\hat{\beta}\big] = \begin{pmatrix} 0.009 & -0.009 & -0.009 \\ -0.009 & 0.018 & 0.009 \\ -0.009 & 0.009 & 0.018 \end{pmatrix}$$

Using these results, we display on Table 3.5 the 99%confidence intervals for the expected log concentration of ammonium in each soil type. We can conclude that the sandstone is significantly (at level 1%) richier in ammonium than the heath soil.

An interest in using the EM algorithm is to check the estimated distribution of the random effects, which here model a spatial effect. Figure 3.17 displays the estimated conditional densities of $Z_j \mid \{\boldsymbol{Y} = \boldsymbol{y}\}$ for the $1 \leq j \leq 9$, corresponding to the 9 plots of the experiment. One can see that the conditional means range from (-0.3) to 0.3, values which can be compared to the estimates of Table 3.5. The conditional standard deviation is the same for all plots as the design of experiments was balanced (20 observations per plot), and is rather small (about 0.05). The inspection of such effect could lead to further investigation to check if these effects are structured spatially.

Finally, we look at the percentage of the variability in ammonium that is explained by the spatial effect (the plot). The residual variance σ^2, estimated at 0.025 (Table 3.4) and the variance due to the plot, γ^2, estimated here at 0.055. Therefore, one can see that the plot is responsible for $0.025/(0.025 + 0.055) \approx 31\%$ of the variability in this dataset.

3.4.7 Conclusion

Linear mixed-effects models are particularly useful in ecology, where data often exhibit hierarchical or nested structures—such as measurements taken across different sites, or individual organisms—making it essential to account for both fixed and random sources of variation. Although one could perform direct likelihood optimization (using the explicit form of the likelihood in Equation (3.33)), using the EM algorithm to estimate parameters in

Parameter	Estimate
β_1	0.491
β_2	-0.361
β_3	0.309
σ^2	0.055
γ^2	0.025

Table 3.4: Ammonium dataset (Example 3.4). Maximum likelihood estimates for parameters of Model ref-mod:lmm). β_1 is the mean $\log_{10}$ concentration of NH_4 on alluvial soils, while β_2 and β_3 are the modification to this reference value due to the heath soil and sandstone soil respectively.

Soil	Expected $\log_{10}NH_4$	99% confidence interval
Alluvial	0.491	[0.246, 0.737]
Heath	0.130	[-0.115, 0.376]
Sandstone	0.801	[0.555, 1.046]

Table 3.5: Ammonium dataset (Example 3.4). Asymptotic 99% confidence intervals for the estimated $\log_{10}(NH_4)$ for each soil type. At risk 1%, we conclude that the sandstone soil has a significantly higher concentration of ammonium than the heath soil.

linear mixed-effects models is often advantageous, as it naturally accommodates the unobserved random effects by iteratively estimating their expected values and updating model parameters, leading to improved numerical stability and convergence in complex or high-dimensional settings.

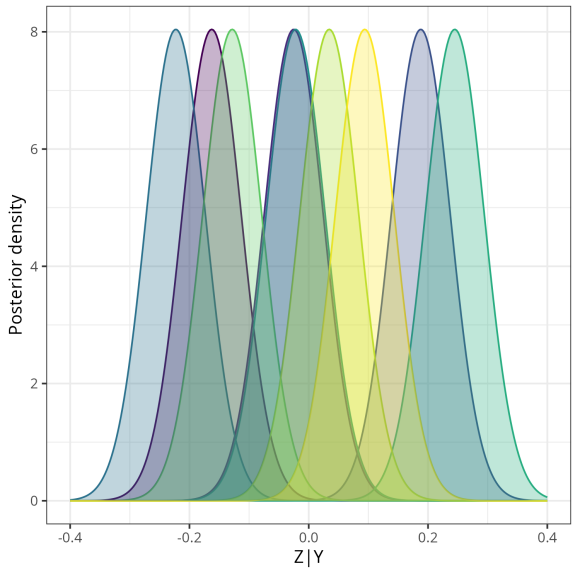


Figure 3.17: Ammonium dataset (Example 3.4). Estimation of the conditional density of the latent variables in the linear mixed model. Each density (color) correspond to a plot in the experiment (and thus to a spatial effect). Positive values correspond to a plot where the ammonium concentration is higher than expected due to the sole effect of the soil.

3.5 Probabilistic principal component analysis

Dimension reduction is a classical task in data analysis and visualization, and Principal Component Analysis (PCA) is one of the most prominent methods for achieving it. Although often approached from a purely operational perspective, dimension reduction implicitly relies on a latent variable representation. It assumes that the observed data have an overly high dimensionality, and that most of the information or variability can be captured by a small set of underlying (latent) variables. In this section, we revisit PCA from the perspective of latent variable modeling."

3.5.1 Data and question

We illustrate the model, its interest and its inference on an ecological dataset gathering morphological traits on fish species in the Gulf of Mexico Villéger et al. [2012].

Dataset 3.5 (Fish species in the Gulf of Mexico Villéger et al. [2012])**.** *Villéger et al. [2012] measured* $p = 16$ *continuous traits among* $n = 47$ *fish species living in the Terminos Lagoon (Gulf of Mexico). Table 3.6 gives the list of the morphological traits; The data are available from the CESTES database [Jeliazkov et al., 2020]. The distribution of the traits among the species displayed in Figure 3.18 shows correlations between the traits, indicating redundancies that may result from few underlying latent morphological characteristics. Because the different traits are measured with different unit, the dataset is classically scaled so to set the variance of each trait to 1.*

logM	Mass	OgSf	Oral gape surface	OgSh	Oral gape shape
OgPo	Oral gape position	GrLg	Gill raker length	GtLg	Gut length
EySz	Eye size	EyPo	Eye position	BdSh	Body transversal shape
BdSf	Body transversal surface	PfPo	Pectoral fin position	PfSh	Pectoral fin aspect ratio
CpHt	Caudal peduncle throttling	CfSh	Caudal fin aspect ratio	FsRt	Fins surface ratio
FsSf	Ratio Fins surface/body size				

Table 3.6: Fish species in the Gulf of Mexico (Example 3.5). Traits recorded by Villéger et al. [2012].

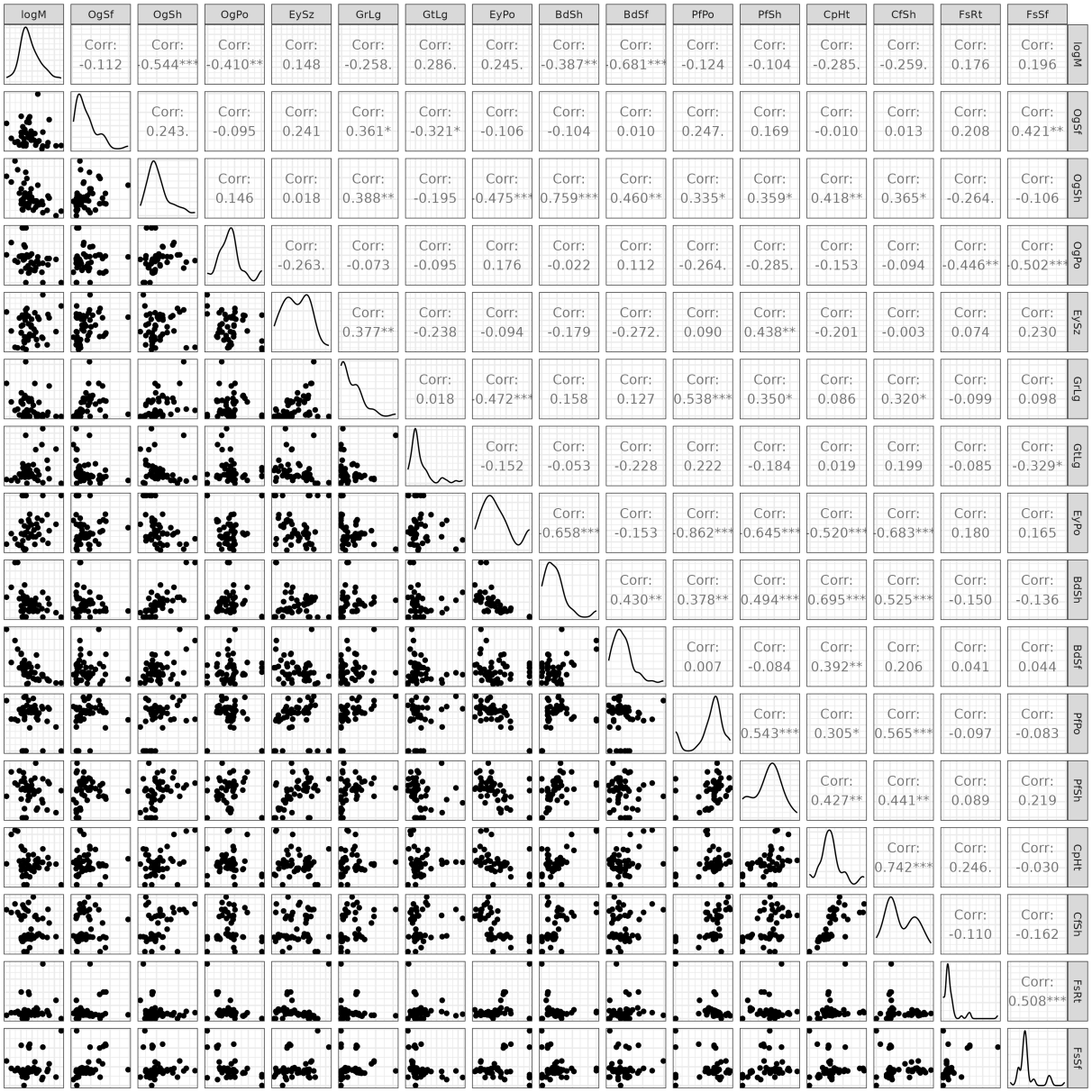


Figure 3.18: Fish species in the Gulf of Mexico (Example 3.5). Distribution of $p = 16$ traits for $n = 47$ fish species from the Terminos Lagoon.

3.5.2 Probabilistic principal component analysis model

The data are a collection of n vector of morphological traits y_i of length p. Rephrased in terms of a latent variable model, dimension reduction amounts at supposing that the observed p variables are actually linear combinations of $q < p$ latent variables, up to some random noise. More specifically to each species i ($1 \le i \le 47$), we associate a q-dimensional random vector Z_i with zero mean and identity variance matrix. y_i is the realisation of Y_i which is assumed to be is the sum of a p-dimensional mean vector μ (set to 0 when the data is centered), plus p linear combinations BZ_i, B being a $p \times q$ matrix $B = [b_{j\ell}]_{1\le j\le p, 1\le \ell\le q}$, plus a noise random vector E_i, with variance $\sigma^2 I_p$. As this formulation is based on a full probabilistic construction, this model is called probabilistic PCA [pPCA Tipping and Bishop, 1999].

Model 3.7 (Probabilistic Principal Component Analysis)**.**

$$Z_i \overset{iid}{\sim} \mathcal{N}_q(0, \mathbf{I}_q), \qquad 1 \le i \le n\,,$$
$$Y_i \mid \{Z_i = z_i\} \overset{ind}{\sim} \mathcal{N}_p\big(\mu + Bz_i, \sigma^2 \mathbf{I}_p\big) \qquad 1 \le i \le n\,.$$

The set of parameters of the pPCA model is hence

$$\theta = (\mu, B, \sigma^2).$$

In the sequel, we denote by $\boldsymbol{Y} = [Y_1^\top \dots Y_n^\top]^\top$ the $n \times p$ observed variable matrix, by $\boldsymbol{y}$ its realisation (the data matrix), and by $\boldsymbol{Z} = [Z_1^\top \dots Z_n^\top]^\top$ the $n \times q$ latent variable matrix.

Remarks.

- *It would be easy to account for available covariates gathered in a vector x_i by replacing the mean vector μ by an observation specific vector $\mu_i = [\mu_{ij}]_{1 \le j \le p}$, taking $\mu_{ij} = x_i^\top \beta_j$ where β_j is the set of regression coefficients for the jth trait in Y_i. For the sake of clarity, we do not develop this model here, but the adaptation is straightforward.*
- *One can notice a resemblance between Model 3.7 and Model 3.5. Indeed, probabilistic PCA can be seen as a generalization of the linear mixed Model 3.5 in two ways:*
  1. *the response variable $\boldsymbol{Y}$ is multivariate,*
  2. *more importantly, the equivalent of the design matrix $\boldsymbol{U}$ in Model 3.5 is B in Model 3.7, but it is unknown.*

Graphical model. The graphical model is the same a this of the mixture Model 3.1, given in Figure 3.2. The random couples $\{(Y_i, Z_i)\}_{1 \le i \le n}$ are independent and identically distributed, as well as the observed vectors $\{Y_i\}_{1 \le i \le n}$.

3.5.3 Complete and marginal log-likelihood

Marginal, joint and conditional distributions. As for any $1 \le i \le n$, the conditional distribution of Y_i given Z_i under Model 3.7 is Gaussian, and depends linearly on Z_i, the marginal distribution of each Y_i is (by Proposition A.6 also Gaussian, and we have:

$$Y_i \overset{\text{iid}}{\sim} \mathcal{N}_p(\mu, \Sigma) \qquad \text{where} \quad \Sigma = BB^\top + \sigma^2 \mathbf{I}_p. \tag{3.38}$$

The joint distribution of each couple (Y_i, Z_i) is a multivariate Gaussian (Proposition A.4):

$$\begin{bmatrix} Y_i \\ Z_i \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \mu \\ 0_q \end{bmatrix}, \begin{bmatrix} \Sigma & B \\ B^\top & \mathbf{I}_q \end{bmatrix} \right). \tag{3.39}$$

Using Proposition A.3 (Appendix A.1.2), the conditional distribution of Z_i given Y_i is

$$Z_i \mid \{Y_i = y_i\} \sim \mathcal{N}\big(B^\top \Sigma^{-1}(y_i - \mu), \mathbf{I}_q - B^\top \Sigma^{-1} B\big). \tag{3.40}$$

Log-likelihood. Because Model (3.7) assumes that the observations (species) are independent, using Equation 3.38, the log-likelihood of the observations writes

$$\log p_\theta(\boldsymbol{y}) = \sum_{i=1}^n \log p_\theta(Y_i) = -\frac{n}{2} \log |\Sigma| - \frac{1}{2} \sum_{i=1}^n \|Y_i - \mu\|^2_{\Sigma^{-1}} - \frac{np}{2} \log 2\pi. \tag{3.41}$$

Complete log-likelihood. The complete log-likelihood derives from the same independence structure and from the conditional distribution of $\boldsymbol{Y} \mid \boldsymbol{Z}$ given by the Model. We get

$$\begin{aligned} \log p_\theta(\boldsymbol{y}, \boldsymbol{Z}) &= \log p_\theta(\boldsymbol{Z}) + \log p_\theta(\boldsymbol{y} \mid \boldsymbol{Z}) \\ &= -\frac{1}{2} \sum_{i=1}^n \|Z_i\|^2 - \frac{np}{2} \log \sigma^2 - \frac{1}{2\sigma^2} \sum_{i=1}^n \|Y_i - \mu - BZ_i\|^2 - \frac{n(p+q)}{2} \log 2\pi. \end{aligned} \tag{3.42}$$

Identifiability. As in the standard geometric PCA, the matrix B is actually given up to a rotation in $\mathbb{R}^q$. Indeed, consider any $q \times q$ orthonormal matrix O (i.e. $O^\top O = I_q$), and define $C = BO$, we have that

$$CC^\top + \sigma^2 \mathbf{I}_p = BOO^\top B^\top + \sigma^2 \mathbf{I}_p = BB^\top + \sigma^2 \mathbf{I}_p = \Sigma,$$

which means that the matrix B can be replaced by the matrix C in Model (3.7), without changing the distribution (3.38) of the observed data $p_\theta(\boldsymbol{y})$. This implies that, without additional constraints, B can only be identified up to an arbitrary rotation.

3.5.4 EM algorithm

Explicit maximum likelihood estimates. As already seen before (for instance in Section 3.4), not all latent variable models require to resort to the EM algorithm. One may easily check that the observed likelihood $\log p_\theta(\boldsymbol{y})$ given in (3.41) is maximal when μ equals the vector sample of mean of $\boldsymbol{y}$:

$$\widehat{\mu} = \frac{1}{n}\boldsymbol{y}^\top \mathbf{1}_n,$$

independently from the variance matrix Σ. The maximization of $\log p_\theta(\boldsymbol{y})$ with respect to B and σ^2 is a bit more involved but turns out to be similar to standard PCA. Namely, we denote by S the empirical covariance matrix of $\boldsymbol{y}$:

$$S = \frac{1}{n}(\boldsymbol{y} - \mathbf{1}_n\widehat{\mu}^\top)^\top(\boldsymbol{y} - \mathbf{1}_n\widehat{\mu}^\top)$$

and by $\lambda = [\lambda_1 \dots \lambda_p]$ the vector of its ordered eigenvalues ($\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p$) and by P the matrix of its eigenvectors, so that $S = P\Lambda P^\top$, with $\Lambda = \mathrm{diag}(\lambda)$. Tipping and Bishop [1999] (Appendix A) show that, for a given dimension q, $\log p_\theta(Y)$ is maximal for

$$\widehat{\sigma}_q^2 = \frac{1}{p-q}\sum_{\ell=q+1}^{p}\lambda_\ell, \qquad \widehat{B}_q = [P_1 \dots P_q]\,\mathrm{diag}(\gamma_{1:q}), \qquad \text{where} \quad \gamma_{1:q} = [\lambda_\ell - \widehat{\sigma}_q^2]_{1\leq\ell\leq q} \tag{3.43}$$

where $P_1, \dots P_q$ are the first q eigenvectors of S (up to any orthonormal rotation). As a consequence, the reconstructed variance matrix $\widehat{\Sigma}_q = \widehat{B}_q\widehat{B}_q^\top + \widehat{\sigma}_q^2\mathbf{I}_p$ has the same first q eigenvalues and eigenvectors as the empirical covariance matrix S, and its last $p-q$ eigenvalues are set to the mean of the last $p-q$ eigenvalues of S.

EM algorithm in the presence of missing data The direct estimation procedure described above is computationally efficient, but does not apply in presence of missing data, namely when for some observations, not all the variables where measured. In this case, the likelihood $\log p_\theta(Y)$ does not have the simple form given in Equation (3.41) any more [see also Josse et al., 2011]. Indeed, denoting $\mathcal{O}_i \subseteq \{1, \dots p\}$ the set of variables measured on species i (the other being missing), the observed vector Y_i is reduced to its components in $\mathcal{O}_i$: $Y_{i,\mathcal{O}_i} = [Y_{ij}]_{j\in\mathcal{O}_i}$. Defining $\mu_{\mathcal{O}_i}$, $B_{\mathcal{O}_i}$ and $\Sigma_{\mathcal{O}_i}$ in the same way, that is:

$$\mu_{\mathcal{O}_i} = \{\mu_j\}_{j\in\mathcal{O}_i}, \qquad B_{\mathcal{O}_i} = \{b_{j\ell}\}_{j\in\mathcal{O}_i, 1\leq\ell\leq q}, \qquad \Sigma_{\mathcal{O}_i} = (B_{\mathcal{O}_i})^\top B_{\mathcal{O}_i} + \sigma^2\mathbf{I}_{\#(\mathcal{O}_i)},$$

the joint distribution of $(Y_{i,\mathcal{O}_i}, Z_i)$ becomes

$$\begin{bmatrix} Y_{i,\mathcal{O}_i} \\ Z_i \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \mu_{\mathcal{O}_i} \\ 0_q \end{bmatrix}, \begin{bmatrix} \Sigma_{\mathcal{O}_i} & B_{\mathcal{O}_i} \\ (B_{\mathcal{O}_i})^\top & \mathbf{I}_q \end{bmatrix}\right)$$

and the conditional distribution of Z_i given the observed (incomplete) $Y_{i,\mathcal{O}_i}$ can still be defined using Proposition A.3 from Appendix A.1.2.

Obviously, the way missing data are spread among the (fictitious) complete $n \times p$ dataset can affect the inference. Intuitively, the identification of the parameter $\theta = (\mu, B)$ for any $1 \leq q \leq p$ requires that each traits j is observed for at least two species i and i' and that each pair of traits (j,k) is observed at least for one species i. In addition, a systematic censoring of certain observations (e.g. always missing small values) may clearly bias the inference. The missing at random assumption [see Rubin, 1976] is sufficient to avoid such biases.

Algorithm 3.6 (EM for the probabilistic PCA)**.** *Starting from* $\theta^{(0)} = \left(\mu^{(0)}, B^{(0)}, (\sigma^2)^{(0)}\right)$ *as defined in Model 3.5, repeat until convergence:*

E step. *Compute:*

$$m_i^{(h)} := \mathbb{E}_{\theta^{(h)}}[Z_i \mid Y_{i,\mathcal{O}_i} = y_{i,\mathcal{O}_i}] = \left(B_{\mathcal{O}_i}^{(h)}\right)^\top \left(\Sigma_{\mathcal{O}_i}^{(h)}\right)^{-1} \left(y_{i,\mathcal{O}_i} - \mu_{\mathcal{O}_i}^{(h)}\right), \tag{3.44}$$

$$Q_i^{(h)} := \mathbb{V}_{\theta^{(h)}}[Z_i \mid Y_{i,\mathcal{O}_i} = y_{i,\mathcal{O}_i}] = I_q - \left(B_{\mathcal{O}_i}^{(h)}\right)^\top \left(\Sigma_{\mathcal{O}_i}^{(h)}\right)^{-1} B_{\mathcal{O}_i}^{(h)}. \tag{3.45}$$

M step. *Update the estimate of θ, setting, for all $1 \le j \le p$*

$$\mu_j^{(h+1)} = \frac{1}{\#(\mathcal{A}_j)} \sum_{i \in \mathcal{A}_j} \left(y_{ij} - B_{j,\mathcal{O}_i}^{(h)} m_i^{(h)} \right), \tag{3.46}$$

$$B_j^{(h+1)} = \left(\sum_{i \in \mathcal{A}_j} (y_{ij} - \mu_j^{(h+1)}) \left(m_i^{(h)} \right)^\top \right) \left(\sum_{i \in \mathcal{A}_i} m_i^{(h)} \left(m_i^{(h)} \right)^\top + Q_i^{(h)} \right)^{-1}, \tag{3.47}$$

$$(\sigma^2)^{(h+1)} = \frac{1}{\#(\mathcal{O})} \sum_{i=1}^{n} \left(\left\| y_{i,\mathcal{O}_i} - \mu_{\mathcal{O}_i}^{(h+1)} - B_{\mathcal{O}_i}^{(h+1)} m_i^{(h)} \right\|^2 + \operatorname{tr}\left(\left(B_{\mathcal{O}_i}^{(h+1)} \right)^\top B_{\mathcal{O}_i}^{(h+1)} Q_i^{(h)} \right) \right), \tag{3.48}$$

where B_j is the jth row of the matrix B, $B_{j,\mathcal{O}_i} = [b_{j\ell}]_{\ell \in \mathcal{O}_i}$ and $\mathcal{A}_j$ is the set of observations for which the jth trait has been recorded: $\mathcal{A}_j = \{i : j \in \mathcal{O}_i\}$.

Proof of Algorithm 3.6

E step. The complete log-likelihood (3.42) can be rewritten as

$$\log p_\theta(\boldsymbol{y}_\mathcal{O}, \boldsymbol{Z}) = -\frac{1}{2} \sum_{i=1}^{n} \|Z_i\|^2 - \frac{1}{2\sigma^2} \sum_{i=1}^{n} \left(-2(y_{i,\mathcal{O}_i} - \mu_{\mathcal{O}_i})^\top B_{\mathcal{O}_i} Z_i + \|B_{\mathcal{O}_i} Z_i\|^2 \right) + \text{cst}, \tag{3.49}$$

where the constant does not depend on $\boldsymbol{Z}$. As the couples $\{(Y_{i,\mathcal{O}_i}, Z_i)\}$ are all independent, the conditional expectation of any function of Z_i given $\boldsymbol{Y}_\mathcal{O} = \boldsymbol{y}_\mathcal{O}$ is the same as this given $Y_{i,\mathcal{O}_i}$. This indicates us which conditional moments of each Z_i we will need to evaluate the objective function $Q(\theta \mid \theta^{(h)})$ at the E step h of Algorithm 2.1. More specifically, given the current estimate $\theta^{(h)}$, we first must evaluate

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

We find the proposed expression in (3.44) as the direct restriction to $\mathcal{O}_i$ of Equation (3.40). The evaluation of $Q(\theta \mid \theta^{(h)})$ also requires to compute $\mathbb{E}_{\theta^{(h)}}\left[\|Z_i\|^2 \mid \boldsymbol{Y}_\mathcal{O} = \boldsymbol{y}_\mathcal{O} \right]$ and $\mathbb{E}_{\theta^{(h)}}\left[\|BZ_i\|^2 \mid \boldsymbol{Y}_\mathcal{O} = \boldsymbol{y}_\mathcal{O} \right]$. Now if we define $Q_i^{(h)}$ as in (3.45) (again, its expression is given by the restriction of (3.40) to the observed data), we have, by Proposition A.1 (Appendix A.1.1)

$$\begin{aligned}
\mathbb{E}_{\theta^{(h)}}\left[\|Z_i\|^2 \mid Y_{i,\mathcal{O}_i} = y_{i,\mathcal{O}_i} \right] &= \left\| m_i^{(h)} \right\|^2 + \operatorname{tr}(Q_i^{(h)}) \\
\mathbb{E}_{\theta^{(h)}}\left[\|BZ_i\|^2 \mid Y_{i,\mathcal{O}_i} = y_{i,\mathcal{O}_i} \right] &= \left\| B m_i^{(h)} \right\|^2 + \operatorname{tr}\left(B^\top B Q_i^{(h)} \right).
\end{aligned}$$

M step. The objective function $Q(\theta \mid \theta^{(h)})$ can be written in different ways to get the update formulas of μ, B and σ^2.

$$\begin{aligned}
Q(\theta \mid \theta^{(h)}) &= \frac{1}{2\sigma^2} \left(\sum_{i=1}^{n} \|\mu_{\mathcal{O}_i}\|^2 - \left(y_{i,\mathcal{O}_i} - B_{\mathcal{O}_i} m_i^{(h)} \right)^\top \mu \right) + \text{cst}(\mu) \\
&= \frac{1}{2\sigma^2} \sum_{i=1}^{n} \left(\left\| y_{i,\mathcal{O}_i} - \mu - B_{\mathcal{O}_i} m_i^{(h)} \right\|^2 + \operatorname{tr}\left((B_{\mathcal{O}_i})^\top B_{\mathcal{O}_i} Q_i^{(h)} \right) \right) - \frac{\#(\mathcal{O})}{2} \log \sigma^2 + \text{cst}(\sigma^2) \\
&= \frac{1}{2\sigma^2} \sum_{i=1}^{n} \left(\left\| B_{\mathcal{O}_i} m_i^{(h)} \right\|^2 + \operatorname{tr}\left((B_{\mathcal{O}_i})^\top B_{\mathcal{O}_i} Q_i^{(h)} \right) - 2(y_{i,\mathcal{O}_i} - \mu_{\mathcal{O}_i})^\top B_{\mathcal{O}_i} m_i^{(h)} \right) + \text{cst}(B)
\end{aligned}$$

where $\text{cst}(\mu)$ means that the remaining term does not depend on μ (resp. B and σ^2). The update formulas are obtained by setting the derivatives of $Q(\theta \mid \theta^{(h)})$ with respect to each parameter to zero. We refer the reader to [Tipping and Bishop, 1999, Appendix B] for the detailed computations that lead to (3.46)-(3.48).

3.5.5 Choosing the dimension of the latent space

A major advantage of the probabilistic version of PCA over standard PCA is that likelihood theory provides access to classical model selection criteria, such as AIC or BIC (see Section 2.4), for selecting the number of latent dimensions q. In the absence of a probabilistic framework, the latent dimension q q is often determined by visually inspecting the spectrum of the empirical covariance matrix.

Since the eigenvectors are orthonormal, the number of independent parameters of Model 3.7 for q latent dimensions is

$$D_q = \underbrace{1}_{\sigma^2} + \underbrace{p}_{\mu} + \underbrace{q}_{\lambda_1,\dots,\lambda_q} + \underbrace{(p-1)+(p-2)+\dots(p-q)}_{P_1,\dots P_q} = 1 + p + q(2p-q+1)/2\,.$$

Whenever the data set has been scaled to set each variance to 1, the sum of the eigenvalues is constraint to be q and the number of independent parameters is decreased by 1: $D_q = p + q(2p-q+1)/2$. The AIC and BIC are then defined as in Chapter 2.

3.5.6 Visualization: shrinkage effect

PCA is most often used for visualization, each observation being associated with a point in a small dimension (typically two). Equation (3.40) provides a prediction of each latent vector Z_i as we may evaluate[5]

$$\begin{aligned}\widehat{m}_i := \mathbb{E}_{\widehat{\theta}}[Z_i \mid Y_i] &= \widehat{B}^\top \widehat{\Sigma}^{-1}(Y_i - \widehat{\mu}) = \operatorname{diag}(\gamma_{1:q})[P_1 \dots P_q]^\top P \operatorname{diag}(\lambda)^{-1} P^\top (Y_i - \widehat{\mu}) \\ &= \operatorname{diag}\left(\left[\frac{\sqrt{\lambda_\ell - \widehat{\sigma}^2}}{\lambda_\ell}\right]_{1\le\ell\le q}\right)[P_1 \dots P_q]^\top (Y_i - \widehat{\mu}),\end{aligned}$$

which is the probabilistic counterpart of the normalized PCA scores

$$\widetilde{Z}_i = \left(\operatorname{diag}([1/\sqrt{\lambda_\ell}]_{1\le\ell\le q}\right)[P_1 \dots P_q]^\top (Y_i - \widehat{\mu}).$$

The comparison of the two shows the so-called *shrinkage* effect induced by the probabilistic modelling. Indeed, for each observation $1 \le i \le n$ and latent dimension $1 \le \ell \le q$, we have that

$$\widehat{m}_{ij}/\widetilde{Z}_{ij} = \sqrt{(\lambda_\ell - \widehat{\sigma}^2)/\lambda_\ell} \quad < 1,$$

which means that the coordinates of $\widehat{m}_i$ are each pushed toward zero, as compared to this of the PCA score vector $\widetilde{Z}_i$. This shrinkage is more severe when λ_ℓ gets smaller, that is when ℓ increases.

3.5.7 Data analysis by PCA

We now apply the Probabilistic PCA analysis to Dataset 3.5 described at the beginning of this section.

Choosing the best latent space dimension The left panel of Figure 3.19 displays the log-likelihood and the AIC and BIC criteria for Example 3.5. While BIC select $q_{BIC} = 4$ latent dimension, AIC does not suggest any significant dimension reduction ($q = 14$), which is consistent with the last remark of Section 2.4.
The middle and right panels of Figure 3.19 compare the $p = 16$ (cumulated) eigenvalues of the empirical matrix S with the (cumulated) eigenvalues of $\widehat{\Sigma}_q$ for $q = 4$ and $q = 14$ dimensions, and illustrates the averaging of the last $p - q$ eigenvalues of S. Because the number of latent dimensions selected by AIC ($q = 14$) is close to the total number of traits ($p = 16$), the corresponding eigenvalues are almost the same as these obtained with the classical PCA.

In the absence of a probabilistic framework, the latent dimension q would have been determined by visually inspecting the spectrum of the empirical covariance matrix, that is, by examining the center or right panel of Figure 3.19. In Example 3.5, the eigenvalues decrease rather smoothly, making this choice far from obvious, especially when considering the cumulative eigenvalues (right panel).

Shrinkage effect When considering $q = q_{BIC}$ latent dimensions for the traits from Example 3.5, the estimated residual variance is $\widehat{\sigma}^2 = 0.378$ and the first (square-rooted) eigenvalues $\sqrt{\lambda_\ell}$, corresponding γ_ℓ and shrinkage coefficients are:

ℓ	1	2	3	4
$\sqrt{\lambda_\ell}$	2.197	1.669	1.455	1.319
γ_ℓ	2.109	1.552	1.319	1.168
$\sqrt{(\lambda_\ell - \widehat{\sigma}^2)/\lambda_\ell}$	0.960	0.930	0.906	0.885

Figure 3.20 illustrates the shrinkage effect. The red dots resulting from the pPCA inference ($\widehat{m}_i$) are closer to zero than the corresponding scores $\widetilde{Z}_i$. The shrinkage is weaker for the first latent dimensions (1 and 2) than for the last ones (3 and 4).

[5] In the following of this section, we dropped the subscript q from estimators of Equation (3.43), for sake of readibility.

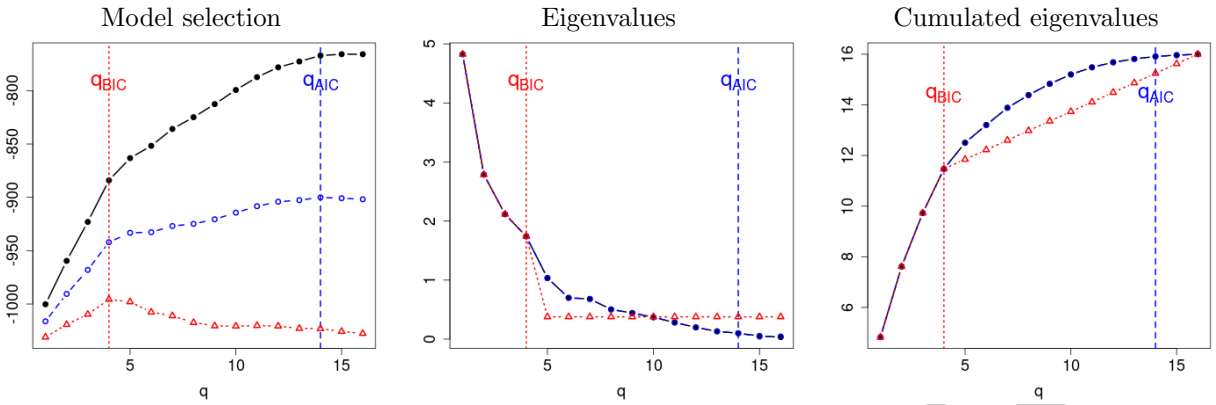


Figure 3.19: Fish species in the Gulf of Mexico (Example 3.5). Left: Log-likelihood (black solid line and dots —•) for Example 3.5 and AIC (blue dashed line and circles - - ∘) end BIC(red dotted line and triangles ···△) criteria, as a function of the latent dimension q. Center: Eigenvalues of the empirical covariance matrix of the same example (black solid line and dots —•) and of its reconstruction with $q = 4$ (red dotted line and triangles ···△) and $q = 14$ (blue dashed line and circles - -∘) latent dimensions. Right: Cumulated eigenvalues of the empirical covariance matrix. Same legend as center panel.

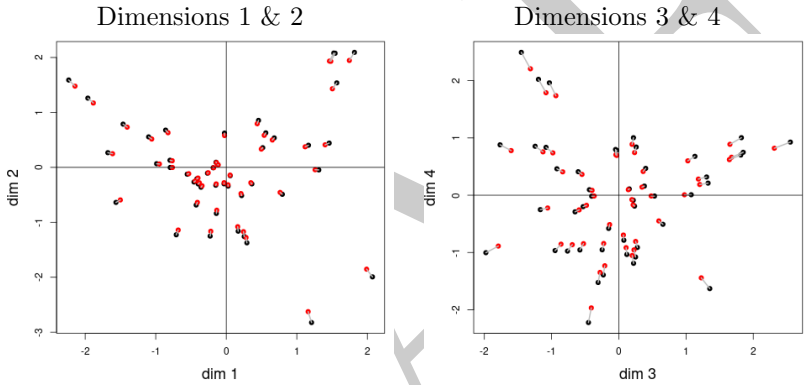


Figure 3.20: Fish species in the Gulf of Mexico (Example 3.5). Illustration of the shrinkage effect. Dots correspond to each of the $n = 47$ species. Black dots = PCA score $\widehat{Z}_{i\ell}$, red dots = conditional expectation of the latent variable $\widehat{m}_i = \mathbb{E}_{\widehat{\theta}}[Z_i \mid Y_i]$ under the pPCA Model 3.7. Left: latent dimensions 1 & 2. Right: latent dimensions 3 & 4.

3.5.8 Imputation of missing data

Another interesting feature of the probabilistic version of PCA is that it provides a prediction for the missing observations. We denote by $\mathcal{M}_i = \{1, \dots p\} \setminus \mathcal{O}_i$ the set of missing variables for observation i and by $Y_{i,\mathcal{M}_i} = \{Y_{ij}\}_{j \in \mathcal{M}_i}$ and by $\boldsymbol{Y}_{\mathcal{M}} = \{Y_{i,\mathcal{M}_i}\}_{1 \le i \le n}$ the whole set of missing data. Taking advantage of the dependence between $Y_{i,\mathcal{O}_i}$ and $Y_{i,\mathcal{M}_i}$ induced by the common latent variable Z_i, $Y_{i,\mathcal{M}_i}$ can be *imputed*, that is predicted conditionally on $Y_{i,\mathcal{O}_i}$. The graphical model of Model (3.7) is then given by Figure (3.21).

Observe that all the triplets $(Z_i, Y_{i,\mathcal{O}_i}, Y_{i,\mathcal{M}_i})$ are independent and that, because $Y_{\mathcal{O}_i}$ and $Y_{i,\mathcal{M}_i}$ are independent given Z_i, the complete likelihood of the 'full' data set $\boldsymbol{Y} = (\boldsymbol{Y}_{\mathcal{O}}, \boldsymbol{Y}_{\mathcal{M}})$ is

$$\log p_\theta(\boldsymbol{Y}, \boldsymbol{Z}) = \log p_\theta(\boldsymbol{Y}_{\mathcal{O}}, \boldsymbol{Y}_{\mathcal{M}}, \boldsymbol{Z}) = \log p_\theta(\boldsymbol{Y}_{\mathcal{O}}, \boldsymbol{Z}) + \log p_\theta(\boldsymbol{Y}_{\mathcal{M}} \mid \boldsymbol{Z})$$

where $\log p_\theta(\boldsymbol{y}_{\mathcal{O}}, \boldsymbol{Z})$ is given by Equation 3.49 and where the last term $\log p_\theta(\boldsymbol{Y}_{\mathcal{M}} \mid \boldsymbol{Z})$ only involves unobserved variable, so that it does no contribute to the inference.

Now we want to predict the missing observations by $\widehat{Y}_{i,\mathcal{M}_i} := \mathbb{E}_{\widehat{\theta}}[Y_{i,\mathcal{M}_i} \mid Y_{i,\mathcal{O}_i} = y_{i,\mathcal{O}_i}]$. To do so, we need to express the joint distribution of $()Y_{i,\mathcal{O}_i}, Y_{i,\mathcal{M}_i}, Z_i)$ according to Model 3.7. We may define $\mu_{\mathcal{M}_i}$, $B_{\mathcal{M}_i}$ and $\Sigma_{\mathcal{M}_i}$ in the same way as $\mu_{\mathcal{O}_i}$, $B_{\mathcal{O}_i}$ and $\Sigma_{\mathcal{O}_i}$. We also need to define $\Sigma^{\mathcal{O}_i,\mathcal{M}_i} = [\sigma_{jk}]_{j \in \mathcal{O}_i, k \in \mathcal{M}}$ and $\Sigma_{\mathcal{M}_i,\mathcal{O}_i} = (\Sigma^{\mathcal{O}_i,\mathcal{M}_i})^\top$.

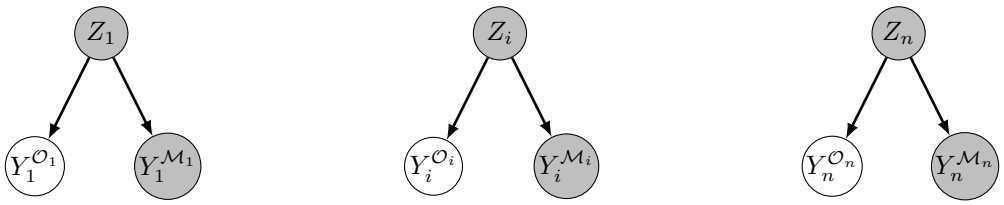


Figure 3.21: Graphical representation of the probabilistic PCA Model 3.7 with missing observations).

The joint distribution of $(Y_{i,\mathcal{O}_i}, Y_{i,\mathcal{M}_i}, Z_i)$ according to Model 3.7 is

$$\begin{bmatrix} Y_{i,\mathcal{O}_i} \\ Y_{i,\mathcal{M}_i} \\ Z_i \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \mu_{\mathcal{O}_i} \\ \mu_{\mathcal{M}_i} \\ 0_q \end{bmatrix}, \begin{bmatrix} \Sigma_{\mathcal{O}_i} & \Sigma^{\mathcal{O}_i,\mathcal{M}_i} & B_{\mathcal{O}_i} \\ \Sigma_{\mathcal{M}_i,\mathcal{O}_i} & \Sigma_{\mathcal{M}_i} & B_{\mathcal{M}_i} \\ (B_{\mathcal{O}_i})^\top & (B_{\mathcal{M}_i})^\top & I_q \end{bmatrix} \right)$$

Hence, based on the estimates of μ, B and Σ, we may predict the missing observations for observation i as

$$\widehat{Y}_{i,\mathcal{M}_i} := \mathbb{E}_{\widehat{\theta}}[Y_{i,\mathcal{M}_i} \mid Y_{i,\mathcal{O}_i} = y_{i,\mathcal{O}_i}] = \widehat{\mu}_{\mathcal{M}_i} + \widehat{\Sigma}_{\mathcal{M}_i,\mathcal{O}_i} \left(\widehat{\Sigma}_{\mathcal{O}_i}\right)^{-1} (y_{i,\mathcal{O}_i} - \widehat{\mu}_{\mathcal{O}_i}) \ ,$$

and estimate their conditional variance as $\mathbb{V}_{\widehat{\theta}}\left[Y_{i,\mathcal{M}_i} \mid Y_{i,\mathcal{O}_i} = y_i^{\mathcal{O}_i}\right] = \widehat{\Sigma}_{\mathcal{M}_i} + \widehat{\Sigma}_{\mathcal{M}_i,\mathcal{O}_i} \left(\widehat{\Sigma}^{\mathcal{O}_i}\right)^{-1} \widehat{\Sigma}^{\mathcal{O}_i,\mathcal{M}_i}$. The variance provides a measure of uncertainty on the prediction.

3.5.9 Conclusion

Probabilistic PCA (PPCA) extends classical PCA by framing it within a probabilistic latent variable model, which provides several advantages. It allows for principled handling of missing data, enables the use of likelihood-based methods for model comparison and selection. Moreover, it naturally integrates with other probabilistic models as will be seen in Chapter 5. Additionally, PPCA provides a clearer interpretation of the assumptions underlying PCA and facilitates extensions to more flexible models such as mixtures or Bayesian variants.

3.6 Conclusion of the chapter

This chapter presents a selection of classical models widely used in ecology, including mixture models, zero-inflated Poisson (ZIP) models, linear mixed-effects (LME) models, and probabilistic principal component analysis (Probabilistic PCA). Each of these models features an explicit latent variable structure, which we exploit using the Expectation-Maximization (EM) algorithm for parameter estimation. This unified EM perspective allows for a coherent treatment of these diverse models and highlights the interpretative power of their latent representations. In contrast, the next chapters introduce models in which the EM algorithm becomes more challenging to formulate due to complex dependency structures between latent variables. These dependencies prevent straightforward factorization of the complete-data likelihood, requiring recurrent formulas or approximations to handle the latent structure effectively.

Chapter 4

Non explicit E step

Contents

The previous chapter was dedicated to models where the conditional distribution of the latent variables given the observations $p_\theta(\boldsymbol{Z} \mid \boldsymbol{Y})$ was obtained very easily, by a direct application of the Bayes formula. In this chapter, we present models such that $p_\theta(\boldsymbol{Z} \mid \boldsymbol{Y})$ can be calculated exactly but by means of recursive formulas. We present here four families of models with Markovian latent structure.

- Section 4.1 is dedicated to hidden Markov models where the latent variable is a Markov chain in discrete time taking a finite number of possible values (the hidden states).
- Section 4.2 presents a hidden Markov model where the latent variable is a Gaussian Markov process in discrete time with infinitely many possible values ($Z_t \in \mathbb{R}^{d_z}$).
- Section 4.3 presents a Markov model in continuous time along the branches of an evolutionary tree. In this case the latent variables are the values of the process for the ancestral species.
- Section 4.4 presents a spatial model where the latent variable is a Markov random field, the space being discretized as a lattice. In each location, the hidden state takes a finite number of possible values.

We shall see that the Markovian assumption is a keystone to design an exact EM algorithm.

4.1 Discrete hidden Markov models

4.1.1 Definitions and properties

A natural question in ecological studies is to identify periods within time series data that reflect different ecological states or processes. Given a time series of ecological observations—such as animal movement, environmental conditions, or species abundance—we aim to detect segments that exhibit similar statistical patterns, potentially corresponding to distinct behavioral modes, seasonal phases, or habitat use. This question is related to time series segmentation, but we assume here that there are K underlying distributions, and each segment belongs to one of them. Therefore, separated segments can belong to a same cluster. Popular probabilistic models for such clustering are discrete hidden Markov models, also known as switching state space models. Figure 4.1 is a typical example of the kind of result we aim for, as it shows the trajectory of a Morus bassanus (d'Entremont et al. [2022]) segmented according to movement speed. This segmentation highlights different behavioral phases of the animal, which is a central goal in the ecological analysis of movement data.

Hidden Markov models (HMM) with discrete states are natural extension of mixture models seen in Section 3.1 for sequential observations. The dependences induces complex dependences and intractable likelihood, but an EM procedure can still be performed. Actually, in the literature (and as we will see below), hidden Markov models are not specific to clustering. Nonetheless, we first present the context of hidden Markov models in the usual discrete case for clustering of time series, *i.e.* when the latent state follows a Markov chain in a discrete space.

Notations We consider an observed series $\boldsymbol{y} = y_{0:T}$ where $y_t \in \mathbb{R}^{d_y}, \forall t$. We assume that these observations are realizations of random variables whose distribution depend on latent variables $Z_{0:T}$. Moreover, we make the following assumptions:

- The stochastic process $(Z_t)_{t\geq 0}$ in $\{1, \dots, K\}$ is an homogeneous[1] Markov chain with initial distribution $\boldsymbol{\omega}_0$ and a transition matrix $\boldsymbol{\Pi}$ of size $K \times K$, i.e.:

$$\begin{aligned} \omega_{0,k} &= \mathbb{P}(Z_0 = k), && \forall k \in \{1, \dots, K\} \\ \pi_{k\ell} &= \mathbb{P}(Z_{t+1} = \ell \mid Z_t = k), && \forall t \in \mathbb{N}, (k, \ell) \in \{1, \dots, K\} \times \{1, \dots, K\} \ . \end{aligned}$$

  $\boldsymbol{\omega}_0$ is such that $\sum_{k=1}^K \omega_{0,k} = 1$ and $\boldsymbol{\Pi}$ is such that, for any $k \in \{1, \dots, K\}$, $\sum_{\ell=1}^K \pi_{k\ell} = 1$. We denote this as

$$(Z_t)_{t\geq 0} \sim \mathsf{MarkovChain}(\boldsymbol{\omega}_0, \boldsymbol{\Pi}).$$

- Conditionally to $(Z_t)_{t\geq 0}$, the observations are realizations of $Y_0, \dots Y_T$ which are independent and follow a certain distribution $p_{k,\theta_{\text{obs}}}$ parametrized by θ_{obs} (this notation is consistent with the previous chapters). For instance, in Example 4.1, we will assume that the emission distribution is Gaussian: $Y_t \mid \{Z_t = k\} \sim \mathcal{N}\big(\mu_k, \sigma_k^2\big)$.

[1] This homogeneity assumption can be relaxed, but is assumed here to keep simple notations.

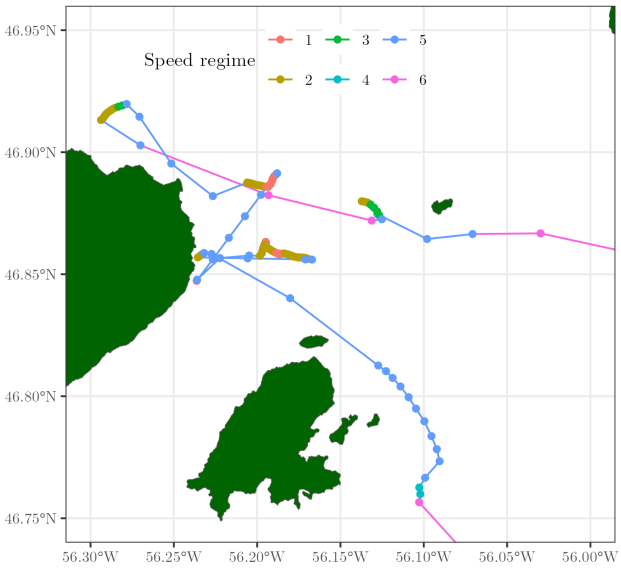


Figure 4.1: Example of GPS track segmented by HMM. Dataset from d'Entremont et al. [2022] and freely available on the online platform for movement data, MoveBank Wikelski et al. [2024].

Model 4.1 (Discrete hidden Markov model)**.**

$$(Z_t)_{t\geq 0} \sim \mathsf{MarkovChain}(\boldsymbol{\omega}_0, \boldsymbol{\Pi})$$
$$Y_t \mid \{Z_t = k\} \overset{iid}{\sim} p_{k,\theta_{obs}}(\cdot), \qquad t \in \{0,\dots,T\}\,,\ k \in \{1,\dots,K\}\,.$$

In this context $\theta = \{\boldsymbol{\omega}_0, \boldsymbol{\Pi}, \theta_{obs}\}$.

Graphical model Figure 4.2 displays the oriented graphical model associated with the joint distribution $p_\theta(\boldsymbol{Y}, \boldsymbol{Z})$ for Model 4.1. One can now see the full dependence between observations, and the independence conditionally to the hidden process. These independence properties are gathered and demonstrated in the Appendix Section A.3.3, Proposition A.14.

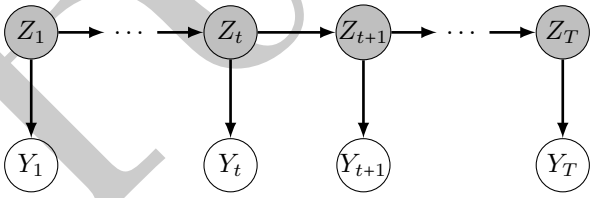


Figure 4.2: Graphical representation of Model 4.1.

Proposition 4.1. *For a hidden Markov model whose DAG is provided in Figure 4.2, the following independence properties hold:*

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

This proposition is proved in the Appendix section page 149.

Filtering and smoothing distributions As we will see in the sections below, inference on hidden Markov models involve two key conditional distributions.

Definition 4.1 (Filtering and smoothing distributions)**.** *For a hidden Markov model whose DAG is provided in Figure 4.2,*

- *The distribution of $Z_{0:T} \mid Y_{0:T}$ is called the **joint smoothing distribution**;*
- *For any $0 \leq t \leq T$:*
  - *The distribution of $Z_t \mid Y_{0:T}$ is called the **marginal smoothing distribution** at time t;*
  - *The distribution of $Z_t \mid Y_{0:t}$ is called the **filtering distribution** at time t.*

Note that at $t = T$, the last two distributions coincides.

4.1.2 Complete and marginal log-likelihoods

Marginal log-likelihood of the observations When the state space of the latent variable is discrete and finite, the log-likelihood of the observation can be written directly:

$$\begin{aligned}\log p_\theta(\boldsymbol{y}) = \log p_\theta(y_{0:T}) &= \log\left(\sum_{z_{0:T}\in\{1,\dots,K\}^{T+1}} p_\theta(y_{0:T}, z_{0:T})\right) \\ &= \log\left(\sum_{z_{0:T}\in\{1,\dots,K\}^{T+1}} \mathbb{P}_\theta(Z_{0:T} = z_{0:t}) p_\theta(y_{0:T} \mid z_{0:T})\right). \end{aligned} \tag{4.1}$$

As we will see below, the inner terms have explicit expressions in the Markov context, however, this direction is an automatic dead end in practice, as the summation involves K^{T+1} terms! Thus, computing the likelihood cannot be done by this brute force technique, and an alternative technique should be found. However, let's notice that[2] (using successive conditioning):

$$p_\theta(y_{0:T}) = p_\theta(y_0) \times \prod_{t=0}^{T-1} p_\theta(y_{t+1} \mid y_{0:t}),$$

and then:

$$\log p_\theta(y_{0:T}) = \log p_\theta(y_0) + \sum_{t=0}^{T} \log p_\theta(y_{t+1} \mid y_{0:t}). \tag{4.2}$$

The following *forward* algorithm shows that all those terms are strongly related, in HMMs, to the filtering distribution, and that they can be computed recursively.

Proposition 4.2 (Forward computation of filtering distributions and log-likelihood in Model 4.1)**.** *For all $0 \leq t \leq T$, let's denote:*

$$\alpha_t(\ell) := \mathbb{P}_\theta(Z_t = \ell \mid Y_{0:t} = y_{0:t}),$$

the filtering distribution at time t. This quantity can be computed for each t using the following algorithm

- ***Initialization*** *For $\ell \in \{1, \dots, K\}$,*

$$\alpha_0(\ell) = \frac{\omega_{0,\ell} p_{\ell,\theta_{obs}}(y_0)}{c_0},$$

  where

$$c_0 = \sum_{\ell=1}^{K} \omega_{0,\ell} p_{\ell,\theta_{obs}}(y_0).$$

- ***Recursion*** *For $\ell \in \{1, \dots, K\}$, and $t \in \{1, \dots, T\}$*

$$\alpha_t(\ell) = \frac{p_{\ell,\theta_{obs}}(y_t) \sum_{k=1}^{K} \pi_{k\ell} \alpha_{t-1}(k)}{c_t},$$

[2] This decomposition of the likelihood of a time series is always true, and does not depend on the hidden Markov model.

where

$$c_t = \sum_{\ell=1}^{K} p_{\ell,\theta_{obs}}(y_t) \sum_{k=1}^{K} \pi_{k\ell}\alpha_{t-1}(k)\,.$$

Moreover, we have that $c_0 = p_\theta(y_0)$ *and, for all* $t \geq 1$, $c_t = p_\theta(y_{t+1} \mid y_{0:t})$. *Then, applying Equation* (4.2) *the log-likelihood expresses as:*

$$\log p_\theta(y_{0:T}) = \sum_{t=0}^{T} \log(c_t)\,.$$

Note that, comparing to (4.1), the complexity is here of order $\mathcal{O}(TK^2)$, and then always feasible for reasonable values of K.

Proof of Proposition 4.2

- At $t = 0$, for $\ell \in \{1, \ldots, K\}$, using the definition of conditional probability:

$$\begin{aligned}\alpha_0(\ell) &= \mathbb{P}_\theta(Z_0 = \ell \mid Y_0 = y_0) \\ &= \frac{p_\theta(y_0 \mid Z_0 = \ell) \times \mathbb{P}_\theta(Z_0 = \ell)}{p_\theta(y_0)} \\ &= \frac{\omega_{0,\ell} p_{\ell,\theta_{\text{obs}}}(y_0)}{c_0}\,.\end{aligned}$$

  In this expression c_0 is the normalizing constant, ensuring that $\sum_{\ell=1}^{K} \alpha_0(\ell) = 1$ (as they are probabilities), hence this results in:

$$c_0 = \sum_{\ell=1}^{K} \omega_{0,\ell} p_{\ell,\theta_{\text{obs}}}(y_0)\,.$$

- For $t \geq 1$, we have:

$$\begin{aligned}\alpha_t(\ell) &= \mathbb{P}_\theta(Z_t = \ell \mid Y_{0:t} = y_{0:t}) \\ &= \frac{p_\theta(y_t, Z_t = \ell \mid y_{0:t-1})}{p(y_t \mid y_{0:t-1})} \\ &= \frac{p_\theta(y_t \mid Z_t = \ell, \cancel{y_{0:t-1}}) \mathbb{P}_\theta(Z_t = \ell \mid y_{0:t-1})}{c_t} \\ &= \frac{p_{\ell,\theta_{\text{obs}}}(y_t) \sum_{k=1}^{K} \mathbb{P}_\theta(Z_t = \ell, Z_{t-1} = k \mid y_{0:t-1})}{c_t} \\ &= \frac{p_{\ell,\theta_{\text{obs}}}(y_t) \sum_{k=1}^{K} \mathbb{P}_\theta(Z_t = \ell \mid Z_{t-1} = k, \cancel{y_{0:t-1}}) \times \mathbb{P}_\theta(Z_{t-1} = k \mid y_{0:t-1})}{c_t} \\ &= \frac{p_{\ell,\theta_{\text{obs}}}(y_t) \sum_{k=1}^{K} \pi_{k\ell}\alpha_{t-1}(k)}{c_t}\,.\end{aligned}$$

  In this computation, we emphasized the use of the conditional independences induced by the graphical model 4.2. Again, c_t being the normalizing constant ensuring $\sum_{\ell=1}^{K} \alpha_t(\ell) = 1$, it implies that:

$$c_t = \sum_{\ell=1}^{K} p_{\ell,\theta_{\text{obs}}}(y_t) \sum_{k=1}^{K} \pi_{k\ell}\alpha_{t-1}(k)\,.$$

Complete log-likelihood If the only goal of inference is to maximize the likelihood, the previous algorithm is sufficient. However the EM algorithm worth to be developed, as it's design is explicit and it is interesting for two main reasons. On the one hand, the computation of the E step relies on a forward backward procedure which is a keystone of many algorithms for sequential data. On the other hand, in many applications (such as the one below), we are interested in the inference of the hidden state, and the forward-backward algorithm computes key quantities for estimators of the hidden state sequence.

In the general case of hidden Markov models (which we will encounter again in Section 4.2) the complete likelihood admits a simple factorization due to the Markov assumption over hidden states.

Proposition 4.3. *For the discrete HMM, the complete log-likelihood is :*

$$\log p_\theta(y_{0:T}, Z_{0:T}) = \sum_{k=1}^{K} Z_{0k} \log \omega_{0,k} + \sum_{t=0}^{T-1} \sum_{k=1}^{K} \sum_{\ell=1}^{K} Z_{tk} Z_{t+1\,\ell} \log \pi_{k\ell} + \sum_{t=0}^{T} \sum_{k=1}^{K} Z_{tk} \log p_{k,\theta_{obs}}(y_t)\,, \tag{4.3}$$

where for all ≥ 0 *and* $k \in \{1, \dots, K\}$, $Z_{tk} = \mathbb{I}_{\{k\}}(Z_t) = \mathbb{I}_{\{Z_t = k\}}$.

Proof of Proposition 4.3

We can use the DAG of the model (Figure 4.2) and factorise the distribution using the parents in the DAG (see Appendix section):

$$p_\theta(\boldsymbol{Y}, \boldsymbol{Z}) = \prod_{t=0}^{T} p_\theta(Z_t \mid pa(Z_t)) p_\theta(Y_t \mid pa(Y_t))$$

Z_0 as no parents. Z_t has for only parent $\{Z_{t-1}\}$, Y_t has $\{Z_t\}$ for unique parent.

$$p_\theta(\boldsymbol{Y}, \boldsymbol{Z}) = p_\theta(Z_0) \prod_{t=1}^{T} p_\theta(Z_t \mid Z_{t-1}) \prod_{t=0}^{T} p_\theta(Y_t \mid Z_t)$$

Then, the log-complete likelihood satisfies:

$$\log p_\theta(y_{0:T}, Z_{0:T}) = \log p_\theta(Z_0) + \sum_{t=0}^{T-1} \log p_\theta(Z_{t+1} \mid Z_t) + \sum_{t=0}^{T} \log p_\theta(y_t \mid Z_t)\,. \tag{4.4}$$

Remark. *Note that this expression does not depend on the discrete nature of the hidden chain or on the nature of observations, but is valid for every joint distribution satisfying the graphical model of Figure 4.2.*

Getting back to the case where the latent variable Z is a discrete Markov chain on $\{1, \dots, K\}$, we obtain the specific formulat for $p_\theta(Z_0)$:

$$p_\theta(Z_0) = \prod_{k=1}^{K} \omega_{0,k}^{Z_{0k}}, \quad \text{where} \quad Z_{0k} = \mathbb{I}_{\{k\}}(Z_0) = \mathbb{I}_{\{Z_0 = k\}}.$$

Similarly :

$$p_\theta(Z_{t+1} \mid Z_t) = \prod_{\ell=1}^{K} \prod_{\ell=1}^{K} \pi_{k\ell}^{Z_{tk} Z_{t+1\,\ell}}$$

$$p_\theta(y_t \mid Z_t) = \prod_{k=1}^{K} (p_{k,\theta_{\text{obs}}}(y_t))^{Z_{tk}},\ t \in \{0, \dots, T\}\ .$$

Finally, combining these three last expressions with Equation (4.4), we obtain the expected result (4.3).

4.1.3 EM algorithm for discrete HMM

In the case of HMMs, the E step involves an iterative procedure that requires careful formulation. To maintain clarity, we gradually introduce $Q\big(\theta \mid \theta^{(h)}\big)$, followed by a detailed explanation of the E step and the M step. The full algorithm is presented at the end of the subsection, once all necessary components have been introduced.

4.1.3.1 Objective function $Q\big(\theta \mid \theta^{(h)}\big)$

For a general HMM, suppose we have a current value $\theta^{(h)}$ then, from (4.4), we can write:

$$\begin{aligned} Q\big(\theta \mid \theta^{(h)}\big) =& \mathbb{E}_{\theta^{(h)}}[\log p_\theta(y_{0:T}, Z_{0:T}) \mid y_{0:T}] \\ =& \mathbb{E}_{\theta^{(h)}}[\log p_\theta(Z_0) \mid y_{0:T}] + \sum_{t=0}^{T-1} \mathbb{E}_{\theta^{(h)}}[\log p_\theta(Z_{t+1} \mid Z_t) \mid y_{0:T}] \\ &+ \sum_{t=0}^{T} \mathbb{E}_{\theta^{(h)}}[\log p_\theta(y_t \mid Z_t) \mid y_{0:T}]\,. \end{aligned} \tag{4.5}$$

In our discrete case context where (4.3) is satisfied, we have:

$$\begin{aligned} Q\big(\theta \mid \theta^{(h)}\big) &= \sum_{k=1}^{K} \underbrace{\mathbb{E}_{\theta^{(h)}}[Z_{0k} \mid Y_{0:T} = y_{0:T}]}_{\tau_{0,k}^{(h)}} \log \omega_{0,k} + \sum_{t=0}^{T-1} \sum_{k=1}^{K} \sum_{\ell=1}^{K} \underbrace{\mathbb{E}_{\theta^{(h)}}[Z_{tk} Z_{t+1\,\ell} \mid Y_{0:T} = y_{0:T}]}_{\xi_{t,k,\ell}^{(h)}} \log \pi_{k\ell} \\ &+ \sum_{t=0}^{T} \sum_{k=1}^{K} \underbrace{\mathbb{E}_{\theta^{(h)}}[Z_{tk} \mid Y_{0:T} = y_{0:T}]}_{\tau_{t,k}^{(h)}} \log p_{k,\theta_{\text{obs}}}(y_t) \\ &= \sum_{k=1}^{K} \tau_{0,k}^{(h)} \log \omega_{0,k} + \sum_{t=0}^{T-1} \sum_{k=1}^{K} \sum_{\ell=1}^{K} \xi_{t,k,\ell}^{(h)} \log \pi_{k\ell} + \sum_{t=0}^{T} \sum_{k=1}^{K} \tau_{t,k}^{(h)} \log p_{k,\theta_{\text{obs}}}(y_t) \end{aligned} \tag{4.6}$$

where

$$\tau_{t,k}^{(h)} := \mathbb{E}_{\theta^{(h)}}[Z_{tk} \mid y_{0:T}] = \mathbb{P}_{\theta^{(h)}}(Z_t = k \mid y_{0:T}),\ t \in \{0, \dots, T\}\,, \tag{4.7}$$

$$\xi_{t,k,\ell}^{(h)} := \mathbb{E}_{\theta^{(h)}}[Z_{tk} Z_{t+1\,\ell} \mid y_{0:T}] = \mathbb{P}_{\theta^{(h)}}(Z_t = k, Z_{t+1} = \ell \mid y_{0:T})\,. \tag{4.8}$$

4.1.3.2 E step

The E step consists in the calculation of the $\tau_{t,k}^{(h)}$'s and $\xi_{t,k,\ell}^{(h)}$'s. Note that:

$$\tau_{t,k}^{(h)} = \sum_{\ell=1}^{K} \xi_{t,k,\ell}^{(h)}, \tag{4.9}$$

and:

$$\begin{aligned} \xi_{t,k,\ell}^{(h)} &= \mathbb{P}_{\theta^{(h)}}(Z_t = k, Z_{t+1} = \ell \mid Y_{0:T} = y_{0:T}) \\ &= \frac{\mathbb{P}_{\theta^{(h)}}(Z_t = k \mid Y_{0:t} = y_{0:t}) \times p_{\theta^{(h)}} y_{(t+1):T}, Z_{t+1} = \ell \mid Z_t = k, Y_{0:t} = y_{0:t}}{p_{\theta^{(h)}}\left(y_{(t+1):T}\right)} \end{aligned}$$

The second term of the numerator can be tweaked to emphasize known quantities.

$$\begin{aligned} p_{\theta^{(h)}}\big(y_{(t+1):T}, Z_{t+1} = \ell \mid Z_t = k, \cancel{Y_{0:t} = y_{0:t}}\big) &= p_{\theta^{(h)}}(Z_{t+1} = \ell, y_{t+1}, y_{t+2:T} \mid Z_t = k) \\ &= p_{\theta^{(h)}}(y_{t+1}, Z_{t+1} = \ell \mid Z_t = k) \times p_{\theta^{(h)}}(y_{t+2:T} \mid \cancel{y_{t+1}}, \cancel{Z_t = k}, Z_{t+1} = \ell) \\ &= p_{\theta^{(h)}}(y_{t+1} \mid \cancel{Z_t = k}, Z_{t+1} = \ell) \times \mathbb{P}_{\theta^{(h)}}(Z_{t+1} = \ell \mid Z_t = k) \times p_{\theta^{(h)}}(y_{t+2:T} \mid Z_{t+1} = \ell) \\ &= p_{\theta^{(h)}}(Z_{t+1} = \ell \mid \cancel{y_{0:t}}, Z_t = k) p_{k,\theta_{\text{obs}}{}^{(h)}}(y_{t+1}) p_{\theta^{(h)}}(y_{t+2:T} \mid Z_{t+1} = \ell) \\ &= p_{\ell,\theta_{\text{obs}}{}^{(h)}}(y_{t+1}) \times \pi_{k\ell}^{(h)} \times p_{\theta^{(h)}}(y_{t+2:T} \mid Z_{t+1} = \ell)\,, \end{aligned}$$

where the conditional independences are consequences of Proposition 4.1. Finally, we obtain:

$$\xi_{t,k,\ell}^{(h)} \propto \mathbb{P}_{\theta^{(h)}}(Z_t = k \mid Y_{0:t} = y_{0:t}) \pi_{k\ell}^{(h)} p_{\ell,\theta_{\text{obs}}{}^{(h)}}(y_{t+1}) p_{\theta^{(h)}}(y_{t+2:T} \mid Z_{t+1} = \ell) \times p_{\ell,\theta_{\text{obs}}{}^{(h)}}(y_{t+1}) \times \pi_{k\ell}^{(h)} \times p_{\theta^{(h)}}(y_{t+2:T} \mid Z_{t+1} = \ell)\,.$$

The first term is simply $\alpha_t^{(h)}(k)$, computed thanks to Proposition 4.2, under the current parameter $\theta^{(h)}$, the two terms $p_{\ell,\theta_{\text{obs}}{}^{(h)}}(y_{t+1})$ and $\pi_{k\ell}^{(h)}$ are quantities given by the model. It remains the term $\beta_{t+1}^{(h)}(\ell) := p_{\theta^{(h)}}(y_{t+2:T} \mid Z_{t+1} = \ell)$ which has not been encountered so far. The normalizing constant is simply obtained summing over all possible pairs of states, to get:

$$\xi_{t,k,\ell}^{(h)} = \frac{\alpha_t^{(h)}(k) \pi_{k\ell}^{(h)} p_{\ell,\theta_{\text{obs}}{}^{(h)}}(y_{t+1}) \beta_{t+1}^{(h)}(\ell)}{\sum_{k=1}^{K} \sum_{\ell=1}^{K} \alpha_t^{(h)}(k) \pi_{k\ell}^{(h)} p_{\ell,\theta_{\text{obs}}{}^{(h)}}(y_{t+1}) \beta_{t+1}^{(h)}(\ell)}\,. \tag{4.10}$$

For a current value of the parameters θ, the $(\beta_t(k))_{t,k}$ can be computed via a backward recursion, as stated by the following proposition:

Proposition 4.4 (Backward recursion)**.** *Consider, for all $k \in \{1, \dots, K\}$:*

$$\beta_T(k) = 1, \tag{4.11}$$
$$\beta_t(k) = p_\theta(y_{t+1:T} \mid Z_t = k), \ t \in \{0, \dots, T-1\} \ . \tag{4.12}$$

For every $t \in \{0, \dots, T-1\}$ and $k \in \{1, \dots, K\}$, the sequence $(\beta_t(k))$ satisfies the following recursion

$$\beta_t(k) = \sum_{\ell=1}^{K} \pi_{k\ell}\, p_{\ell,\theta_{obs}}(y_{t+1})\, \beta_{t+1}(\ell) \, . \tag{4.13}$$

Proof of Proposition 4.4

Again, this is done by successive conditionning, and using conditional independence properties. For $t \in \{0, \dots, T-1\}$:

$$\begin{aligned}
\beta_t(k) = p_\theta(y_{t+1:T} \mid Z_t = k) &= \sum_{\ell=1}^{K} p_\theta(y_{t+1:T}, Z_{t+1} = \ell \mid Z_t = k) = \sum_{\ell=1}^{K} p_\theta(y_{t+1}, y_{t+2:T}, Z_{t+1} = \ell \mid Z_t = k) \\
&= \sum_{\ell=1}^{K} \mathbb{P}_\theta(Z_{t+1} = \ell \mid Z_t = k) p_\theta(y_{t+1} \mid Z_{t+1} = \ell, Z_t = k) p_\theta(y_{t+2:T} \mid y_{t+1}, Z_{t+1} = \ell, Z_t = k) \\
&= \sum_{\ell=1}^{K} \mathbb{P}_\theta(Z_{t+1} = \ell \mid Z_t = k) p_\theta(y_{t+1}, \mid Z_{t+1} = \ell) p_\theta(y_{t+2:T} \mid Z_{t+1} = \ell) \\
&= \sum_{\ell=1}^{K} \pi_{k\ell}\, p_{\ell,\theta_{\text{obs}}}(y_{t+1})\, \beta_{t+1}(\ell) \, .
\end{aligned}$$

Note that the initialization given by (4.11) is rather arbitrary, but ensures that (4.13) holds at time $T-1$.

Remark (Log-sum-exp trick)**.** *The direct computation of the secquence of $\beta_t(k)$ given by Proposition 4.4 is prone to numerical instability when p is large. Indeed, $\beta_t(k)$ tends quickly towards 0 when T is large and t decreases to 0. To avoid numerical problems, a classical workaround is the work in the logarithmic scale, computing the sequence of $\log(\beta_t(k))$. We therefore rewrite Proposition 4.4, for all $k \in \{1, \dots, K\}$:*

$$\begin{aligned}
\log(\beta_T(k)) &= 0, \\
\log(\beta_t(k)) &= \log\left(\sum_{\ell=1}^{K} \pi_{k\ell}\, p_{\ell,\theta_{obs}}(y_{t+1})\, \beta_{t+1}(\ell) \right) \\
&= \log\left(\sum_{\ell=1}^{K} \exp\left(\log(\pi_{k\ell}) + \log(p_{\ell,\theta_{obs}}(y_{t+1})) + \log(\beta_{t+1}(\ell)) \right) \right) .
\end{aligned}$$

Written this way, the recursion is still prone to the same instability as before, as we now exponentiate terms that are susceptible to be strongly negative. Howevere, we can use the log-sum-exp trick to avoid any numerical problem. Using the fact that for any reals terms $r_1, \dots r_K$, defining $m := \max_{\ell \in \{1,\dots,K\}} r_\ell$, one has:

$$\log\left(\sum_{\ell=1}^{K} \exp(r_k) \right) = m + \log\left(\sum_{\ell=1}^{K} \exp(r_k - m) \right) ,$$

we can rewrite the recursion as:

$$\log(\beta_t(k)) = m_{t+1} + \log\left(\sum_{\ell=1}^{K} \exp\left(\log(\pi_{k\ell}) + \log(p_{\ell,\theta_{obs}}(y_{t+1})) + \log(\beta_{t+1}(\ell) - m_{t+1}) \right) \right) ,$$

where:

$$m_{t+1} = \max_{\ell \in \{1,\dots,K\}} \log(\pi_{k\ell}) + \log(p_{\ell,\theta_{obs}}(y_{t+1})) + \log\left(\beta_{t+1}(\ell)\right) \, .$$

As at least one term in the sum is equal to $\exp(0)$, this provides a numerically stable recursion, which is often used in practice.

This backward recursion enables to perform the E step of the algorithm through a forward backward procedure:

Proposition 4.5 (Forward-backward algorithm for E step in discrete HMM)**.** *Let* $\tau_{t,k}^{(h)}$ *and* $\xi_{t,k,\ell}^{(h)}$ *be defined as in* (4.7) *and* (4.8)*, and* $\alpha_t^{(h)}(k)$ *and* $\beta_t^{(h)}(k)$ *be the quantities of Propositions 4.2 and 4.4* [a]*, then*[b]

$$\xi_{t,k,\ell}^{(h)} = \frac{\alpha_t^{(h)}(k)\,\pi_{k\ell}^{(h)}\,p_{\ell,\theta_{obs}{}^{(h)}}(y_{t+1})\,\beta_{t+1}^{(h)}(\ell)}{\sum_{j=1}^{K}\alpha_t^{(h)}(j)\beta_t^{(h)}(j)} \tag{4.14}$$

$$\tau_{t,k}^{(h)} = \frac{\alpha_t^{(h)}(k)\beta_t^{(h)}(k)}{\sum_{\ell=1}^{K}\alpha_t^{(h)}(\ell)\beta_t^{(h)}(\ell)}\,. \tag{4.15}$$

[a] Computed with $\theta^{(h)}$.

[b] Again, all these expressions are, in practice, implemented in the logarithmic scale, using the log-sum-exp trick to compute the sums at denominators.

Proof of Proposition 4.5

The numerator of (4.14) is only (4.10) in which we plugged $\alpha_t^{(h)}(j)$ and $\beta_{t+1}^{(h)}(k)$, while the denominator is obtained by seeing that the sum over ℓ in the denominator of (4.10) simplifies thanks to (4.13). Finally (4.15) is obtained by summing (4.14) over ℓ (and, again, using (4.13)).

Therefore, the E step is done. Again, note that this step consisted in computing probabilities (and not complicated expectations). The resulting function $Q\big(\theta \mid \theta^{(h)}\big)$ is then a finite sum involving these probabilities.

4.1.3.3 M step

Once the E step is performed using Proposition 4.5, we want to update $\Big\{\boldsymbol{\omega}_0^{(h)}, \mathbf{\Pi}^{(h)}, \theta_{\text{obs}}{}^{(h)}\Big\}$.

Initial distribution $\boldsymbol{\omega}_0$ In our context where we only consider one time series, the initial distribution is of little interest. However, it still can be estimated. It is worth noting here that the estimation of the initial distribution does not only involve y_0, but all observations (as the posterior distribution of Z_0 depends on all observations). We want to maximize (4.6) with respect to $\boldsymbol{\omega}_0$, with the constraint that $\sum_{j=1}^{K}\omega_{0,j} = 1$. Using Lagrange multipliers, we then want to cancel the gradient, for some $\lambda \in \mathbb{R}$,

$$\nabla_{\boldsymbol{\omega}_0,\lambda}\underbrace{\left(Q\big(\theta \mid \theta^{(h)}\big) - \lambda\left(\sum_{k=1}^{K}\omega_{0,k} - 1\right)\right)}_{:=g_0(\boldsymbol{\omega}_0,\lambda)} = \begin{pmatrix} \frac{\tau_{0,1}^{(h)}}{\omega_{0,1}} - \lambda \\ \vdots \\ \frac{\tau_{0,K}^{(h)}}{\omega_{0,K}} - \lambda \\ \sum_{k=1}^{K}\omega_{0,k} - 1 \end{pmatrix}.$$

Therefore, noticing that $\sum_{k=1}^{K}\tau_{0,k}^{(h)} = 1$,

$$\nabla_{\boldsymbol{\omega}_0,\lambda} g_0(\boldsymbol{\omega}_0^{(h+1)}, \lambda) = 0 \Leftrightarrow \begin{pmatrix} \omega_{0,1}^{(h+1)} \\ \vdots \\ \omega_{0,K}^{(h+1)} \\ \lambda \end{pmatrix} = \begin{pmatrix} \tau_{0,1}^{(h)} \\ \vdots \\ \tau_{0,K}^{(h)} \\ 1 \end{pmatrix}.$$

Note that during the E step, $\boldsymbol{\omega}_0^{(h)}$ is only used to compute $(\alpha_0^{(h)}(k))_{k\in\{1,\dots,K\}}$, but it is updated in the M using *all observations*, (as $(\tau_{0,k}^{(h)})_{k\in\{1,\dots,K\}}$ is computed using $(\beta_0(k))_{k\in\{1,\dots,K\}}^{(h)}$, which implies, through recursions, all observations).

Transition matrix $\mathbf{\Pi}$ The maximization of $Q\big(\theta \mid \theta^{(h)}\big)$ with respect to $\mathbf{\Pi}$ is made under K constraints, which are that the K columns of $\mathbf{\Pi}$ sum to one. The derivative with respect to the components of this matrix only concerns the terms of the second line of (4.6). Therefore,

$$\nabla_{\mathbf{\Pi},\lambda}\underbrace{\left(Q\big(\theta \mid \theta^{(h)}\big) - \sum_{k=1}^{K}\lambda_k\left(\sum_{\ell=1}^{K}\pi_{k\ell}-1\right)\right)}_{:=g(\mathbf{\Pi},\lambda_{1:K})} = \begin{pmatrix} \frac{1}{\pi_{11}}\sum_{t=0}^{T-1}\xi_{t,1,1}^{(h)} - \lambda_1 \\ \vdots \\ \frac{1}{\pi_{1K}}\sum_{t=0}^{T-1}\xi_{t,1,K}^{(h)} - \lambda_1 \\ \frac{1}{\pi_{21}}\sum_{t=0}^{T-1}\xi_{t,2,1}^{(h)} - \lambda_2 \\ \vdots \\ \frac{1}{\pi_{KK}}\sum_{t=0}^{T-1}\xi_{t,K,K}^{(h)} - \lambda_K \\ \sum_{\ell=1}^{K}\pi_{1\ell}-1 \\ \vdots \\ \sum_{\ell=1}^{K}\pi_{K\ell}-1 \end{pmatrix}.$$

Therefore,

$$\begin{aligned}
\nabla_{\mathbf{\Pi},\lambda} g(\mathbf{\Pi}^{(h+1)},\lambda_{1:K}) = 0 \Leftrightarrow & \begin{cases} \sum_{t=0}^{T-1}\xi_{t,k,\ell}^{(h)} &= \pi_{k\ell}^{(h)}\lambda_k, \ k\in\{1,\dots,K\} \\ \sum_{\ell=1}^{K}\sum_{t=0}^{T-1}\xi_{t,k,\ell}^{(h)} &= \sum_{\ell=1}^{K}\pi_{k,\ell}^{(h)}\lambda_j \\ \sum_{\ell=1}^{K}\pi_{k,\ell}^{(h)} &= 1 \end{cases}, \text{ for all } j\in\{1,\dots,K\}. \\
\Leftrightarrow & \begin{cases} \lambda_k &= \sum_{k=1}^{K}\sum_{t=0}^{T-1}\xi_{t,k,\ell}^{(h)} \\ \pi_{k,\ell}^{(h)} &= \frac{1}{\lambda_k}\sum_{t=0}^{T-1}\xi_{t,k,\ell}^{(h)}, \ k\in\{1,\dots,K\} \end{cases}, \text{ for all } k\in\{1,\dots,K\}. \\
\Leftrightarrow & \pi_{k,\ell}^{(h)} = \frac{1}{\sum_{t=0}^{T-1}\tau_{t,k}^{(h)}}\sum_{t=0}^{T-1}\xi_{t,k,\ell}^{(h)}, \quad \text{for all } k\in\{1,\dots,K\}, \ell\in\{1,\dots,K\},
\end{aligned}$$

where the last simplification was made using (4.9). This expression is actually rather intuitive as it gives the expected number of pairs of state (k,ℓ) divided by the expected number of state k during the T first times.

Observation parameters $\theta_{\mathbf{obs}}$ It remains to derive the parameters of the observation model. The computation of course depends on the chosen distribution. In the case of multivariate normal distributions, it would have exactly the same expressions as the ones of Gaussian mixture (3.7) and (3.8) (replacing the sum from one 1 to n by a sum from 0 to T). In Section 4.1.6, we will provide formulas for updates in the univariate case of Example 4.1.

$$\mu_k^{(h+1)} = \frac{1}{N_k^{(h)}}\sum_{t=0}^{T}\tau_{t,k}^{(h)}y_t\,, \qquad \sigma_k^{2,(h+1)} = \frac{1}{N_k^{(h)}}\sum_{t=0}^{T}\tau_{t,k}^{(h)}\left(y_t-\mu_k^{(h+1)}\right)^2$$

where

$$N_k^{(h)} = \sum_{t=0}^{T}\tau_{t,k}^{(h)}\,. \tag{4.16}$$

4.1.3.4 EM algorithm for the HMM model

Combining the E step and M step developed before, we obtain the following algorithm

Algorithm 4.1 (EM for the Gaussian HMM)**.** *Starting from $\theta^{(0)}$ as defined in Model 4.1, repeat until convergence:*

E step. *Perform the forward-backward procedure of Proposition 4.5*

$$\begin{aligned}
\tau_{t,k}^{(h)} &:= \mathbb{P}_{\theta^{(h)}}(Z_t = k \mid y_{0:T}), \ t\in\{0,\dots,T\} \\
\xi_{t,k,\ell}^{(h)} &:= \mathbb{P}_{\theta^{(h)}}(Z_t = k, Z_{t+1} = \ell \mid y_{0:T}).
\end{aligned}$$

and set $N_k^{(h)} = \sum_{t=0}^{T}\tau_{t,k}^{(h)}$.

M step. *Update the parameters*

$$\omega_{0,k}^{(h)} = \tau_{0,k}^{(h)}$$
$$\pi_{k,\ell}^{(h)} = \frac{1}{\sum_{t=0}^{T-1} \tau_{t,k}^{(h)}} \sum_{t=0}^{T-1} \xi_{t,k,\ell}^{(h)}, \quad \textit{for all } k \in \{1,\dots,K\}\ ,\ \ell \in \{1,\dots,K\}$$
$$\mu_k^{(h+1)} = \frac{1}{N_k^{(h)}} \sum_{t=0}^{T} \tau_{t,k}^{(h)} y_t\,, \qquad \sigma_k^{2,(h+1)} = \frac{1}{N_k^{(h)}} \sum_{t=0}^{T} \tau_{t,k}^{(h)} \left(y_t - \mu_k^{(h+1)}\right)^2$$

4.1.4 Inference of the hidden states

Once parameter estimation is performed and has provided an estimate $\widehat{\theta}$, the estimation of hidden states can be done in different ways. The first natural possibility is to consider the state j that maximizes $\widehat{\tau}_{t,j}$, and therefore, for each time t, we set:

$$\widehat{Z}_t = \underset{k\in\{1,\dots,K\}}{\arg\max}\ \widehat{\tau}_{t,k}\,. \tag{4.17}$$

This MAP estimator considers the maximal probability of the marginal distribution of the sequence at time t. Another approach is to consider the joint sequence that maximizes the joint probability, which is then by:

$$\widehat{Z}_{0:T} = k_{0:T}^* = \underset{k_0,\dots,k_T}{\arg\max}\, \mathbb{P}_{\widehat{\theta}}(Z_0 = k_0, \dots, Z_T = k_T \mid y_{0:t})\,. \tag{4.18}$$

Despite involving a maximum on discrete set having K^{T+1} elements, this estimator can be computed in an efficient way via the Viterbi algorithm.

Algorithm 4.2 (Viterbi algorithm to get the best posterior sequence)**.**

Forward. *Let's define* $0 \leqslant t \leqslant T$*, and* $1 \leq k \leq K$ *Formally:*

$$\delta_0(k) = \mathbb{P}_{\widehat{\theta}}(Z_0 = k \mid Y_0 = y_0)$$
$$\delta_t(k) = \max_{k_{0:(t-1)}}\ \mathbb{P}_{\widehat{\theta}}\big(Z_{0:(t-1)} = k_{0:(t-1)}, Z_t = k \mid Y_{0:t} = y_{0:t}\big) \qquad \textit{for } t \geq 1\,.$$

Then, $\delta_t(k)$ *can be computed recursively: for* $t \geq 1$*:*

$$\delta_t(k) = p_{\widehat{\theta}}(y_t \mid Z_t = k) \times \max_{1\leq\ell\leq K} \{\widehat{\pi}_{\ell k} \times \delta_{t-1}(\ell)\}\ .$$

Backward. *Moreover,*

$$k_{0:T}^* = \underset{k_0,\dots,k_T}{\arg\max}\, \mathbb{P}_{\widehat{\theta}}(Z_0 = k_0, \dots, Z_T = k_T \mid Y_{0:T} = y_{0:T})$$

can be obtained recursively by

$$k_T^* = \underset{1\leq\ell\leq K}{\arg\max}\, \delta_T(\ell)$$
$$k_{t-1}^* = \underset{1\leq\ell\leq K}{\arg\max}\, \widehat{\pi}_{\ell k_t^*} \delta_{t-1}(\ell)\,, \ \textit{for } t = T-1, \dots 0\,.$$

The take home message here is that the Viterbi algorithm allows to compute the estimator of Equation (4.18) in an efficient manner. When the user is interested in trajectories (or joint distribution) of the hidden states, and not only in the marginal values of the hidden states, this estimator might be better suited than the MAP of Equation (4.17). We here present the proof which is, to the best of our knowledge, rarely written. It is rather technical but, in our opinion, is a good help to understand "why" the Viterbi algorithm works.

Proof of Algorithm 4.2

Forward. The recursion over $\delta_t(k)$ comes from the fact that[a] it can be decomposed into two terms:

$$\begin{aligned}\delta_t(k) &= \max_{k_{0:(t-1)}} p_{\widehat{\theta}}(y_{0:t}, Z_{0:(t-1)} = k_{0:(t-1)}, Z_t = k)\\ &= \max_{k_{0:(t-1)}} \mathbb{P}_{\widehat{\theta}}\big(Z_{0:(t-1)} = k_{0:(t-1)}, Z_t = k\big) p_{\widehat{\theta}}(y_{0:t} \mid Z_{0:(t-1)} = k_{0:(t-1)}, Z_t = k)\end{aligned}$$

Each term of the product can now be reformulated:

$$\begin{aligned}\mathbb{P}_{\widehat{\theta}}\big(Z_{0:(t-1)} = k_{0:(t-1)}, Z_t = k\big) =& \mathbb{P}_{\widehat{\theta}}(Z_t = k \mid Z_{t-1} = k_{t-1}) \times \mathbb{P}_{\widehat{\theta}}\big(Z_{0:(t-2)} = k_{0:(t-2)}, Z_{t-1} = k_{t-1}\big)\\ =& \pi_{k_{t-1}k} \times \mathbb{P}_{\widehat{\theta}}\big(Z_{0:(t-2)} = k_{0:(t-2)}, Z_{t-1} = k_{t-1}\big)\\ p_{\widehat{\theta}}(y_{0:t} \mid Z_{0:(t-1)} = k_{0:(t-1)}, Z_t = k) =& p_{\widehat{\theta}}(y_t \mid Z_{0:(t-1)} = k_{0:(t-1)}, Z_t = k, y_{0:(t-1)})\\ & \times p_{\widehat{\theta}}(y_{0:(t-1)} \mid Z_{0:(t-1)} = k_{0:(t-1)}, Z_t = k)\\ =& p_{\widehat{\theta}}(y_t \mid Z_t = k) \times p_{\widehat{\theta}}(y_{0:(t-1)} \mid Z_{0:(t-1)} = k_{0:(t-1)})\\ =& p_{\widehat{\theta}}(y_t \mid Z_t = k) \times p_{\widehat{\theta}}(y_{0:(t-1)} \mid Z_{0:(t-1)} = k_{0:(t-1)})\,.\end{aligned}$$

As a consequence, if we rename k_{t-1} by ℓ:

$$\begin{aligned}\delta_t(k) &= \max_{k_{0:(t-1)}} \pi_{k_{t-1}k} \times \mathbb{P}_{\widehat{\theta}}\big(Z_{0:(t-2)} = k_{0:(t-2)}, Z_{t-1} = k_{t-1}\big) p_{\widehat{\theta}}(y_t \mid Z_t = k) p_{\widehat{\theta}}(y_{0:(t-1)} \mid Z_{0:(t-1)} = k_{0:(t-1)})\\ &= \max_{k_{0:(t-2)},\ell} \pi_{\ell k} \mathbb{P}_{\widehat{\theta}}\big(Z_{0:(t-2)} = k_{0:(t-2)}, Z_{t-1} = \ell\big) p_{\widehat{\theta}}(y_t \mid Z_t = k) p_{\widehat{\theta}}(y_{0:(t-1)} \mid Z_{0:(t-2)} = k_{0:(t-2)}, Z_{t-1} = \ell)\\ &= p_{\widehat{\theta}}(y_t \mid Z_t = k)\\ &\quad \times \max_{\ell \in \{1,\dots,K\}} \left\{ \pi_{\ell k} \max_{k_{0:(t-2)}} p_{\widehat{\theta}}(y_{0:(t-1)} \mid Z_{0:(t-2)} = k_{0:(t-2)}, Z_{t-1} = \ell) \mathbb{P}_{\widehat{\theta}}\big(Z_{0:(t-2)} = k_{0:(t-2)}, Z_{t-1} = \ell\big) \right\}\\ &= p_{\widehat{\theta}}(y_t \mid Z_t = k) \max_{\ell \in \{1,\dots,K\}} \{\widehat{\pi}_{\ell k} \times \delta_{t-1}(\ell)\}\end{aligned}$$

And we obtain the forward formula for $\delta_t(k)$.

Backward. To prove that $k_T^* = \arg\max_{\ell \in \{1,\dots,K\}} \delta_T(\ell)$, let's introduce the function

$$\begin{array}{rccl} f_T: & \{1,\dots,K\} & \to & [0,1] \\ & \ell & \mapsto & \max\limits_{k_{0:T-1}} \mathbb{P}_{\widehat{\theta}}(Z_{0:T-1} = k_{0:T-1}, Z_T = \ell \mid Y_{0:T} = y_{0:T})\,, \end{array}$$

which gives, for any $\ell \in \{1,\dots,K\}$, the probability (conditionnaly to the observations) of the best sequence that ends by $Z_T = \ell$. Therefore, by definition, it reaches its maximum at k_T^*, the last term of $k_{0:T}^* = \arg\max_{k_{0:T}} \mathbb{P}_{\widehat{\theta}}(Z_{0:T} = k_{0:T} \mid y_{0:T})$. In other words,

$$k_T^* = \arg\max_{\ell \in \{1,\dots,K\}} f_T(\ell)\,.$$

Moreover:

$$\begin{aligned}\arg\max_{\ell \in \{1,\dots,K\}} f_T(\ell) &= \arg\max_{\ell \in \{1,\dots,K\}} \left(\max_{k_{0:T-1}} \mathbb{P}_{\widehat{\theta}}(Z_{0:T-1} = k_{0:T-1}, Z_T = \ell \mid Y_{0:T} = y_{0:T}) \right)\\ &= \arg\max_{\ell \in \{1,\dots,K\}} \left(\frac{1}{p_{\widehat{\theta}}(y_{0:T})} \times \max_{k_{0:T-1}} p_{\widehat{\theta}}(y_{0:T}, Z_{0:T-1} = k_{0:T-1}, Z_T = \ell) \right)\\ &= \arg\max_{\ell \in \{1,\dots,K\}} \delta_T(\ell)\,,\end{aligned}$$

where the last two simplifications hold as $p_{\widehat{\theta}}(y_{0:T})$ depend neither on $k_{0:T-1}$ nor ℓ, which proves the initial step of the backward procedure.

Let's now show that $k_{T-1}^* = \arg\max_{\ell \in \{1,\dots,K\}} \widehat{\pi}_{\ell k_T^*} \delta_{T-1}(\ell)$. Let's now define:

$$\begin{array}{rccl} f_{T-1}: & \{1,\dots,K\} & \to & [0,1] \\ & \ell & \mapsto & \max\limits_{k_{0:T-2}} \mathbb{P}_{\widehat{\theta}}(Z_{0:T-2} = k_{0:T-2}, Z_{T-1} = \ell, Z_T = k_T^* \mid Y_{0:T} = y_{0:T})\,, \end{array}$$

which gives, for any $\ell \in \{1, \dots, K\}$, the probability (conditionnaly to the observations) of the best sequence that ends by $Z_{T-1} = \ell$ and $Z_T = k_T^*$. Again, we have by definition:

$$k_{T-1}^* = \underset{\ell \in \{1,\dots,K\}}{\arg\max}\, f_{T-1}(\ell)\,.$$

Moreover:

$$\begin{aligned}
\underset{\ell \in \{1,\dots,K\}}{\arg\max}\, f_{T-1}(\ell) &= \underset{\ell \in \{1,\dots,K\}}{\arg\max} \left(\max_{k_{0:T-2}} \mathbb{P}_{\widehat{\theta}}(Z_{0:T-2} = k_{0:T-2}, Z_{T-1} = \ell, Z_T = k_T^* \mid Y_{0:T} = y_{0:T}) \right) \\
&= \underset{\ell \in \{1,\dots,K\}}{\arg\max} \left(\frac{1}{p_{\widehat{\theta}}(y_{0:T})} \times \max_{k_{0:T-2}} p_{\widehat{\theta}}(y_{0:T}, Z_{0:T-2} = k_{0:T-2}, Z_{T-1} = \ell, Z_T = k_T^*)) \right) \\
&= \underset{\ell \in \{1,\dots,K\}}{\arg\max} \left(p_{k_T^*,\widehat{\theta}}(y_T) \mathbb{P}_{\widehat{\theta}}(Z_T = k_T^* \mid Z_{T-1} = \ell) \times \max_{k_{0:T-2}} p_{\widehat{\theta}}(y_{0:T-1}, Z_{0:T-2} = k_{0:T-2}, Z_{T-1} = \ell) \right) \\
&= \underset{\ell \in \{1,\dots,K\}}{\arg\max} \left(\widehat{\pi}_{\ell k_T^*} \times \delta_{T-1}(\ell) \right),
\end{aligned}$$

where the term $p_{k_T^*,\widehat{\theta}}(y_T)$ vanishes in the last line as it does not depend on ℓ.

And we can back propagate this reasoning for every k_t^*.

[a] Again, using the conditional independences given by Figure 4.2

4.1.5 Selecting the number of hidden states

All previous computations were performed for a fixed number of components K. The choice of K is made following the exact same logic as in section 3.1 of Chapter 3. Note that, the number of observation parameters is $K \times \dim(\theta_{\text{obs}})$. For the latent Markov chain, because of the sum constraints, $\boldsymbol{\omega}_0$ and $\boldsymbol{\Pi}$ involve respectively $K-1$ and $K(K-1)$ free parameters (the rows of $\boldsymbol{\Pi}$ sum to 1). A s a consequence:

$$D_K = K\dim(\theta_{\text{obs}}) + K - 1 + K(K-1) = K\dim(\theta_{\text{obs}}) + (K+1)(K-1)\,.$$

4.1.6 Analysis of animal movement with HMM

4.1.6.1 Data and question

A common field of application of discrete hidden Markov models is the study of GPS data in ecology. A classical experiment in ecology is to mark animals with GPS tag in order to follow their displacements over time. Natural questions arise over the behavior of the animal, and therefore an early goal of such studies was whether different behaviors can be infered from movement data (see Morales et al. [2004] for the pioneer work in the domain).

Dataset 4.1 (Movement of *Morus bassanus* in Newfoundland)**.** *We illustrate the use of hidden Markov models to infer animal behavior from GPS data with data extracted from d'Entremont et al. [2022] and freely available on the online platform for movement data, MoveBank [Wikelski et al., 2024]. The authors tagged several gannets in southern Newfoundland during multiple years. We focus on a specific indivual for which we have 18 days of uninterrupted tracking, with a time step of roughly 15 minutes between two GPS locations. When a time step was missing, linear interpolation was performed (which was necessary for about 100 time steps). Overall, we focus on a trajectory of 1732 successive positions. For each pair of successive positions, the length of the step (in meters) separating those points was measure, resulting in a time series of 1731 successive step lengths. As we are mainly interested in orders of magnitude, we take the log in base 10 of this measure (therefore 1 corresponds to 10 meters, 2, to 100 meters, and so on...). Our goal is to identify distinct clusters in terms of step lengths during the trajectory.*

In this chapter, we focus on a univariate descriptor of the trajectory. In movement ecology, the most classical approach consists in taking two descriptors, the step length, and the turning angle. We refer to Etienne and Gloaguen [2022] for a description of these different descriptors, and the common related parametric probabilistic distributions.

t	Longitude	Latitude	log10(Step length)
2022-07-23 13:22:00	-54.187	46.815	2.723
2022-07-23 13:37:00	-54.191	46.811	2.723
2022-07-23 13:52:00	-54.195	46.807	2.201
2022-07-23 14:07:00	-54.197	46.807	2.085
2022-07-23 14:22:00	-54.199	46.807	1.966

Table 4.1: GPS data for *Morus bassanus* (data from d'Entremont et al. [2022])

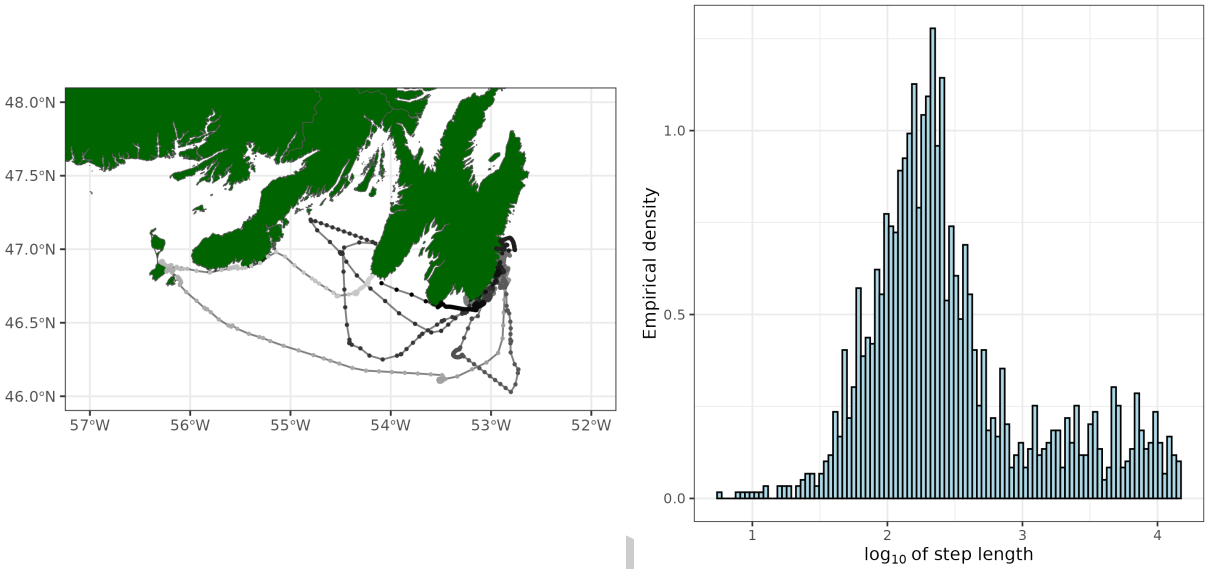


Figure 4.3: *Left*: Observed GPS track of a *Morus bassanus* during 18 days in south-east Newfoundland. The grey scale on the track corresponds to different times, with light gray indicating the beginning and black marking the end of the trajectory. *Right*: Empirical distribution of the logarithm (base 10) of step length (measured in meters) between two positions.

4.1.6.2 Infer animal behaviours through a hidden Markov model

Observation model In order to cluster the different observations through the observed times, we will fit a hidden Markov model like the one of Model (4.1). Our observations $y_{0:T}$ ($T = 1731$) are the $\log_{10}$ of the distance (in meters) covered by the bird in 15 minutes. We assume that this process consists in K different processes, consisting in different speed regimes, relative to the behaviour of the bird. We assume that for a given time t, within a regime, the observations are the realisation of a Gaussian distribution with a specific mean and specific variance for each cluster. Formally, we set that:

$$Y_t \mid \{Z_t = k\} \sim \mathcal{N}\big(\mu_k, \sigma_k^2\big)\,.$$

Inference The EM algorithm is performed as described in Section 4.1.1. Once the E step is performed, the M step has an analytical solution here (due to the simple observation distribution considered). More specically, at iteration h, for each $1 \leqslant j \leqslant K$, the mean and variance of the corresponding cluster is updated by:

$$\mu_j^{(h+1)} = \frac{1}{N_j^{(h)}} \sum_{t=0}^{T} \tau_{t,k}^{(h)} y_t\,,$$

$$\sigma_j^{2,(h+1)} = \frac{1}{N_k^{(h)}} \sum_{t=0}^{T} \tau_{t,k}^{(h)} \left(y_t - \mu_k^{(h+1)}\right)^2$$

where quantities $\tau_{t,k}^{(h)}$ are obtained by Proposition 4.5, and $N_k^{(h)} = \sum_{t=0}^{T} \tau_{t,k}^{(h)}$. Once convergence is reached, the best sequence of states for the estimated parameters is computed thanks to the Viterbi algorithm of Proposition 4.2.

Results

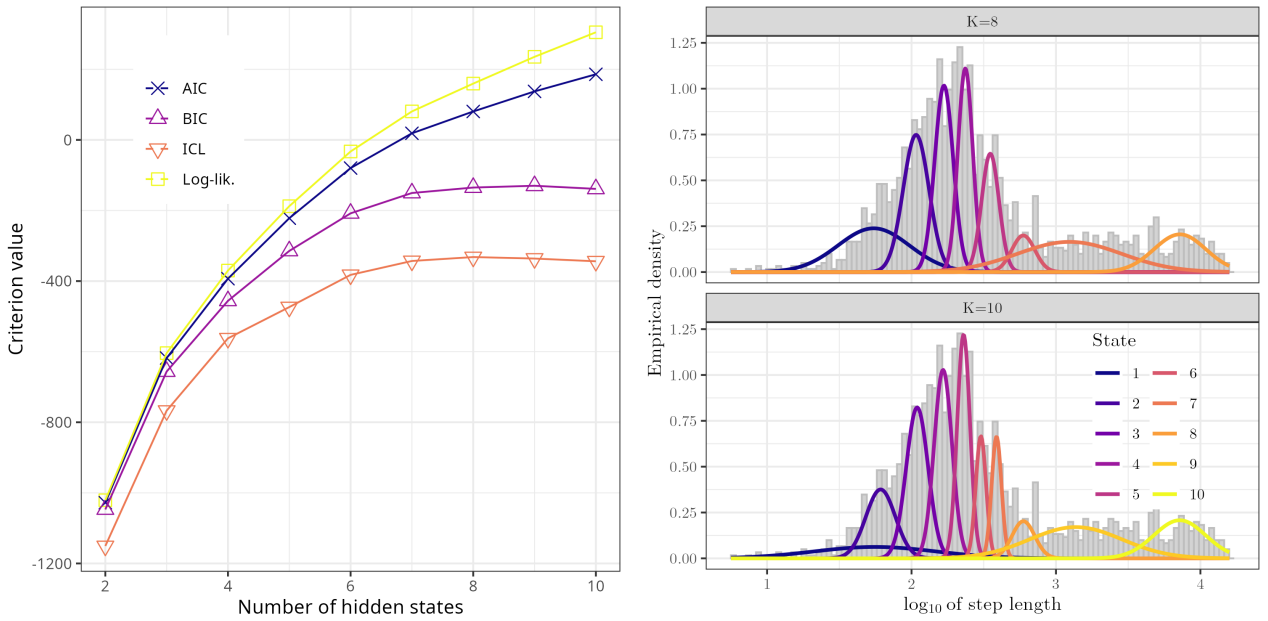


Figure 4.4: *Left*: Evolution of model selection criterions. Each criterion is the best result among 15 results of the EM algorithm, starting from 15 different starting points. *Right*: fit of the distribution for $K = 8$ and $K = 10$.

Choosing the number of components The proposed procedure was performed for $K \in \{2, \dots, 10\}$, and choosing the best final point among 15 random starting points. For the 9 models, we compute the AIC, BIC and ICL, together with the log-likelihood and show the results on the left hand side of Figure 4.4.

One can notice that relying on the AIC would lead to a more complex model than on the ICL (8 hidden states model) or the BIC (9 hidden states). The right hand side of Figure 4.4 shows the estimated observation distribution[3] in each state, for $K = 8$ and $K = 10$. One can see that those two models seem overcomplex, considering the global distribution of observations. This topic of potential overestimation of number of states in parametric HMM is largely discussed in Pohle et al. [2017], where the authors advocate for a model selection also guided by interpretation. This apparent overestimation of the number of components could also be induced by the misspecification of the observation conditional distribution (here, supposed to be gaussian). Specifically, assuming the same type distribution (here Gaussian) for each state might be unrealistic. A workaround to avoid this choice would be a non parametric framework for the observation distribution. Recent works focus on the choice of K in this context (see du Roy de Chaumaray et al. [2022] for instance).

In the following, we depict the results for $K = 6$, following an heuristic elbow rule on the curve of ICL of Figure 4.4.

Clustering results Focusing on the distribution for $K = 6$ on Figure 4.5, we can interpret the estimated speed regimes. Two extremes regimes (1 and 6) depicts the small displacement (less than 100 meters in 15 minutes) and large displacement (around 10 km in 15 minutes). Regime 5 will depict the displacement of about 1 km in 15 minutes, while regimes 2, 3, 4 will depict 3 different regimes between 130 and 350 meters, which represent the vast majority of observed steps.

An interesting feature is the alternance in the Markov chain between these states. The estimation of the transition matrix is provided on the left panel of Figure 4.6. One can interpret from the high values on the diagonal that every speed regime is rather persistent (the bird tends to stay in a given speed regime for more than one time step and that all transitions between states are not possible. The right hand side of Figure 4.6 shows a piece of the trajectory, with colors corresponding to the states sequence estimated using the Viterbi algorithm. This pictures also allows to observe the succession of speed regimes.

[3]The weight of each state is computed as its weight in the stationnary distribution induced by the transition matrix $\mathbf{\Pi}$.

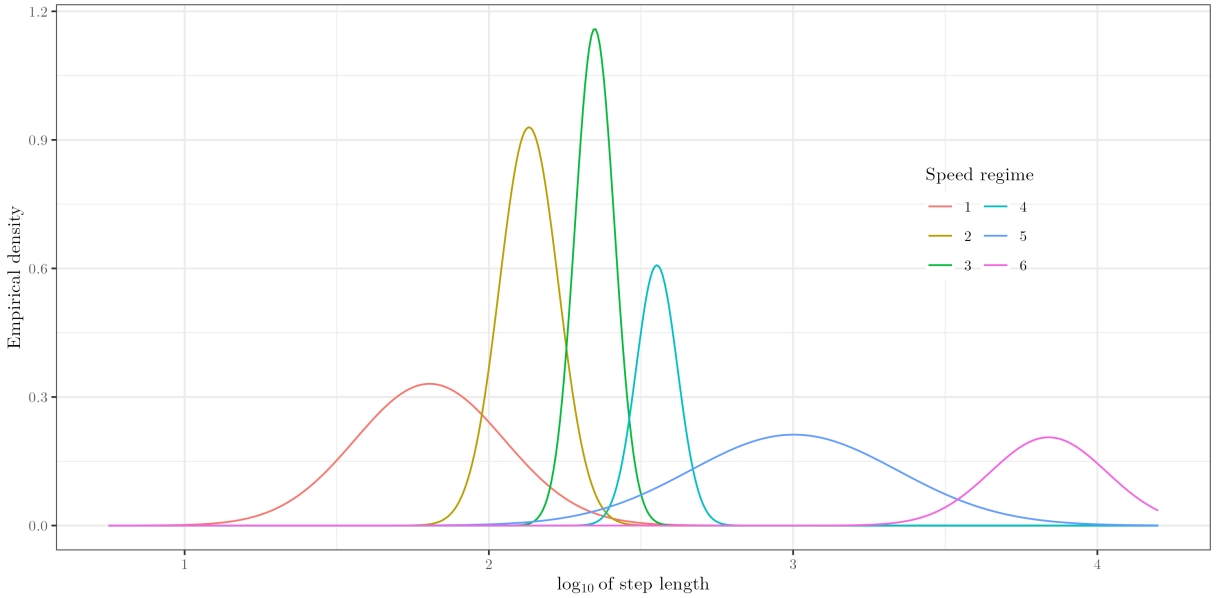


Figure 4.5: Distributions of the speed regime for $\widehat{K} = 6$.

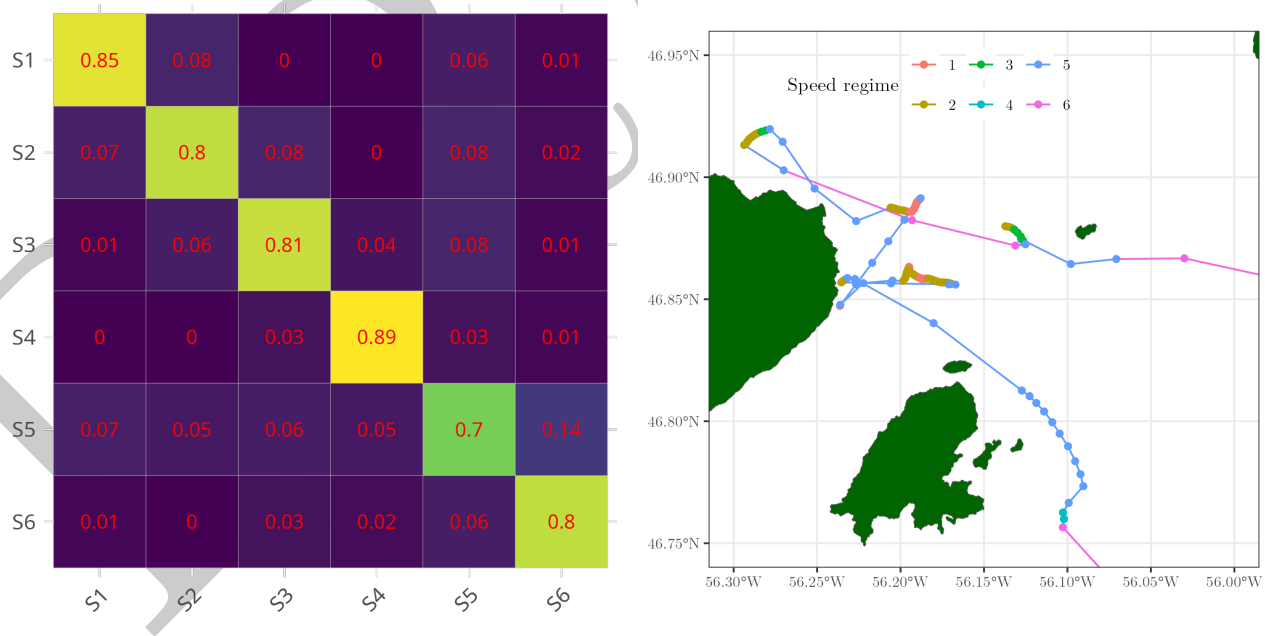


Figure 4.6: *Left*: estimated transition matrix $\widehat{\mathbf{\Pi}}$ between the 6 speed regimes (1 being the slowest, 6 the fastest). *Right*: focus on a specific part of the trajectory, colored by the estimated hidden speed regime (obtained using the Viterbi algorithm).

4.1.7 A discrete HMM to infer the genetic structure of a population

Hidden Markov Models (HMMs) are well-suited for modeling sequential data and have found powerful applications in population genetics. Although originally designed for temporal observations, HMMs can be naturally applied to genomic data by treating chromosomal position as the sequential axis. In this context, HMMs are used to model the genetic structure of populations by capturing patterns of ancestry, linkage disequilibrium, and recombination along the genome. For example, they can identify segments inherited from different ancestral populations or detect regions under selection, reflecting the structured mosaic nature of genomes shaped by evolutionary processes.

In this section, we consider the analysis of genetic diversity within a population of individuals from a same species, as introduced in Section 3.3. We only present here a generalisation of the mixture model presented in Section 3.3, with no further illustration.

4.1.7.1 A model for genetic data

The mixture model presented in Model 3.3 assumes that the whole genome of each individual originates from one single founding population (Assumption 1). Although this assumption makes sense for neighbor loci along the genome, because of possible past crossings between sub-population, it does probably not hold at the whole genome scale. Hidden Markov models introduced in Section 4.1.1 provide a natural framework to allow each loci of each individual to originate from its own founding population, while accounting for the dependency between neighbour loci.

Model 4.2 (HMM for the genetic structure of a population)**.** *We use same notations for the observed data as in Model 3.3 and now denote by Z_{ij} the population of origin of locus j for individual i. The hidden Markov model states that each $Z_i = (Z_{ij})_{1\le j\le p}$ forms a Markov chain and that the hidden path $\{Z_i\}_{1\le i\le n}$ are all independent:*

$$\begin{aligned} Z_i &\overset{iid}{\sim} \mathsf{MarkovChain}(\boldsymbol{\omega}_0, \boldsymbol{\Pi}), && 1 \le i \le n\,, \\ Y_{ij} \mid \{Z_{i,j} = z_{i,j}\} &\overset{ind}{\sim} \mathrm{Cat}\left(\boldsymbol{\phi}_{\boldsymbol{z_{ij}j}}\right) && 1 \le i \le n, 1 \le j \le n\,. \end{aligned}$$

where, for $1 \le k \le K$ $\boldsymbol{\phi_{kj}}$ is given in (3.24) *of Model 3.3. The parameter γ has the same interpretation as in Model 3.3 and*

- $\boldsymbol{\omega}_0 = [\omega_{0k}]_{1\le k\le K}$ *is the vector of probabilities for the first locus of each individual to come from a given population,*
- $\boldsymbol{\Pi} = [\pi_{k\ell}]_{1\le k,\ell\le K}$ *is the transition matrix:* $\pi_{k\ell} = \mathbb{P}(Z_{i,j+1} = \ell \mid Z_{i,j} = k)$,

which leads to

$$\theta = (\boldsymbol{\omega}_0, \boldsymbol{\Pi}, \gamma)\,.$$

The latent variables are the locus-specific memberships Z_{ij}, the observed variables are the genotypes Y_i.

Remarks.

- *Note that Model 3.3 considered that we only observed one trajectory along time (noted t). Here, we observe n trajectories of the HMM ($i = 1, \dots, n$) and the observations are not along time (t) but along the place on the genome (j). Another difference is that the emission distribution ϕ_{kj} depends on the location along the genome.*
- *When the loci are spread over several chromosome, the latent path along each chromosome can be assumed to be independent. Also, the latent Markov chain is not necessarily homogeneous, and the transition probabilities may account for the distance separating each locus from the next. Note that Model 4.2 does not apply to Taita thrush dataset from Example 3.3, because of the very small number of markers ($p = 7$) and because their respective locations along the genome are unknown.*

Assumptions. As compared to the mixture Model 3.3, the hidden Markov Model 4.2 does not make Assumption 1 (unique population of origin), but still relies on Assumption 3 (Hardy-Weinberg principle). As for Assumption 2 (conditional independence of the genotypes at each locus), it still holds but each locus has its own population of origin and the population of origin of neighbour loci are likely to be the same.

4.1.7.2 Inference

Graphical model and likelihood. Because Model 4.2 is a regular HMM, its graphical model is the one depicted in Figure 4.2 and its complete log-likelihood is a sum over all individuals of complete log-likelihood as given Equation (4.4). More specifically,

$$\log p_\theta(\boldsymbol{y}, \boldsymbol{Z}) = \sum_{i=1}^{n} \left(\log \boldsymbol{\omega}_{0,Z_{i,1}} + \sum_{j=1}^{p-1} \log \pi_{Z_{i,j},Z_{i,j+1}} + \sum_{j=1}^{p} \log \phi_{Z_{i,j},j}(y_{ij}) \right). \tag{4.19}$$

EM algorithm.

E step. Because of the similarity between Equations (4.4) and (4.19), the quantities to be evaluated at the E step, for each individual i, are therefore the same as these defined in Equations (4.7) and (4.8), so the forward-backward recursion given by Proposition 4.5 applies individual by individual.

M step. The update formulas for $\boldsymbol{\omega}_0$ and $\boldsymbol{\Pi}$ are the same as in Section 4.1.1 and the update formulas for the allelic frequencies content in γ can be easily adapted from these given in Section 3.3.

Model selection. As for the mixture Model 3.3, the number of founding populations K from the hidden Markov Model 4.2 can be determined using the BIC or ICL criteria presented in Section 2.4.

4.2 Continuous HMM for correction of animal location (PG)

While discrete HMMs have proven effective for modeling latent state dynamics such as behavioral states in animal movement, they are limited by their inherently categorical state structure. This discrete-state framework is well-suited for systems characterized by abrupt transitions between qualitatively distinct states. However, many ecological processes are better represented by continuous latent dynamics governed by linear or approximately linear processes.

Continuous HMMs provide a natural extension in such contexts, offering a principled approach for estimating unobserved continuous states from noisy observations under the assumptions of linearity and Gaussian noise. The Kalman filtering [4] Kalman [1960] is widely applied in ecology when that observations are corrupted by noise, which can, for instance, come from observation error in the device. As a common example are animal tracking data (coming from GPS, ARGOS, or any telemetry device) where the observed times series often come with an error. A challenge is therefore to retrieve the actual position of animal from the corrupted observations.

This algorithm relies on a probabilistic model that is again a latent variable model and can be seen as sequential version of the linear mixed model or an HMM with continuous latent space model (see Roweis and Ghahramani [1999] for a unifying point of view on these models).

4.2.1 Data and question

Dataset 4.2 (Correction of GPS observations)**.** *In movement ecology, animal tracking using GPS devices is prone to measurement errors. A challenge is therefore to remove this error (often called noise in signal processing) from the observed trajectory. We illustrate this approach using a dataset from Silva et al. [2014], and available on the MoveBank website, consisting in trajectories of fin whales (Balaenoptera physalus) in the North Atlantic ocean. The data then consists in bidimensional time series, which is plotted on Figure 4.7*

4.2.2 The linear Gaussian hidden Markov model

Notations We consider an observed series $\boldsymbol{y} = \{y_t\}_{0 \leq t \leq T}$ where $y \in \mathbb{R}^{d_y}$. We assume that these observations are realizations of random variables whose distribution depend on latent variables $Z_{0:T}$. Moreover, we make the following assumptions:

- The stochastic process $\{Z_t\}_{t \in \mathbb{N}}$ in $\mathbb{R}^{d_z}$ is an homogeneous[5] Markov process having the following properties:
  - It's initial distribution is a Gaussian distribution in $\mathbb{R}^{d_z}$ with mean μ_0 and variance V_0;

[4] or, as we will see later, the Kalman smoother

[5] This homogeneity assumption can be relaxed, but is assumed here to keep simple notations.

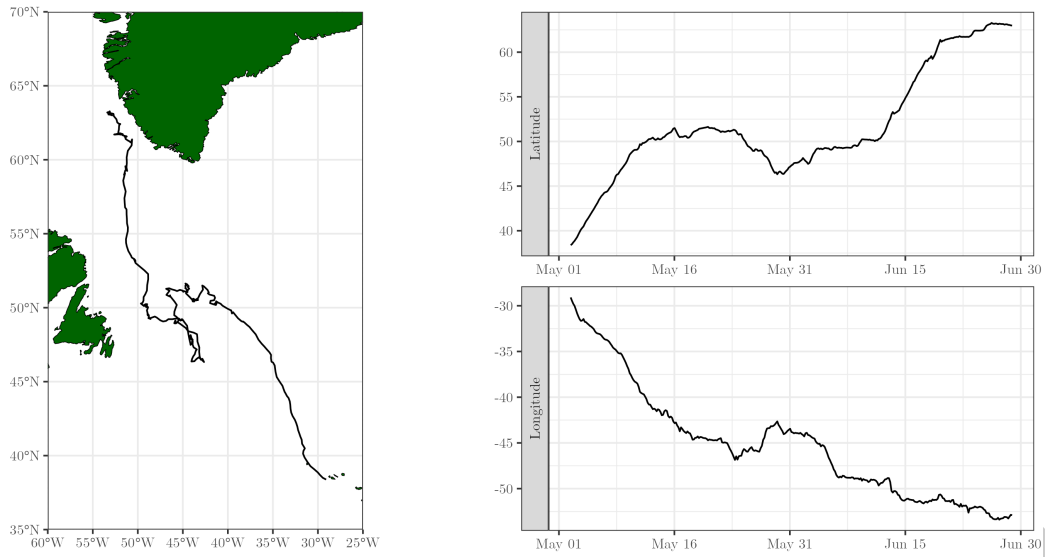


Figure 4.7: *Left:* Observed trajectory of a fin whale (*Balaenoptera physalus*). *Right:* Representation as bidimensional time series. Data extracted from Silva et al. [2014] and available on `movebank.org`.

  - The conditional distribution of Z_{t+1} knowing Z_t is also Gaussian, whose variance does not depend on Z_t and mean depends linearly on Z_t.

- Conditionally to $\{Z_t\}_{t\in\mathbb{N}}$, observations $Y_0, \dots Y_T$ are independent and follow a Gaussian distribution whose variance does not depend on Z_t and mean depends linearly on Z_t.

Model 4.3 (Linear Gaussian hidden Markov model)**.**

$$
\begin{aligned}
Z_0 &\sim \mathcal{N}_{d_z}(\mu_0, V_0) \\
Z_t \mid Z_{t-1} &= BZ_{t-1} + E_t \text{ where } E_t \overset{ind}{\sim} \mathcal{N}_{d_z}(0, \Omega) \ , \quad E_t \perp\!\!\!\perp Z_{t-1} \ t \in \mathbb{N}^* \\
Y_t \mid Z_t &\sim \mathcal{N}_{d_y}(AZ_t, \Sigma), \ t \in \mathbb{N},
\end{aligned}
$$

where A and B are matrices of dimension $d_y \times d_z$ and $d_z \times d_z$. In this context $\theta_{obs} = \{A, \Sigma\}$, $\theta_{lat} = \{\mu_0, V_0, \Omega, B\}$, $\theta = \{\theta_{lat}, \theta_{obs}\}$.

Graphical model. The DAG for Model 4.3 is the same as that of the HMM model (Figure 4.2). One can now see the full dependence between observations, and the independence conditionally to the hidden process.

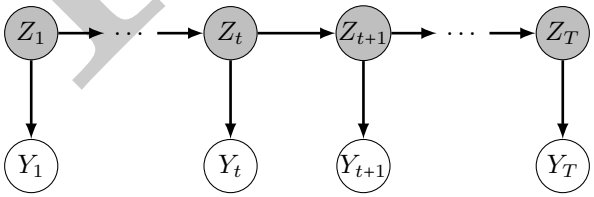


Figure 4.8: Graphical representation of Model 4.3.

Identifiability. Similarly to probabilistic PCA, Model 4.3 is not statistically identifiable, as two different parameter can lead to two different models. Indeed, as Ω is a (real) symetric matrix, it exists a rotation matrix R and a diagonal matrix D such that $\Omega = RDR^\top$. Therefore, let a model be parameterized as Model 4.3, the model defined by

$$
\begin{aligned}
\tilde{Z}_t \mid \tilde{Z}_{t-1} &= D^{-\frac{1}{2}} P^\top B P D^{-\frac{1}{2}} \tilde{Z}_{t-1} + \tilde{E}_t \text{ where } \tilde{E}_t \overset{\text{ind}}{\sim} \mathcal{N}_{d_z}(0, \mathbf{I}) \ , \quad E_t \perp\!\!\!\perp Z_{t-1} \ t \in \mathbb{N}^* \\
Y_t \mid Z_t &\sim \mathcal{N}_{d_y}\left(APD^{\frac{1}{2}} \tilde{Z}_t, \Sigma\right), \ t \in \mathbb{N},
\end{aligned}
$$

leads to the same distribution of observations. Therefore, one needs to pose constraint on either A, B or Ω, for instance, as above, that $\Omega = \mathbf{I}$.

4.2.3 Marginal and complete log-likelihoods of the linear Gaussian HMM

Marginal distribution of observations The following proposition states that marginally $Y_{0:T}$ is a multivariate Gaussian vector.

Proposition 4.6. *Let $Y_{0:T}$ be a distributed as Model 4.3. Then*

$$\begin{aligned} Z_{0:T} &\sim \mathcal{N}_{d_z \times (T+1)}(\mu_{0:T}, \Omega_{0:T}) \,, \\ Y_{0:T} \mid Z_{0:T} &\sim \mathcal{N}_{d_y \times (T+1)}((I_{T+1} \otimes A) Z_{0:T}, I_{T+1} \otimes \Sigma) \,. \end{aligned}$$

where $\otimes$ is the Kronecker product of matrices[a]. Thus, the marginal distribution of $Y_{0:T}$ is fully known and is a Gaussian distribution:

$$Y_{0:T} \sim \mathcal{N}_{d_y \times (T+1)}\big((I_{T+1} \otimes A)\mu_{0:T}, I_{T+1} \otimes \Sigma + (I_{T+1} \otimes A)\Omega_{0:T}(I_{T+1} \otimes A)^\intercal\big) \,. \tag{4.20}$$

[a] $I_{T+1} \otimes A$ is the block-diagonal matrix where the diagonal is composed of $T+1$ times the matrix A

Proof of Proposition 4.6

Let's first consider the sequence $Z_{0:T}$. We can rewrite the second line of Model 4.3 as:

$$Z_t \mid Z_{t-1} \sim \mathcal{N}_{d_z}(BZ_{t-1}, \Omega) \; t \in \mathbb{N}^*$$

By Proposition A.6, we can deduce the distribution of Z_1. Its marginal distribution is a Gaussian distribution with:

$$\mathbb{E}[Z_1] = B\mu_0, \qquad \mathbb{V}[Z_1] = V_0 + B\Omega B^\intercal =: V_1 \,.$$

By induction, we have that for every $t \in \mathbb{N}^*$:

$$\mathbb{E}[Z_t] = B^t\mu_0, \qquad \mathbb{V}[Z_t] = V_{t-1} + B\Omega B^\intercal =: V_t \,.$$

where B^t is the matrix B to the power t. Moreover, noticing that for all $t \in \mathbb{N}^*$

$$Z_t = BZ_{t-1} + E_t \,,$$

where E_t is a Gaussian noise (with mean 0 and variance Ω) independent of Z_{t-1}, we have that:

$$\mathbb{C}\mathrm{ov}\,(Z_{t-1}, Z_t) = \mathbb{C}\mathrm{ov}\,(Z_{t-1}, BZ_{t-1} + E_t) = \mathbb{V}[Z_{t-1}]B^\intercal + \mathbb{C}\mathrm{ov}\,(Z_{t-1}, E_t) = V_{t-1}B^\intercal \,.$$

More generally, for $(t, h) \in \mathbb{N}^2$, the covariance between Z_t and Z_{t+h} is given by:

$$C_{t,h} := \mathbb{C}\mathrm{ov}\,(Z_t, Z_{t+h}) = V_t(B^h)^\intercal \,.$$

Thus, we completely characterized the distribution of $Z_{0:T}$, and we have:

$$Z_{0:T} \sim \mathcal{N}_{d_z \times (T+1)}(\mu_{0:T}, \Omega_{0:T}) \,,$$

where

$$\mu_{0:T} = \begin{pmatrix} \mu_0 \\ B\mu_0 \\ \vdots \\ B^T\mu_0 \end{pmatrix} \qquad \text{and} \qquad \Omega_{0,T} = \begin{pmatrix} V_0 & C_{0,1} & \dots & C_{0,T} \\ C_{0,1}^\intercal & V_1 & \dots & C_{1,T-1} \\ \vdots & & \ddots & \vdots \\ C_{0,T}^\intercal & \dots & \dots & V_T \end{pmatrix} .$$

Thus, as observations are indepenndent conditionally to latent variables, Model 4.3 is equivalent to:

$$Z_{0:T} \sim \mathcal{N}_{d_z \times (T+1)}(\mu_{0:T}, \Omega_{0:T}),$$
$$Y_{0:T} \mid Z_{0:T} \sim \mathcal{N}_{d_y \times (T+1)}((I_{T+1} \otimes A) Z_{0:T}, I_{T+1} \otimes \Sigma).$$

We deduce the marginal distribution of $Y_{0:T}$ from Proposition A.6.

Marginal log-likelihood As a consequence of Proposition 4.6, the marginal log-likelihood of the observations $\boldsymbol{y} = \{y_t\}_{0 \leq t \leq T}$ can theoretically be computed directly for any set of parameters. Moreover, Proposition A.5 also gives us the conditional distribution of $Z_{0:T}|Y_{0:T}$. However, computing both the likelihood and the conditional distribution implies to inverse large matrices –of dimension $d_y \times (T+1)$ or $d_z \times (T+1)$– which would lead to numerical instability for large T. This justifies alternative computation of these two quantities. As in Section 4.1.1, there is an efficient way of computing the log-likelihood using an iterative *forward* process named the Kalman filter algorithm, whereas the conditional distribution will be efficiently compute through an iterative *backward* process in the next distribution.

Proposition 4.7 (Kalman filter and log-likelihood in Model 4.3.)**.** *Under Model 4.3, the filtering distributions (Definition 4.1) are:*

- *For* $t = 0$, $Z_0 \mid Y_0 \sim \mathcal{N}_{d_z}\left(\mu_{0|0}, \Omega_{0|0}\right)$ *where:*

$$\begin{cases} \mu_{0|0} &= \mu_0 + K_0 (Y_0 - A\mu_0) \\ \Omega_{0|0} &= (I - K_0 A) V_0 \end{cases} \quad \text{where} \quad K_0 := V_0 A^\top (\Sigma + A V_0 A^\top)^{-1}. \tag{4.21}$$

  The matrix K_0 is called the Kalman gain and quantifies the reduction in variance (our gain) by including Y_0 in the prediction of Z_0.

- *For* $0 \leq t \leq T-1$, $Z_{t+1} \mid Y_{0:t+1} \sim \mathcal{N}_{d_z}\left(\mu_{t+1|t+1}, \Omega_{t+1|t+1}\right)$ *where:*

$$\begin{cases} \mu_{t+1|t+1} &= B\mu_{t|t} + K_{t+1}(Y_{t+1} - AB\mu_{t|t}) \\ \Omega_{t+1|t+1} &= (I - K_{t+1}A) W_{t+1} \end{cases} \text{where} \begin{cases} W_{t+1} &:= \Omega + B\Omega_{t|t}B^\top \\ K_{t+1} &:= W_{t+1}A^\top(\Sigma + AW_{t+1}A^\top)^{-1} \end{cases} \tag{4.22}$$

Moreover, we have:

$$Y_0 \sim \mathcal{N}_{d_y}\left(A\mu_0, \Sigma + AV_0A^\top\right) \tag{4.23}$$
$$Y_{t+1} \mid Y_{0:t} \sim \mathcal{N}_{d_y}\left(AB\mu_{t|t}, \Sigma + AW_{t+1}A^\top\right) \qquad 0 \leq t \leq T-1, \tag{4.24}$$

which allows the computation of the log-likelihood through Equation (4.2)*.*

Proof of Proposition 4.7

First, at $t = 0$, we have, by definition of Model 4.3:

$$Z_0 \sim \mathcal{N}_{d_z}(\mu_0, V_0), \tag{4.25}$$
$$Y_0 \mid Z_0 \sim \mathcal{N}_{d_y}(AZ_0, \Sigma). \tag{4.26}$$

Equation (4.21) is then a direct application of Proposition A.5.

Now, let's suppose we have computed $(\mu_{t|t}, \Omega_{t|t})$ for $t \geqslant 0$. We consider the following intermediate model for the triplet $(Z_t, Z_{t+1}, Y_{t+1}) \mid \{Y_{0:t} = y_{0:t}\}$:

$$Z_t \mid \{Y_{0:t} = y_{0:t}\} \sim \mathcal{N}_{d_z}\left(\mu_{t|t}, \Omega_{t|t}\right), \tag{4.27}$$
$$Z_{t+1} \mid \{Z_t = z_t, Y_{0:t} = y_{0:t}\} \sim \mathcal{N}_{d_z}(Bz_t, \Omega), \tag{4.28}$$
$$Y_{t+1} \mid \{Z_{t+1} = z_{t+1}, Y_{0:t} = y_{0:t}\} \sim \mathcal{N}_{d_y}(Az_{t+1}, \Sigma). \tag{4.29}$$

Notice that:

- this right hand sides of (4.28) and (4.29) actually do not depend on $y_{0:t}$, which is a consequence of

the conditional dependence structure of the initial model. Nonetheless, we kept the dependence on the left hand side to highlight that this model is conditional to $Y_{0:t}$.

- for $t = 0$, it is a modified version of Model 4.3 where the initial distribution of Z_0 has been updated according to y_0.

Thanks to Proposition A.6, we can compute the marginal $Z_{t+1} \mid Y_{0:t}$ and we have:

$$Z_{t+1} \mid \{Y_{0:t} = y_{0:t}\} \sim \mathcal{N}_{d_z}\left(B\mu_{t|t}, \Omega + B\Omega_{t|t}B^\top\right).$$

This distribution (that does depends on $Y_{0:t}$, thanks to marginalization) is often called the *posterior predictive* distribution, and this intermediate step is often referred to as the *predict step* of the Kalman filter. Then, as $Y_{t+1} \mid Z_{t+1} \overset{\text{Law}}{=} Y_{t+1} \mid (Z_{t+1}, Z_t, Y_{0:t})$, our intermediate model is equivalent to:

$$Z_{t+1} \mid \{Y_{0:t} = y_{0:t}\} \sim \mathcal{N}_{d_z}\left(B\mu_{t|t}, \Omega + B\Omega_{t|t}B^\top\right), \tag{4.30}$$

$$Y_{t+1} \mid \{Z_{t+1}\} \sim \mathcal{N}_{d_y}(AZ_{t+1}, \Sigma). \tag{4.31}$$

Equation (4.22) is again the straightforward application of Proposition A.5 to this model.

Turning to Equation (4.23), the marginal distribution of Y_0 is obtained using the Model (4.25)-(4.26) and Proposition A.6.

Finally, the distribution of $Y_{t+1} \mid \{Y_{0:t} = y_{0:t}\}$ is obtained thanks to Proposition A.6, as it is the marginal distribution in the Model given by Equations (4.30)-(4.31). This results in Equation (4.24) and concludes the proof.

Complete log-likelihood It was already stated in Section 4.1.1, Equation (4.4), that the complete log-likelihood satisfies:

$$\begin{aligned}\log p_\theta(\boldsymbol{y}, \boldsymbol{Z}) = \log p_\theta(y_{0:T}, Z_{0:T}) &= \log p_\theta(Z_0) + \log p_\theta(y_0 \mid Z_0) + \sum_{t=0}^{T-1} \left(\log p_\theta(y_{t+1} \mid Z_{t+1}) + \log p_\theta(Z_{t+1} \mid Z_t)\right) \\ &= \log p_\theta(Z_0) + \sum_{t=0}^{T-1} \log p_\theta(Z_{t+1} \mid Z_t) + \sum_{t=0}^{T} \log p_\theta(y_t \mid Z_t).\end{aligned}$$

In the context of linear Gaussian HMM, all the involved terms will be quadratic forms in $(Z_t)_{0\leqslant t\leqslant T}$, resulting in:

$$\begin{aligned}\log p_\theta(y_{0:T}, Z_{0:T}) = &-\frac{1}{2}\log|V_0| - \frac{1}{2}(Z_0 - \mu_0)^\top V_0^{-1}(Z_0 - \mu_0) \\ &- \frac{T-1}{2}\log|\Omega| - \frac{1}{2}\sum_{t=0}^{T-1}(Z_{t+1} - BZ_t)^\top \Omega^{-1}(Z_{t+1} - BZ_t) \\ &- \frac{T}{2}\log|\Sigma| - \frac{1}{2}\sum_{t=0}^{T}(y_t - AZ_t)^\top \Sigma^{-1}(y_t - AZ_t) + \text{cst}.\end{aligned} \tag{4.32}$$

4.2.4 EM algorithm for the linear Gaussian HMM

Once again, we slowly detail each step of the EM and provide the complete EM at the end of the section.

4.2.4.1 Objective function $Q\left(\theta \mid \theta^{(h)}\right)$

For a current parameter $\theta^{(h)}$, we have to compute from (4.32)

$$Q\left(\theta \mid \theta^{(h)}\right) = \mathbb{E}_{\theta^{(h)}}\left[\log p_\theta(Y_{0:T}, Z_{0:T}) \mid \boldsymbol{Y} = \boldsymbol{y}\right],$$

where $\boldsymbol{Y} = Y_{0:T}$ and $\boldsymbol{y} = y_{0:T}$. Notice that it has the same form as Equation (4.5) in the discrete HMM case. Developing for this specific case, it results in:

$$
\begin{aligned}
Q\big(\theta \mid \theta^{(h)}\big) =& \text{cst} - \frac{1}{2}\left(\log|V_0| + (T-1)\log|\Omega| + T\log|\Sigma| + \mu_0^\top V_0^{-1}\mu_0 + \sum_{t=0}^{T} y_t^\top \Sigma^{-1} y_t\right) \\
&- \mu_0^\top V_0^{-1}\mathbb{E}_{\theta^{(h)}}[Z_0 \mid \boldsymbol{Y} = \boldsymbol{y}] - \sum_{t=0}^{T} y_t^\top \Sigma^{-1} A \mathbb{E}_{\theta^{(h)}}[Z_t \mid \boldsymbol{Y} = \boldsymbol{y}] \\
&- \sum_{t=0}^{T-1} \operatorname{tr}\left(\Omega^{-1}B \times \mathbb{E}_{\theta^{(h)}}\left[Z_t Z_{t+1}^\top \mid \boldsymbol{Y} = \boldsymbol{y}\right]\right) \\
&- \frac{1}{2}\operatorname{tr}\left(\left(V_0^{-1} + A^\top\Sigma^{-1}A\right)\mathbb{E}_{\theta^{(h)}}\left[Z_0 Z_0^\top \mid \boldsymbol{Y} = \boldsymbol{y}\right]\right) - \frac{1}{2}\operatorname{tr}\left(\left(\Omega^{-1} + A^\top\Sigma^{-1}A\right) \times \mathbb{E}_{\theta^{(h)}}\left[Z_T Z_T^\top \mid \boldsymbol{Y} = \boldsymbol{y}\right]\right) \\
&- \frac{1}{2}\sum_{t=1}^{T-1} \operatorname{tr}\left(\left(\Omega^{-1} + A^\top\Sigma^{-1}A + B^\top\Omega^{-1}B\right) \times \mathbb{E}_{\theta^{(h)}}\left[Z_t Z_t^\top \mid \boldsymbol{Y} = \boldsymbol{y}\right]\right).
\end{aligned}
$$

The key takeaway from this discussion is that the objective function depends on the marginal smoothing distributions (see Definition 4.1), specifically on their mean (line 2 of the equation above) and second-order moment (lines 3 to 5). Specifically, remembering that:

$$
\begin{aligned}
\mathbb{E}_{\theta^{(h)}}\left[Z_t Z_t^\top \mid \boldsymbol{Y} = \boldsymbol{y}\right] &= \mathbb{V}_{\theta^{(h)}}[Z_t \mid \boldsymbol{Y} = \boldsymbol{y}] + \mathbb{E}_{\theta^{(h)}}[Z_t \mid \boldsymbol{Y} = \boldsymbol{y}]\mathbb{E}_{\theta^{(h)}}[Z_t \mid \boldsymbol{Y} = \boldsymbol{y}]^\top, \\
\mathbb{E}_{\theta^{(h)}}\left[Z_t Z_{t+1}^\top \mid \boldsymbol{Y} = \boldsymbol{y}\right] &= \mathbb{C}\text{ov}_{\theta^{(h)}}\left(Z_t, Z_{t+1} \mid \boldsymbol{Y} = \boldsymbol{y}\right) + \mathbb{E}_{\theta^{(h)}}[Z_t \mid \boldsymbol{Y} = \boldsymbol{y}]\mathbb{E}_{\theta^{(h)}}[Z_{t+1} \mid \boldsymbol{Y} = \boldsymbol{y}]^\top,
\end{aligned}
$$

we have to compute (for any $t \in \{0, \dots, T\}$ and a current parameter θ):

$$
\mu_{t|T}^{(h)} := \mathbb{E}_{\theta^{(h)}}[Z_t \mid \boldsymbol{Y} = \boldsymbol{y}] \qquad \text{Smoothing mean,} \tag{4.33}
$$

$$
\Omega_{t|T}^{(h)} := \mathbb{V}_{\theta^{(h)}}[Z_t \mid \boldsymbol{Y} = \boldsymbol{y}] \qquad \text{Smoothing variance.} \tag{4.34}
$$

Moreover, for $t \in \{0, \dots, T-1\}$, we need

$$
\Omega_{t,t+1|T}^{(h)} := \mathbb{C}\text{ov}_{\theta^{(h)}}\left(Z_t, Z_{t+1} \mid \boldsymbol{Y} = \boldsymbol{y}\right) \qquad \text{Smoothing auto-covariance.} \tag{4.35}
$$

These quantities can be computed efficiently using the following Kalman smoother, which builds on the Kalman filter from Proposition 4.7 and proceeds via a backward recursion.

Proposition 4.8 (Kalman smoother)**.** *Under Model 4.3, the smoothing distributions, i.e., the distributions of $Z_t \mid Y_{0:T}$ (for $0 \leqslant t \leqslant T$) are given by Gaussian distributions with mean $\mu_{t|T}$ and variance $\Omega_{t|T}$ such that:*

- *For $t = T$, $\mu_{T|T}$ and $\Omega_{T|T}$ are given by the filtering distribution at time T (given by Proposition 4.7);*
- *For $T - 1 \geqslant t \geqslant 0$:*

$$
\begin{cases} \mu_{t|T} &= \mu_{t|t} + \widetilde{K}_t\left(\mu_{t+1|T} - B\mu_{t|t}\right) \\ \Omega_{t|T} &= \Omega_{t|t} + \widetilde{K}_t\left(\Omega_{t+1|T} - W_{t+1}\right)\widetilde{K}_t^\top, \end{cases} \qquad \text{where} \qquad \widetilde{K}_t = \Omega_{t|t}B^\top W_{t+1}^{-1}, \tag{4.36}
$$

and the other quantities ($\mu_{t|t}$ and W_{t+1}) are defined in Proposition 4.7. Moreover, we have that:

$$
\Omega_{t,t+1|T} = \widetilde{K}_t \Omega_{t+1|T}. \tag{4.37}
$$

Proof of Proposition 4.8

For $t = T$, the expression of $\mu_{T|T}$ and $\Omega_{T|T}$ are indeed given by the Kalman filter, as the marginal filtering and smoothing distributions coincides at time T (see Definition 4.1).

For $T - 1 \geqslant t \geqslant 0$ will then proceed by a backward induction. Working with densities, we have, for $0 \leqslant t \leqslant T - 1$,

$$
\begin{aligned}
p_\theta(z_t \mid y_{0:T}) &= \int_{\mathbb{R}^{d_z}} p_\theta(z_t, z_{t+1} \mid y_{0:T}) \mathrm{d}z_{t+1} \\
&= \int_{\mathbb{R}^{d_z}} p(z_t \mid z_{t+1}, y_{0:T}) p_\theta(z_{t+1} \mid y_{0:T}) \mathrm{d}z_{t+1}.
\end{aligned}
$$

Taking advantage of conditional independence, we have:

$$p_\theta(z_t \mid y_{0:T}) = \int_{\mathbb{R}^{d_z}} p_\theta(z_t \mid z_{t+1}, y_{0:t}) p_\theta(z_{t+1} \mid y_{0:T}) \mathrm{d}z_{t+1} \,. \tag{4.38}$$

We recognize in the integral the distribution of $Z_t | \{Z_{t+1} = z_{t+1}, Y_{0:t} = y_{0:t}\}$ that can be obtained using Proposition A.5 on Equations (4.27) and (4.28):

$$Z_t \mid \{Z_{t+1} = z_{t+1}, Y_{0:t} = y_{0:t}\} \sim \mathcal{N}_{d_z}\left(\mu_{t|t} + \widetilde{K}_t \left(z_{t+1} - B\mu_{t|t}\right), \left(I - \widetilde{K}_t B\right)\Omega_{t|t}\right), \tag{4.39}$$

where we defined

$$\widetilde{K}_t = \Omega_{t|t} B^\top (\Omega + B\Omega_{t|t}B^\top)^{-1} = \Omega_{t|t} B^\top W_{t+1}^{-1} \,,$$

which can be computed after the forward pass. Equation (4.39) defines the backward analogous of the predict step. It is known in the hidden Markov literature (see Cappé et al. [2005], for instance) as the *backward kernel*. This backward model defines a new model to move backward from the smoothing distribution at time T. This model is defined by the following equations, for every $0 \leqslant t \leqslant T-1$:

$$Z_{t+1} \mid \{\boldsymbol{Y} = \boldsymbol{y}\} \sim \mathcal{N}_{d_z}\left(\mu_{t+1|T}, \Omega_{t+1|T}\right), \tag{4.40}$$

$$Z_t \mid \{Z_{t+1} = z_{t+1}, \boldsymbol{Y} = \boldsymbol{y}\} \sim \mathcal{N}_{d_z}\left(\mu_{t|t} + \widetilde{K}_t \left(Z_{t+1} - B\mu_{t|t}\right), \left(I - \widetilde{K}_t B\right)\Omega_{t|t}\right), . \tag{4.41}$$

Then, the marginal distribution of Z_t in this model admits (4.38) as probability density function. This distribution is again given by Proposition A.6:

$$Z_t \mid \{\boldsymbol{Y} = \boldsymbol{y}\} \sim \mathcal{N}_{d_z}\left(\mu_{t|t} + \widetilde{K}_t \left(\mu_{t+1|T} - B\mu_{t|t}\right), \left(I - \widetilde{K}_t B\right)\Omega_{t|t} + \widetilde{K}_t \Omega_{t+1|T} \widetilde{K}_t^\top\right).$$

Noticing that:

$$B\Omega^\top = W_{t+1}\widetilde{K}_t^\top,$$

we have that $\Omega_{t|T} = \Omega_{t|t} + \widetilde{K}_t \left(\Omega_{t+1|T} - W_{t+1}\right)\widetilde{K}_t^\top$, which proves (4.36).

Finally, Proposition A.4 gives the joint distribution of $(Z_t, Z_{t+1}) \mid \{\boldsymbol{Y} = \boldsymbol{y}\}$ (for $0 \leq t \leq T-1$), and in particular:

$$\mathbb{C}\mathrm{ov}_\theta\left(Z_t, Z_{t+1} \mid \boldsymbol{Y} = \boldsymbol{y}\right) = \widetilde{K}_t \Omega_{t+1|T} \,,$$

which proves (4.37) and concludes the proof.

4.2.4.2 EM algorithm

Algorithm 4.3 (EM step for Model 4.3)**.** *Starting from $\theta^{(h)}$, the update $\theta^{(h+1)}$ is obtained by:*

- *<u>E step</u> Computing $\left(\mu_{t|T}^{(h)}, \Omega_{t|T}^{(h)}\right)_{0\leqslant t\leqslant T}$ and $\left(\Omega_{t,t+1|T}^{(h)}\right)_{0\leqslant t\leqslant T-1}$ using Proposition 4.8 with parameter $\theta^{(h)}$ and setting:*

$$S_{t|T}^{(h)} = \Omega_{t|T}^{(h)} + \mu_{t|T}^{(h)}\left(\mu_{t|T}^{(h)}\right)^\top, \qquad C_{t,t+1|T}^{(h)} = \Omega_{t,t+1|T}^{(h)} + \mu_{t|T}\left(\mu_{t+1|T}^{(h)}\right)^\top.$$

- *<u>M step</u> Setting:*

$$\begin{aligned}
B^{(h+1)} &= \left(\sum_{t=0}^{T-1} C_{t,t+1|T}^{(h)}\right)^\top \times \left(\sum_{t=0}^{T-1} S_{t|T}^{(h)}\right)^{-1} \\
\Omega^{(h+1)} &= \frac{1}{T}\sum_{t=0}^{T-1}\left(S_{t+1|T}^{(h)} + B^{(h+1)} S_{t|T}^{(h)} (B^{(h+1)})^\top - \left(B^{(h+1)} C_{t,t+1|T}^{(h)}\right)^\top - B^{(h+1)} C_{t,t+1|T}\right) \\
A^{(h+1)} &= \left(\sum_{t=0}^{T} Y_t \left(\mu_{t|T}^{(h)}\right)^\top\right) \times \left(\sum_{t=0}^{T} S_{t|T}^{(h)}\right)^{-1} \\
\Sigma^{(h+1)} &= \frac{1}{T+1}\sum_{t=0}^{T}\left(Y_t Y_t^\top - A^{(h+1)} S_{t|T}^{(h)} \left(A^{(h+1)}\right)^\top - \left(A^{(h+1)} \mu_{t|T}^{(h)} Y_t^\top\right)^\top - A^{(h+1)} \mu_{t|T}^{(h)} Y_t^\top\right).
\end{aligned}$$

Proof of Algorithm 4.3

The quantities required for the E-step were justified in the previous paragraph. We now turn to the M-step, which we proceed to establish. We begin by noting that

$$\nabla_\theta Q\big(\theta \mid \theta^{(h)}\big) = \nabla_\theta \mathbb{E}_{\theta^{(h)}}[\log p_\theta(Y_{0:T}, Z_{0:T}) \mid \boldsymbol{Y} = \boldsymbol{y}] = \mathbb{E}_{\theta^{(h)}}[\nabla_\theta \log p_\theta(Y_{0:T}, Z_{0:T}) \mid \boldsymbol{Y} = \boldsymbol{y}],$$

since the integration step does note depend on θ. It remains to compute the gradient with respect to the different elements of θ. For sake of simplicity, we focus on the updates for Σ, Ω, A, and B, which represent the main parameters of interest. [a] Using rules of derivatives with respect to matrices[b], we have that:

$$\begin{aligned}
\frac{\partial \log p_\theta(Y_{0:T}, Z_{0:T})}{\partial \Omega} &= -\frac{1}{2}\left(T + \Omega^{-1}\sum_{t=0}^{T-1}(Z_{t+1} - BZ_t)(Z_{t+1} - BZ_t)^\top\right)\Omega^{-1} \\
\frac{\partial \log p_\theta(Y_{0:T}, Z_{0:T})}{\partial B} &= -\Omega^{-1}\sum_{t=0}^{T-1}(Z_{t+1} - BZ_t)Z_t^\top \\
\frac{\partial \log p_\theta(Y_{0:T}, Z_{0:T})}{\partial \Sigma} &= -\frac{1}{2}\left((T+1) + \Sigma^{-1}\sum_{t=0}^{T}(Y_t - AZ_t)(Y_t - AZ_t)^\top\right)\Sigma^{-1} \\
\frac{\partial \log p_\theta(Y_{0:T}, Z_{0:T})}{\partial A} &= -\Sigma^{-1}\sum_{t=0}^{T}(Y_t - AZ_t)Z_t^\top .
\end{aligned}$$

Thus, $\nabla_\theta Q\big(\theta \mid \theta^{(h)}\big) = 0$ leads to the updates:

$$B^{(h+1)} = \left(\sum_{t=0}^{T-1}\mathbb{E}_{\theta^{(h)}}\big[Z_{t+1}Z_t^\top \mid \boldsymbol{Y} = \boldsymbol{y}\big]\right) \times \left(\sum_{t=0}^{T-1}\mathbb{E}_{\theta^{(h)}}\big[Z_t Z_t^\top \mid \boldsymbol{Y} = \boldsymbol{y}\big]\right)^{-1}, \tag{4.42}$$

$$\begin{aligned}
\Omega^{(h+1)} &= \frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}_{\theta^{(h)}}\Big[\big(Z_{t+1} - B^{(h+1)}Z_t\big)\big(Z_{t+1} - B^{(h+1)}Z_t\big)^\top \mid \boldsymbol{Y} = \boldsymbol{y}\Big], \qquad (4.43)\\
&= \frac{1}{T}\sum_{t=0}^{T-1}\Big(\mathbb{E}_{\theta^{(h)}}\big[Z_{t+1}Z_{t+1}^\top \mid \boldsymbol{Y} = \boldsymbol{y}\big] + B^{(h+1)}\mathbb{E}_{\theta^{(h)}}\big[Z_t Z_t^\top \mid \boldsymbol{Y} = \boldsymbol{y}\big](B^{(h+1)})^\top \\
&\quad - \mathbb{E}_{\theta^{(h)}}\big[Z_{t+1}Z_t^\top \mid \boldsymbol{Y} = \boldsymbol{y}\big](B^{(h+1)})^\top - B^{(h+1)}\mathbb{E}_{\theta^{(h)}}\big[Z_t Z_{t+1}^\top \mid \boldsymbol{Y} = \boldsymbol{y}\big]\Big),
\end{aligned}$$

$$A^{(h+1)} = \left(\sum_{t=0}^{T} Y_t\mathbb{E}_{\theta^{(h)}}\big[Z_t^\top \mid \boldsymbol{Y} = \boldsymbol{y}\big]\right) \times \left(\sum_{t=0}^{T}\mathbb{E}_{\theta^{(h)}}\big[Z_t Z_t^\top \mid \boldsymbol{Y} = \boldsymbol{y}\big]\right)^{-1}, \tag{4.44}$$

$$\begin{aligned}
\Sigma^{(h+1)} &= \frac{1}{T+1}\sum_{t=0}^{T}\mathbb{E}_{\theta^{(h)}}\Big[\big(Y_t - A^{(h+1)}Z_t\big)\big(Y_t - A^{(h+1)}Z_t\big)^\top \mid \boldsymbol{Y} = \boldsymbol{y}\Big], \qquad (4.45)\\
&= \frac{1}{T+1}\sum_{t=0}^{T}\Big(Y_t Y_t^\top - A^{(h+1)}\mathbb{E}_{\theta^{(h)}}\Big[Z_t Z_t^\top \mid \boldsymbol{Y} = \boldsymbol{y}\Big](A^{(h+1)})^\top \\
&\quad - Y_t\mathbb{E}_{\theta^{(h)}}\Big[Z_t^\top \mid \boldsymbol{Y} = \boldsymbol{y}\Big](A^{(h+1)})^\top - A^{(h+1)}\mathbb{E}_{\theta^{(h)}}[Z_t \mid \boldsymbol{Y} = \boldsymbol{y}]Y_t^\top\Big).
\end{aligned}$$

The proof is concluded by replacing the expectations by there computations from the E step.

[a] In practice, μ_0 and V_0 are often supposed to be known or related to Ω and B.
[b] That can be found in Petersen and Pedersen [2008].

4.2.5 Conclusion

In this section, we have highlighted the strong relationship between Kalman filtering, a well-established method in signal processing, and the Expectation-Maximization (EM) algorithm used for parameter estimation in Linear Gaussian Hidden Markov Models, which are commonly applied in ecology. To obtain the maximum likelihood estimates of the parameters, the EM algorithm can be employed, albeit with the cost of an iterative E-step, which corresponds to a Kalman smoother.

4.3 Latent variable models based on phylogenetics trees for evolution

4.3.1 Context and motivation

Evolution lies at the heart of many conceptual frameworks in biology and ecology. Numerous models have been developed to describe how species' characteristics evolve over time. These evolutionary models typically rely on a phylogenetic tree, which represents the sequence of speciation events leading to the currently observed species, known as the *extant* species. Two commonly studied types of characteristics are: (i) quantitative traits (such as body weight, gestation length, etc.), and (ii) genomic sequences (for a given gene).

The aim of this section is to introduce two classical evolutionary models and show how the general inference methodology described in the book applies. As with most models, the critical component is the E step and we will see how a recursion (similar to forward-backward recursion presented in Section 4.1.1) allows to achieve it in an efficient manner. While we introduce both models and detail their inference procedures—specifically the E-step—we do not include any biological illustrations at this stage.

The data at hand typically consists of,

- the traits or sequences $(y_i)_{1\leq i\leq n}$ observed for n extant species and,
- the phylogenetic tree describing the common past evolution that relates the extant species.

The left panel of Figure 4.9 gives an example of a phylogenetic tree: it encodes both how species (or ancestor) derive from each other (topology of the tree) and the time that separates each (extant or ancestral) species from its direct ancestor (branch lengths). Each extant species is associated with a terminal node of the tree while each ancestral species corresponds to an internal nodes.

Definition 4.2. *The root of the tree, located at the top, is called the Most Recent Common Ancestor (MRCA) of the n extant species.*

Assuming each speciation event results in exactly two descendant species, the tree includes $n-1$ ancestral species (labelled $j = n+1, \dots 2n-1$). In the following, we may replace the index $2n-1$ with 'MRCA' to emphasize its role as the root. We further denote by $pa(k)$ the (unique) parent of node k (either internal or terminal). The branch lengths d_k ($1 \leq k \leq 2n-2$) refer to the evolutionary time that separates species k from its parent $pa(k)$.

For any $k = 1, \dots, 2n-2$, let h_k stand for the length of the path from the root (MRCA) to node k

Depending on the problem, branch lengths d_k may either be provided or require estimation.

Latent variables and graphical model Let us denote by $\{Z_i\}_{n+1\leq i\leq 2n-1}$ the traits or sequences corresponding to ancestral species, which are not directly observed. These variables are latent and are related to the observed traits of the extant (currently living) species, denoted by $\{Y_i\}_{1\leq i\leq n}$. The relationship between ancestral and extant traits is defined by a statistical model, whose structure reflects the dependencies encoded in the phylogenetic tree. The graphical model associated with this evolutionary process mirrors the topology of the phylogenetic tree, as illustrated in the right panel of Figure 4.9. In this graph, the edges represent evolutionary relationships and are directed forward in time, from ancestral species to their descendants.

4.3.2 Two models of evolution for quantitative traits and genetic sequences

In this section, we address in parallel models for quantitative traits (continuous variables) and for genetic sequences (discrete variables). To avoid duplicating each formula for the two settings, we adopt a slight abuse of notation: for a continuous variable Z, we denote by $p_\theta(Z = z)$ its probability density evaluated at z. Unless otherwise specified, all marginalization and integration formulas will be written as integrals; in the discrete case, these should be understood as finite sums.

4.3.2.1 Gaussian models for traits evolution.

The most widely used model for studying the evolution of a continuous trait is the branching Brownian motion [see, e.g. Felsenstein, 1985, Bastide et al., 2024]. The available data consist of

- the observed values of the continuous trait for n extant species $\boldsymbol{y} = (y_1, \dots, y_n)$ and

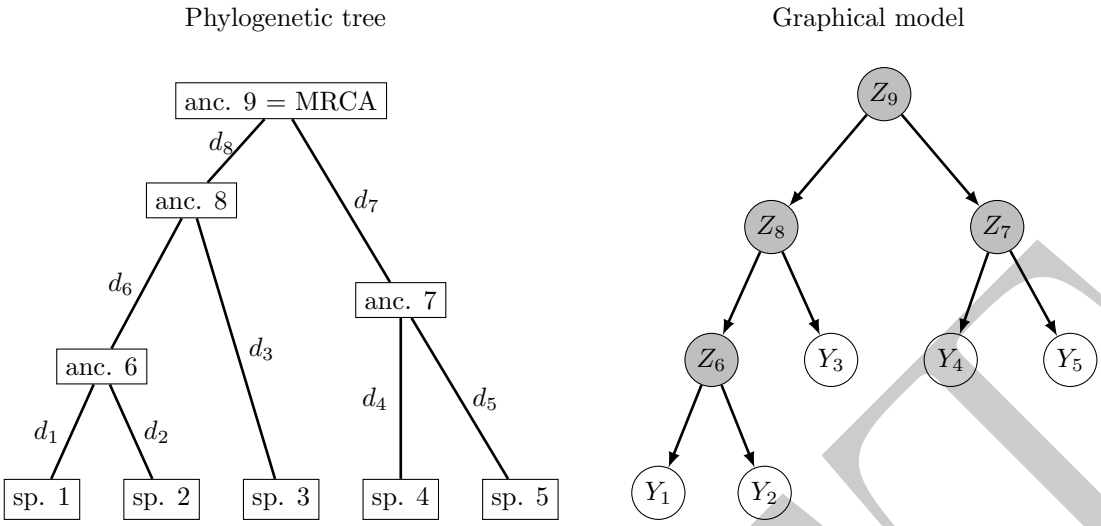


Figure 4.9: Evolution model: Fictitious example with 5 extant species. *Left*: Phylogenetic tree. 'sp. 1' to 'sp. 5' = extant species, 'anc. 6' to 'anc. 9' = ancestral species. The branch lengths d_i $(1 \le i \le 2n-2)$ refers to the evolutionary time that separates species i from its the parent $pa(i)$. *Right*: Graphical model. The variables $(Y_i)_{1\le i\le 5}$ are observed, whereas the variables $(Z_j)_{6\le j\le 9}$ are latent.

- a phylogenetic tree as depicted in Figure 4.9. The tree provides information about both its topology and the branch lengths $(d_1, \dots, d_{2n-2})$.

Model 4.4 (Branching Brownian motion for quantitative traits)**.** *Starting from the trait value of the MRCA*

$$Z_{MRCA} = Z_{2n-1} = \mu,$$

the value of the traits evolves as a Brownian motion with variance σ^2 down each branch of the tree and two independent paths start along each new branch below each internal node.
The difference between the value of trait at any internal or terminal node $k = 1, \dots 2n-2$ the value of the traits of its parent has then a Gaussian distribution with mean 0 and variance $d_k\sigma^2$:

$$\forall 1 \le i \le n: \quad Y_i - Z_{pa(i)} \sim \mathcal{N}\big(0, d_i\sigma^2\big), \qquad \forall n+1 \le j \le 2n-2: \quad Z_j - Z_{pa(j)} \sim \mathcal{N}\big(0, d_j\sigma^2\big).$$

The values of the traits $Y_{1:n}$ at the leaves (or terminal nodes) are observed, whereas these at the internal nodes $Z_{n+1:2n-1}$ are latent.

For each species, the centring of the distribution of the trait difference reflects the inheritance of the ancestor's trait. Because the phylogenetic tree is supposed to be known, the model's parameters hence reduce to

$$\theta = (\mu, \sigma^2).$$

Remark. *Note that here, there is no natural split between θ_{lat} and θ_{obs}. Besides, the latent variables are continuous random variables.*

Figure 4.10 gives an example of a path of Model 4.4, consistent with the tree given in Figure 4.9. The value at the MRCA can be assumed to be fixed ($Z_{MRCA} = \mu$) or random ($Z_{MRCA} \sim \mathcal{N}\big(\mu, \gamma^2\big)$). In the latter case, the model parameter becomes $\theta = (\mu, \gamma^2, \sigma^2)$.

Remark. *Observe that, because the extant species are all contemporary, the branch lengths need to satisfy the ultrametric constraint, which is that the time that separates each of them from the MRCA is the same. For example, in Figure 4.9, the total height h of the tree is*

$$h = d_8 + d_6 + d_1 = d_8 + d_6 + d_2 = d_8 + d_3 = d_7 + d_4 = d_7 + d_5.$$

4.3.2.2 Continuous time Markov models for evolution of sequences

Sequence evolution models aim at understanding the diversity of the gene sequences that are shared (or *homologous*) across a set of n extant species. The data consists of the n sequences $(Y_i)_{1\le i\le n}$, each of length

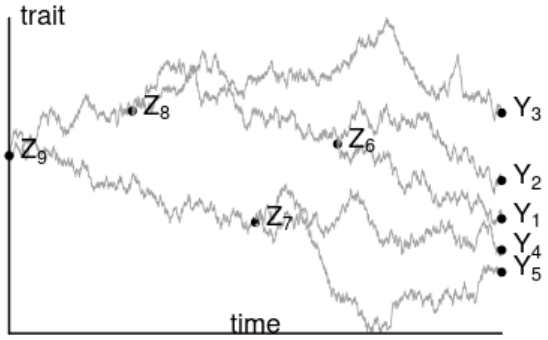


Figure 4.10: Branching Brownian model for quantitative traits. Example of a branching path consistent with the topology and the branch lengths of the evolution tree from Figure 4.9.

p: $Y_i = (Y_{i1}, \dots Y_{ip})$, and the topology of phylogenetic tree that relates the n species. The branch lengths are usually left unknown, and therefore need to to be estimated.

When dealing with genomic sequences, the sites Y_{ij} take values in the alphabet $\mathcal{A} = \mathtt{a}, \mathtt{g}, \mathtt{c}, \mathtt{t}$. In contrast, for protein sequences, the alphabet corresponds to the 20 standard amino acids, so $|\mathcal{A}| = 20$. Under the most classical models, the p sites of the sequence are supposed to evolve independently according to a branching continuous time Markov process with rate matrix (infinitesimal generator) Q.
The reader may refer to Karlin and Taylor [1981] for a general introduction to continuous time Markov processes and to Felsenstein [1981], Durrett [2008], or Pardoux [2024] for their application to sequence evolution.

In this setting, we denote by $\exp(A)$ the matrix exponential of A, that is, for a square matrix A: $\exp(A) = \sum_{k\geq 0} A^k/k!$.

Model 4.5 (Sequence evolution)**.** *Starting from the sequence $Z_{2n-1} = Z_{MRCA}$ of the MRCA, each site of the sequence evolves independently as a continuous time Markov process with rate matrix $Q = [\lambda_{ab}]_{a,b\in\mathcal{A}}$ down each branch of the tree and two independent paths start along each new branch below each internal node.*
The conditional distribution of each site of sequence of any internal or terminal node is hence:

$$\forall 1 \leq i \leq n, \forall 1 \leq u \leq p: \qquad \mathbb{P}\big(Y_{iu} = b \mid Z_{pa(i)u} = a\big) = [\exp(d_i Q)]_{ab},$$
$$\forall n+1 \leq j \leq 2n-2, \forall 1 \leq u \leq p: \qquad \mathbb{P}\big(Z_{ju} = b \mid Z_{pa(j)u} = a\big) = [\exp(d_j Q)]_{ab}.$$

The sequences $Y_{1:n}$ at the leaves (or terminal nodes) are observed, whereas these at the internal nodes $Z_{n+1:2n-1}$ are latent.

Similarly to Model 4.4, the root sequence can be assumed either fixed ($Z_{MRCA} = z_{MRCA}$) or random. In the latter case it is often assumed that the entire process evolves in a stationary regime, meaning that each site $Z_{MRCA,u}$ of the root sequence is drawn from the stationary distribution ν of the rate matrix Q ($\nu^\top Q = 0$).
The topology of the tree is usually supposed to be known (or to be estimated by another method) and the unknown parameters are the branch lengths and the rate matrix, that is:

$$\theta = (Q, (d_k)_{1\leq k\leq 2n-2})\,.$$

A wide variety of versions of Model 4.5 exist, depending on the parametrization of the rate matrix Q. For example, the Jukes and Cantor model [JC model, see Jukes and Cantor, 1969] assumes that the mutations between all nucleotides occur at the same rate, whereas the two-parameter model proposed by Kimura [1981] (K2) makes a distinction between *transitions* (i.e. either $\mathtt{a} \leftrightarrow \mathtt{g}$ or $\mathtt{c} \leftrightarrow \mathtt{t}$) and *transversions* (all other mutations). Ordering the nucleotides as $\{\mathtt{a}, \mathtt{g}, \mathtt{c}, \mathtt{t}\}$, the two respective rate matrices are

$$Q_{JC} = \begin{bmatrix} - & \alpha & \alpha & \alpha \\ \alpha & - & \alpha & \alpha \\ \alpha & \alpha & - & \alpha \\ \alpha & \alpha & \alpha & - \end{bmatrix}, \qquad Q_{K2} = \begin{bmatrix} - & \alpha & \beta & \beta \\ \alpha & - & \beta & \beta \\ \beta & \beta & - & \alpha \\ \beta & \beta & \alpha & - \end{bmatrix} \tag{4.46}$$

so the actual set of parameters is $\theta = (\alpha, (d_k)_{1\leq k\leq 2n-2})$ for the JC model and $(\alpha, \beta, (d_k)_{1\leq k\leq 2n-2})$ for the K2

model. Under the JC model the transition matrix after a time t has a simple form:

$$[\exp(tQ)]_{ab} = \begin{cases} (1+3\mathrm{e}^{-4\alpha t})/4 & \text{if } a=b, \\ (1-\mathrm{e}^{-4\alpha t})/4 & \text{otherwise.} \end{cases}$$

Hence $[\exp(tQ)]_{aa}$ decreases from 1 to 1/4, as t goes from 0 to infinity, whereas, when $a \neq b$, $[\exp(tQ)]_{ab}$ increases from 0 to 1/4. As a consequence, under the JC evolution model, all transition probabilities tend to 1/4, meaning that the transition process is asymptotically uniform.

Reversibility. For identifiability reasons, sequence evolution models such as Model 4.5 are most often required to be time-reversible, which means that the probability to move from state (nucleotide) a to state b within time t, is the same as the probability to move from state b to state a within the same time. This implies that $p_\theta(Z_j = z' \mid Z_{pa(j)} = z) = p_\theta(Z_{pa(j)} = z \mid Z_j = z')$. Note that, for the trait evolution Model 4.4, the same property actually already holds . Indeed, because the Gaussian distribution is symmetric, we have that $p_\theta(Z_j = z' \mid Z_{pa(j)} = z) = \phi\big(z'; z, d_j\sigma^2\big) = \phi\big(z; z', d_j\sigma^2\big)$ where $z \mapsto \phi\big(z; m, \sigma^2\big)$ is the Gaussian density.

4.3.3 Likelihood functions for the two evolution models

Complete likelihood Because of the tree shape of the graphical model, the complete likelihood of a model such as Models 4.4 or 4.5, can be written as product of conditional distributions, starting from the root, that is:

$$p_\theta(Y_{1:n}, Z_{n+1:2n-1}) = p_\theta(Z_{2n-1}) \prod_{j=n+1}^{2n-2} p_\theta(Z_j \mid Z_{pa(j)}) \prod_{i=1}^{n} p_\theta(Y_i \mid Z_{pa(i)}). \tag{4.47}$$

The conditional distributions obviously depends on the model. For example:

$$\text{under Model 4.4}: \qquad p_\theta(Z_j \mid Z_{pa(j)}) = \phi\big(Z_j; Z_{pa(j)}, d_j\sigma^2\big), \tag{4.48}$$

$$\text{under Model 4.5}: \qquad p_\theta(Z_j \mid Z_{pa(j)}) = \prod_{u=1}^{p} [\exp(d_j Q)]_{Z_{pa(j),u}, Z_{j,u}}, \tag{4.49}$$

and respectively for $p_\theta(Y_i \mid Z_{pa(i)})$.

Marginal likelihood We can write the marginal distribution of any variable Z_j or Y_i by iterating the conditional distributions (4.48).

Proposition 4.9 (Marginal distributions of the variables)**.** *For any $j = 1, \ldots, 2n-2$, let h_j stand for the length of the path from the root (MRCA) to node j [a]. Under Models 4.4 and 4.5 the marginal distribution of each node Z_j for $j = n+1, \ldots, 2n-1$ or Y_i for $i = 1, \ldots, n$ is:*

$$p_\theta(Z_j) = \int p_\theta(Z_{MRCA} = z) p_\theta(Z_j \mid Z_{MRCA} = z)\, dz. \tag{4.50}$$

where

$$\textit{under Model 4.4}: \qquad p_\theta(Z_j \mid Z_{MRCA} = z) = \phi\big(Z_j; z, h_i\sigma^2\big), \tag{4.51}$$

$$\textit{under Model 4.5}: \qquad p_\theta(Z_j \mid Z_{MRCA} = z) = \prod_{u=1}^{p} [\exp(h_i Q)]_{z_u, Z_{j,u}},$$

and $p_\theta(Z_{MRCA} = z)$ is a Dirac distribution if Z_{MRCA} is assumed to be fixed or

$$\textit{under Model 4.4}: \qquad Z_{MRCA} \sim \mathcal{N}\big(\mu, \gamma^2\big),$$

$$\textit{under Model 4.5}: \qquad Z_{MRCA,u} \sim \nu$$

if Z_{MRCA} is a random variable.

[a] In Figure 4.9: $h_6 = d_6 + d_8$ and $h_4 = d_7 + d_4$

As a consequence, we have an explicit expression for each $p_\theta(y_i)$. However, theses expressions do not allow for the calculation of the joint likelihood of all the observations $p_\theta(y_{1:n})$, due to the dependencies between the latent

variables:

$$\begin{aligned} p_\theta(y_{1:n}) &= \int\cdots\int p_\theta(z_{2n-1}) p_\theta(y_i \mid Z_{pa(i)} = z_{pa(i)})\, \mathrm{d}z_{n+1} \dots\, \mathrm{d}z_{2n-1} \\ &= \int\cdots\int p_\theta(z_{2n-1}) \prod_{j=n+1}^{2n-2} p_\theta(z_j \mid Z_{pa(j)} = z_{pa(j)}) \prod_{i=1}^{n} p_\theta(y_i \mid Z_{pa(i)} = z_{pa(i)})\, \mathrm{d}z_{n+1} \dots\, \mathrm{d}z_{2n-1}, \end{aligned} \tag{4.52}$$

which involves the evaluation of a $(n-1)$-dimensional integral. [Felsenstein, 1981] proposed an efficient way to evaluate the marginal likelihood of the observed data $p_\theta(y_{1:n})$ (Equation 4.52), taking advantage of the tree structure of the graphical model.

Definition 4.3. *Let* $\mathsf{sub}(k)$ *be the set of all observed variables located downward the (internal or terminal) node* k*. As an example, in Figure 4.9:* $y_{\mathsf{sub}(8)} = \{y_1, y_2, y_3\}$ *and* $y_{\mathsf{sub}(4)} = \{y_4\}$*. For any tree,*

$$y_{\mathsf{sub}(MRCA)} = \{y_i\}_{1 \le i \le n}.$$

In addition, each internal node j *has two direct offsprings. Let* $\mathfrak{L}(j)$ *denote the left one and* $\mathfrak{R}(j)$ *the right one. Then:*

$$y_{\mathsf{sub}(j)} = y_{\mathsf{sub}(\mathfrak{L}(j)),\mathsf{sub}(\mathfrak{R}(j))}.$$

Because each node is independent from the top of the tree, given its parent, a pruning principle can be applied recursively to the whole tree, from the leaves to the root.

Proposition 4.10 (Felsenstein's tree pruning algorithm)**.** *Let* $\ell_j(z)$ *denote by the likelihood of the observed variables downstream* j *conditional on* $Z_j = z$*.*

$$\ell_j(z) = p_\theta(y_{\mathsf{sub}(j)} \mid Z_j = z) \tag{4.53}$$

Then $\ell_j(z)$ *can be computed iteratively by the following formula:*

$$\ell_j(z) = \left(\int p_\theta(Z_{\mathfrak{L}(j)} = u \mid Z_j = z) \ell_{\mathfrak{L}(j)}(u)\, du \right) \left(\int p_\theta(Z_{\mathfrak{R}(j)} = v \mid Z_j = z) \ell_{\mathfrak{R}(j)}(v)\, dv \right). \tag{4.54}$$

As a consequence,

$$p_\theta(y_{1:n}) = p_\theta(y_{\mathsf{sub}(MRCA)}) = \int p_\theta(Z_{MRCA} = z) \ell_{MRCA}(z)\, dz. \tag{4.55}$$

Proof of Proposition 4.10

$$\begin{aligned} \ell_j(z) =& p_\theta(y_{\mathsf{sub}(j)} \mid Z_j = z) \\ =& \int_{u,v} p_\theta(y_{\mathsf{sub}(j)} \mid Z_{\mathfrak{L}(j)} = u, Z_{\mathfrak{R}(j)} = v, \cancel{Z_j = z}) p_\theta(Z_{\mathfrak{L}(j)} = u, Z_{\mathfrak{R}(j)} = v \mid Z_j = z)\, \mathrm{d}u\, \mathrm{d}v \end{aligned}$$

using the fact $y_{\mathsf{sub}(j)} = y_{\mathsf{sub}(\mathfrak{L}(j)),\mathsf{sub}(\mathfrak{R}(j))}$. So

$$\begin{aligned} \ell_j(z) =& \int_{u,v} p_\theta(Z_{\mathfrak{L}(j)} = u \mid Z_j = z) p_\theta(Z_{\mathfrak{R}(j)} = v \mid Z_j = z) p_\theta(y_{\mathsf{sub}(\mathfrak{L}(j)),\mathsf{sub}(\mathfrak{R}(j))} \mid Z_{\mathfrak{L}(j)} = u, Z_{\mathfrak{R}(j)} = v)\, \mathrm{d}u\, \mathrm{d}v \\ =& \int_{u,v} p_\theta(Z_{\mathfrak{L}(j)} = u \mid Z_j = z) p_\theta(Z_{\mathfrak{R}(j)} = v \mid Z_j = z) p_\theta(y_{\mathsf{sub}(\mathfrak{L}(j))} \mid Z_{\mathfrak{L}(j)} = u) p_\theta(y_{\mathsf{sub}(\mathfrak{R}(j))} \mid Z_{\mathfrak{R}(j)} = v)\, \mathrm{d}u\, \mathrm{d}v \end{aligned}$$

using the conditional independence

$$= \int p_\theta(Z_{\mathfrak{L}(j)} = u \mid Z_j = z) \ell_{\mathfrak{L}(j)}(u) p_\theta(Z_{\mathfrak{R}(j)} = v \mid Z_j = z) \ell_{\mathfrak{R}(j)}(v)\, \mathrm{d}u\, \mathrm{d}v\,.$$

We obtain the expected result (4.54). The last formula comes from the fact that $y_{1:n} = y_{\mathsf{sub}(MRCA)}$.

Thus, all functions $\ell_j(z)$ can be evaluated recursively for $j = n+1 \dots 2n-1$ to compute eventually $p_\theta(y_{1:n})$.

For the tree from Figure 4.9, this gives

$$
\begin{aligned}
& p_\theta(y_{1:n}) = p_\theta(y_{\mathsf{sub}(9)}) = \int p_\theta(Z_9 = z)\ell_9(z)\,\mathrm{d}z \\
\text{where} \quad & \ell_9(z) = \left(\int p_\theta(Z_8 = u \mid Z_9 = z)\ell_8(u)\,\mathrm{d}u\right) \times \left(\int p_\theta(Z_7 = v \mid Z_9 = z)\ell_7(v)\,\mathrm{d}v\right), \\
& \ell_7(z) = p_\theta(y_4 \mid Z_7 = z) \times p_\theta(y_5 \mid Z_7 = z), \\
& \ell_8(z) = p_\theta(y_4 \mid Z_8 = z) \times \left(\int p_\theta(Z_6 = u \mid Z_8 = z)\ell_6(u)\,\mathrm{d}u\right), \\
\text{and} \quad & \ell_6(z) = p_\theta(y_1 \mid Z_6 = z) \times p_\theta(y_2 \mid Z_6 = z).
\end{aligned}
$$

4.3.4 E step for the tree based evolution models

We may now present the E step of the EM algorithm that enables to estimate the parameter θ of both Models 4.4 and 4.5. As in most other models presented in this book, the M step does not raise specific difficulties and can be achieved either in close form or via numerical optimization, so we do not provide the explicit update formulas for the parameter.

Because the complete likelihood given in (4.47) only involves pairs of variables of the form $(Z_{pa(j)}, Z_j)$, the E step, aims at evaluating the conditional distributions $p_\theta(Z_j \mid Y_{1:n} = y_{1:n})$ and $p_\theta(Z_{pa(j)}, Z_j \mid Y_{1:n} = y_{1:n})$. This can be done thanks to an upward-downward recursion along the tree, which is similar to the forward-backward recursion already described for hidden Markov models (HMM: see Section 4.1.1, Proposition 4.5). This similarity stems from the fact that the graphical model of a HMM given in Figure 4.2 is itself a tree, Z_1 being the root and $Y_{1:n}$ being the terminal nodes.

Let us denote the conditional distributions of interest by

$$
\begin{array}{ll}
\alpha_j(z) & = p_\theta(Z_j \mid Y_{\mathsf{sub}(j)} = y_{\mathsf{sub}(j)}) \\
\tau_j(z) & = p_\theta(Z_j = z \mid Y_{1:n} = y_{1:n}) \\
\xi_j(u,z) & = p_\theta(Z_{pa(j)} = u, Z_j = z \mid Y_{1:n} = y_{1:n}).
\end{array}
\tag{4.56}
$$

Proposition 4.11 (Upward-downward recursion for E step in evolution models)**.** *The conditional distributions $\alpha_j(z)$, $\tau_j(z)$ and $\xi_j(u,z)$ defined in (4.56) can be computed as follows.*

Upward recursion: *starting from $\alpha_i(y) = \delta_{\{y_j\}}(y)$ for $i = 1, \dots n$, iterate for $j = n+1$ to $j = 2n-1$:*

$$\alpha_j(z) \propto p_\theta(Z_j = z)\ell_j(z).$$

Downward recursion: *starting from $\tau_{MRCA}(z) = \tau_{2n-1}(z) = \alpha_{2n-1}(z)$, iterate down the branches of the tree:*

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

where

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

The proof is provided in Appendix B.1. Notably, all integrals involved in the upward-downward recursion from Proposition 4.11 simplify to finite sums in the case of sequences (Model 4.5) or have explicit forms when applied to Gaussian models, such as in Model 4.4.

4.3.5 The special case of Gaussian models

The EM algorithm is actually not required to make the inference of certain trait evolution models. Indeed, under Model 4.4, assuming that the trait value of the MRCA is fixed to $Z_{MRCA} = \mu$, it is easy to check all other traits values have a joint multivariate Gaussian distribution.

Using the distributions given in (4.50) and (4.51), we see that the mean of each latent (Z_j) or observed (Y_i) trait is μ (as the motion has do drift) and that its variance is proportional to the length (h_j or $h_i = h$) of the path that separates its from the MRCA. Furthermore, using the properties of the Brownian motion, it comes that the covariance between two nodes is proportional to the length of the common path that separates them from the MRCA.

Overall, denoting $Y = [Y_{1:n}]^\intercal$ and $Z = [Z_{n+1:2n-2}]^\intercal$, we have

$$\begin{bmatrix} Y \\ Z \end{bmatrix} \sim \mathcal{N}\left(\mu \begin{bmatrix} 1_n \\ 1_{n-2} \end{bmatrix}, \sigma^2 \underbrace{\begin{bmatrix} D_{YY} & D_{YX} \\ D_{YX}^\intercal & D_{YY} \end{bmatrix}}_{D} \right). \tag{4.57}$$

In the example of Figure 4.9, the matrix D is

$$D = \left[\begin{array}{ccccc|ccc} h & d_8+d_6 & d_8 & 0 & 0 & d_8+d_6 & 0 & d_8 \\ & h & d_8 & 0 & 0 & d_8+d_6 & 0 & d_8 \\ & & h & 0 & 0 & d_8+d_6 & 0 & d_8 \\ & & & h & d_7 & 0 & d_7 & 0 \\ & & & & h & 0 & d_7 & 0 \\ \hline & & & & & h_6 & 0 & d_8 \\ & & & & & & h_7 & 0 \\ & & & & & & & h_8 \end{array}\right]$$

where h is the height of the tree and both D_{YY} and DXX are symmetric matrices.

Hence, under Model 4.4, we have a direct access to the marginal likelihood of the observed traits $y = [y_1 \dots y_n]^\intercal$:

$$p_\theta(y) = \phi\big(y; \mu 1_n, \sigma^2 D_{YY}\big),$$

so (i) the MLE of μ and σ^2 can obtained by a direct maximization of $\log p_\theta(y)$ and (ii) combining the joint distribution (4.57) with Proposition A.3 from Appendix A.1.2, we also have access to the conditional distribution of the ancestors' traits $Z_{n+1:2n-2}$:

$$(Z \mid Y = y) \sim \mathcal{N}\big(\mu 1_{n-1} + D_{YX}^\intercal D_{YY}^{-1}(y - \mu 1_n), \sigma^2 \left(D_{XX} - D_{YX}^\intercal D_{YY}^{-1} D_{YX}\right)\big).$$

The joint distribution (4.57) can be easily be adapted to the case where the root value Z_{MRCA} is itself random and Gaussian ($Z_{MRCA} \sim \mathcal{N}\big(\mu, \gamma^2\big)$ by simply adding the root Z_{MRCA} as a new coordinate of the vector Z and γ^2 to all variances and covariances.

Such close form results obviously provide a comfortable setting to work with. Still, when dealing with a large number of species n (say $n > 1000$ as in microbial ecology), one may prefer to avoid the inversion of the matrix D_{YY} when inferring the ancestral traits and resort algorithmic simplifications similar to the EM algorithm presented above [see Ho and Ané, 2014].

Remark. *The factorizations used in Proposition 4.11 are valid thanks to the Markovian structure of the Models 4.4 and 4.5, which ensures conditional independances. In Models 4.4 and 4.5, the conditional moments required for the E step have close form expressions because the emission distribution is respectively Gaussian and multinomial. The existence of an exact EM for these two models therefore results from both a tree-structure Markovian dependency structure and specificities of the emission distribution.*

4.3.6 Conclusion

In this section, we explored the application of the Expectation-Maximization (EM) algorithm to evolutionary models, where the E-step is enhanced by the Felsenstein algorithm. This approach takes advantage of the phylogenetic tree structure to efficiently compute the conditional distribution of the latent variables. By incorporating the tree's topology, the Felsenstein algorithm significantly reduces the computational complexity, enabling the EM algorithm to perform more efficiently in large-scale evolutionary studies. This method highlights the synergy between statistical inference and evolutionary biology, offering a robust framework for parameter estimation in models that account for both phylogenetic relationships and latent evolutionary dynamics.

4.4 Composite likelihood: application to spatial data (SR)

As said in Introduction chapter 1, we chose not to consider spatial models in this book, as it requires very specific methodological developments. Still, because ecological observations are often associated with a spatial structure, we briefly present here an emblematic model that involves both a spatial dependency structure and a latent variable structure: the hidden Markov random field model. Because of the complex dependency structure induced by this model (each site influences its neighbour, and conversely), its inference is often carried via a so-called composite likelihood approach, rather than via maximum likelihood. Taking the opportunity of this model, the aim of this section is to show how the general EM Algorithm 2.1 introduced for maximum likelihood inference can be adapted to composite likelihood inference.

4.4.1 Data and question

A typical dataset would consist in the observation of the abundance of, say, a plant species accross a series of sites arranged in a lattice (i.e. a regular grid). Denoting by $i = 1 \dots n$ the row of the lattice and by $j = 1 \dots m$ its row, each observation site is located as $s = (i, j)$. We denote the observed abundance at site s as Y_s. Assuming a spatial heterogeneity of the environment, the goal is to classify the sites into a finite set of categories $1, \dots K$, corresponding to different environment types, each being more or less favourable to the species.

4.4.2 The hidden Markov random field model

An easy way to account for some spatial dependency structure in the data is to encode it in a latent random field $\boldsymbol{Z}$, such as a Gibbs random field. We only consider a simple version of Gibbs field, namely the (simple) Potts model.

Definition 4.4 (Potts model). *A random field $\boldsymbol{Z} = (Z_{s=(i,j)})_{1\le i\le n, 1\le j\le m} \in \{1, \dots K\}^{nm}$ follows a Potts model with parameters $\alpha = (\alpha_1 \dots \alpha_K) \in \mathbb{R}^K$ (such that $\sum_{k=1}^K e^{\alpha_k} = 1$) and $\beta \in \mathbb{R}$: $\boldsymbol{Z} \sim Potts(\alpha, \beta)$, iff*

$$\mathbb{P}\{\boldsymbol{Z} = \boldsymbol{z}\} = \frac{1}{C(\alpha, \beta, nm, K)} \exp\left(\sum_{s=1}^{nm}\sum_{k=1}^{K} \alpha_k z_{sk} + \beta \sum_{s\sim s'}\sum_{k=1}^{K} z_{sk} z_{s'k}\right)$$

where

- *Z_{sk} is the indicator variable $Z_{sk} = \mathbb{I}\{Z_s = k\}$,*
- *$s = (i, j) \sim s' = (i', j')$ means that the sites s and s' are neighbors: $|i - i| = 1$ or $|j - j'| = 1$, but not both.*
- *$C(\alpha, \beta, nm, K)$ is the normalizing constant ensuring that $\sum_{\boldsymbol{z}\in\{1,\dots K\}^{nm}} \mathbb{P}\{\boldsymbol{Z} = \boldsymbol{z}\} = 1$.*

In Definition 4.4, the parameter α_k controls the marginal probability for each site to be in state k and the parameter β controls the strength of the interaction between neighbor sites: neighbor sites tend to be in the same state when $\beta > 0$, whereas the tend to be in different states when $\beta < 0$. $\beta = 0$ yields an absence of spatial dependency. Heterogeneous versions of the Potts model involve state-specific interaction parameters $\beta_{k\ell}$.
An important feature of the Potts model is that each site is independent from all others, given its neighbors. Another important feature is that there is no close-form expression of the normalizing constant C, making its evaluation intractable even for moderately large dimensions of the lattice.

Hidden Markov random field model.

Model 4.6 (Hidden Markov random field model with Poisson emission). *Let*

$$\boldsymbol{Z} = (Z_{s=(i,j)})_{1\le i\le n, 1\le j\le m},$$
$$\boldsymbol{Y} = (Y_{s=(i,j)})_{1\le i\le n, 1\le j\le m},$$

be respectively the set of latent variables and observed abundances at each site. We assume:

$$\boldsymbol{Z} \sim Potts(\alpha, \beta),$$
$$Y_s \mid \{Z_s = z_s\} \overset{ind}{\sim} \mathcal{P}(\lambda_{z_s}) \qquad 1 \le s \le n \times m.$$

Denoting $\lambda = (\lambda_k)_{1\le k\le K}$, the set of intensities, the set of parameters of the hidden Markov random field Model 4.6 is

Graphical model. The graphical model associated with Model 4.6 is displayed in Figure 4.11. Observe that, the dependency structure of the hidden field $\boldsymbol{Z}$ is not tree-shaped, as opposed to hidden Markov models (Figure 4.2) or evolution models (Figure 4.9).

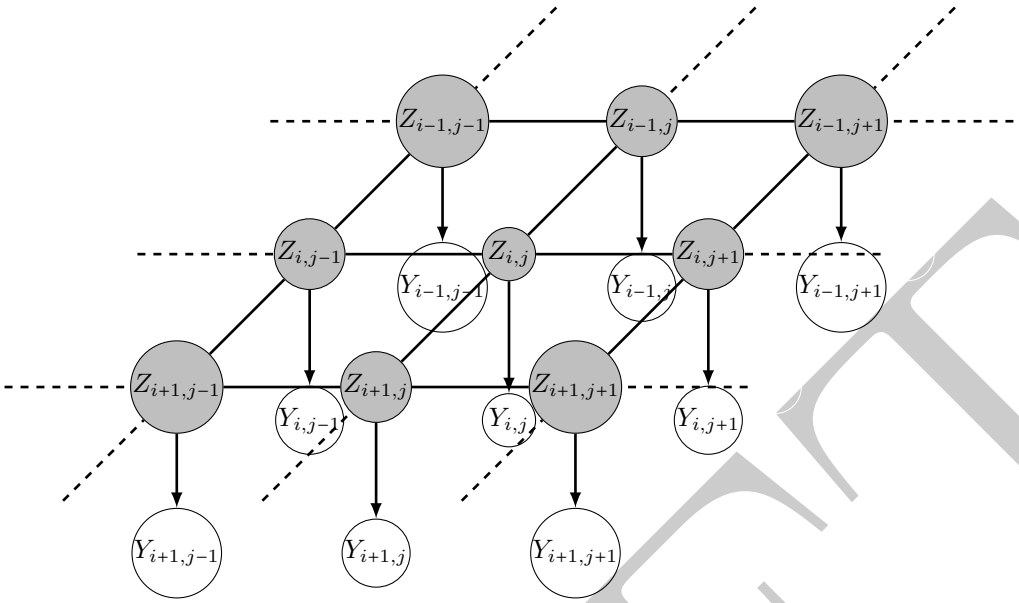


Figure 4.11: Graphical representation of the hidden Markov random field model (Model 4.6).

4.4.3 Likelihood and composite likelihood for the hidden Markov random field

The likelihood of $\boldsymbol{y}$ for Model 4.6 has the following expression:

$$\begin{aligned} p_\theta(\boldsymbol{y}) &= \sum_{\boldsymbol{z}\in\{1,\dots K\}^{nm}} \mathbb{P}(\boldsymbol{Z}=\boldsymbol{z}) \prod_{s=1}^{nm} \mathcal{P}(y_s;\lambda_{z_s}) \\ &= \frac{1}{C(\alpha,\beta,nm,K)} \sum_{\boldsymbol{z}\in\{1,\dots K\}^{nm}} \exp\left(\sum_{s=1}^{nm}\sum_{k=1}^{K} \alpha_k z_{sk} + \beta \sum_{s\sim s'}\sum_{k=1}^{K} z_{sk}z_{s'k}\right) \prod_{s=1}^{nm} \mathcal{P}(y_s;\lambda_{z_s}). \end{aligned}$$

However, $C(\alpha,\beta,nm,K)$ having no explicit expression and depending on θ, $p_\theta(\boldsymbol{y})$ can not be computed explicitly and can not be optimized. In addition, because of the two-dimensional structure of the hidden field $\boldsymbol{Z}$, exact recursions such as these presented in Sections 4.1.1 or 4.3 do not apply. The maximization of the (log-)likelihood can therefore not be achieved directly. An classical alternative then consists in dealing with a modified version of the likelihood.

Composite likelihood. The name 'composite likelihood' (or 'pseudo-likelihood' or 'quasi-likelihood') refers to a wide series of methods that have been developed for the inference of models with complex dependency structure [see Varin et al., 2011, for a general presentation]. Considering a model parametrized by θ and involving N non-independent variable $\boldsymbol{Y} = (Y_1, \dots Y_N)$, a composite log-likelihood is a (possibly weighted) sum of the marginal log-likelihoods of subsets of variables $Y_1, \dots Y_N$.

More specifically, let $\{C_b\}_{1\leq b\leq B}$ be a set of subsets (also named 'blocks') of $\{1, \dots N\}$, such that their union is $\{1, \dots N\}$, denoting $\boldsymbol{Y}_{C_b} = (Y_s)_{s\in C_b}$ and $\boldsymbol{y}_{C_b} = (y_s)_{s\in C_b}$, the composite log-likelihood based on the blocks $C_1, \dots C_B$, is defined as

$$c\ell_\theta(\boldsymbol{y}) = \sum_{b=1}^{B} w_b \log p_\theta\left(\boldsymbol{y}_{C_b}\right). \tag{4.58}$$

Note that the intersections between the blocks are not required to be empty (and are not in general). The corresponding composite likelihood writes

$$\prod_{b=1}^{B} \left[p_\theta\left(\boldsymbol{y}_{C_b}\right)\right]^{w_b}.$$

The maximum composite likelihood estimator is then defined as

$$\widehat{\theta}_{CL} = \arg\max_\theta c\ell_\theta(\boldsymbol{y}).$$

Importantly, under fairly general conditions, $\widehat{\theta}_{CL}$ enjoys similar properties as the maximum likelihood estimator $\widehat{\theta}_{ML} = \arg\max_\theta \log p_\theta(\boldsymbol{y})$. In particular, $\widehat{\theta}_{CL}$ is consistent, asymptotically Gaussian and its asymptotic variance is given by the so-called Godambe matrix [see Varin et al., 2011].

As said before, the blocks (C_b) often overlap. Introducing this redundancy is intentional: overlapping blocks can help capture more of the dependence structure without needing to model the full joint distribution. However, some precautions must be taken in, for instance by weighting the composite terms to avoid over-counting certain parts of the data (w_b in the definition).

Remark. *Composite likelihood inference is obviously not limited to spatial data: the combination of log-likelihood evaluated evaluated over blocks of observation may be considered as soon as the dependency structure of the model is too involved. For example, it is used by Ambroise and Matias [2012] for the inference of a stochastic block model (see Section 5.2) as an alternative to a variational approximation (see Section 5.1).*

Pseudo likelihood for the Potts hidden Markov random field. In the case of Model 4.6, the blocks are often chosen as the pairs of neighbour sites:

$$\{C_b\}_{1\leq b\leq B} = \{(s,s') : s \sim s'\}.$$

There are actually $B = 2mn - n - m$ such pairs and, setting all the weights to $w_b = 1$, this yields

$$c\ell_\theta(\boldsymbol{y}) = \sum_{b=1}^{B} \log p_\theta\left(\boldsymbol{y}_{C_b}\right) = \sum_{s\sim s'} \log p_\theta(y_s, y_{s'}). \tag{4.59}$$

Under the stationary regime of the Potts model from Definition 4.4 (neglecting boundary effects), the joint distribution of two neighbor sites s and s' is given by

$$\omega_{k\ell} = \mathbb{P}_\theta(Z_s = k, Z_{s'} = \ell) = \frac{1}{C(\alpha,\beta,2,K)} e^{\alpha_k} e^{\alpha_\ell} e^{\beta \mathbb{I}_{\{k=\ell\}}} = \frac{1}{C(\alpha,\beta,2,K)} a_k a_\ell b^{\mathbb{I}_{\{k=\ell\}}}, \tag{4.60}$$

where $a_k = e^{\alpha_k}$, $\sum_{k=1}^K a_k = 1$ and $b = \mathrm{e}^\beta$. In addition, $C(\alpha,\beta,2,K) = 1 + (b-1)\sum_{1\leq k\leq K} a_k^2$. Indeed,

$$\begin{aligned}
C(\alpha,\beta,2,K) &= \sum_{1\leq k,\ell\leq K} a_k a_\ell b^{\mathbb{I}_{\{k=\ell\}}} &&= \sum_{1\leq k\leq K} a_k \left[\left(\sum_{1\leq \ell\leq K, \ell\neq k} a_\ell\right) + a_k b\right] \\
&= \sum_{1\leq k\leq K} a_k \left[\left(\sum_{1\leq \ell\leq K} a_\ell - a_k\right) + a_k b\right] &&= \sum_{1\leq k\leq K} a_k \left[\underbrace{\sum_{1\leq \ell\leq K} a_\ell}_{=1} + a_k(b-1)\right] \\
&= 1 + (b-1)\sum_{1\leq k\leq K} a_k^2.
\end{aligned}$$

Hence, under Model 4.6, the marginal log-likelihood of a couple of neighbor sites is

$$\log p_\theta(y_s, y_{s'}) = \log\left(\sum_{k=1}^K \sum_{\ell=1}^K \omega_{k\ell}\, \mathcal{P}(y_s;\lambda_k)\, \mathcal{P}(y_{s'};\lambda_\ell)\right),$$

where the associated couple of latent variables $Z_s, Z_{s'}$ can only take K^2 different values, instead of K^{nm} for the whole latent field $\boldsymbol{Z}$.

4.4.4 EM algorithm for composite likelihood inference.

We now demonstrate how an EM algorithm can be constructed to perform maximum composite likelihood inference in the setting where each individual observation Y_s is associated with a single latent variable Z_s. In this case, the entire latent variable vector $\boldsymbol{Z}$ can be partitioned into blocks $C_1, \ldots, C_B$ in the same manner as the set of observed variables $\boldsymbol{Y}$. We may the denote $\boldsymbol{Z}_{C_b} = (Z_s)_{s\in C_b}$.

Because decomposition from Proposition 2.2 applies to each term of the sum of Equation (4.59), we have:

$$c\ell_\theta(\boldsymbol{y}) = \sum_{b=1}^{B} \mathbb{E}_\theta\big[\log p_\theta(\boldsymbol{y}_{C_b}, \boldsymbol{Z}_{C_b}) \mid \boldsymbol{Y}_{C_b} = \boldsymbol{y}_{C_b}\big] - \mathbb{E}_\theta\big[\log p_\theta(\boldsymbol{Z}_{C_b} \mid \boldsymbol{Y}_{C_b} = \boldsymbol{y}_{C_b}) \mid \boldsymbol{Y}_{C_b} = \boldsymbol{y}_{C_b}\big],$$

which suggests the following adaptation of the EM algorithm to composite likelihood inference.

Algorithm 4.4 (EM for composite likelihood)**.** *Repeat until convergence:*

E step: *given the current estimate* $\theta^{(h)}$ *of* θ*, for each block* $1 \le b \le B$*, compute* $p_{\theta^{(h)}}(\boldsymbol{Z}_{C_b} \mid \boldsymbol{Y}_{C_b} = \boldsymbol{y}_{C_b})$*, or at least all the quantities needed to evaluate*

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

M step: *update the estimate of* θ *as*

$$\theta^{(h+1)} = \arg\max_{\theta \in \Theta} \sum_{b=1}^{B} Q_{C_b}(\theta \mid \theta^{(h)}).$$

The EM algorithm (Algorithm 4.4) for composite likelihood inference shares a similar property to that established in Proposition 2.3 for the standard EM algorithm (Algorithm 2.1) used in maximum likelihood inference.

Proposition 4.12 (EM for composite likelihood)**.** *The sequence* $(\theta^{(h)})_{h \ge 0}$ *defined by the EM Algorithm 4.4 is such that:*

$$c\ell_{\theta^{(h+1)}}(\boldsymbol{y}) \ge c\ell_{\theta^{(h)}}(\boldsymbol{y}), \qquad \forall h \ge 0.$$

Proof of Proposition 4.12

The proof follows the same line as this of Proposition 2.3: because $\theta^{(h+1)}$ maximizes $\sum_{b=1}^{B} Q_{C_b}(\theta \mid \theta^{(h)})$, we have that

$$\begin{aligned} 0 &\le \sum_{b=1}^{B} Q_{C_b}(\theta^{(h+1)} \mid \theta^{(h)}) - \sum_{b=1}^{B} Q_{C_b}(\theta^{(h)} \mid \theta^{(h)}) = \sum_{b=1}^{B} \mathbb{E}_{\theta^{(h)}}\left[\log\left(\frac{p_{\theta^{(h+1)}}\left(\boldsymbol{y}_{C_b}, \boldsymbol{Z}_{C_b}\right)}{p_{\theta^{(h)}}\left(\boldsymbol{y}_{C_b}, \boldsymbol{Z}_{C_b}\right)}\right) \,\middle|\, \boldsymbol{Y}_{C_b} = \boldsymbol{y}_{C_b}\right] \\ &\le \sum_{b=1}^{B} \log\left(\mathbb{E}_{\theta^{(h)}}\left[\frac{p_{\theta^{(h+1)}}\left(\boldsymbol{y}_{C_b}, \boldsymbol{Z}_{C_b}\right)}{p_{\theta^{(h)}}\left(\boldsymbol{y}_{C_b}, \boldsymbol{Z}_{C_b}\right)} \,\middle|\, \boldsymbol{Y}_{C_b} = \boldsymbol{y}_{C_b}\right]\right) \end{aligned}$$

thanks to Jensen's inequality. Then, applying the same simplification as in the proof of Proposition 2.3 within each block, we get for each $b \in \{1, \dots B\}$ that

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

Hence, summing over all the blocks, we get

$$\begin{aligned} 0 &\le \sum_{b=1}^{B} \left(Q_{C_b}(\theta^{(h+1)} \mid \theta^{(h)}) - Q_{C_b}(\theta^{(h)} \mid \theta^{(h)})\right) \\ &\le \sum_{b=1}^{B} \left(\log p_{\theta^{(h+1)}}\left(\boldsymbol{y}_{C_b}\right) - \log p_{\theta^{(h)}}\left(\boldsymbol{y}_{C_b}\right)\right) = c\ell_{\theta^{(h+1)}}(\boldsymbol{y}) - c\ell_{\theta^{(h)}}(\boldsymbol{y}), \end{aligned}$$

which concludes the proof.

Case of the hidden Markov random field. Going back to Model 4.6 and the pairs of neighbor sites as blocks, we have for block $C_b = (s, s')$:

$$\log p_\theta\left(\boldsymbol{y}_{C_b}, \boldsymbol{Z}_{C_b}\right) = \sum_{k=1}^{K} \sum_{\ell=1}^{K} \boldsymbol{Z}_{sk} Z_{s'\ell} \left(\log \omega_{k\ell} + \log \mathcal{P}(y_s; \lambda_k) + \log \mathcal{P}(y_{s'}; \lambda_\ell)\right)$$

which gives

$$Q_{C_b}(\theta \mid \theta^{(h)}) = \sum_{k=1}^{K} \sum_{\ell=1}^{K} \tau^{(h)}_{ss',k\ell} \left(\log \omega_{k\ell} + \log \mathcal{P}(y_s; \lambda_k) + \log \mathcal{P}(y_{s'}; \lambda_\ell)\right).$$

The E step of Algorithm 4.4 then consists in the evaluation, for each block $C_b = (s, s')$, of the joint conditional probability

$$\tau^{(h)}_{ss',k\ell} := \mathbb{P}_{\theta^{(h)}}(Z_s = k, Z_{s'} = \ell \mid Y_s = y_s, Y_{s'} = y_{s'}) = \frac{\omega^{(h)}_{k\ell}\mathcal{P}(y_s; \lambda^{(h)}_k)\mathcal{P}(y_{s'}; \lambda^{(h)}_\ell)}{\sum_{u=1}^K \sum_{v=1}^K \omega^{(h)}_{uv}\mathcal{P}(y_s; \lambda^{(h)}_u)\mathcal{P}(y_{s'}; \lambda^{(h)}_v)},$$

which makes the E step fully explicit.

As for the M step, the update formulas for the Poisson parameters follows as

$$\lambda^{(h+1)}_k = \sum_{b=1}^B \left(\sum_{\ell=1}^K \tau^{(h)}_{ss',k\ell} y_s + \tau^{(h)}_{ss',\ell k} y_{s'} \right) \Bigg/ \sum_{b=1}^B \left(\sum_{\ell=1}^K \tau^{(h)}_{ss',k\ell} + \tau^{(h)}_{ss',\ell k} \right).$$

The update for the hidden Markov random field parameters a and b is more involved because of the form of Equation (4.60). Let us rewrite $\sum_{C_b} Q_{C_b}(\theta \mid \theta^{(h)})$ as a function of a and b:

$$\begin{aligned}
\sum_{b=1}^B Q_{C_b}(\theta \mid \theta^{(h)}) &= \sum_{s \sim s'} \sum_{k,\ell=1}^K \tau^{(h)}_{ss',k\ell} \log \omega_{k\ell} + C \\
&= \sum_{s \sim s'} \sum_{k,\ell=1}^K \tau^{(h)}_{ss',k\ell} \left[\log a_k + \log a_\ell + \mathbb{I}_{\{k=\ell\}} \log b - \log\left(1 + (b-1)\sum_{k=1}^K a_k^2\right)\right] + C \\
&= \sum_{k,\ell=1}^K \sum_{s \sim s'} \tau^{(h)}_{ss',k\ell} \left[\log a_k + \log a_\ell + \mathbb{I}_{\{k=\ell\}} \log b\right] - \log\left(1 + (b-1)\sum_{k=1}^K a_k^2\right) + C.
\end{aligned}$$

While the zeros of the partial derivatives cannot be determined analytically, they can be approximated numerically.

4.4.5 Conclusion

In models such as the Potts Hidden Markov Model, the full likelihood is intractable due to the presence of an unknown normalizing constant, making direct application of the EM algorithm unfeasible. In such cases, we showed that we can resort to a composite likelihood approach, which provides a practical alternative and fits naturally within a standard EM framework. For more background on composite likelihood methods, we advise the review by Varin et al. [2011].

Chapter 5

Deterministic approximation of the E step

Contents

So far, we have considered models in which the distribution $p_\theta(\boldsymbol{Z} \mid \boldsymbol{Y})$ either has a known form (e.g., discrete or Gaussian, as discussed in Chapter 3) or possesses moments that can be computed exactly (Chapter 4). However, in many situations, neither the distribution itself nor its moments have a known or tractable form, making the integration of the complete log-likelihood in the E-step infeasible. This typically occurs when

the observation model follows a complex distribution[1] or when $\mathcal{Z}$ is a very large space, and no conditional independence between the Z_i's can be exploited to reduce the computational effort required to evaluate the moments of $p_\theta(\boldsymbol{Z} \mid \boldsymbol{Y})$. In such cases, one can resort to a deterministic approximation of this distribution within a carefully chosen family of distributions. Resorting to such an approximation is known as variational inference (see Blei et al. [2017] for a didactic review). The adjective variational refers to the fact that the approximation is optimal in a sense to be specified. In the context of this book, the combination of such an approximation with the EM algorithm leads to the Variational Expectation-Maximization algorithm (referred to as VEM hereafter).

In Section 5.1, we introduce the variational EM methodology and discuss its implications for inference—for instance for model selection.

In the subsequent sections, we present three families of probabilistic models whose complexity necessitates alternatives to the classical EM algorithm for parameter inference. We illustrate the use of the variational EM algorithm with the following three models:

1. The Stochastic Block Model (SBM, Section 5.2), which is used to model network data, such as ecological networks;
2. The Poisson Log-Normal Model (PLN, Section 5.3), which is used to model dependent count data, such as species abundances in biodiversity studies;
3. The Variational Autoencoder (VAE, Section 5.4), a nonlinear dimension reduction method based on neural networks, in the same spirit as PCA.

5.1 Variational version of the EM algorithm

We consider the setting in which the E step is infeasible due to the intractable or unknown form of $p_\theta(\boldsymbol{Z} \mid \boldsymbol{Y} = \boldsymbol{y})$. The principle of variational inference is to replace this complex distribution with a simpler one, chosen from a predefined, tractable parametric family. More precisely, we select the member of this simpler family that is closest to the true distribution, according to a specified measure of similarity between distributions. Among the various possible metrics, the most commonly used is the Kullback–Leibler (KL) divergence.

5.1.1 The Kullback-Leibler divergence

Let q and p be two probability density functions[2] on $\mathbb{R}^{d_z}$. The Kullback–Leibler (KL) divergence from q to p is defined as:

$$\mathsf{KL}\left[q(\boldsymbol{Z}) \parallel p(\boldsymbol{Z})\right] = \mathbb{E}_q\left[\log \frac{q(\boldsymbol{Z})}{p(\boldsymbol{Z})}\right] = \int_{\mathbb{R}^{d_z}} \log\left(\frac{q(z)}{p(z)}\right) q(z)\, \mathrm{d}z\,.$$

Note that q and p do not play symmetric roles in this definition. In fact, in general:

$$\mathsf{KL}\left[q(\boldsymbol{Z}) \parallel p(\boldsymbol{Z})\right] \neq \mathsf{KL}\left[p(\boldsymbol{Z}) \parallel q(\boldsymbol{Z})\right].$$

The KL divergence is therefore not symmetric and thus does not qualify as a distance in the strict mathematical sense. However, it does satisfy a key property of distances: non-negativity.

Using Jensen's inequality (Lemma 2.1) applied to the negative logarithm, we obtain:

$$\mathsf{KL}\left[q(\boldsymbol{Z}) \parallel p(\boldsymbol{Z})\right] = -\mathbb{E}_q\left[\log \frac{p(\boldsymbol{Z})}{q(\boldsymbol{Z})}\right] \geq -\log \mathbb{E}_q\left[\frac{p(\boldsymbol{Z})}{q(\boldsymbol{Z})}\right] = -\log \int_{\mathbb{R}^{d_z}} p(z)\, \mathrm{d}z = -\log(1) = 0\,.$$

Moreover, the KL divergence equals zero if and only if $q = p$ almost everywhere. Thus, the KL divergence is strictly positive unless the two distributions are identical.

5.1.2 The variational EM

In our context, where $p(\boldsymbol{Z} \mid \boldsymbol{Y} = \boldsymbol{y})$ is intractable, the EM algorithm cannot be applied directly, as the update of $\theta^{(h)}$ at step h requires maximizing the expectation of a function with respect to $p_{\theta^{(h)}}(\boldsymbol{Z} \mid \boldsymbol{Y} = \boldsymbol{y})$.

The idea of the VEM algorithm is to instead maximize the expectation of the same function, but with respect to a "close enough" distribution for which this expectation is tractable. In practice, this variational distribution

[1] Readers familiar with Bayesian statistics might think of situations where no conjugacy exists between the prior and the observation model.

[2] In the following, p will represent the true distribution of $\boldsymbol{Z} \mid \boldsymbol{Y} = \boldsymbol{y}$, and q will denote its simpler approximation.

is chosen from a fixed parametric family—such as Gaussian distributions—which we denote by $\mathcal{Q}$. The closeness of the variational distribution to $p_{\theta^{(h)}}(\boldsymbol{Z} \mid \boldsymbol{Y} = \boldsymbol{y})$ is measured using the Kullback–Leibler divergence.

At each iteration h, this approach requires finding, given the current parameter $\theta^{(h)}$, the best approximation to $p_{\theta^{(h)}}(\boldsymbol{Z} \mid \boldsymbol{Y} = \boldsymbol{y})$ within the family $\mathcal{Q}$. This step is known as the variational E-step.

Algorithm 5.1 (Variational EM)**.** *Starting from an initial guess $\theta^{(0)}$, repeat until convergence:*

VE step: *Given the current estimate $\theta^{(h)}$ of θ, and a family of distributions $\mathcal{Q}$, compute:*

$$q^{(h)}(\boldsymbol{Z}) \quad = \quad \arg\min_{q \in \mathcal{Q}} \mathsf{KL}\left[q(\boldsymbol{Z}) \parallel p_{\theta^{(h)}}(\boldsymbol{Z} \mid \boldsymbol{Y} = \boldsymbol{y})\right].$$

M step: *Update the estimate of θ as*

$$\theta^{(h+1)} \quad = \quad \arg\max_{\theta} \mathbb{E}_{q^{(h)}}[\log p_\theta(\boldsymbol{y}, \boldsymbol{Z})].$$

Although the substitution made in the E-step may appear intuitive, a natural question arising from Algorithm 5.1 concerns the feasibility of the minimization it involves, since it depends on the intractable posterior distribution $p_{\theta^{(h)}}(\boldsymbol{Z} \mid \boldsymbol{Y} = \boldsymbol{y})$. Nevertheless, as the next section will show, the algorithm remains operational despite this apparent difficulty.

5.1.3 The VEM as an alternating optimization of a lower bound of the likelihood

Interestingly, the VEM algorithm described above can be interpreted as a maximization procedure of a lower bound on the likelihood $\log p_\theta(\boldsymbol{y})$. Indeed, for every θ, the Kullback-Leibler divergence can be rewritten as:

$$\begin{aligned}\mathsf{KL}\left[q(\boldsymbol{Z}) \parallel p_\theta(\boldsymbol{Z} \mid \boldsymbol{Y} = \boldsymbol{y})\right] &= \mathbb{E}_q\left[\log \frac{q(\boldsymbol{Z})}{p_\theta(\boldsymbol{Z} \mid \boldsymbol{Y} = \boldsymbol{y})}\right] \\ &= \mathbb{E}_q[\log q(\boldsymbol{Z})] - \mathbb{E}_q[\log p_\theta(\boldsymbol{Z}, \boldsymbol{y})] + \mathbb{E}_q[\log p_\theta(\boldsymbol{y})].\end{aligned}$$

Note that the first term is, by definition, the negative Shannon entropy of q, denoted $\mathsf{Ent}_q[\boldsymbol{Z}]$, and the last term is $\log p_\theta(\boldsymbol{y})$, as it does not depend on q. Therefore, we obtain:

$$\log p_\theta(\boldsymbol{y}) = \mathsf{KL}\left[q(\boldsymbol{Z}) \parallel p_\theta(\boldsymbol{Z} \mid \boldsymbol{Y} = \boldsymbol{y})\right] + \mathbb{E}_q[\log p_\theta(\boldsymbol{Z}, \boldsymbol{y})] + \mathsf{Ent}_q[\boldsymbol{Z}]. \tag{5.1}$$

Equation (5.1) has two important implications:

1. Since the Kullback-Leibler divergence is always non-negative, the sum of the last two terms on the right-hand side of Equation (5.1) provides a lower bound for the log-likelihood. Because the log-likelihood is sometimes[3] referred to as the evidence, we define, for q, θ, and $\boldsymbol{y}$, the Evidence Lower Bound, or ELBO:

$$\mathsf{ELBO}(q, \theta, \boldsymbol{y}) = \mathbb{E}_q[\log p_\theta(\boldsymbol{y}, \boldsymbol{Z})] + \mathsf{Ent}_q[\boldsymbol{Z}] \leq \log p_\theta(\boldsymbol{y}). \tag{5.2}$$

   and we obtain:

$$\mathsf{KL}\left[q(\boldsymbol{Z}) \parallel p_{\theta^{(h)}}(\boldsymbol{Z} \mid \boldsymbol{Y} = \boldsymbol{y})\right] + \mathsf{ELBO}(q, \theta, \boldsymbol{y}) = \log p_\theta(\boldsymbol{y})$$

2. As $\log p_\theta(\boldsymbol{y})$ does not depend on q, minimizing $\mathsf{KL}\left[q(\boldsymbol{Z}) \parallel p_{\theta^{(h)}}(\boldsymbol{Z} \mid \boldsymbol{Y} = \boldsymbol{y})\right]$ (as in the VE step) is then equivalent to maximizing $\mathsf{ELBO}(q, \theta, \boldsymbol{y})$, or, equivalently:

$$\arg\min_{q \in \mathcal{Q}} \mathsf{KL}\left[q(\boldsymbol{Z}) \parallel p_{\theta^{(h)}}(\boldsymbol{Z} \mid \boldsymbol{Y} = \boldsymbol{y})\right] = \arg\max_{q \in \mathcal{Q}} \mathsf{ELBO}(q, \theta, \boldsymbol{y}). \tag{5.3}$$

   This equality is of first interest as the maximization of the ELBO does not require to know $p_{\theta^{(h)}}(\boldsymbol{Z} \mid \boldsymbol{Y} = \boldsymbol{y})$, which makes Algorithm 5.1 operational.

The VE step is done by maximizing, with respect to q, the evidence lower bound. However, one can notice that the M step also maximizes the ELBO, with respect to θ this time, as the entropy term in Equation (5.2) does not depend on θ:

$$\arg\max_{\theta} \mathbb{E}_{q^{(h)}}[\log p_\theta(\boldsymbol{y}, \boldsymbol{Z})] = \arg\max_{\theta} \mathsf{ELBO}(q, \theta, \boldsymbol{y}).$$

Algorithm 5.1 can be reformulated as:

[3] Particularly in Bayesian statistics

Algorithm 5.2 (Variational EM -reformulation)**.** *Starting from an initial guess $\theta^{(0)}$, repeat until convergence:*

VE step: *Given the current estimate $\theta^{(h)}$ of θ, and a family of distributions $\mathcal{Q}$, compute:*

$$q^{(h)}(\boldsymbol{Z}) \quad = \quad \arg\max_{q \in \mathcal{Q}} \mathsf{ELBO}(q, \theta, \boldsymbol{y}) \, .$$

M step: *Update the estimate of θ as*

$$\begin{aligned} \theta^{(h+1)} &= \arg\max_{\theta \in \Theta} \mathsf{ELBO}(q^{(h)}, \theta, \boldsymbol{y}) \\ &= \arg\max_{\theta} \mathbb{E}_{q^{(h)}}[\log p_\theta(\boldsymbol{y}, \boldsymbol{Z})] \, . \end{aligned}$$

Thus the variational algorithm is actually an algorithm to maximize jointly in q and θ the ELBO by successively maximizing in q then in θ. This process is known as alternating optimization [Bezdek and Hathaway, 2003]. The VEM in this last version is operational since $\mathcal{Q}$ is chosen such that $\mathsf{ELBO}(q, \theta, \boldsymbol{y})$ can be computed in a close form.

The following proposition shows that each step of the VEM algorithm increases the ELBO.

Proposition 5.1. *The sequence of $(q^{(h)}, \theta^{(h)})$ defined by the VEM algorithm 5.1 is such that:*

$$\mathsf{ELBO}(q^{(h+1)}, \theta^{(h+1)}, \boldsymbol{y}) \geq \mathsf{ELBO}(q^{(h)}, \theta^{(h)}, \boldsymbol{y}).$$

Proof of Proposition 5.1

Because

$$q^{(h+1)}(\boldsymbol{Z}) = \arg\min_{q \in \mathcal{Q}} \mathsf{KL}\left[q(\boldsymbol{Z}) \parallel p_{\theta^{(h)}}(\boldsymbol{Z} \mid \boldsymbol{Y} = \boldsymbol{y})\right] = \arg\max_{q \in \mathcal{Q}} \mathsf{ELBO}(q, \theta^{(h)}, \boldsymbol{y})$$

we have that

$$\mathsf{ELBO}(q^{(h+1)}, \theta^{(h)}, \boldsymbol{y}) \geq \mathsf{ELBO}(q^{(h)}, \theta^{(h)}, \boldsymbol{y}),$$

and because

$$\theta^{(h+1)} = \arg\max_{\theta} \mathbb{E}_{q^{(h+1)}}[\log p_\theta(\boldsymbol{y}, \boldsymbol{Z})] = \arg\max_{\theta \in \Theta} \mathsf{ELBO}(q^{(h+1)}, \theta, \boldsymbol{y})$$

we have that

$$\mathsf{ELBO}(q^{(h+1)}, \theta^{(h+1)}, \boldsymbol{y}) \geq \mathsf{ELBO}(q^{(h+1)}, \theta^{(h)}, \boldsymbol{y}).$$

Proposition 5.1 parallels Proposition 2.3: the VEM algorithm enjoys the same property for the $\mathsf{ELBO}(q, \theta, \boldsymbol{y})$ as the EM algorithm does for the observed data likelihood $\log p_\theta(\boldsymbol{y})$. Notably, the proof holds when the posterior $p_\theta(\boldsymbol{Z} \mid \boldsymbol{Y} = \boldsymbol{y})$ is tractable, that is, when the variational distribution satisfies $q^{(h)}(\boldsymbol{Z}) = p_{\theta^{(h)}}(\boldsymbol{Z} \mid \boldsymbol{Y} = \boldsymbol{y})$. Hence, the proof of Proposition 5.1 also provides an alternative proof of Proposition 2.3.

In other respects, it is important to emphasize that we no longer maximize the likelihood directly, which raises questions regarding the theoretical properties of the obtained estimator. They will be discussed in Subsection 5.1.6.

Alternative formulation for the ELBO As the ELBO is now our objective function, it might be insightful to have another point of view on its meaning by rewriting in the following way:

$$\begin{aligned} \mathsf{ELBO}(q, \theta, \boldsymbol{y}) &= \mathbb{E}_q[\log p_\theta(\boldsymbol{y}, \boldsymbol{Z})] + \mathsf{Ent}_q[\boldsymbol{Z}] \\ &= \mathbb{E}_q[\log p_\theta(\boldsymbol{y} \mid \boldsymbol{Z})] + \mathbb{E}_q[\log p_\theta(\boldsymbol{Z})] - \mathbb{E}_q[\log q(\boldsymbol{Z})] \\ &= \mathbb{E}_q[\log p_\theta(\boldsymbol{y} \mid \boldsymbol{Z})] - \mathsf{KL}\left[q(\boldsymbol{Z}) \parallel p_\theta(\boldsymbol{Z})\right] \end{aligned} \tag{5.4}$$

In this alternative writing we can recognize the first term as a measure of the quality of reconstruction of the model p_θ, whereas the second term can be seen as a regularization penalty, *i.e.*, q will be forced to stick to the marginal (or prior) distribution of Z. This formulation is the most frequent one in machine learning applications.

5.1.4 The mean field approximation

The regular E step has been transformed into a variational E step (VE step), the aim of which is to find the best approximation of the posterior distribution within a certain class of distributions $\mathcal{Q}$. The choice of $\mathcal{Q}$ is crucial and results from a balance between the quality of the approximation (requiring $\mathcal{Q}$ to be as large as possible) and the computational burden (requiring $\mathcal{Q}$ to be as small as possible). Choosing $\mathcal{Q}$ leads to two main problems: fixing the class of parametric distributions for $q \in \mathcal{Q}$ (for instance, Gaussian distributions), and choosing the structure of dependence for the variables with distribution q (for example, whether the Gaussian distribution is diagonal).

In terms of dependence structure, the simplest family is the one where every random vectors has independent components, namely, the class of factorized distributions:

$$\mathcal{Q}_{\text{fact}} = \left\{ q : q(\boldsymbol{Z}) = \prod_{i}^{d_z} q_i(Z_i) \right\}. \tag{5.5}$$

This class of approximate distributions yields a mean-field approximation, because of the following property.

Proposition 5.2 (Mean-field approximation)**.** *The optimal distribution*

$$\hat{q}(\boldsymbol{z}) = \underset{q \in \mathcal{Q}_{fact}}{\arg\min}\, \mathsf{KL}\left[q(\boldsymbol{Z}) \parallel p_\theta(\boldsymbol{Z} \mid \boldsymbol{Y} = \boldsymbol{y})\right]$$

satisfies, for each latent variable Z_i,

$$\hat{q}_i(z_i) \propto \exp\left(\mathbb{E}_{\hat{q}_{\setminus i}}[\log p_\theta(\boldsymbol{Y}, z_i, \boldsymbol{Z}_{\setminus i})]\right), \tag{5.6}$$

where $\boldsymbol{Z}_{\setminus i}$ stands for the set of all latent variables, except Z_i: $\boldsymbol{Z}_{\setminus i} = (Z_j)_{j \neq i}$ and $\hat{q}_{\setminus i}$ is the approximate distribution of all Z_j's, except Z_i:

$$\hat{q}_{\setminus i}(\boldsymbol{z}_{\setminus i}) = \prod_{j \neq i} \hat{q}_j(z_j).$$

Proof of Proposition 5.2

As stated by Equation (5.3), minimizing the Kullback-Liebler divergence is equivalent to maximizing the ELBO given by Equation (5.2). Moreover, under the factorization (5.5), the ELBO satisfies:

$$\mathsf{ELBO}(q, \theta, \boldsymbol{y}) = \mathbb{E}_q[\log p_\theta(\boldsymbol{y}, \boldsymbol{Z})] - \mathbb{E}_{q_i}\left[\log \prod_{i=1}^{d_z} q_i(Z_i)\right],$$

where we rewrote the entropy term under the factorized distribution assumption. Now, let's consider a fixed index $1 \leq i \leq d_z$, then, note that:

$$\begin{aligned}\mathsf{ELBO}(q, \theta, \boldsymbol{y}) &= \mathbb{E}_{q_i}\left[\mathbb{E}_{q_{\setminus i}}\left[\log \frac{p_\theta(\boldsymbol{y}, Z_i, \boldsymbol{Z}_{\setminus i})}{\prod_{j=1, j\neq i}^{d_z} q_j(Z_j)}\right]\right] - \mathbb{E}_{q_i}[\log(q_i(Z_i))], \\ &= -\mathbb{E}_{q_i}\left[\log \frac{q_i(Z_i)}{\exp\left(\mathbb{E}_{q_{\setminus i}}\left[\log \frac{p_\theta(\boldsymbol{y}, Z_i, \boldsymbol{Z}_{\setminus i})}{\prod_{j=1, j\neq i}^{d_z} q_j(Z_j)}\right]\right)}\right].\end{aligned}$$

Now, let's consider[a]

$$\tilde{q}_i(z_i) = \frac{\exp\left(\mathbb{E}_{q_{\setminus i}}\left[\log \frac{p_\theta(\boldsymbol{y}, z_i, \boldsymbol{Z}_{\setminus i})}{\prod_{j=1, j\neq i}^{d_z} q_j(Z_j)}\right]\right)}{\int \exp\left(\mathbb{E}_{q_{\setminus i}}\left[\log \frac{p_\theta(\boldsymbol{y}, z, \boldsymbol{Z}_{\setminus i})}{\prod_{j=1, j\neq i}^{d_z} q_j(Z_j)}\right]\right) \mathrm{d}z} \propto \exp\left(\mathbb{E}_{q_{\setminus i}}[\log p_\theta(\boldsymbol{y}, z_i, \boldsymbol{Z}_{\setminus i})]\right),$$

which defines a probability distribution function for Z_i. Thus, we have for every $1 \leq i \leq d_z$:

$$\mathsf{ELBO}(q, \theta, \boldsymbol{y}) = -\mathsf{KL}\left[q_i(Z_i) \parallel \tilde{q}_i(Z_i)\right] + \text{cst},$$

where the cst term does not depend on q_i. Then, if all other components of $q = (q_1, \ldots, q_{d_z})$ are fixed, replacing $q_i(z_i)$ by $\tilde{q}_i(z_i)$ would yield a null KL divergence, that is, a maximal value for the negative

divergence, and then a greater ELBO. This implies that:

$$\hat{q}_i(z_i) \propto \exp\left(\mathbb{E}_{\hat{q}_{\setminus i}}[\log p_\theta(\boldsymbol{Y}, z_i, \boldsymbol{Z}_{\setminus i})]\right) ,$$

which concludes the proof.

[a] Note that the integral at the denominator has a finite value, as it is upper bounded by $p(\boldsymbol{y})$ (by Jensen inequality).

Remark. *A more generic but more mathematically involved way of proving this result is to due to Euler and Lagrange, and uses the calculus of variations, which explains the term variational inference.*

However, note that the argument of the proposed proof naturally leads to an iterative algorithm to obtain $\hat{q}(\boldsymbol{z})$*. Starting from* $q^{(0)} = \left(q_1^{(0)}, \ldots, q_{d_z}^{(0)}\right)$*, and setting, sequentially, for* $t \geq 0$ *and all* $1 \leq i \leq d_z$*:*

$$q_i^{(t+1)} \propto \mathrm{e}^{\mathbb{E}_{q_{\setminus i}^{(t)}}[\log p_\theta(\boldsymbol{y}, z_i, \boldsymbol{Z}_{\setminus i})]} ,$$

where, in this context:

$$q_{\setminus i}^{(t)} = (q_1^{(t+1)}, \ldots, q_{i-1}^{(t+1)}, q_{i+1}^{(t)}, \ldots, q_{d_z}^{(t)}) ,$$

leads to a sequence $q^{(t)}$ *that increases the ELBO, which is a bounded function. Then, by a fixed point argument, any limit*[4] $\hat{q}(\boldsymbol{z})$ *of this sequence satisfies Equation* (5.6)*. This algorithm is called the coordinate ascent variational inference (or CAVI, see [Bishop, 2006, Chapter 10]).*

5.1.5 Variational versions of BIC and ICL

As discussed in Section 2.4, model selection is primarily based on the log-likelihood evaluated at the maximum likelihood estimator (MLE). This raises the question of how model selection can be performed when variational inference is used – a method that, as will be seen in the rest of the chapter, is sometimes unavoidable. The common practice in such cases is to replace each intractable quantity with its corresponding variational approximation.

Now let us denote $\mathrm{BIC}_{\mathsf{V}}$ and $\mathrm{ICL}_{\mathsf{V}}$ the variational versions of BIC and ICL, $\hat{q}$ the best variational approximation of $p_{\widehat{\theta}_{\mathsf{V}}}(\boldsymbol{Z} \mid \boldsymbol{Y} = \boldsymbol{y})$, and $\hat{\boldsymbol{z}}_{\mathsf{V}} = \arg\max_z \hat{q}(\boldsymbol{z})$, we have, for a model m

$$\begin{aligned}
\mathrm{BIC}_{\mathsf{V}}(m) &= \mathsf{ELBO}(\hat{q}, \widehat{\theta}_{\mathsf{V}}, \boldsymbol{y}) - \mathrm{pen}(m) \\
\mathrm{ICL}_{\mathsf{V},1}(m) &= \log p_{\widehat{\theta}_{\mathsf{V}}}(\boldsymbol{y}, \hat{\boldsymbol{z}}_{\mathsf{V}}) - \mathrm{pen}(m) \\
\mathrm{ICL}_{\mathsf{V},2}(m) &= \mathsf{ELBO}(\hat{q}, \widehat{\theta}_{\mathsf{V}}, \boldsymbol{y}) - \mathsf{Ent}_{\hat{q}}[\boldsymbol{Z}] - \mathrm{pen}(m) .
\end{aligned}$$

The BIC and ICL criteria will be detailed in the following sections for various models.

5.1.6 Conclusion

Variational approximations provide a convenient way to deal with models where the conditional distribution $p_\theta(Z \mid Y)$ is intractable. Still, it must be reminded that the associated VEM Algorithm 5.1 aims at maximizing the ELBO (5.2) and not the genuine marginal log-likelihood $\log p_\theta(Y)$.

An important consequence is that the resulting estimates are not maximum likelihood estimates and do not enjoy the general properties of maximum likelihood theory. For example, their consistency, or asymptotic normality is not guaranteed in general, and uncertainty measures (e.g. confidence intervals) are often not available. In practice, one often has to assess the accuracy of variational estimates with careful simulation studies and to resort to resampling or jackknife post-processing to have access to the variability of the estimates.

Variational inference is one example of the no-free-lunch principle: what we gain in terms of model complexity and computational efficiency is partially lost in terms of statistical guaranties.

5.2 Network analysis with SBM

In this section, we introduce a probabilistic model for analyzing ecological networks: the stochastic block model (SBM).

A network is considered unipartite when it represents interactions within a single group of species. For example, food webs are directed unipartite networks (see Aubert et al. [2022] for a more detailed introduction).

[4] Which would depend on $q^{(0)}$.

In contrast, a network is said to be bipartite when it describes interactions between two distinct types of entities—such as fungi and trees, which will be discussed later. A common example of a bipartite network is a plant-pollinator system.

We begin by presenting the Stochastic Block Model (SBM) for unipartite networks, along with the Variational Expectation-Maximization (VEM) algorithm used to estimate its parameters and a criterion for selecting the number of blocks. The extension of this model to bipartite networks is discussed at the end of the section (Section 5.2.7; technical details for this case are provided in the Appendix C.2.

5.2.1 Network data and question

The study of inter-species relations within an ecosystem has been an important theme in ecology since the publication of Charles Elton's seminal work on food chains Elton [1927]. Relationships between species can be represented as a network of interactions, with nodes taking the place of biological entities (generally species) of interest and edges (or links) representing the interactions in question. The analysis of ecological networks is the focus of much recent ecological literature and has enjoyed an upsurge in interest in the last 20 years, notably through the work of Dunne et al. [2002]. Networks of interactions may be studied at microscopic level (e.g. microbial network) or at higher levels (e.g. plant-pollinator relationships). They may represent a wide variety of interaction types, such as mutualism, competition or predation. The main ecological assumption is that all species do not play the same role in the network and analyzing the structure of a network may be crucial for understanding the organization of an ecosystem, notably via the extraction of its organizational structure in summary form. Two types of approach may be used in this context. The first family of approaches are descriptive, based on different metrics used to characterize properties at node level or across the whole network. The second family consists of detecting clusters of species with similar interaction properties in an agnostic approach, in that it does not aim to highlight a particular structure, but rather to cluster nodes which behave in the same way in the network. Many probabilistic models have been proposed to model the heterogeneity of connexions in a network. In this book, we intentionnally focus on the Stochastic Block Model (SBM). Let's consider this example from Vacher et al. [2008].

Dataset 5.1. ***(Host-host network)*** *The presence of 157 parasitic fungal species was collected on 51 tree taxa, based on a long-term (1972–2005) survey of forest health at the regional scale. In a first approach we can build the unipartite tree-tree network by counting for any pair of trees (i, j) the number of fungus species they share. In that case, the network is said to be weighted since the number of shared species represents a strength of interactions. Let y_{ij} denote the number of shared species between trees i and j. The matrix $\boldsymbol{y} = \{y_{ij}\}_{1 \leq i,j \leq n}$ is referred as the adjacency matrix.*

From network to adjacency matrix Network data can be encoded in various format. Either it is provided as a pair of edges $((1,2) - (1,3)...)$ with possibly weights on each edge if the relation corresponds to a count. If the network involves a small number of nodes, one can encode the network in a matrix as described before. Figure 5.1 is a graphical representation of the weighted tree-tree network and its adjacency matrix from Vacher et al. [2008]. Our goal is then to perform clustering of trees, such that trees in the same clusters play the same role in the network.

5.2.2 The stochastic block model (SBM)

Let us consider a unipartite network of n nodes encoded in its adjacency matrix. For any pair of nodes (i, j), let

$$\begin{cases} y_{ij} \neq 0 & \text{if } i \text{ and } j \text{ are interaction} \\ y_{ij} = 0 & \text{if } i \text{ and } j \text{ are not in interaction.} \end{cases}$$

The stochastic block model (SBM) is a probabilistic model aiming at clustering nodes with respect to their behavior in the network. Introduced in the field of sociology by Snijders and Nowicki [1997], this model assumes that nodes are divided into latent blocks (groups, clusters, etc.) containing entities with similar connection profiles. SBM has been shown effective for identifying clusters of nodes or blocks which play the same role in a network, irrespective of the type of structure in question. In the spirit of mixture models (Model 3.1, for instance), blocks are introduced as latent variables in both models.

Model 5.1. *(SBM) Let $(Y_{ij})_{i,j=1,\ldots,n}$ be the adacency of a non-oriented network involving n nodes. In the SBM, the nodes are assumed to be divided into K latent groups (blocks). Let Z_i be the categorical random*

variable encoding this clustering:

$$Z_i \overset{iid}{\sim} \text{Cat}(\boldsymbol{\omega}) \tag{5.7}$$

where $\boldsymbol{\omega} = (\omega_1, \ldots, \omega_K) \in [0,1]^K$, $\sum_{k=1}^K \omega_k = 1$. *Then, conditionally to the latent variables* $\boldsymbol{Z}$, *the nodes are connected independently:*

$$Y_{ij} \mid \{Z_i = k, Z_j = \ell\} \overset{ind}{\sim} \mathcal{F}(\alpha_{k\ell}) \qquad \forall (i,j) \mid j < i \tag{5.8}$$

- *If the network represents binary interactions then* $\mathcal{F}(\alpha_{k\ell})$ *is the Bernoulli distribution:*

$$Y_{ij} \mid \{Z_i = k, Z_j = \ell\} \overset{ind}{\sim} \mathcal{B}ern(\alpha_{k\ell}).$$

- *In case where the network is weighted,* $\mathcal{F}(\alpha_{k\ell})$ *must be adapted. For instance, in case where the strengh is a count, one can chose the Poisson*

$$Y_{ij} \mid \{Z_i = k, Z_j = \ell\} \overset{ind}{\sim} \mathcal{P}(\alpha_{k\ell})$$

or the Negative Binomial distribution [Mariadassou et al., 2010].

Note that is general, the self-interactions are not considered so the $(Y_{ii})_{i=1,\ldots,n}$ are not modelized.

Remark (Oriented networks)**.** *As stated earlier, we develop the calculus only for non-oriented networks, meaning that* $y_{ij} = y_{ji}$ *for any pair of nodes. As a result, only the entries* Y_{ij} *for* $j < i$ *need to be modeled (see Equation* (5.8)*), and the parameters* $\alpha_{k\ell}$ *must satisfy the symmetry condition* $\alpha_{k\ell} = \alpha_{\ell k}$*. This type of network is well-suited for representing symmetric relationships, such as co-occurrence. In contrast, for oriented relationship such as predation, symmetry does not hold (*$y_{ij} \neq y_{ji}$*), and the entire matrix (excluding the diagonal) must be modeled.*

DAG of the SBM For non oriented networks, the DAG is represented in Figure 5.2.

Parameters The parameters of the model are

$$\theta_{\text{lat}} = \boldsymbol{\omega} = \{\omega_k\}_{k=1,\ldots,K} \quad \text{and} \quad \theta_{\text{obs}} = \{\alpha_{k\ell}\}_{k \geq \ell;\, k,\ell=1,\ldots,K},$$

where $\boldsymbol{\omega}$ denotes the vector of block proportions, and the $\alpha_{k\ell}$'s are the connection parameters. The α_{kk} represent intra-block connection parameters, while the $\alpha_{k\ell}$ for $k \neq \ell$ correspond to inter-block connections. For non-oriented networks, $\alpha_{k\ell} = \alpha_{\ell k}$, which does not hold in the case of oriented networks.

Inclusion of covariates Note that, in some cases, network data is supplemented with covariates defined on pairs of nodes. For example, in the tree-tree dataset 5.1, each pair of trees is associated with three types of distances: geographic, genetic, and taxonomic. Let us denote by $\boldsymbol{x}_{ij} \in \mathbb{R}^d$ the vector of covariates associated with the pair (i,j). In such cases, one may consider using these covariates to explain the connection mechanism between nodes. A possible model specification is the following:

$$\mathbb{E}[Y_{ij} \mid Z_i = k, Z_i = \ell] = g^{-1}(\alpha_{k\ell} + \boldsymbol{x}_{ij}^\top \boldsymbol{\beta}) \tag{5.9}$$

where $\boldsymbol{x}_{ij}^\top$ is the transposed vector of $\boldsymbol{x_{ij}}$, g is the link function and $\boldsymbol{\beta}$ is the vector of parameters related to the covariates. Note that when the model includes a single block ($K = 1$), it reduces to a generalized linear model. If the covariates do not fully account for the heterogeneity in $\boldsymbol{y}$, then using multiple blocks ($K > 1$) allows the model to capture some of the remaining structured variability that is not explained by the covariates.

5.2.3 Marginal and complete likelihoods for the SBM

Complete log-likelihood Denoting $\boldsymbol{y} = \{y_{ij}\}_{1\leq j<i\leq n}$ the observations, $\boldsymbol{Z} = \{Z_i\}_{1\leq i\leq n}$ the whole set of latent variables, the complete likelihood of Model 5.1 is

$$\begin{aligned}\log p_\theta(\boldsymbol{y}, \boldsymbol{Z}) &= \log p_\theta(\boldsymbol{Z}) + \log p_\theta(\boldsymbol{y} \mid \boldsymbol{Z}) \\ &= \sum_{i=1}^{n}\sum_{k=1}^{K} Z_{ik}\log\omega_k + \sum_{\substack{i=1,j=1\\ i>j}}^{n}\sum_{k,\ell=1}^{K} Z_{ik}Z_{j\ell}\log f(y_{ij};\alpha_{k\ell}) \qquad (5.10)\\ &= \sum_{i=1}^{n}\sum_{k=1}^{K} Z_{ik}\log\omega_k + \frac{1}{2}\sum_{\substack{i=1,j=1\\ i\neq j}}^{n}\sum_{k,\ell=1}^{K} Z_{ik}Z_{j\ell}\log f(y_{ij};\alpha_{k\ell}) \qquad (5.11)\end{aligned}$$

where Z_{ik} is the usual notation defined in Equation (2.5), and $f(y;\alpha)$ is the density of $\mathcal{F}(\alpha)$. The $\frac{1}{2}$ is there for non oriented network and takes into account the symmetry in y for non-oriented networks. It would disappear for an oriented network.

Log-likelihood The marginal (observed) log-likelihood expression can be written directly by integration over the complete log-likelihood, resulting to:

$$\log p_\theta(\boldsymbol{y}) = \sum_{\boldsymbol{z}\in\{1,\dots,K\}^n} \log p_\theta(\boldsymbol{y}, \boldsymbol{z}) \qquad (5.12)$$

which can be burdensome as soon as n and K increase. Indeed the number of terms in the sum becomes very large. Therefore, direct maximisation is impossible, and one wants to adopt the EM algorithm approach. It is worth noting (as in previous models), that the EM algorithm is also attractive as it would provide the distribution of the latent variables given the observations, *i.e.* of the clusters of the nodes which can be of great interest from the application point of view.

In order to apply to EM algorithm, one has to take the expectation of Equation (5.11), and then should be able to compute the two expectations:

$$\begin{aligned}\mathbb{E}_\theta[Z_{ik} \mid \boldsymbol{Y} = \boldsymbol{y}] &= \mathbb{P}_\theta(Z_i = k \mid \boldsymbol{Y} = \boldsymbol{y}), && 1\leq i\leq n,\ 1\leq k\leq K\\ \mathbb{E}_\theta[Z_{ik}Z_{j\ell} \mid \boldsymbol{Y} = \boldsymbol{y}] &= \mathbb{P}_\theta(Z_i = k, Z_j = \ell \mid \boldsymbol{Y} = \boldsymbol{y}) && 1\leq j<i\leq n,\ 1\leq k,\ell\leq K\,.\end{aligned}$$

However, a quick look at the DAG of the SBM model (Figure 5.2) shows that once conditioned by the observations $\boldsymbol{y}$, the latent variables are not independent anymore and form a clique. So, these two quantities can only be computed by integration of the whole distribution $p_\theta(\boldsymbol{Z} \mid \boldsymbol{Y} = \boldsymbol{y})$ which is computationally too heavy to be done at each step E of the EM. In order to tackle this issue, a solution is to resort to a variational version of the EM algorithm presented in Section 5.1.

5.2.4 VEM algorithm for the SBM

Having a look at the graphical model on Figure 5.2, the necessity of approximating the posterior comes from the fact the latent variables are all dependant conditionally on the observations. Consequently, in order to decrease the complexity, we assume that $q \in \mathcal{Q}_{\text{fact}}$ where $\mathcal{Q}_{\text{fact}}$ was defined by Equation (5.5), *i.e.* it is the a product of n independent categorical distributions:

$$q(\boldsymbol{Z}) = \prod_{i=1}^{n} q_i(Z_i) = \prod_{i=1}^{n}\prod_{k=1}^{K} \tau_{ik}^{Z_{ik}} \qquad (5.13)$$

where $\tau_{ik} = \mathbb{P}_q(Z_i = k) = \mathbb{E}_q[Z_{ik}]$.

VE step for the SBM The choice, for Model 5.1, of the variational family given by Equation (5.13) leads to an operational VEM algorithm, as stated in the following proposition.

Proposition 5.3. *In the SBM, with the approximate conditional distribution q chosen in $\mathcal{Q}_{\text{fact}}$, the solution of the VE step satisfies the fixed-point relation*

$$\tau_{ik} \propto \omega_k \prod_{\substack{j=1\\ j\neq i}}^{n}\prod_{\ell=1}^{K} f(y_{ij};\alpha_{k\ell})^{\tau_{j\ell}}\,, \qquad (5.14)$$

where $\tau_{ik} = \mathbb{E}_q[Z_{ik}]$.

Proof of Proposition 5.3

From Section 5.1.3, we know that the VE step consists in maximizing the evidence lower bound which is equal to:

$$\begin{aligned}\mathsf{ELBO}(q,\theta,\boldsymbol{y}) &= \mathbb{E}_q[\log p_\theta(\boldsymbol{y},\boldsymbol{Z})] + \mathsf{Ent}[q(\boldsymbol{Z})]\,.\\ &= \sum_{i=1}^{n}\sum_{k=1}^{K}\mathbb{E}_q[Z_{ik}]\log\omega_k + \frac{1}{2}\sum_{\substack{i,j=1\\ i\neq j}}^{n}\sum_{k,\ell=1}^{K}\mathbb{E}_q[Z_{ik}Z_{j\ell}]\log f(y_{ij};\alpha_{k\ell}) + \mathsf{Ent}_q[\boldsymbol{Z}]\,.\end{aligned}$$

Using the expression of q in Equation (5.13), we have $q(\boldsymbol{Z}) = \prod_{i=1}^{n}\prod_{k=1}^{K}\tau_{ik}^{Z_{ik}}$ which implies

$$\mathsf{Ent}[q(\boldsymbol{Z})] = \sum_{i=1}^{n}\mathsf{Ent}[q_i(Z_i)] = -\sum_{i=1}^{n}\sum_{k=1}^{K}\tau_{ik}\log\tau_{ik}\,.$$

Moreover, by independence, for any $i > j$,

$$\mathbb{E}_q[Z_{ik}Z_{j\ell}] = \mathbb{E}_q[Z_{ik}]\mathbb{E}_q[Z_{j\ell}] = \tau_{ik}\tau_{j\ell}$$

which leads to

$$\mathsf{ELBO}(q,\theta,\boldsymbol{y}) = \sum_{i=1}^{n}\sum_{k=1}^{K}\tau_{ik}\log\omega_k + \frac{1}{2}\sum_{\substack{i,j=1\\ i\neq j}}^{n}\sum_{k,\ell=1}^{K}\tau_{ik}\tau_{j\ell}\log f(y_{ij};\alpha_{k\ell}) - \sum_{i=1}^{n}\sum_{k=1}^{K}\tau_{ik}\log\tau_{ik} \tag{5.15}$$

We then have to maximize the lower bound with respect to the τ_{ik}'s, subject to the constraint $\sum_{k=1}^{K}\tau_{ik} = 1$ for all $i \in \{1,\dots,n\}$. The derivative with respect to τ_{ik} is zero if and only if [a]:

$$\log\omega_k + \sum_{\substack{j=1\\ j\neq i}}^{n}\sum_{\ell=1}^{K}\tau_{j\ell}\log f(y_{ij};\alpha_{k\ell}) - \log\tau_{ik} - 1 - \lambda_i = 0\,,$$

where λ_i is the Lagrange multiplier for the constraint $\sum_{k=1}^{K}\tau_{ik} = 1$, which proves the proposition.

[a] Note that in Equation (5.15), each term τ_{ik} appears in the sums in i and j and $y_{ij} = y_{ji}$ so the 1/2 disappears.

Remark. *In the SBM, denoting* $\boldsymbol{Z}_{\setminus i} = \{Z_j, j \neq i\}$*, we have*

$$\begin{aligned}\mathbb{P}_\theta(Z_i = k \mid \boldsymbol{Y} = \boldsymbol{y}, \boldsymbol{Z}_{\setminus i}) = \frac{\mathbb{P}_\theta(Z_i = k, \boldsymbol{Y} = \boldsymbol{y} \mid \boldsymbol{Z}_{\setminus i})}{p(\boldsymbol{y} \mid \boldsymbol{Z}_{\setminus i})} &= \frac{p(y \mid Z_i = k)\mathbb{P}_\theta(Z_i = k)}{p(\boldsymbol{y} \mid \boldsymbol{Z}_{\setminus i})}\\ &\propto \omega_k \prod_{\substack{j=1\\ j\neq i}}^{n}\prod_{\ell=1}^{K} f(y_{ij};\alpha_{k\ell})^{Z_{j\ell}}.\end{aligned}$$

Indeed, Z_i *impacts the distributions of all the interactions between* i *and any other node* $j \neq i$*. Comparing this equation with Equation* (5.14)*, the mean-field approximation can be viewed as a simple plug-in of the (approximate) mean* $\tau_{j\ell} = \mathbb{E}_q[Z_{j\ell}]$ *in place of* $Z_{j\ell}$*. This justifies the mean field denomination: when considering one individual, the other elements of the field (i.e. the other individuals) are set to their respective means.*

M step for the SBM

Proposition 5.4. *In the SBM with approximate conditional distribution* q *chosen in* $\mathcal{Q}_{fact}$ *and such that the emission distribution belongs to the exponential family:*

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

where $S(y)$ is the vector of the sufficient statistics, the solution of the M step of the VEM is:

$$\widehat{\omega}_k = \frac{\sum_{i=1}^n \tau_{ik}}{n} \tag{5.16}$$

$$\nabla b(\widehat{\alpha}_{k\ell}) = \frac{\sum_{i\neq j=1}^n \tau_{ik}\tau_{j\ell} S(y_{ij})}{\sum_{i\neq j=1}^n \tau_{ik}\tau_{j\ell}} \tag{5.17}$$

In particular, if the emission distribution is $\mathcal{B}ern(\alpha_{k\ell})$ for binary interactions or $\mathcal{P}(\alpha_{k\ell})$ for weighted interactions then

$$\widehat{\alpha}_{k\ell} = \frac{\sum_{i\neq j=1}^n \tau_{ik}\tau_{j\ell} y_{ij}}{\sum_{i\neq j=1}^n \tau_{ik}\tau_{j\ell}}$$

Proof of Proposition 5.4

The proof relies on the resolution of the following equation to (θ, λ) where λ is the Lagrange multiplier for the constraint $\sum_{k=1}^K \omega_k = 1$:

$$\nabla_{\theta,\lambda}\left[\sum_{i=1}^n\sum_{k=1}^K \tau_{ik}\log\omega_k + \frac{1}{2}\sum_{i\neq j=1}^n\sum_{k,\ell=1}^K \tau_{ik}\tau_{j\ell}\log f(y_{ij};\alpha_{k\ell}) + \lambda\left(\sum_{k=1}^K \omega_k - 1\right)\right] = 0.$$

If considering ω_k, we obtain:

$$\omega_k = \frac{\sum_{i=1}^n \tau_{ik}}{n}$$

Then, assuming that the emission distribution belongs to the exponential family, we get:

$$\sum_{\substack{i,j=1\\ i\neq j}}^n \tau_{ik}\tau_{j\ell}\left[S(y_{ij}) - \nabla b(\alpha_{kl})\right] = 0$$

and Equation (5.17) follows. Using the results given in the Appendix section A.2.1, the Bernoulli distribution can be put in the exponential formulation as follows :

$$\log f(y,p) = y\log(p) + (1-y)\log(1-p) = y\log\frac{p}{1-p} + \log(1-p)\,.$$

Then, by identification, $t(y) = y$, $\alpha = \log\frac{p}{1-p}$ and $p = \frac{e^\alpha}{1+e^\alpha}$ and $b(\alpha) = -\log(1-p) = \log(1+e^\alpha)$. Finally :

$$b'(\alpha) = \frac{e^\alpha}{1+e^\alpha} = p\,.$$

Applying the general formula (5.17) leads to the final formula of Proposition 5.4.

Remark. *If we include covariates in the model (Model (5.9)), then $\log f(y_{ij};\alpha_{k\ell})$ must be adapted. For a Poisson SBM model with covariates, it is:*

$$\log f(y_{ij};\alpha_{k\ell}) = -\exp\{\alpha_{k\ell} + \boldsymbol{x}_{ij}^\top\boldsymbol{\beta}\} + y_{ij}(\alpha_{k\ell} + \boldsymbol{x}_{ij}^\top\boldsymbol{\beta}) - \log(y_{ij}!). \tag{5.18}$$

In that case, the M step is not explicit any more and one must resort to a numerical optimization.

Theoretical guarantees As mentioned in Section 5.1.6, variational estimates are often lacking of theoretical guarantees. The Bernoulli SBM Model 5.1 is one example where such guatanties do exist. A series of results obtained by Celisse et al. [2012], Bickel et al. [2013] or Mariadassou and Matias [2015], proved their consistency and their asumtotic normality.

5.2.5 Choosing the number of blocks

In the case of the SBM 5.1, the number of locks K has to be selected where $K \in \{1,\dots,K_{\max}\}$ with $K_{\max}$ the maximum number of blocks considered. The main issue here is that the marginal distribution of each $\boldsymbol{y} = \{y_{ij}\}_{i\neq j=1}^n$ cannot be evaluated, and even if it could, the edges are dependent on one another. As a result, deriving the BIC criterion is not straightforward.

Dimension of the model Going back the Bayesian framework from Model 2.4, we may adopt the ICL point of view from Biernacki et al. [2000] and consider the Laplace approximation of the integral

$$\int p(\boldsymbol{y}, \boldsymbol{z}, \theta, K)\, \mathrm{d}\theta = \int p(\boldsymbol{y}, \mid \boldsymbol{z}, \theta, K) p(z, \mid \theta, K) p(\theta \mid K)\, \mathrm{d}\theta.$$

Reminding that the parameter set of Model 5.1 is $\theta = (\boldsymbol{\omega}, \alpha)$ and assuming that $\boldsymbol{\omega}$ and α are independent conditional on K, i.e.:

$$p(\boldsymbol{\omega}, \alpha \mid K) = p(\boldsymbol{\omega} \mid K) p(\alpha \mid K)\,,$$

we may prove (see Appendix C.1 for the binary or Poisson SBM) that

$$\log\left(\int p(\boldsymbol{y}, \boldsymbol{z}, \theta, K)\, \mathrm{d}\theta\right) = \log p_{\widehat{\theta}_K}(\boldsymbol{y}, \boldsymbol{z}) - \mathrm{pen}(K) + O_n(1), \tag{5.19}$$

where

$$\mathrm{pen}(K) = \frac{1}{2}\left(\frac{K(K+1)}{2} \log\left(\frac{n(n-1)}{2}\right) + (K-1)\log n\right). \tag{5.20}$$

Observe that the penalty $\mathrm{pen}(K)$ is composed of two terms. The first one results from the integration with respect to α, which involves $K(K+1)/2$ independent parameters and rules the conditional distribution of the $n(n-1)/2$ edges Y_{ij}. The second one results from the integration with respect to $\boldsymbol{\omega}$, which involves $(K-1)$ independent parameters and rules the conditional distribution of the n latent variables Z_i.

Remark. *The penalty of Equation* (5.20) *is adapted to non oriented networks. In case of an oriented network, it becomes:*

$$\mathrm{pen}(K) = \frac{1}{2}\left(K^2 \log\left(n^2 - n\right) + (K-1)\log n\right).$$

Expressions of the ICL's The ICL criterion relies on the dominant term of Equation (5.19), where the complete log-likelihood $\log p_\theta(\boldsymbol{y}, \boldsymbol{z})$ has a close form given in Equations (5.10) and (5.11). Still, because the latent variable $\boldsymbol{Z}$ is not observed, we need to resort to the variational approximation either to approximate its conditional mode (for $\mathrm{ICL}_{\mathsf{V},1}$), or to integrate with respect to its approximate conditional distribution (for $\mathrm{ICL}_{\mathsf{V},2}$):

$$\begin{aligned}
\mathrm{ICL}_{\mathsf{V},1}(K) &= \log p_{\widehat{\theta}_K}(\boldsymbol{y}, \hat{\boldsymbol{z}}_{\mathsf{V}}) - \mathrm{pen}(K) \\
\mathrm{ICL}_{\mathsf{V},2}(K) &= \mathbb{E}_{\hat{q}}\left[\log p_{\widehat{\theta}_K}(\boldsymbol{y}, \boldsymbol{z})\right] - \mathrm{pen}(K) \\
&= \mathsf{ELBO}(\hat{q}, \widehat{\theta}_K, \boldsymbol{y}) + \mathsf{Ent}_{\hat{q}}[\boldsymbol{Z}] - \mathrm{pen}(K)
\end{aligned}$$

where:

$$\mathsf{Ent}_{\hat{q}}[\boldsymbol{Z}] = \sum_{i=1}^{n} \mathsf{Ent}_{\hat{q}_i}[Z_i] = -\sum_{i=1}^{n}\sum_{k=1}^{K} \widehat{\tau}_{ik} \log \widehat{\tau}_{ik}$$

and

$$\hat{\boldsymbol{z}}_{\mathsf{V}} = (\hat{z}_{\mathsf{V},i}, \ldots, \hat{z}_{\mathsf{V},i}) \qquad \text{with} \qquad \hat{z}_{\mathsf{V},i} = \arg\max_{z_i} \hat{q}_i(z_i),\ 1 \le i \le n\,.$$

The inference of the SBM (i.e. parameter estimation, clustering for a fixed number of blocks and selection of a number of blocks using the ICL criterion) is implemented in R using on the R package `blockmodels` [Léger, 2016] (part of the collection of package `sbm`). This algorithm implements a forward backward research of the number of blocks.

5.2.6 Analysis of the tree-tree parasite network with SBM

Poisson SBM We analyse the tree-tree dataset with the SBM model. Remind that this network is a weighted symmetric network where y_{ij} is the number of shared parasite fungi. We model it with the Poisson-SBM model:

$$Y_{ij} \mid \{Z_i = k, Z_j = \ell\} \overset{\text{ind}}{\sim} \mathcal{P}(\alpha_{k\ell})\,.$$

We performed the inference with the `sbm` R package. In Figure 5.3, we plot the ICL as a function of K. The criterion is maximal for $\widehat{K} = 6$ blocks. For $K = 6$, we plot the reordered adjacency matrix where the tree species

are gathered by $\widehat{Z}_i$ (left panel of Figure 5.4). The right panel of Figure 5.4 displays the expected number of shared parasites for any pairs of trees, where the trees have been reordered so that the blocks stand together:

$$\widehat{\mathbb{E}}[Y_{ij}] = \sum_{k,\ell=1}^{K} \widehat{\tau}_{ik}\widehat{\tau}_{j\ell}\widehat{\alpha}_{k\ell}$$

Theses matrices following the clusters highlight a structure in the number of shared parasites. As an example, block 1 gathers the trees that share together a high number of fungi while block 6 is composed of trees that do not share any fungi with other trees. Block 1 share fungi with block 3 while block 2 share its parasites with block 1 but not block 3.

Distance covariates As said before, this dataset also provides three distances calculated for each pair of trees (i, j), namely the genetic (gen_{ij}), taxonomic (taxo_{ij}) and geographic (geo_{ij}) distances.All of them have been standardized. We set the following model:

$$\begin{aligned} Y_{ij} \mid \{Z_i = k, Z_j = \ell\} &\sim \mathcal{P}(\alpha_{ijk\ell}) \\ \alpha_{ijk\ell} &= \exp\{\alpha_{k\ell} + \beta_1 \cdot \text{gen}_{ij} + \beta_2 \cdot \text{taxo}_{ij} + \beta_3 \cdot \text{geo}_{ij}\} \\ Z_i &\sim \mathcal{C}at(\pi_1, \ldots, \pi_K). \end{aligned}$$

Note that when the model includes a single block ($K = 1$), it reduces to a generalized linear model. If the covariates do not fully account for the heterogeneity in $\boldsymbol{y}$, then using multiple blocks ($K > 1$) allows the model to capture some of the remaining structured variability that is not explained by the covariates. We infer the parameters with the same R package. When considering covariates as explanatory variables, the blocks account for the residual variability. As a result, their interpretation becomes more challenging, and their signification cannot be visualized simply by reordering the nodes according to the blocks. Another approach is to fit the Poisson SBM and then use the covariates to explain the resulting blocks.

gen	taxo	geo
$\widehat{\beta}_1$	$\widehat{\beta}_2$	$\widehat{\beta}_3$
0.20	−2.06	−0.36

The taxonomic distance has the strongest effect: two species with a small taxonomic distance share a high number of fungi. [5]

5.2.7 Extension to bipartite networks

Bipartite networks are particularly valuable in ecology because they provide a natural framework for representing interactions between two distinct groups of organisms, such as plants and their pollinators, or trees and fungi.

Dataset 5.2. ***(Tree-fungi network)*** *We consider again the data of Example 5.1. This time, we consider the data a bipartite network, showed in Figure 5.6, where an edge links a fungus i to a tree j if this fungus was observed on this tree. The size of each node is proportional to its number of connections. The same information can be encoded in the bi-adjacency matrix represented in Figure 5.5: the trees are in row and the fungi are in columns. $y_{ij} = 1$ if fungus species j has been observed of tree species i.*

Our objective is to make clusters of both trees and fungi based on their interactions between each other.

A stochastic block model for bipartite networks The SBM was extented to bipartite networks by introucing a bi-clustering of the row nodes and the column nodes. The resulting model is the Latent block model (or biSBM, as bipartite SBM, Larremore et al., 2014) and was introduced by Govaert and Nadif [2003].

Model 5.2. ***(LBM)*** *Let $(Y_{ij})_{i=1,\ldots,n,j=1,\ldots,p}$ be the incidence matrix of a bipartite network involving n nodes of type 1 (fungi in Example 5.2) and p nodes of type 2 (trees in Example 5.2).*

In the LBM, the nodes of type 1 (respectively of type 2) are assumed to be divided into K (respectively L) latent groups (blocks) . Let Z_i be the categorical random variable encoding this clustering of the nodes of

[5] Note that one can also perform covariate selection using the ICL criterion in order to choose the most appropriate subset of covariates.

type 1 and W_j for type 2 :

$$\begin{aligned} Z_i &\overset{iid}{\sim} \text{Cat}\left(\boldsymbol{\omega}^{(1)}\right) \quad \forall i = 1, \dots, n \\ W_j &\overset{iid}{\sim} \text{Cat}\left(\boldsymbol{\omega}^{(2)}\right) \quad \forall j = 1, \dots, p \end{aligned} \tag{5.21}$$

where

$$\boldsymbol{\omega}^{(1)} = (\omega_1^{(1)}, \dots, \omega_K^{(1)}) \in [0,1]^K, \qquad \sum_{k=1}^{K} \omega_k^{(1)} = 1$$

$$\boldsymbol{\omega}^{(2)} = (\omega_1^{(2)}, \dots, \omega_L^{(2)}) \in [0,1]^L, \qquad \sum_{\ell=1}^{L} \omega_\ell^{(2)} = 1\,.$$

Then, conditionally to the latent variables $\boldsymbol{Z}$ and $\boldsymbol{W}$, the nodes are connected independently:

$$Y_{ij} \mid \{Z_i = k, W_j = \ell\} \overset{ind}{\sim} \mathcal{F}(\alpha_{k\ell}) \qquad \forall (i,j) \in \{1, \dots, n\} \times \{1, \dots, p\} \tag{5.22}$$

Remark. *As for the SBM model, the emission distribution $\mathcal{F}(\alpha_{k\ell})$ should be chosen in order to be adapted to the types of interactions (binary or weighted).*

Statistical inference The parameters of the model are $\theta = (\boldsymbol{\omega}^{(1)}, \boldsymbol{\omega}^{(2)}, (\alpha_{k\ell})_{k,\ell=1,\dots,K})$, $\boldsymbol{\omega}^{(1)}$ being the vector of the row block proportions while $\boldsymbol{\omega}^{(2)}$ is the vector of the column block proportions. The $\alpha_{k\ell}$ are the inter block connection parameters. We do not talk about intra-block connection parameters proportions since the row and column blocks are defined on different types of nodes (plants/pollinators for instance, or trees/fungi).

The parameters are estimated by the VEM algorithm. The numbers of row and column blocks K, L are selected using the ICL_V criterion. In this case the penalty term is:

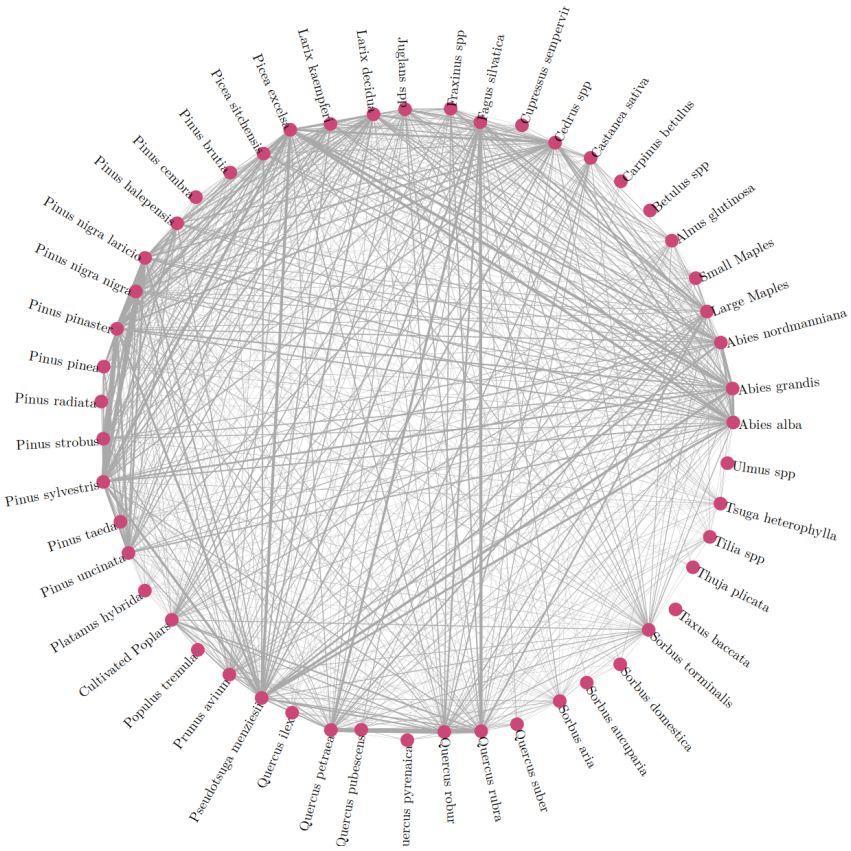
$$\text{pen}(K, L) = \frac{1}{2} \{KL \log(np) + (K-1) \log n + (L-1) \log p\}\,.$$

The technical details are provided in the Appendix section C.2.

Analysis of the tree-fungi dataset Fitting this model on the data with the sbm-package, we obtain $\widehat{K} = 4$ and $\widehat{L} = 4$. The same bi adjacency matrix reordered per blocks is represented in Figure 5.7. We observe an interesting chequerboard pattern. The first block of trees is parasitized by the first three blocks of fungi while the second block is parasitized by blocks 2 and 4 and not at all by block 3. The third block of trees has the same pattern as block 1 but with a lower intensity (less fungi species). The last block of trees has very few fungi.

5.2.8 Conclusion on SBM

In conclusion, Stochastic Block Models (SBMs) offer a versatile and powerful framework for modeling complex networks. Their hierarchical formulation allows for extensive flexibility and adaptability, enabling a wide range of model extensions to suit specific structural assumptions or data characteristics. However, this flexibility comes at the cost of increased inference complexity, primarily due to the intricate dependencies among latent variables. In this case, we have to resort to the variational version of the EM algorithm.


Abies alba
Abies grandis
Abies nordmanniana
Large Maples
Small Maples
Alnus glutinosa
Betulus spp
Carpinus betulus
Castanea sativa
Cedrus spp
Cupressus sempervirens
Fagus silvatica
Fraxinus spp
Juglans spp
Larix decidua
Larix kaempferi
Picea excelsa
Picea sitchensis
Pinus brutia
Pinus cembra
Pinus halepensis
Pinus nigra laricio
Pinus nigra nigra
Pinus pinaster
Pinus pinea
Pinus radiata
Pinus strobus
Pinus sylvestris
Pinus taeda
Pinus uncinata
Platanus hybrida
Cultivated Poplars
Populus tremula
Prunus avium
Pseudotsuga menziesii
Quercus ilex
Quercus petraea
Quercus pubescens
Quercus pyrenaica
Quercus robur
Quercus rubra
Quercus suber
Sorbus aria
Sorbus aucuparia
Sorbus domestica
Sorbus torminalis
Taxus baccata
Thuja plicata
Tilia spp
Tsuga heterophylla
Ulmus spp


Trees
Trees
Abies alba
Abies grandis
Abies nordmanniana
Large Maples (Acer platanoides
Small Maples (Acer campestre,
Alnus glutinosa
Betulus spp (Betula pendula, B
Carpinus betulus
Castanea sativa
Cedrus spp (Cedrus atlantica,
Cupressus sempervirens
Fagus silvatica
Fraxinus spp (Fraxinus angusti
Juglans spp (Juglans nigra, Ju
Larix decidua
Larix kaempferi
Picea excelsa
Picea sitchensis
Pinus brutia (Pinus brutia, Pi
Pinus cembra
Pinus halepensis
Pinus nigra laricio
Pinus nigra nigra
Pinus pinaster
Pinus pinea
Pinus radiata
Pinus strobus
Pinus sylvestris
Pinus taeda
Pinus uncinata
Platanus hybrida
Cultivated Poplars (Populus tr
Populus tremula
Prunus avium
Pseudotsuga menziesii
Quercus ilex
Quercus petraea
Quercus pubescens
Quercus pyrenaica
Quercus robur
Quercus rubra
Quercus suber
Sorbus aria
Sorbus aucuparia
Sorbus domestica
Sorbus torminalis
Taxus baccata
Thuja plicata
Tilia spp (Tilia platiphyllos,
Tsuga heterophylla
Ulmus spp (Ulmus minor, Ulmus
120
0
5
10
15
20

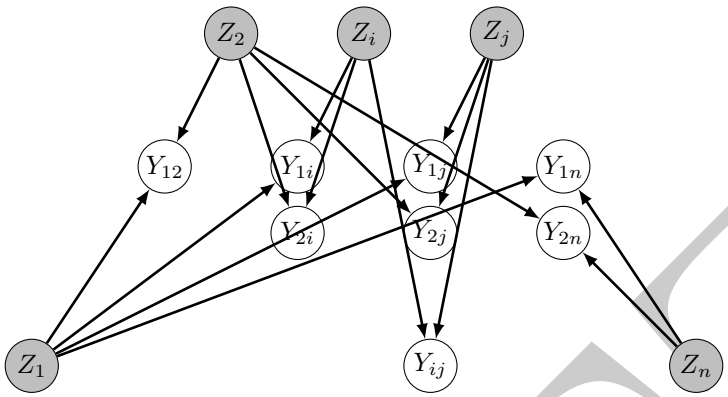


Figure 5.2: Graphical representation of the SBM defined in Equations (5.7) and (5.8).

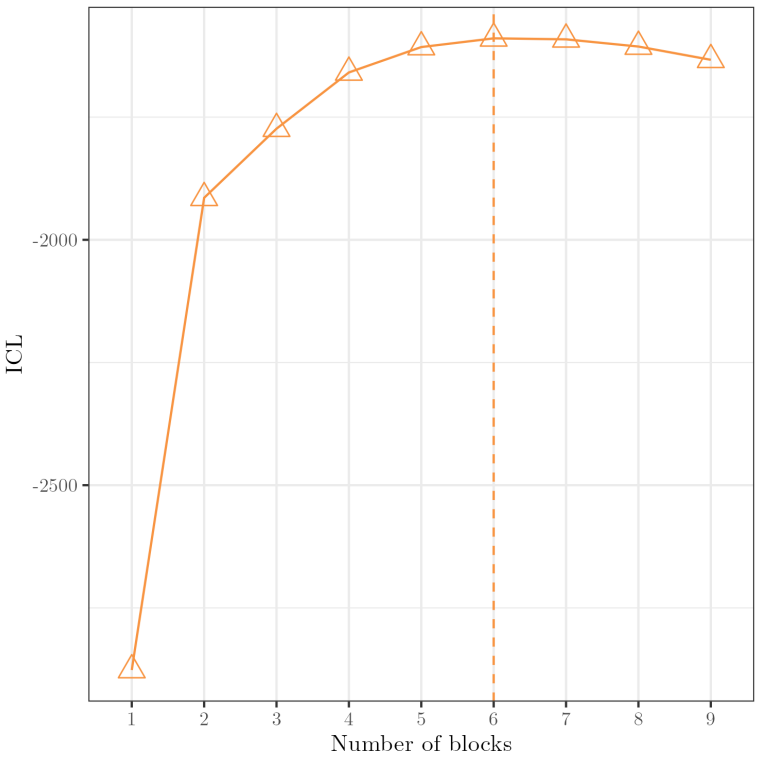


Figure 5.3: Tree-tree network. ICL as a function of K.

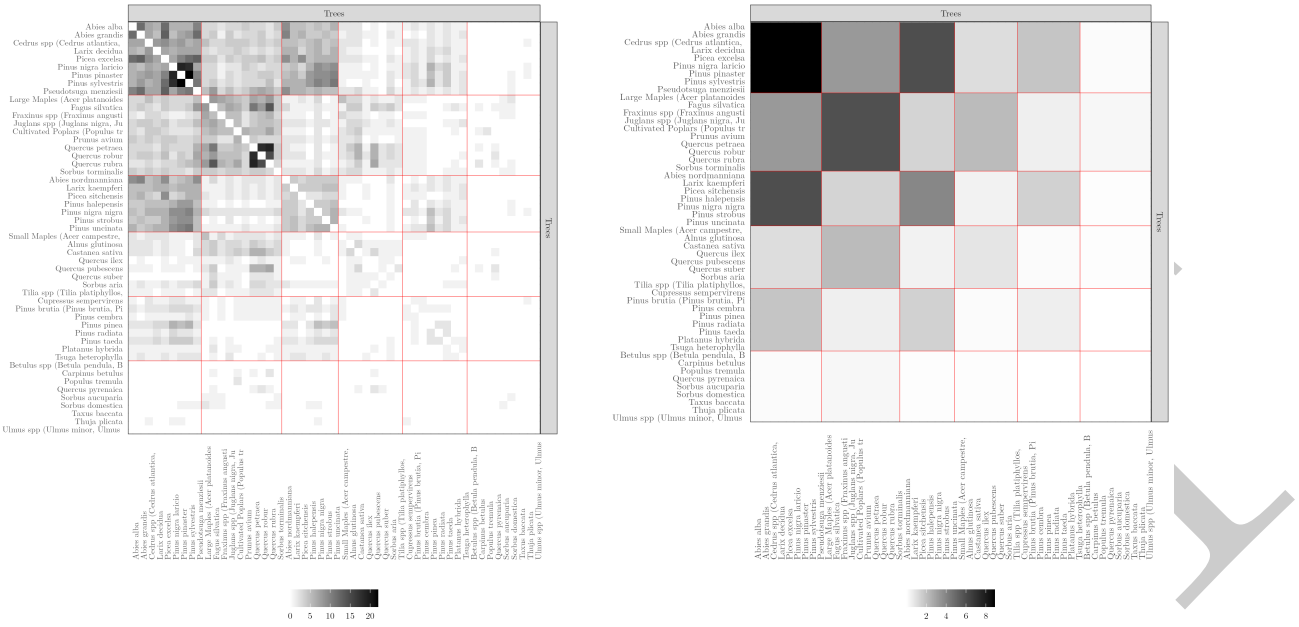


Figure 5.4: Tree-tree network. Left: reordered adjacency matrix where the trees have been gathered by blocks. Right: expected number of shared fungi $\sum_{k,\ell=1}^{K} \widehat{\tau}_{ik}\widehat{\tau}_{j\ell}\widehat{\alpha}_{k\ell}$ where the trees have been gathered by blocks. estimated

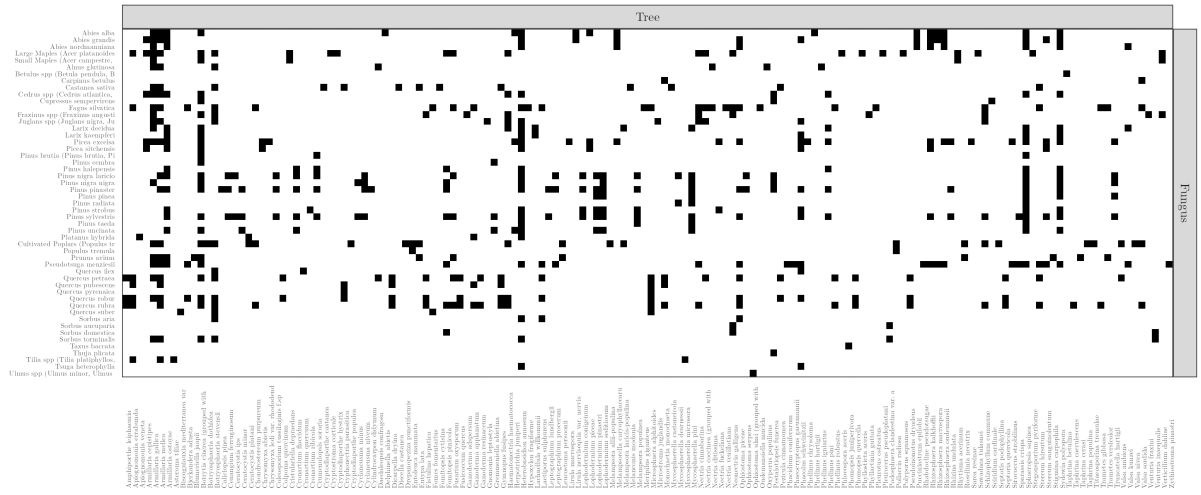


Figure 5.5: Bipartite tree-fungi networks Vacher et al. [2008]. An edge links i to j if fungus j is observed on tree i. The size of each node is proportional to its number of connections.

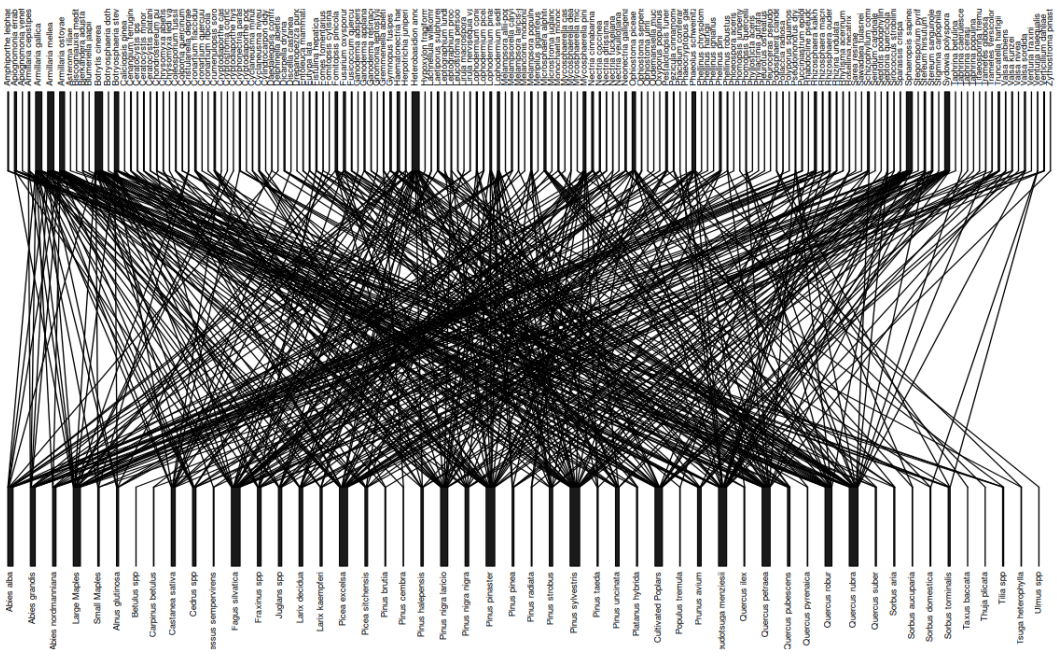


Figure 5.6: Bipartite tree-fungi network Vacher et al. [2008]. An edge links i to j if fungus j is observed on tree i. The size of each node is proportional to its number of connections.

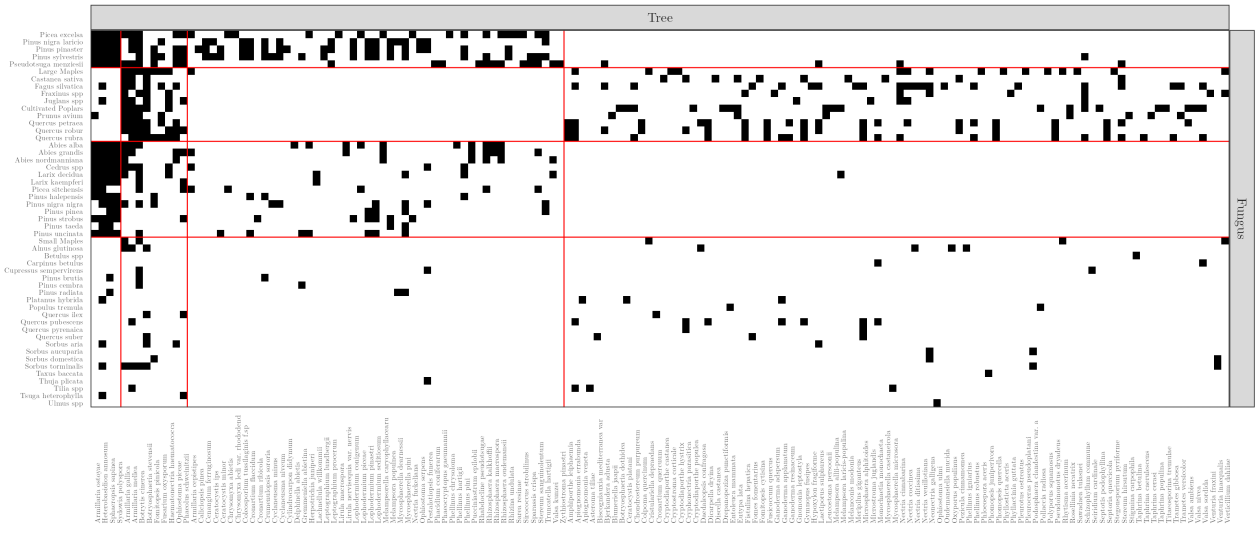


Figure 5.7: Tree-fungi network. Bi adjacency Reordered incidence matrix where the trees and fungi have been re-ordered following their blocks.

5.3 Joint species distribution models

5.3.1 Data and question

Joint species distribution models (JSDMs) have the same aim as species distribution models (SDM) already introduced in Section 3.2, but consider several species at once. The data from Example 3.2 are actually part of a larger study where few tens of fish species were studied.

Dataset 5.3 (Fish species in the Barents sea)**.** *Fossheim et al. [2006] measured the abundance of $p = 30$ fish species in $n = 89$ stations of the Barents. The capture protocol was the same for all species and stations and 4 descriptors (latitude, longitude, depth and temperature of the water) were recorded in each stations (the data are available from the* `PLNmodels` *R package [Chiquet et al., 2021]).*

The objective of statistical modelling here is to infer the effect of environment on the abundance of species, but also to detect potential 'interactions'[6] that may exist between the species. Many model for modelling joint species distributions have been proposed in the literature, adapted to different kind of observations (presence/absence and abundance data [Harris, 2015, Ovaskainen et al., 2017] or focusing on abundance data [Popovic et al., 2018, 2019]) Most of them rely, in one way or another, on a latent layer to account for species interactions. The interested reader might read Ovaskainen and Abrego [2020] for a general introduction written for ecologists.

5.3.2 The PLN model

We present here the multivariate Poisson log-normal model, first proposed by Aitchison and Ho [1989], which can be seen as a multivariate mixed Poisson regression model. The data set is typically organized as follows: consider n independent sites ($1 \leq i \leq n$) and p species ($1 \leq j \leq p$). The environmental descriptors of each sites i are gathered in a d-dimensional vector x_i and the abundance Y_{ij} is the number of individuals from species j captured in station i. The environmental descriptors are gathered into a $n \times d$ matrix X (matrix of covariates). The p-dimensional vector $Y_i = [Y_{i1} \dots Y_{ip}]^\top$ is the abundance vector in site i; all abundances are gathered into a $n \times p$ matrix $\boldsymbol{Y}$ (the abundance matrix).

The Poisson log-normal (PLN) model poses that

- a p-dimensional latent random Gaussian vector Z_i with distribution $\mathcal{N}(0_p, \Sigma)$ is drawn independently for each site $1 \leq i \leq n$;
- the observed abundance Y_{ij} is drawn independently from all others, conditionally on the Z_i's, with distribution $\mathcal{P}(\exp(x_i^\top \beta_j + Z_{ij}))$.

Model 5.3 (Poisson log-normal model)**.**

$$Z_i \overset{iid}{\sim} \mathcal{N}_p(0_p, \Sigma)\,,$$
$$Y_{ij} \mid \{Z_{ij} = z_{ij}\} \overset{ind}{\sim} \mathcal{P}(\exp(x_i^\top \beta_j + z_{ij})).$$

The d-dimensional vector of regression coefficients β_j contains the specific effects of of the d descriptors on species j. Theses effects are sometimes named abiotic. The vectors β_j can be gathered into the $d \times p$ matrix $B = [\beta_1 \dots \beta_p]$. The $p \times p$ variance matrix Σ encodes the interactions between the p species: these effects are sometimes named biotic. The parameters of Model 5.3 are hence

$$\theta = (B, \Sigma).$$

As already seen for the zero-inflated Poisson (ZIP) (Model 3.2), when the experimental protocol varies from one species to another or from one site to another, this heterogeneity can be accounted for by adding an offset term o_{ij} in the conditional distribution of the corresponding abundance: $Y_{ij} \sim \mathcal{P}(\exp(o_{ij} + x_i^\top \beta_j + Z_{ij}))$. For the sake of clarity, and because it is not required to deal with Example 5.3, we drop this term is the sequel of this section.

The distribution of $Y_i, 1 \leq i \leq n$, the i-th row of $\boldsymbol{Y}$, is called the Poisson log-normal distribution.

[6] The use of the word *interaction* is very debated in ecology. We use it in a broad sense, to be precised by the model itself.

Definition 5.1 (Multivariate Poisson log-normal distribution)**.** *Let $Y = [Y_j]_{1\leq j\leq p}$ be a random vector from $\mathbb{N}^p$, $\mu = [\mu_j]_{1\leq j\leq p}$ a vector from $\mathbb{R}^p$ and $\Sigma = [\sigma_{jk}]_{1\leq j,k\leq p}$ a real positive definite $p \times p$ matrix. $\boldsymbol{Y}$ is said to have a multivariate Poisson log-normal distribution with mean vector μ and variance matrix Σ:, denoted*

$$Y \sim \mathcal{PLN}(\mu, \Sigma) ,$$

if, for all integer-valued vector y from $\mathbb{N}^p$,

$$\begin{aligned}\mathbb{P}(Y = y) &= \mathbb{P}(Y_1 = y_1, \dots Y_p = y_p) \\ &= \int \cdots \int_{\mathbb{R}^p} \phi(z; 0_p, \Sigma) \prod_{j=1}^{p} \exp\left(-e^{\mu_j + z_j}\right) \frac{\left(e^{\mu_j + z_j}\right)^{y_j}}{y_j!} \, dz ,\end{aligned}$$

where $z \mapsto \phi(z; 0_p, \Sigma)$ stands for the density of the multivariate normal distribution with zero mean and covariance matrix Σ.

Proposition 5.5 (Moments of the PLN distribution)**.** *If $Y = [Y_j]_{1\leq j\leq p} \sim \mathcal{PLN}(\mu, \Sigma)$, we have that*

$$\begin{aligned}\mathbb{E}[Y_j] &= e^{\mu_j + \sigma_{jj}/2} , \\ \mathbb{V}[Y_j] &= \mathbb{E}[Y_j] + \mathbb{E}[Y_j]^2 \times \left(e^{\sigma_{jj}} - 1\right) \\ \mathbb{C}\mathrm{ov}(Y_j, Y_k) &= \mathbb{E}[Y_j] \times \mathbb{E}[Y_k] \times \left(e^{\sigma_{jk}} - 1\right) , \quad \textit{for } k \neq j .\end{aligned}$$

The proof of Proposition 5.5 is not given here: it only relies on classical (but tedious) integrations with respect to the multivariate Gaussian density. This proposition has several interesting consequences.

1. The mean of each coordinate Y_j does not only depend on the corresponding mean parameter μ_j, but also on the corresponding variance σ_{jj}.
2. Because the variance of each coordinate of Y exceeds its mean ($\mathbb{V}[Y_j] \geq \mathbb{E}[Y_j]$), the PLN distribution is over-dispersed as compared to the Poisson distribution. This property is desirable as, in ecology, counts and abundances are often observed to be more dispersed than predicted by a simple Poisson model.
3. The covariance $\mathbb{C}\mathrm{ov}(Y_j, Y_k)$ has the same sign as the corresponding entry σ_{jk} of the covariance matrix Σ (and is zero when σ_{jk} is zero). This makes the interpretation of Σ easier, as it encodes the nature of the correlation (negative, positive or null) that exists between the coordinates of the vector Y (*i.e.* the columns of $\boldsymbol{Y}$).

Alernative formulation of Model 5.3 Based on Definition 5.1, Model 5.3 is equivalent to

$$\forall 1 \leq i \leq n, \quad Y_i \overset{\text{ind}}{\sim} \mathcal{PLN}(\mu_i, \Sigma) , \qquad \text{where } \mu_i = x_i^\top B . \tag{5.23}$$

The PLN model therefore states that the abundance vector Y_i are independent, with site specific mean vectors μ_i but common variance matrix Σ. Hence, this model assumes that the influence of the environment on each species (encoded in the term $x_i^\top \beta_j$) is site-specific, whereas the biotic interactions (encoded in Σ) are the same in all sites.

Graphical model. Although the purpose of the PLN model is very different from those of mixture models described in Section 3.1, the graphical model of Model 5.3 is the same as this given in Figure 3.2: one single (multivariate) latent variable Z_i controls the distribution of each observed (multivariate) response Y_i. An important consequence is that the moments of the vector Z_i conditionally on Y are the same as these conditionally on Y_i: for any function f, $\mathbb{E}[f(Z_i) \mid \boldsymbol{Y}] = \mathbb{E}[f(Z_i) \mid Y_i]$.

5.3.3 Log-likelihoods

Marginal log-likelihood The combination of Equation (5.23) and Definition 5.1 yields the observed log-likelihood $\log p_\theta(\boldsymbol{Y})$, where $\boldsymbol{y}$ denotes the observed abundance matrix:

$$\log p_\theta(\boldsymbol{y}) = \sum_{i=1}^{n} \log \left(\int_{\mathbb{R}^p} \phi(z_i; 0_p, \Sigma) \prod_{j=1}^{p} \exp\left(-e^{\mu_{ij} + z_{ij}}\right) \frac{\left(e^{\mu_{ij} + z_{ij}}\right)^{y_{ij}}}{y_{ij}!} \, \mathrm{d}z_i \right) .$$

Clearly, $\log p_\theta(\boldsymbol{y})$ does not have a closed-form expression, and its evaluation involves a multivariate integral over a potentially p-dimensional space, which can quickly become computationally infeasible.

Complete log-likelihood As with many of the models encountered thus far, the complete log-likelihood $\log p_\theta(\boldsymbol{Y}, \boldsymbol{Z})$ is more tractable, as it does not involve any integration:

$$\begin{aligned}\log p_\theta(\boldsymbol{Y}, \boldsymbol{Z}) &= \log p_\theta(\boldsymbol{Z}) + \log p_\theta(\boldsymbol{Y} \mid \boldsymbol{Z}) \\ &= -\frac{n}{2}\log|\Sigma| - \frac{1}{2}\sum_{i=1}^{n}\|Z_i\|^2_{\Sigma^{-1}} + \sum_{i=1}^{n}\sum_{j=1}^{p}\left(Y_{ij}\left(x_i^\top\beta_j + Z_{ij}\right) - \mathrm{e}^{x_i^\top\beta_j + Z_{ij}}\right) \\ &\quad - \frac{np}{2}\log(2\pi) - \sum_{i=1}^{n}\sum_{j=1}^{p}\log(Y_{ij}!)\,..\end{aligned}$$

Observe that the last two terms are constant with respect to the parameter $\theta = (B, \Sigma)$.

5.3.4 Variational EM algorithm

5.3.4.1 Objective function

Using the independence between the sites (or, which is equivalent, the graphical model from Figure 3.2), we see that for an current guess $\theta^{(h)}$ the conditional expectation $\mathbb{E}_{\theta^{(h)}}[\log p_\theta(\boldsymbol{y}, \boldsymbol{Z}) \mid \boldsymbol{Y} = \boldsymbol{y}]$ is

$$\begin{aligned}Q\big(\theta \mid \theta^{(h)}\big) &= \text{cst} - \frac{n}{2}\log|\Sigma| - \frac{1}{2}\sum_{i=1}^{n}\mathbb{E}_{\theta^{(h)}}\Big[\|Z_i\|^2_{\Sigma^{-1}} \mid Y_i\Big] \\ &\quad + \sum_{i=1}^{n}\sum_{j=1}^{p}\Big(Y_{ij}\left(x_i^\top\beta_j + \mathbb{E}_{\theta^{(h)}}[Z_{ij} \mid Y_i]\right) - \mathrm{e}^{x_i^\top\beta_j}\mathbb{E}_{\theta^{(h)}}\Big[\mathrm{e}^{Z_{ij}} \mid Y_i\Big]\Big).\end{aligned}$$

To derive an EM algorithm, as defined in Section 2.2, we are therefore left with evaluating the conditional moments:

$$\mathbb{E}_{\theta^{(h)}}[Z_{ij} \mid Y_i], \qquad \mathbb{E}_{\theta^{(h)}}\Big[\|Z_i\|^2_{\Sigma^{-1}} \mid Y_i\Big], \qquad \mathbb{E}_{\theta^{(h)}}\Big[\mathrm{e}^{Z_{ij}} \mid Y_i\Big].$$

Let us now examine the distribution of $Z_i \mid Y_i$. Using Bayes' formula, we can easily see that[7], for every $z_i \in \mathbb{R}^p$:

$$p(z_i \mid Y_i = y_i) = \frac{1}{C_i} \times \exp\left(-\|z_i\|^2_{\Sigma^{-1}} - \sum_{j=1}^{p}\mathrm{e}^{\mu_{ij}+z_{ij}} + \sum_{j=1}^{p} y_{ij}\left(\mu_{ij} + z_{ij}\right)\right), \tag{5.24}$$

where:

$$C_i = \int_{\mathbb{R}^p}\exp\left(-\|u\|^2_{\Sigma^{-1}} - \sum_{j=1}^{p}\mathrm{e}^{\mu_{ij}+u_j} + \sum_{j=1}^{p} y_{ij}\left(\mu_{ij} + u_j\right)\right)\mathrm{d}u\,.$$

This form does not belong to any known family of distributions, and as a result, none of the expectations above can be computed, even when $\theta^{(h)}$ is known. Hence, the PLN model provides another example of a case where a regular EM algorithm cannot be defined due to the intractability of the conditional distribution $p_\theta(\boldsymbol{z} \mid \boldsymbol{Y} = \boldsymbol{y})$. As before, we address this issue by defining a variational version of the EM algorithm, in which the conditional distribution is approximated.

5.3.4.2 Variational family

Since the sites are independent under Model 5.3, we only need to consider the conditional distribution of each latent vector Z_i given the corresponding count vector Y_i. However, because the conditional distribution $p_\theta(Z_i \mid Y_i)$ given in Equation (5.24) is intractable, we resort to variational inference. To this end, we must choose an appropriate approximation class $\mathcal{Q}$. Given that the Z_i are random vectors in $\mathbb{R}^p$ and that $p_\theta(Z_i \mid Y_i)$ has a unique mode, a natural choice for $\mathcal{Q}$ is the set of multivariate Gaussian distributions, that is,

$$\mathcal{Q}_{PLN} = \left\{q : q = \bigotimes_{i=1}^{n} q_i, q_i = \mathcal{N}_p(m_i, S_i)\right\}.$$

The mean vectors $m_i = [m_{ij}]_{1\leq j\leq p}$ and the $p \times p$ variance matrices S_i are the variational parameters of the approximation. We gather them into $M = [m_1 \ \ldots \ m_n]^\top$ and $S = (S_i)_{1\leq i\leq n}$, respectively.

[7] All constants (i.e., quantities that do not depend on z_i) from both the Gaussian and Poisson log-normal distributions are absorbed into the constant C_i

Remark. *The product over the sites in the definition of $\mathcal{Q}_{PLN}$ results from the PLN Model 5.3 itself, so no approximation lies there. The approximation comes from the fact that the conditional distribution of each Z_i given Y_i is actually not Gaussian (see Equation* (5.24)*).*

Lower bound of the log-likelihood. If $q \in \mathcal{Q}_{PLN}$, we know from Proposition A.7 (Appendix A.1.2) that

$$\mathsf{Ent}_q[\boldsymbol{Z}] = \sum_{i=1}^{n} \mathsf{Ent}_{q_i}[\boldsymbol{Z}_i] = \frac{np}{2}(1 + \log(2\pi)) + \frac{1}{2}\sum_{i=1}^{n} \log|S_i|\,.$$

and from Proposition 5.5 that

$$\begin{aligned}\mathbb{E}_q[\log p_\theta(\boldsymbol{y}, \boldsymbol{Z})] = &-\frac{n}{2}\log|\Sigma| - \frac{1}{2}\sum_{i=1}^{n}\Big(\|m_i\|^2_{\Sigma^{-1}} + \mathrm{tr}(\Sigma^{-1}S_i)\Big) \\ &+ \sum_{i=1}^{n}\sum_{j=1}^{p}\left(y_{ij}\left(x_i^\top\beta_j + m_{ij}\right) - A_{ij}\right) - \frac{np}{2}\log(2\pi) - \sum_{i=1}^{n}\sum_{j=1}^{p}\log(y_{ij}!),\end{aligned}$$

where we defined

$$A_{ij} = \exp(\mu_{ij} + m_{ij} + s_{ij}/2)\,,$$

with s_{ij} the jth diagonal element of S_i. Observe that, according to Proposition 5.5, A_{ij} is the variational approximation of the expected count Y_{ij}, that is

$$A_{ij} = \mathbb{E}_{q_i}[Y_{ij}]\,.$$

The combination of the two preceding equations yields the $\mathsf{ELBO}(q, \theta, \boldsymbol{y})$ which will serve as the basis for variational inference in the PLN model:

$$\begin{aligned}\mathsf{ELBO}(q, \theta, \boldsymbol{y}) = &-\frac{n}{2}\log|\Sigma| - \frac{1}{2}\sum_{i=1}^{n}\Big(\|m_i\|^2_{\Sigma^{-1}} + \mathrm{tr}(\Sigma^{-1}S_i)\Big) \\ &+ \sum_{i=1}^{n}\sum_{j=1}^{p}\left(y_{ij}\left(x_i^\top\beta_j + m_{ij}\right) - A_{ij}\right) + \frac{1}{2}\sum_{i=1}^{n}\log|S_i| \\ &- \sum_{i=1}^{n}\sum_{j=1}^{p}\log(y_{ij}!) + \frac{np}{2},\end{aligned}$$

where, once again, the last two terms depend neither on q nor on θ. It is useful to note that, up to an additive constant, this function can be expressed as a sum of the n functions: $\forall\ 1 \le i \le n$,

$$\mathcal{L}(q_i, \theta, y_i) = -\frac{1}{2}\log|\Sigma| - \frac{1}{2}\Big(\|m_i\|^2_{\Sigma^{-1}} + \mathrm{tr}(\Sigma^{-1}S_i)\Big) + \sum_{j=1}^{p} y_i^\top(\mu_i + m_i) - \sum_{j=1}^{p} A_{ij} + \frac{1}{2}\sum_{i=1}^{n}\log|S_i|\,,$$

where the dependence in q_i is seen in the terms m_i and S_i (as q_i belongs to the parametric Gaussian family), and we denoted $\mu_i = Bx_i$. We therefore have:

$$\mathsf{ELBO}(q, \theta, \boldsymbol{y}) = \sum_{i=1}^{n}\mathcal{L}(q_i, \theta, y_i) + \mathrm{cst}\,. \tag{5.25}$$

5.3.4.3 Variational E step for the PLN model

The VE step consists of maximizing (5.25) with respect to the variational distribution q_i, *i.e.* with respect to the parameters $(m_{1:n}, S_{1:n})$. Thanks to the additive decomposition of the objective, we can therefore update each component by setting:

$$(m_i^{(h+1)}, S_i^{(h+1)}) = \underset{m_i, S_i}{\arg\max}\,\mathcal{L}(q_i, \theta^{(h)}, y_i)\,.$$

We have that:

$$\begin{aligned}\nabla_{m_i}\mathcal{L}(q_i, \theta^{(h)}, y_i) &= -\Sigma^{-1,(h)}m_i - A_i^{(h)}, \\ \nabla_{S_i}\mathcal{L}(q_i, \theta^{(h)}, y_i) &= -\frac{1}{2}\Sigma^{-1,(h)} - \mathrm{diag}(A_i^{(h)}) + \frac{1}{2}S_i^{-1},\end{aligned}$$

where we used useful results on derivatives of matrices (see Petersen and Pedersen [2008]), and $\mathrm{diag}(v)$ denotes the diagonal matrix with diagonal entries given by the vector v. The reader should keep in mind that $A_i^{(h)}$ depends non-linearly on m_i and the diagonal elements of S_i. Therefore, setting these derivatives to zero is not trivial. However, this can be achieved via gradient ascent, and since the Kullback-Leibler divergence is convex in (m_i, S_i) [see Chiquet et al., 2018, Lemmas 1 and 2], the unique global solution $(\widetilde{m}_i, \widetilde{s}_i)$ will be attained, which completes the VE step.

5.3.4.4 M step for the PLN model

We now derive the update formulas for the model parameter $\theta = (B, \Sigma)$, which maximize the ELBO given in Equation (5.25), using the current variational distribution $q^{(h+1)}$. For this purpose, it is convenient to consider the gradient of the ELBO with respect to the inverse of Σ, denoted by Ω, commonly referred to as the precision matrix. We have:

$$\partial_\Omega \mathsf{ELBO}(q, \theta, \boldsymbol{y}) = \left(n\Sigma - M^\top M - S^+\right)/2\,,$$

where $S^+ = \sum_{i=1}^n S_i$. The update formula for Σ is obtained by setting this gradient to zero for the current values of M and S, that is

$$\Sigma^{(h+1)} = \frac{1}{n}\left((M^{(h+1)})^\top M^{(h+1)} + (S^{(h+1)})^+\right)\,,$$

which is positive definite because the $S_i^{(h+1)}$ are all positive definite.

The update of B has no close form, but one may observe that, for each $1 \leq j \leq p$, we have

$$\begin{aligned}\arg\max_{\beta_j} \mathsf{ELBO}(q, \theta, \boldsymbol{Y}) &= \arg\max_{\beta_j} \sum_{i=1}^n \left(Y_{ij}(x_i^\top \beta_j + m_{ij}) - \mathrm{e}^{x_i^\top \beta_j + m_{ij} + s_{ij}/2} - \log(Y_{ij}!)\right) \\ &= \arg\max_{\beta_j} \sum_{i=1}^n \left(Y_{ij}(x_i^\top \beta_j + o_{ij}) - \mathrm{e}^{x_i^\top \beta_j + o_{ij}} - \log(Y_{ij}!)\right)\end{aligned}$$

with $o_{ij} = m_{ij} + s_{ij}/2$. We may then recognize the last expression as the log-likelihood of a Poisson regression model for the counts Y_{ij} with covariates x_i and offset terms o_{ij}, as seen in Section 3.2, Equation (3.13). Again, thanks to the general properties of generalized linear models (see Proposition A.10, Appendix A.2.1), we know that this log-likelihood is concave so any standard optimization procedure or dedicated function can be used to update each vector of regression coefficients β_j as

$$\beta_j^{(h+1)} = \arg\max_{\beta_j} \mathsf{ELBO}(q^{(h+1)}, \theta, \boldsymbol{Y})\,.$$

As recalled in Section 5.1.6, the VEM algorithm provides estimates of both the regression coefficient matrix B and the covariance matrix Σ, but the (asymptotic) variances of these estimates remain unknown. To estimate the variance of each parameter estimator, we could resort to the jackknife procedure [Efron and Stein, 1981], which also allows deriving pseudo-test statistics to assess the effects of each environmental covariate on each species. Note that this approach is computationally demanding, as the VEM algorithm must be run multiple times.

5.3.4.5 Covariates selection for the PLN model

In this context, model selection consists of choosing among the available covariates. The models to be compared thus correspond to the collection of all subsets of these covariates, so that $\#(\mathcal{M}) = 2^d$. Since the sites are assumed independent under any model $m \in \mathcal{M}$, the joint distribution of the species counts $Y_i = (Y_{i1}, \ldots, Y_{ip})$ is given by Equation (5.23). From this, we can apply the variational model selection criterion introduced in Section 5.1.5 by deriving the corresponding quantities.

First, the number of free parameters in the model is given by:

$$D_m = pd_m + \frac{p(p+1)}{2}\,,$$

where d_m stands for the number of covariates involved in model m, so pd_m is the number of regression parameters in B and $p(p+1)/2$ is the number of independent covariance parameters in Σ. Therefore, the penalty for both BIC_V and ICL_V is given by:

$$\mathrm{pen}(m) = D_m \times \frac{\log(n)}{2}\,.$$

Now, the variational family was chosen such that the final variational posterior is given by a product of normal probability density functions:

$$\hat{q}(\boldsymbol{Z}) = \prod_i^n \mathcal{N}_p\left(\widehat{m}_i, \widehat{S}_i\right)\,,$$

where $\widehat{m}_i$ and $\widehat{S}_i$ are the final variational parameters. From this, we can easily compute:

- the variational maximum a posteriori $\hat{\boldsymbol{z}}_\mathsf{V}$ which is simply $= (\widehat{m}_1, \ldots, \widehat{m}_n)$;
- the entropy of $\hat{q}(\boldsymbol{Z})$, which is, by property of Gaussian distributions (and independance of sites):

$$\mathsf{Ent}_{\hat{q}}[\boldsymbol{Z}] = \frac{1}{2}\sum_{i=1}^n \left(\det \widehat{S}_i + p(1 + \log(2\pi))\right)\,.$$

Plugging these in the formulas of Section 5.1.5, one can compute the three proposed model selection criteria.

5.3.5 Analyzing the fish abundances in the Barents sea with PLN

We now return to the fish abundance data described in Example 5.3. The objective is to study both the effects of the four environmental covariates (latitude, longitude, depth, and temperature) on the $p = 30$ fish species, and the dependency structure between the species after accounting for environmental factors.

Effects of the environmental covariates. We first consider the full model, including the four covariates plus an intercept. For ease of interpretation, the four covariates were standardized to have zero mean and unit standard deviation. This transformation allows the regression coefficients $\beta_{\ell j}$ to be comparable across covariates as well as across species. The left panel of Figure 5.8 shows that longitude has the least pronounced effect on species abundances, whereas latitude and temperature may have a dramatic impact on the population size of some species.

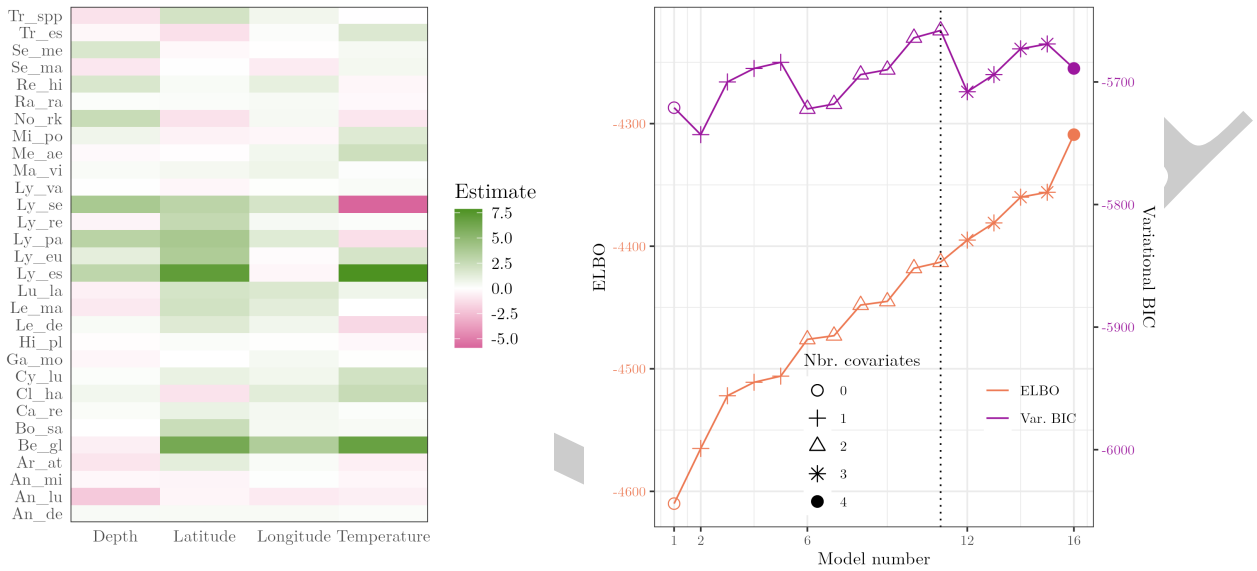


Figure 5.8: Barents fish species (Example 5.3). *Left*: Estimated regression coefficient $\widetilde{\beta}_{\ell j}$ the four covariates (columns) on the 30 species (rows). Green: positive effect, Red: negative. *Right*: ELBO and variational BIC for the 16 models. The models are numbered by their ELBO values, the dot shapes give the number of covariates. The vertical line gives the best model according to variational BIC.

Selecting a subset of covariates To further investigate the main environmental drivers of fish species assemblages in the Barents Sea, we compare the $2^4 = 16$ possible models corresponding to all combinations of the four covariates (ranging from none to all included). An intercept is retained in all models to account for the heterogeneity in the mean abundance of the species. The right panel of Figure 5.8 displays the optimized ELBO and the $\mathrm{BIC}_{\mathsf{V}}$ for each model (defined in Section 5.1.5), while Table 5.1 presents detailed results for the top three models, as well as the null model (no covariates) and the full model (all covariates). We observe that both the null and full models yield poor $\mathrm{BIC}_{\mathsf{V}}$ values—either due to poor fit (low ELBO for the null model) or excessive complexity (high D_m for the full model).

The best model according to $\mathrm{BIC}_{\mathsf{V}}$ includes two covariates: latitude and depth, indicating these as the primary environmental drivers among the available variables (the best single covariate is latitude). Note that the second-best model includes temperature and depth, and the correlation coefficient between temperature and latitude is -0.79, suggesting a confounding effect between these two covariates.

Observe that, in this case, we intentionally use the $\mathrm{BIC}_{\mathsf{V}}$ criterion rather than the $\mathrm{ICL}_{\mathsf{V}}$ criterion, as we don't intend to penalize the conditional entropy of latent variables. We will generally prefer the $\mathrm{ICL}_{\mathsf{V}}$ criterion for models, such as SBM and LBM (see Sections 5.2 and C.2), for which it is of interest to obtain a clear classification.

Between-species covariance structure. Now, we can examine the estimates of the between-species covariance matrix Σ. We first observe that the diagonal elements of Σ range from 0.7 to 14.8 (median = 4.5) under the null model, and markedly decrease under the full model (ranging from 0.06 to 5.0, median = 1.7). This indicates

Model		Covariates	ELBO	D_m	pen(m)	BIC_V
11		Latitude, Depth	-4415	555	1246	-5660
10		Depth, Temperature	-4418	555	1246	-5663
15		Latitude, Depth, Temperature	-4354	585	1313	-5667
1	(null)	$\varnothing$	-4615	495	1111	-5726
16	(full)	Latitude, Longitude, Depth, Temperature	**-4307**	615	1380	-5687

Table 5.1: Barents fish species (Example 5.3). ELBO, dimension D_m, BIC penalty pen(m) and variational BIC for the three best models according to BIC_V and for the null and full models. The model's number is the same as on Figure 5.8.

that environmental variation substantially contributes to the variation in species abundances (recall that, under the PLN Model 5.3, the random effect governed by Σ has an exponential impact on the mean abundance).

Figure 5.9 displays the estimated species-species correlations encoded by Σ under three different models. By construction, these correlations capture 'interactions' (in a broad sense) that are not explained by the covariates included in the regression part of the PLN model. Consequently, the correlations observed under the null model mainly reflect the effects of environmental variation (which the model does not account for) rather than specific interactions between pairs of species. Notably, many strong correlations observed under the null model tend to disappear when covariates are included. For example, the correlation between the species `Lu_la` (*Lumpenus lampraetaeformis*, first row) and `Ly_se` (*Lycodes seminudus*, near the center), as well as between `Le_ma` (*Leptoclinus maculatus*, also near the center), weakens substantially once covariates are considered.

However, for both the best and full models, strong correlations—either positive or negative—remain between certain blocks of species. For instance, the top-left and bottom-right corners of the correlation matrix reveal two groups of species exhibiting strong positive correlations within each group, suggesting a tendency to co-occurrence not explained by the included covariates, and negative correlations between groups, indicating potential avoidance. These estimated patterns may reflect particular biological interactions between species or responses to unmeasured environmental drivers not included in the study.

5.3.6 Conclusion about PLN model

Poisson Log-Normal (PLN) models provide a flexible framework for modeling count data with complex dependencies, by introducing latent Gaussian variables to capture overdispersion and correlation. While the variational inference procedure, particularly the VE step, involves intricate dependencies and nonlinearities, the convexity properties ensure that efficient optimization algorithms converge to a unique solution. This balance of model expressiveness and tractable inference makes PLN models well-suited for a wide range of applications.

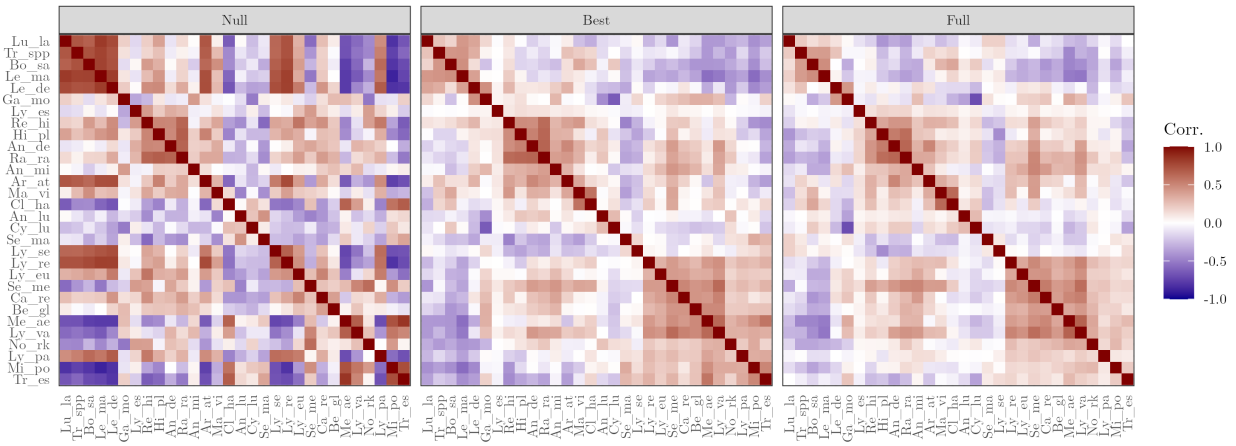


Figure 5.9: Barents fish species (Example 5.3). Correlation matrix associated with the estimated species-species covariance matrix Σ under the null, best (according to the variational BIC) and full models.

5.4 Variational (probabilistic) autoencoders

In this section, we discuss an interesting bridge that can be built between the VEM approach introduced in Section 5.1 and a popular class of models in machine learning: the variational autoencoders (VAEs), first introduced by Kingma [2013] (see also Kingma and Welling, 2019). VAEs are now widely used for probabilistic generative modeling. We illustrate this connection using the Poisson log-normal (PLN) model introduced in Model 5.3.

5.4.1 Probabilistic decoders

A recent application of latent variable models is probabilistic generative modeling. The goal of generative modeling is to learn a probabilistic distribution whose samples resemble the original dataset[8]. Among the methods developed for this task, a popular approach assumes that complex high-dimensional observations $Y \in \mathbb{R}^{d_y}$ are generated from lower-dimensional latent variables $Z \in \mathbb{R}^{d_z}$, which are typically assumed to follow a standard multivariate Gaussian distribution. The relationship between Y and Z is then modeled through:

- a parametrized non-linear function $g_{\theta_g} : \mathbb{R}^{d_z} \mapsto \mathbb{R}^{d_y}$, called the *decoder* (or generative model), which typically defines the expected value of $Y \mid Z$. This function is often implemented as a neural network with parameters (weights and biases) denoted by θ_g;
- a probability distribution $p_{\theta_{\text{obs}}}$, parametrized by θ_{obs}, which defines the noise model for Y.

All these elements lead to the following model.

Model 5.4 (Probabilistic decoder)**.** *The probabilistic decoder is defined as the following latent space model:*

$$Z_i \overset{iid}{\sim} \mathcal{N}(0, \mathbf{I}_{d_z}),$$
$$Y_i \mid \{Z_i = z_i\} \overset{ind}{\sim} p_{\theta_{obs}}\left(\cdot; g_{\theta_g}(z_i)\right) \qquad \forall i = \{1, \ldots, n\},$$

where $g_{\theta_g} : \mathbb{R}^{d_z} \mapsto \mathbb{R}^{d_y}$ is a neural network.

[8] A classical application is generating new images that resemble those in an existing image database.

The most common choice for $p_{\theta_{\text{obs}}}$ is a Gaussian distribution, due to its natural connection with the classical L_2 loss used in machine learning[9]: $Y = g_{\theta_g}(Z) + \mathcal{N}_{d_y}(0, \theta_{\text{obs}})$ ($\mathbb{E}[Y \mid Z = z] = g_{\theta_g}(z)$ and the parameter θ_{obs} corresponds to its variance). However, one can observe that the multivariate Poisson log-normal (PLN) model (Model 5.3) also fits within this generative framework, with $d_y = d_z$, and $p_{\theta_{\text{obs}}}$ being the probability mass function of a Poisson distribution with mean $g_{\theta_g}(z)$ satisfying (in the case without offset):

$$g_{\theta_g}(z) = \exp\left(x^\top \beta + \Sigma^{\frac{1}{2}} z\right).$$

Therefore, in this context, $\theta_g = \{\beta, \Sigma\}$ (as we suppose x is a known vector of covariates), and there is no θ_{obs}.

In general, denoting $\theta = \{\theta_{\text{obs}}, \theta_g\}$, the likelihood is, for a set of observations $\boldsymbol{y}$:

$$\log p_\theta(\boldsymbol{y}) = \log \int_z p_{\theta_{\text{obs}}}\left(\boldsymbol{y}; g_{\theta_g}(\boldsymbol{z})\right) p(\boldsymbol{z}) d\boldsymbol{z}$$

where $p(\boldsymbol{z})$ is the density of a standard multivariate Gaussian distribution. Because of the non linearity of g_{θ_g} this likelihood is generally be intractable. Applying the EM algorithm would require to have an explicit expression of $p_{\theta_g}(\boldsymbol{Z} \mid \boldsymbol{Y} = \boldsymbol{y})$, which is also non-explicit for the same reasons. Therefore, as for the PLN model, a common alternative is to perform variational inference.

5.4.2 From VEM to variational autoencoders

In this variational context, the conditional distribution of each $Z_i \mid \{Y_i = y_i\}$ is approximated by a Gaussian distribution:

$$Z_i \mid Y_i \overset{\text{approx.}}{\sim} \mathcal{N}_{d_z}(m_i, S_i),$$

and we denote $q_i(z)$ the corresponding p.d.f. Because of the independence (assumed by Model 5.4), we have that $p_\theta(\boldsymbol{Z} \mid \boldsymbol{Y} = \boldsymbol{y}) = \prod_{i=1}^n p_\theta(Z_i \mid Y_i = y_i)$ and we saw in Section 5.3.4 that the VE step then boils down, for each site i, to minimize of the Küllback-Leibler divergence $\mathsf{KL}\left[q_i(Z_i) \parallel p_\theta(Z_i \mid y_i)\right]$ which is a function, say $\mathcal{L}_i$ of $(m_i, S_i, y_i, \theta_{\text{obs}}, \theta_g)$. The VE step then requires to solve the n minimization problems:

$$(y_i, \theta_{\text{obs}}, \theta_g) \longrightarrow (\widetilde{m}_i, \widetilde{S}_i) = \arg\min_{m;S} \mathcal{L}_i(m, S, y_i, \theta_{\text{obs}}, \theta_g),$$

which is computationally demanding when n is large.

A possible alternative to avoid n minimizations is to learn a function $e_{\theta_e}(y)$ parameterized by θ_e (which is specific to each $\theta_{\text{obs}}, \theta_g$, in general), such that for every $1 \leq i \leq n$:

$$e_{\theta_e}(y_i) \simeq \arg\min_{m;S} \mathcal{L}_i(m, S, y_i, \theta_{\text{obs}}, \theta_g). \tag{5.26}$$

This idea of approximating the functional that links y_i to the conditional moments of the corresponding latent vector Z_i is called *amortization*. This name comes from the fact that when n is large, optimizing the ELBO with respect to θ_e may become cheaper (hence, amortized) than making n separate optimizations independently. In other words, for close observations y and y', the amortization scheme will assume that the result of the optimization will be related by the function e_{θ_e}.

The function e_{θ_e}, called an *encoder*[10] in the machine learning literature, is generally a neural network whose weights and biases are given by θ_e. The most common structure is to assume that the encoder outputs two components, *i.e.*

$$e_{\theta_e}(y_i) = \begin{pmatrix} e_{\theta_e,1}(y_i) \\ e_{\theta_e,2}(y_i) \end{pmatrix}$$

and to set $m_i = e_{\theta_e,1}(y_i)$ and $S_i = e_{\theta_e,2}(y_i)$ in the posterior distribution given by (5.26).

In order to link mean and variance, and reduce the number of parameters, an additional common practice is to set:

$$e_{\theta_e,1}(y) = h_1 \circ f(y) \quad \text{and} \quad e_{\theta_e,2}(y) = h_2 \circ f(y),$$

where f is a shared function and h_1 and h_2 are specific functions (giving results in the proper space). Such an encoder can be depicted as shown in Figure 5.11. As the model assumes that the data emerges from the latent space using the decoder, the global framework for estimation in Model 5.4 is known as a *variational autoencoder*, based on the idea that the data "autoencodes" themselves. The complete representation of the variational autoencoder framework is shown on Figure 5.10.

[9] More precisely, a typical loss function for a given latent variable z_i and an observed sample y_i is the reconstruction error $\|y_i - g_{\theta_g}(z_i)\|^2$, which also appears in the p.d.f. of a Gaussian distribution.

[10] As it encodes the data in the latent space.

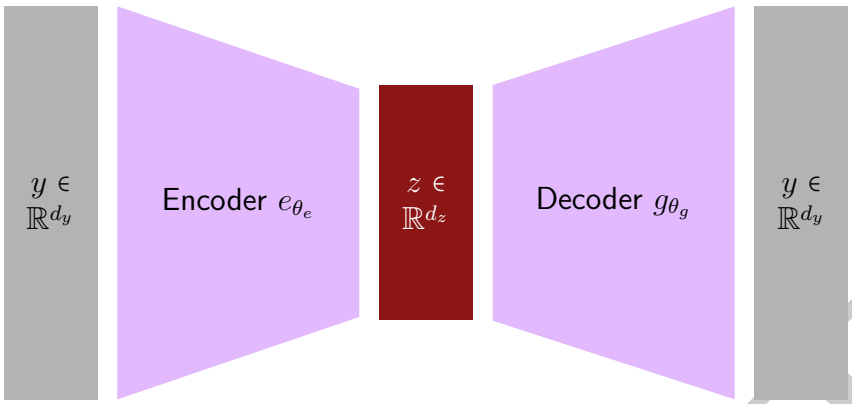


Figure 5.10: Schematic representation of the variational autoencoder framework for variational inference in Model 5.4. The VE step can be seen as learning the encoding part while the M step can be seen as learning the decoding part.

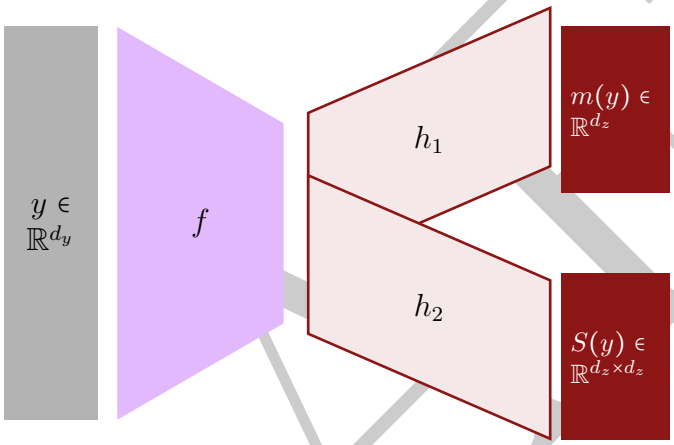


Figure 5.11: Schematic representation of the encoder function to be learnt in the VE step for Model 5.4.

5.4.3 Maximization of the ELBO for variational autoencoders

Overall, the ELBO then becomes a function of θ_g, θ_e and θ_{obs}. Indeed, the variational distribution is given by:

$$q_{\theta_e}(\boldsymbol{z}) = \prod_{i=1}^{n} \phi(z_i; e_{\theta_e,1}(y_i), e_{\theta_e,2}(y_i))\,,$$

where $\phi(\cdot; m, S)$ is the probability distribution function of a Gaussian vector with mean m and covariance matrix S. Using the formulation of Equation (5.4) the ELBO becomes

$$\mathsf{ELBO}(q_{\theta_e}, \theta_g, \theta_{\text{obs}}, \boldsymbol{y}) = \mathbb{E}_{q_{\theta_e}}\left[\sum_{i=1}^{n} \log p_{\theta_{\text{obs}}}\left(y_i; g_{\theta_g}(Z_i)\right)\right] - \mathsf{KL}\left[q_{\theta_e}(\boldsymbol{Z}) \,\|\, \phi(\boldsymbol{Z}; 0, \mathbf{I}_{d_z})\right]. \tag{5.27}$$

The VEM algorithm then corresponds to alternate optimization of the ELBO with respect to θ_e (VE step) and $(\theta_g, \theta_{\text{obs}})$ (M step[11]). Both maximizations are usually done a using gradient algorithm, which requires the computation of the gradient of the ELBO with respect to $\theta = \{\theta_e, \theta_g, \theta_{\text{obs}}\}$.

Note, however, that whereas the Kullback-Leibler divergence on the right hand side has an explicit expression (being the divergence between two Gaussian distributions), the expectation with respect to q_{θ_e} (left term) has no closed form in general due to the highly complex non linear mapping g_{θ_g}. Therefore, nor the ELBO neither its gradient can be computed. In practice, this problem is solved using Monte-Carlo methods. First, note that if $U_i \overset{\text{iid}}{\sim} \mathcal{N}_{d_z}(0, \mathbf{I}_{d_z})$ $(1 \le i \le n)$, then:

$$Z_i \mid \{Y_i = y_i\} \overset{\text{Law}}{=} e_{\theta_e,1}(y_i) + (e_{\theta_e,2}(y_i))^{1/2} U_i\,.$$

[11] Note that in most machine learning applications for generative modelling, there is no θ_{obs} to estimate, as potential noise parameters are supposed to be known.

This writing, known in machine learning as the *reparametrization trick*, enables to express the expectations in (5.27) as expectations with respect to U_i whose distributions do not depend on θ. Therefore, the gradient of the expectation becomes the expectation of the gradient, resulting in:

$$\begin{aligned}\nabla_\theta \mathsf{ELBO}(q_{\theta_e}, \theta_g, \theta_{\text{obs}}, \boldsymbol{y}) = \sum_{i=1}^{n} \mathbb{E}_{U_i}\Big[\nabla_\theta \log p_{\theta_{\text{obs}}}\Big(y_i; g_{\theta_g}\Big(e_{\theta_e,1}(y_i) + (e_{\theta_e,2}(y_i))^{1/2} U_i\Big)\Big)\Big] \\ - \nabla_\theta \,\mathsf{KL}\left[q_{\theta_e}(\boldsymbol{Z}) \parallel \phi(\boldsymbol{Z}; 0, \mathbf{I}_{d_z})\right].\end{aligned}$$

Again, the second term is easy to compute. Nonetheless, the expectations remain intractable, and are, in practice, estimated by Monte Carlo sampling, which allows to obtain an unbiased estimate of the gradient, and then to perform stochastic gradient ascent.

To summarize, on Figure 5.10, the (V)E step consists in maxizing with respect to θ_e, and then learning the left hand side mapping, while the M step maximizes with respect to θ_g, and then learns the right hand side mapping.

5.5 Conclusion of the chapter

In this chapter, we present the principle of VEM and its practical relevance through its application to several models in ecology. Despite its empirical success and computational efficiency, the method remains only partially understood from a theoretical perspective, with relatively few rigorous results available.

Other strategies based on deterministic approximations of the conditional distribution $p_\theta(Z \mid Y = y)$ can also be considered but have not been presented here. For instance, Expectation Propagation (EP) would approximate $p_\theta(Z \mid Y = y)$ by moment matching. While EP may yield more accurate approximations than the variational approximation, its applicability is more limited (since the moment of the true distribution must be known), and its implementation can be technically demanding [see Minka and Lafferty, 2012, for an example of application]. When the posterior $p_\theta(Z \mid Y = y)$ is unimodal and approximately Gaussian, the Laplace EM method replaces the exact E-step with a Gaussian approximation [first paper by Steele, 1996].

All these methods perform deterministic approximations of the conditional distribution in order to replace the exact E step of the EM. Obviously, an another approach is to propose Monte Carlo approximation of the E step resulting into a Monte Carlo EM or variants. This approach presented in the next chapter has found the widest use in practice. It offers a flexible and general way to approximate the E-step using samples, making it applicable to a broad range of models.

Bibliography

J. Aitchison and C.H Ho. The multivariate Poisson-log normal distribution. Biometrika, 76(4):643–653, 1989.

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

C. Ambroise and C. Matias. New consistent and asymptotically normal parameter estimates for random-graph mixture models. Journal of the Royal Statistical Society: Series B, 74(1):3–35, 2012.

Theodore Wilbur Anderson. An introduction to multivariate statistical analysis, volume 2. Wiley New York, 1958.

Julie Aubert, Pierre Barbillon, Sophie Donnet, and Vincent Miele. Using latent block models to detect structure in ecological networks. Statistical Approaches for Hidden Variables in Ecology, pages 117–134, 2022.

P. Bastide, M. Mariadassou, and S. Robin. Models and Methods for Biological Evolution: Mathematical Models and Algorithms to Study Evolution, chapter Evolutionary Models of Continuous Traits, pages 27–38. John Wiley & Sons, 2024.

J. C Bezdek and R. J Hathaway. Convergence of alternating optimization. Neural, Parallel & Scientific Computations, 11(4):351–368, 2003.

P. Bickel, D. Choi, X. Chang, and H. Zhang. Asymptotic normality of maximum likelihood and its variational approximation for stochastic blockmodels. The Annals of Statistics, 41(4):1922 – 1943, 2013. doi: 10.1214/13-AOS1124. URL `https://doi.org/10.1214/13-AOS1124`.

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.

Christopher Bishop. Pattern recognition and machine learning, volume 4. Springer, 2006.

David M Blei, Alp Kucukelbir, and Jon D McAuliffe. Variational inference: A review for statisticians. Journal of the American statistical Association, 112(518):859–877, 2017.

O. Cappé, E. Moulines, and T. Rydén. Inference in Hidden Markov Models. Springer Series in Statistics. Springer, 2005. ISBN 9780387402642. URL `https://books.google.fr/books?id=-3_A3_l1yssC`.

A. Celisse, J.-J. Daudin, and L. Pierre. Consistency of maximum-likelihood and variational estimators in the stochastic block model. Electronic Journal of Statistics, 6(none):1847 – 1899, 2012. doi: 10.1214/12-EJS729. URL `https://doi.org/10.1214/12-EJS729`.

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. Variational inference for probabilistic Poisson PCA. The Annals of Applied Statistics, 12(4):2674–2698, 2018.

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`.

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

M. Delattre, M. Lavielle, and M-A Poursat. A note on bic in mixed-effects models. Electronic Journal of Statistics, 8:456–475, 2014.

Maud Delattre and Estelle Kuhn. Computing an empirical fisher information matrix estimate in latent variable models through stochastic approximation. Computo, 2023.

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.

Kyle JN d'Entremont, Gail K Davoren, Carolyn J Walsh, Sabina I Wilhelm, and William A Montevecchi. Intra-and inter-annual shifts in foraging tactics by parental northern gannets morus bassanus indicate changing prey fields. Marine Ecology Progress Series, 698:155–170, 2022.

Marie du Roy de Chaumaray, Salima El Kolei, and Matthieu Marbac. Estimation of the order of non-parametric hidden markov models using the singular values of an integral operator. arXiv preprint arXiv:2210.03559, 2022.

Jennifer A Dunne, Richard J Williams, and Neo D Martinez. Food-web structure and network theory: the role of connectance and size. Proceedings of the National Academy of Sciences, 99(20):12917–12922, 2002.

Richard Durrett. Probability models for DNA sequence evolution, volume 2. Springer, 2008.

B. Efron and C. Stein. The jackknife estimate of variance. The Annals of Statistics, pages 586–596, 1981.

Charles Elton. Animal ecology. Sidgwick and Jackson, London, page 10, 1927.

Marie-Pierre Etienne and Pierre Gloaguen. Trajectory reconstruction and behavior identification using geolocation data. Statistical Approaches for Hidden Variables in Ecology, pages 1–25, 2022.

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

J. Felsenstein. Evolutionary trees from DNA sequences: a maximum likelihood approach. J. Mol. Evol., 17(6): 368–376, 1981.

J. Felsenstein. Phylogenies and the Comparative Method. The American Naturalist, 125(1):pp. 1–15, January 1985.

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.

Gérard Govaert and Mohamed Nadif. Clustering with block mixture models. Pattern Recognition, 36(2):463 – 473, 2003. ISSN 0031-3203. Biometrics.

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

D. J Harris. Generating realistic assemblages with a joint species distribution model. Methods in Ecology and Evolution, 6(4):465–473, 2015.

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.

L.s.T. Ho and C. Ané. A linear-time algorithm for gaussian and non-gaussian trait evolution models. Systematic biology, 63(3):397–408, 2014.

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.

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.

T.H. Jukes and C.R. Cantor. Evolution of Protein Molecules. New York: Academic Press, 1969.

Rudolph Emil Kalman. A new approach to linear filtering and prediction problems. 1960.

S. Karlin and H.E. Taylor. A second course in stochastic processes. Elsevier, 1981.

M. Kimura. Estimation of evolutionary distances between homologous nucleotide sequences. Proceedings of the National Academy of Sciences, 78(1):454–458, 1981.

Diederik P Kingma. Auto-encoding variational bayes. arXiv preprint arXiv:1312.6114, 2013.

D.P Kingma and M. Welling. An introduction to variational autoencoders. Foundations and Trends® in Machine Learning, 12(4):307–392, 2019.

Daniel B Larremore, Aaron Clauset, and Abigail Z Jacobs. Efficiently inferring community structure in bipartite networks. Physical Review E, 90(1):012805, 2014.

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.

Jean-Benoist Léger. Blockmodels: A r-package for estimating in latent block model and stochastic block model, with various probability functions, with or without covariates. arXiv: Computation, 2016. URL https://api.semanticscholar.org/CorpusID:88515012.

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.

M. Mariadassou and C. Matias. Convergence of the groups posterior distribution in latent or stochastic block models. Bernoulli, 21(1):537–573, 2015.

Mahendra Mariadassou, Stéphane Robin, and Corinne Vacher. Uncovering latent structure in valued graphs: A variational approach. The Annals of Applied Statistics, 4(2):715–742, 06 2010.

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.

Thomas P. Minka and John Lafferty. Expectation-propogation for the generative aspect model, 2012. URL https://arxiv.org/abs/1301.0588.

Juan Manuel Morales, Daniel T Haydon, Jacqui Frair, Kent E Holsinger, and John M Fryxell. Extracting more out of relocation data: building movement models as mixtures of random walks. Ecology, 85(9):2436–2445, 2004.

O. Ovaskainen, G. Tikhonov, A. Norberg, F. Guillaume Blanchet, L. Duan, D. Dunson, T. Roslin, and N. Abrego. How to make more out of community data? a conceptual framework and its implementation as models and software. Ecology Letters, 20(5):561–576, 2017.

Otso Ovaskainen and Nerea Abrego. Joint species distribution modelling: With applications in R. Cambridge University Press, 2020.

É Pardoux. Models and Methods for Biological Evolution: Mathematical Models and Algorithms to Study Evolution, chapter Models of Sequences and Discrete Traits Evolution, pages 27–38. John Wiley & Sons, 2024.

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.

Jennifer Pohle, Roland Langrock, Floris M Van Beest, and Niels Martin Schmidt. Selecting the number of states in hidden markov models: pragmatic solutions illustrated using animal movement. Journal of Agricultural, Biological and Environmental Statistics, 22:270–293, 2017.

G. C Popovic, F. KC Hui, and D. I Warton. A general algorithm for covariance modeling of discrete data. Journal of Multivariate Analysis, 165:86–100, 2018.

G. C. Popovic, D. I. Warton, F. J. Thomson, F. K. C. Hui, and A. T. Moles. Untangling direct species associations from indirect mediator species effects with graphical models. Methods in Ecology and Evolution, 10(9):1571–1583, 2019. doi: 10.1111/2041-210X.13247.

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/.

S. Roweis and Z. Ghahramani. A unifying review of linear gaussian models. Neural Computation, 11(2): 305–345, 02 1999. ISSN 0899-7667. doi: 10.1162/089976699300016674. URL https://doi.org/10.1162/089976699300016674.

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.

G Sellan, FQ Brearley, R Nilus, J Titin, and N Majalap-Lee. Differences in soil properties among contrasting soil types in northern borneo. Journal of Tropical Forest Science, 33(2):191–202, 2021.

Mónica A. Silva, Ian Jonsen, Deborah J. F. Russell, Rui Prieto, Dave Thompson, and Mark F. Baumgartner. Assessing performance of bayesian state-space models fit to argos satellite telemetry locations processed with kalman filtering. PLOS ONE, 9(3):1–13, 03 2014. doi: 10.1371/journal.pone.0092277. URL https://doi.org/10.1371/journal.pone.0092277.

Tom A. B. Snijders and Krzysztof Nowicki. Estimation and prediction for stochastic blockmodels for graphs with latent block structure. J. Classification, 14(1):75–100, 1997. ISSN 0176-4268.

B. M. Steele. A modified em algorithm for estimation in generalized mixed models. Biometrics, 52(4):1295–1310, 1996. ISSN 0006341X, 15410420. URL http://www.jstor.org/stable/2532845.

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

Corinne Vacher, Dominique Piou, and Marie-Laure Desprez-Loustau. Architecture of an antagonistic tree/fungus network: The asymmetric influence of past evolutionary history. PLOS ONE, 3(3):1–10, 03 2008. doi: 10.1371/journal.pone.0001740. URL https://doi.org/10.1371/journal.pone.0001740.

C. Varin, N. Reid, and D. Firth. An overview of composite likelihood methods. Statistica Sinica, 21:5–42, 2011.

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.

M Wikelski, SC Davidson, and R Kays. Movebank: archive, analysis and sharing of animal movement data., 2024. URL www.movebank.org.

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.

Appendix A

Some classical technical results

Contents

This chapter presents key technical results and properties. While these results are commonly found in standard statistics textbooks, we include them here to ensure the document is self-contained. For completeness, we also provide detailed proofs of these results.

A.1 Multivariate distributions

While univariate distributions are well known and widely studied, multivariate distributions are less familiar and often require the use of matrix algebra for their manipulation. To ensure a clear understanding of the material presented in this book, we review general results on the well-known multivariate Gaussian model.

A.1.1 General properties

Proposition A.1. *Let U be a random real vector of size n with mean vector $\mathbb{E}[U] = \mu$ and variance matrix $\mathbb{V}[U] = \Sigma$. Let B be any $p \times n$ matrix, then*

$$\mathbb{E}[BU + \mu_0] = B\mu + \mu_0, \qquad \mathbb{V}[BU + \mu_0] = \mathbb{V}[BU] = B\Sigma B^\top.$$

Let A be a symmetric matrix, denoting $\|U\|_A^2 = U^\top AU$, we have

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

Proof of Proposition A.1

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

$$\mathbb{E}[V] = \mathbb{E}[BU] = B\mu, \qquad \mathbb{V}[V] = \mathbb{V}[BU] = B\Sigma B^\top, \qquad \mathbb{E}\Big[\|U\|_A^2\Big] = \mathbb{E}\big[V^\top V\big].$$

Furthermore, by definition of the variance, we have that

$$\mathbb{E}\big[V^\top V\big] = \mathbb{E}[V]^\top \mathbb{E}[V] + \mathrm{tr}(\mathbb{V}[V]) = \mu^\top B^\top B\mu + \mathrm{tr}(B\Sigma B^\top),$$

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

A.1.2 Multivariate Gaussian distribution

Definition A.1. *A random vector*

$$X = \begin{pmatrix} X_1 \\ \vdots \\ X_p \end{pmatrix} \in \mathbb{R}^p$$

is a Gaussian vector if every non-zero linear combination of its components is a Gaussian random variable. Formally, for all $u = (u_1, \ldots, u_d)' \neq \mathbf{0}$, we have

$$u^\top X \sim \mathcal{N}(\mu_u, \sigma_u^2)\,,$$

where μ_u and σ_u^2 are the expectation and variance (depending on u) of the linear combination. For a Gaussian vector, we write:

$$\mu := \mathbb{E}[X]\,,$$
$$\mathbb{V}[X] := \mathbb{E}\big[(X - \mathbb{E}[X])(X - \mathbb{E}[X])^\top\big]\,.$$

We then use the notation:

$$X \sim \mathcal{N}_p(\mu, \Sigma)\,.$$

We first state the following facts about the Gaussain vector, also known as multivariate Gaussian distribution.

Proposition A.2. *Let $U \sim \mathcal{N}_p(\mu, \Sigma)$. We have the following properties.*

- ***Characteristic function**: The characteristic function of U is given by, for $u \in \mathbb{R}^p$:*

$$\phi_U(u) = \exp\left(iu^\top \mu - \frac{1}{2} u^\top \Sigma u\right).$$

  In particular, a Gaussian vector is defined even if Σ cannot be inverted.

- ***Probability density function** If Σ^{-1} exists, then U admits a probability density function given by, for $u \in \mathbb{R}^p$:*

$$\phi(u; \mu, \Sigma) = \frac{1}{(2\pi)^{p/2}\,|\Sigma|^{1/2}} \exp\left\{-\frac{1}{2}\|u - \mu\|^2_{\Sigma^{-1}}\right\}. \tag{A.1}$$

- ***Linear combination of Gaussian vectors**. Let and $\sim \mathcal{N}_p(\mu_V, \Sigma_V)$ such that U and V are independent, and α and β be two real numbers, then:*

$$\alpha U + \beta V \sim \mathcal{N}_p\big(\alpha\mu + \beta\mu_V, \alpha^2\Sigma + \beta^2\Sigma_V\big)\,. \tag{A.2}$$

- ***Linear transformation of a Gaussian vector**. Let A be a $q \times p$ matrix and $m \in \mathbb{R}^q$. Let's define $V := AU + m$ then:*

$$V \sim \mathcal{N}_q\big(A\mu + m, A\Sigma A^\top\big)\,. \tag{A.3}$$

Proof of Proposition A.2

Proofs of these results can be found in numerous references, such as in Anderson [1958].

We now state useful results linking conditional, marginal, and joint distributions of Gaussian vectors, that will be used all along this book.

Proposition A.3. *Let (U, V) be a couple of real random vectors in $\mathbb{R}^{d_U}$ and $v \in \mathbb{R}^{d_V}$ 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 $\mathbb{C}\mathrm{ov}(U, V) = \Sigma_{UV} = \Sigma_{VU}^{\intercal} = \mathbb{C}\mathrm{ov}(V, U)^{\intercal}$ and joint multivariate Gaussian distribution:*

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

Then

- *the marginal distribution of U and V are given by:*

$$U \sim \mathcal{N}_{d_U}(\mu_U, \Sigma_{UU})$$
$$V \sim \mathcal{N}_{d_V}(\mu_V, \Sigma_{VV});$$

- *the conditional distribution of $U \mid V$ is Gaussian[a] and*

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

[a] The distribution of $V \mid U$ is of course obtained by interverting the matrices.

The proof (and those of the following Propositions) will require the following useful linear algebra results to manipulate inverse of matrices:

Lemma A.1 (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 Lemma A.1

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

Proof of Proposition A.3

The result on the marginal distributions is simply obtained by applying Equation (A.3). For instance, for the marginal of U. Let's consider the diagonal matrix:

$$A = \mathrm{diag}\begin{pmatrix} d_U \text{ times } d_V \text{ times} \\ 1, \ldots, 1, 0, \ldots, 0 \end{pmatrix}.$$

Then, $U = A \times \begin{pmatrix} U \\ V \end{pmatrix}$ and, by Equation (A.3):

$$U \sim \mathcal{N}_{d_U}(\mu_U, \Sigma_{UU}) .$$

Now let's prove the result on the conditional distribution. Let's remark from Equation (A.1) that if $X \sim \mathcal{N}_d(m, S)$, such that $P =: S^{-1}$ exists, then, it's probability density function satisfies, for $x \in \mathbb{R}^d$

$$\log p(x) = -\frac{1}{2}(x-m)^\top P(x-m) + \text{cst} = -\frac{1}{2}x^\top Px + x^\top Pm + \text{cst} , \tag{A.4}$$

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.4) is the one of a Gaussian vector with mean m and variance $S = P^{-1}$. The technique of the proof, well known by Bayesian users as *completing the square*, consists in identifying such form for the law of $U|V$.

Let's consider the matrix Σ of our Proposition, and denote $\Lambda = \Sigma^{-1}$. By Lemma A.1, we have:

$$\Lambda = \begin{bmatrix} \Lambda_{UU} & \Lambda_{UV} \\ \Lambda_{VU} & \Lambda_{VV} \end{bmatrix} ,$$

where[b]

$$\begin{aligned} \Lambda_{UU} &= \left(\Sigma_{UU} - \Sigma_{UV}\Sigma_{VV}^{-1}\Sigma_{VU}\right)^{-1} \\ \Lambda_{UV} &= -\Lambda_{UU}\Sigma_{UV}\Sigma_{VV}^{-1} \\ \Lambda_{VU} &= \Lambda_{UV}^\top . \end{aligned}$$

Now, let's recall that our target density $p(u|v)$ satisfies, for $u \in \mathbb{R}^{d_U}$ and $v \in \mathbb{R}^{d_V}$

$$\log\left(p(u \mid v)\right) = \log\left(\frac{p(u,v)}{p(v)}\right) = \log\left(p(u,v)\right) + \text{cst} .$$

Moreover, $\log(p(u,v))$ is given by :

$$\begin{aligned} \log(p(u,v)) &= -\frac{1}{2}\begin{pmatrix} u^\top & v^\top \end{pmatrix}\begin{bmatrix} \Lambda_{UU} & \Lambda_{UV} \\ \Lambda_{VU} & \Lambda_{VV} \end{bmatrix}\begin{pmatrix} u \\ v \end{pmatrix} + \begin{pmatrix} u^\top & v^\top \end{pmatrix}\begin{pmatrix} \Lambda_{UU}\mu_U + \Lambda_{UV}\mu_V \\ \Lambda_{VU}\mu_U + \Lambda_{VV}\mu_V \end{pmatrix} \\ &= -\frac{1}{2}\left(u^\top\Lambda_{UU}u + u^\top\Lambda_{UV}v\right) + v^\top\Lambda_{VU}u + u^\top\left(\Lambda_{UU}\mu_U + \Lambda_{UV}\mu_V\right) \\ &\quad -\frac{1}{2}v^\top\Lambda_{VV}v + v^\top\left(\Lambda_{VU}\mu_U + \Lambda_{VV}\mu_V\right) \end{aligned}$$

Note that the last row does not depend on u and thus is a constant when considering $p(u \mid v)$. Moreove, note that $v^\top\Lambda_{VU}u$ is a scalar, thus equal to its transpose, and thus equal to $u^\top\Lambda_{UV}v$. This leads to

$$\begin{aligned} \log\left(p(u \mid v)\right) &= -\frac{1}{2}u^\top\Lambda_{UU}u + u^\top\left(\Lambda_{UU}\mu_U + \Lambda_{UV}(\mu_V - v)\right) \\ &= -\frac{1}{2}u^\top\Lambda_{UU}u + u^\top\Lambda_{UU}\left(\mu_U + \Lambda_{UU}^{-1}\Lambda_{UV}(\mu_V - v)\right) . \end{aligned}$$

Then we recognize an expression as in Equation (A.4), *i.e.* the probability distribution of a Gaussian vector with:

$$\begin{aligned} \mathbb{V}[U \mid V] &= \Lambda_{UU}^{-1} = \Sigma_{UU} - \Sigma_{UV}\Sigma_{VV}^{-1}\Sigma_{VU} , \\ \mathbb{E}[U \mid V] &= \mu_U + \Lambda_{UU}^{-1}\Lambda_{UV}(\mu_V - V) \\ &= \mu_U + \Sigma_{UV}\Sigma_{VV}^{-1}(V - \mu_V) , \end{aligned}$$

where we used the expression of Λ_{UU} and Λ_{UV} to conclude.

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

[b] As Λ_{VV} is not needed in the subsequent computations, we spare its expression to the brave reader that came until this point.

Let's now consider two random variables U and V whose distribution satisfy:

$$\begin{aligned} U &\sim \mathcal{N}_{d_U}(\mu, \Omega) \\ V \mid U &= \mathcal{N}_{d_V}(AU + b, \Sigma), \end{aligned} \tag{A.5}$$

where A is a $d_V \times d_U$ matrix. Then, the law of (U, V), $U \mid V$ and V are given by the three following propositions.

Proposition A.4 (Joint distribution of (U, V))**.**

$$\begin{pmatrix} U \\ V \end{pmatrix} \sim \mathcal{N}_{d_U + d_V}\left(\begin{pmatrix} \mu \\ A\mu + b \end{pmatrix}, \begin{pmatrix} \Omega & \Omega A^\top \\ A\Omega & \Sigma + A\Omega A^\top \end{pmatrix}\right).$$

Proposition A.5 (Conditional distribution of $U \mid V$)**.** *Writing $\Lambda = \Omega^{-1}$ and $\Gamma = \Sigma^{-1}$, we have:*

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

Proposition A.6 (Marginal distribution of V)**.**

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

Proof of Proposition A.4

We start by writing the joint p.d.f. of (U, V). Let's denote $\Lambda = \Omega^{-1}$ and $\Gamma = \Sigma^{-1}$.

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

Now, let's stop here. Denoting $x = (u, v)^\top$, we again (almost) recognize the form of Equation (A.4), 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

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

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 A.1 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]:

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

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

Proof of Proposition A.5

The first expression is a direct application of Proposition A.3 to the Gaussian vector given by Proposition A.4. The second expression comes from the Woodbury identity of Lemma A.1. We have that (retransforming precision matrices into variances):

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

And therefore:

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

Proof of Proposition A.6

This is simply obtained by applying the result on the marginals of Proposition A.3 to the element V of Proposition A.4. Note that on other way to prove this is to remark that the second line of Equation (A.5) is equivalent to state that:

$$U = AV + b + E\,,$$

where $E \sim \mathcal{N}_{d_V}(0, \Sigma)$ and is independent of V. Thus, the results directly comes from Equation (A.2).

We finally state two useful results for some computations made in this book.

Proposition A.7. *Let $U \sim \mathcal{N}_p(\mu, \Sigma)$, then*

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

Proof of Proposition A.7

The entropy of a random variable U with density p is $-\mathbb{E}_p[\log p(U)]$, so, taking $U \sim \mathcal{N}_p(\mu, \Sigma)$ and the expression of the pdf for the multivariate Gaussian of Equation (A.1), we have

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

The result follows from Proposition A.1 because $\mathbb{E}[U-\mu] = 0_p$ and $\mathrm{tr}(\Sigma\Sigma^{-1}) = \mathrm{tr}(I_p) = p$.

Proposition A.8 (Moment of order 3 of a Gaussian vector)**.** *Let $W \sim \mathcal{N}_d(m, S)$: Then, for any triplet of indices $1 \le i, j, k \le d$:*

$$\mathbb{E}[W_i W_j W_k] = m_i m_j m_k + m_i S_{jk} + m_j S_{ik} + m_k S_{ij}\,.$$

Proof of Proposition A.8

Let's consider $Z = W - m$, which is a centered Gaussian vector, *i.e.*, for any $1 \le i, j \le d$

$$\begin{aligned}\mathbb{E}[Z_i] &= 0\\ \mathbb{E}[Z_i Z_j] &= \mathbb{V}[Z_{i,j}] = V_{ij}\,.\end{aligned}$$

Now, note that Z and $-Z$ are identically distributed. Then for any triplet of indices $1 \le i, j, k \le d$:

$$\mathbb{E}[Z_i Z_j Z_k] = -\mathbb{E}[(-Z_i) Z_j Z_k] = -\mathbb{E}[Z_i Z_j Z_k] .$$

And therefore $\mathbb{E}[Z_i Z_j Z_k] = 0$. Moreover, we have:

$$\begin{aligned}
\mathbb{E}[W_i W_j W_k] =& \mathbb{E}[(Z_i + m_i)(Z_j + m_j)(Z_k + m_k)] \\
=& m_i m_j m_k + \mathbb{E}[Z_i Z_j Z_k] + m_i \mathbb{E}[Z_j Z_k] + m_j \mathbb{E}[Z_i Z_k] + m_k \mathbb{E}[Z_i Z_j] \\
&+ m_i m_j \mathbb{E}[Z_k] + m_i m_k \mathbb{E}[Z_j] + m_j m_k \mathbb{E}[Z_k] \\
=& m_i m_j m_k + m_i S_{jk} + m_j S_{ik} + m_k S_{ij} .
\end{aligned}$$

A.2 Exponential family and generalized linear models

We begin by recalling fundamental properties of distributions from the exponential family.

A.2.1 The natural exponential family

The natural exponential family is a class of probability distributions that includes common examples such as the normal, Bernoulli, binomial, Poisson, and gamma distributions, among others. What these distributions have in common is that they can be written in a specific exponential form, allowing for a unified presentation of results.

Definition A.2. *The distribution $f(\,.\,;\gamma)$ with support $\mathcal{Y}$ belongs to exponential family with parameter $\gamma \in \mathbb{R}^d$ if*

$$f(y;\gamma) = \exp[\gamma^\top S(y) - a(y) - b(\gamma)], \qquad \forall y \in \mathcal{Y}, \tag{A.6}$$

where $S(y) \in \mathbb{R}^d$ is the vector of the sufficient statistics , $\gamma \mapsto b(\gamma) \in \mathbb{R}$ is a differentiable function. If, in addition, $y \mapsto S(y)$ is the identity then γ is called the canonical parameter (or natural parameter)

Remark. *Note that another formulation used in Chapter ?? is:*

$$f(y;\theta) = \exp\left[-\psi(\theta) + \langle S(y), \phi(\theta)\rangle\right]$$

which is equivalent with $\gamma = \phi(\theta)$ and $\psi(\theta) = a(y) - b(\gamma)$.

Examples

- If $f(\,.\,;\lambda)$ is the exponential distribution $\mathcal{E}(\lambda)$, then:

$$\begin{aligned} f(y;\lambda) = \lambda e^{-\lambda y} &= \exp[-\lambda y + \log\lambda] \\ &= \exp[\gamma y - (-\log(-\gamma))] \end{aligned}$$

  $\gamma = -\lambda$ is the canonical parameter, and $b(\gamma) = -\log(-\gamma)$.

- If $f(\,.\,;p)$ is the Bernoulli distribution $\mathcal{B}\text{ern}(p)$, then:

$$\begin{aligned} f(y;p) = p^y(1-p)^{1-y} &= \exp[y\log p + (1-y)\log(1-p)] \\ &= \exp[\log\left(\frac{p}{1-p}\right) y + \log(1-p)] \end{aligned}$$

  $\gamma = \log\left(\frac{p}{1-p}\right)$ is the canonical parameter. Moreover, $p = \frac{e^\gamma}{1+e^\gamma}$ et $b(\gamma) = -\log(1-p) = -\log\left(\frac{1}{1+e^\gamma}\right) = \log(1+e^\gamma)$.

- If $f(\,.\,;\mu)$ is the Poisson distribution $\mathcal{P}(\mu)$, then:

$$f(y;\mu) = e^{-\mu}\frac{\mu^y}{y!} = \exp[y\log\mu - \mu - \log(y!)]$$

  $\gamma = \log\mu$ is the canonical parameter and $b(\gamma) = \mu = \exp(\gamma)$.

Moments in the exponential family The parameter γ is linked to the moments of the sufficient statistics as follows:

Proposition A.9.

$$\mathbb{E}[S(Y)] = \nabla_\gamma b(\gamma), \qquad \mathbb{V}[S(Y)] = \mathbf{H}_\gamma\, b(\gamma)$$

where ∇ and $\mathbf{H}$ are respectively the gradient and Hessian of b. In addition, if γ is the canonical parameter, then

$$\mathbb{E}[Y] = \nabla_\gamma b(\gamma), \qquad \mathbb{V}[Y] = \mathbf{H}_\gamma\, b(\gamma) \tag{A.7}$$

Proof of Proposition A.9

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. $\mathbf{H}_\theta \psi(\theta)$ is the Hessian matrix i.e. the matrix of the second derivatives of ϕ : $\mathbf{H}_\theta(\psi(\theta)) = \mathbf{J}_\theta(\nabla_\theta \psi(\theta))$.

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).

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

that is

$$0 = [\mathbb{E}[S(y)] - b'(\gamma)] \underbrace{\int f_Y(y;\gamma,\phi)\,\mathrm{d}y}_{=1}.$$

Differentiating one more time, we get

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

Maximum likelihood estimation in the exponential family

Proposition A.10. *For an iid sample $\boldsymbol{Y} = (Y_1, \dots Y_n)$ of $f(\cdot;\gamma)$, the log-likelihood of $\boldsymbol{y}$ realisation of $\boldsymbol{Y}$ is:*

$$\log p_\gamma(\boldsymbol{y}) = \log p_\gamma(y_1, \dots y_n) = \sum_{i=1}^n \left[\gamma^\top S(y_i) - a(y_i) - b(\gamma) \right]. \tag{A.8}$$

$\gamma \mapsto \log p_\gamma(\boldsymbol{y})$ is a concave function w.r.t. γ

Proof of Proposition A.10

The expression of the likelihood is easy to write. On the one hand, the Hessian matrix of the log-likelihood (A.8) is $-n\,\mathbf{H}\, b(\gamma)$. On the other hand, Proposition A.9 states that $\mathbf{H}\, b(\gamma)$ 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.

Proposition A.11. *The MLE $\widehat{\gamma} = \arg\max_\gamma \log p(\boldsymbol{y})$ satisfies*

$$\nabla_\gamma b(\widehat{\gamma}) = \frac{1}{n} \sum_{i=1}^n S(Y_i) =: \overline{S}(Y).$$

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

Proof of Proposition A.11

The derivative of $\gamma \mapsto \log p_\gamma(y_1, \dots y_n)$ w.r.t. γ is

$$\nabla_\gamma \log p_\gamma(y_1, \dots y_n) = \sum_{i=1}^{n} S(Y_i) - n\nabla_\gamma b(\gamma).$$

Setting it to zero gives the result.

A.2.2 Generalized linear models

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.

Definition A.3. *The generalized linear is given by a probability distribution for Y_i with density $p_{\gamma_i}(y)$ and a function g called the link function such that*

$$g(\mathbb{E}[Y_i]) = x_i^\top \beta\,.$$

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.

Choosing the link function g Any bijection from the space of $\mathbb{E}[Y]$ into $\mathbb{R}$ can be chosen as a link function g. However, in the case where $p_\gamma(y)$ belongs to the exponential family, we choose as link function the function that transforms the expectation $\mathbb{E}[Y]$ into the canonical parameter i.e.

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

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.7, 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

- If $y_i \in \{0, 1\}$, it is natural to model y_i using the Bernoulli distribution:

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

  Let $g : [0, 1] \to \mathbb{R}$ be a bijective function. We define:

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

  - Noticing that this distribution belongs to the exponential family and using the previous calculus, we get $b(\gamma_i) = \log(1 + e^{\gamma_i})$ and so

    $$\mathbb{E}[Y_i] = b'(\gamma_i) = \frac{e^{x_i^\top \beta}}{1 + e^{x_i^\top \beta}}$$

    i.e. b' is the inverse logit function, $g = \text{logit}$.
  - Alternatively, one can choose g^{-1} as the cumulative distribution function (CDF) of the standard normal distribution, since it is a bijection from $\mathbb{R}$ to $[0, 1]$, in which case $g = \text{probit}$.

- If $y_i \in \mathbb{N}$, we may model y_i using the Poisson distribution: $Y_i \sim \mathcal{P}(\mu_i)$. In this case, the distribution belongs to the exponential family with $b(\gamma_i) = b'(\gamma) = e^\gamma$. $\mathbb{E}[Y_i] = e^{x_i^\top \beta} \in \mathbb{R}^{*+}$.

A.3 Graphical models

In this section we only provide a few notions on graphical models. The reader may refer to the book by Lauritzen [1996] for more formalism and theory.

A graphical model is a representation of a probabilistic model as a graph, which encodes its conditional independence structure. In this graph, nodes represent random variables, and edges signify the conditional independence relations between these variables. The graphical model hence provides a global picture of the dependence structure. For example, it allows to see if 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 $\mathcal{G}$ is made of a set of nodes (or vertices) $\mathcal{V} = \{1, \dots N\}$ and a set of edges $\mathcal{E}$, which are pairs of nodes. Formally, a graph is therefore defined as

$$\mathcal{G} = (\mathcal{V}, \mathcal{E}), \qquad \mathcal{E} \subset \mathcal{V} \times \mathcal{V}.$$

In a directed graph, the edges have directions, meaning that the edge $(1,2)$ is different from the edge $(2,1)$. The edges are therefore represented as arrows, the edge $(1,2)$ being represented as $1 \to 2$. Figure A.1 gives an example of a directed graph.

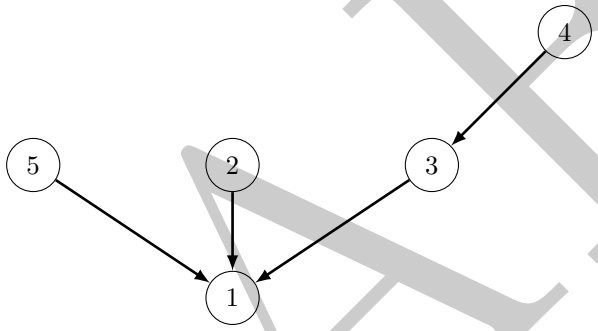


Figure A.1: An example of directed acyclic graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ with vertices $\mathcal{V} = \{1,2,3,4,5\}$ and edges $\mathcal{E} = \{(5,1),(2,1),(3,1),(4,3)\}$

Definition A.4 (Vocabulary for directed graphs)**.**

- *If there is an arrow from U to V then U is said to be a parent for V. We denote by $pa(V,\mathcal{G})$ the set of all the parents of V in graph $\mathcal{G}$.*
- *A directed path between two nodes is an ordered set of nodes connected by edges all with the same direction linking one node to the other as a chain.*
- *U is an ancestor for V is there is a directed path from U to V or equivalently, V is a descendant of U. We denote by $desc(U,\mathcal{G})$ the set of descendant of U in the directed graph $\mathcal{G}$.*
- *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.*

Remark. *Obviously, if $\mathcal{G}$ is a DAG, a parent of node cannot be one of its descendant:*

$$\forall V \in \mathcal{V}: \qquad pa(V,\mathcal{G}) \cap desc(V,\mathcal{G}) = \varnothing.$$

In Figure A.1, the parents of node 1 are 3, 4 and 5 so $pa(1,\mathcal{G}) = \{5,2,3\}$, the only parent of 3 is 4 so $pa(3,\mathcal{G}) = \{4\}$ and 5 has no parent, so $pa(5,\mathcal{G}) = \varnothing$. In the same figure, $(4,3,1)$ forms a directed path from 4 to 1: $4 \to 3 \to 1$.

Figure A.2 gives an example of a cycle. The graph represented in Figure A.1 has no cycle and is therefore a DAG. Directed graphical models are supposed to DAGs.

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

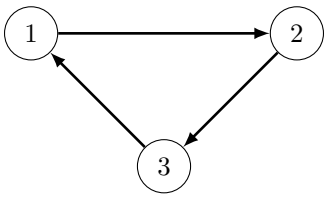


Figure A.2: An example of directed graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ with vertices $\mathcal{V} = \{1, 2, 3\}$ and edges $\mathcal{E} = \{(1, 2), (2, 3), (3, 1)\}$ forming a cycle.

A.3.2 DAGs and probability

Let us consider a set of variables $\mathcal{U} = \{U_i\}_{1:N}$ and a DAG $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ where $\mathcal{V} = \{1, \dots, N\}$. The distribution P on $\mathcal{V}$ is said to be factorized with respect to $\mathcal{G}$ if the joint distribution $p(U_{1:N})$ can be written as

$$p(U_{1:N}) = \prod_{i=1}^{N} p(U_i \mid (U_j)_{j \in pa(i,\mathcal{G})}).$$

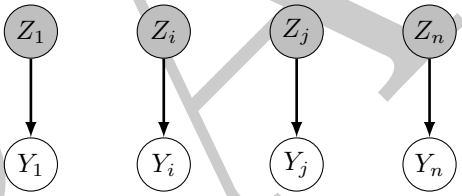
In the sequel we will often abuse the notations and identify the node i with the corresponding variable U_i.

Examples of DAG and corresponding joint probability.

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

$$p(U_1, U_2, U_3, U_4, U_5) = p(U_1 \mid U_2, U_3, U_5)p(U_2)p(U_3 \mid U_4)p(U_4)p(U_5).$$

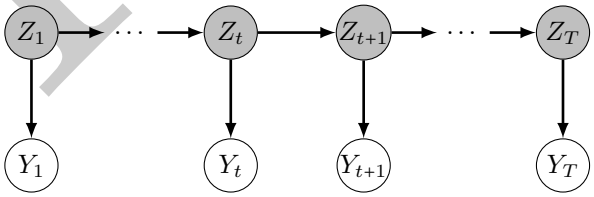
- The graphical model associated with the mixture model (Section 3.1) and the the zero-inflated Poisson model (Section 3.2) is the following DAG:



  meaning that:

$$p(Y_{1:n}, X_{1:n}) = \left(\prod_{i=1}^{n} p(Z_i)\right)\left(\prod_{i=1}^{n} p(Y_i \mid Z_i)\right).$$

- The graphical model associated with the Hidden Markov model (Section 4.1.1) is the following DAG:



  meaning that:

$$p(Y_{1:n}, Z_{1:n}) = p(Z_1)\left(\prod_{t=2}^{n} p(Z_t \mid Z_{t-1})\right)\left(\prod_{t=1}^{n} p(Y_t \mid Z_t)\right).$$

- Obviously, no distribution $p(U_1, U_2, U_3)$ could be factorized with respect to the DAG from Figure A.2, as the factorization $p(U_1 \mid U_3)p(U_2 \mid U_1)p(U_3 \mid U_2)$ makes no sense.

Proposition A.12. *Let $p(U_1, \dots U_n)$ be a joint distribution factorized with respect to the directed graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ with $\mathcal{V} = (U_1, \dots U_N)$.*
Define the set $\overline{desc}(V)$ of non-descendants of node $V \in \mathcal{V}$ as the set of the nodes that neither its descendants,

nor its parents (nor itself):

$$\overline{desc}(V) = \mathcal{V} \setminus (\{V\} \cup pa(V, \mathcal{G}) \cup desc(V, \mathcal{G}))\,.$$

Then, conditionally on its parents, a node is independent from its non-descendants

$$\forall V \in \mathcal{V}: \qquad V \perp\!\!\!\perp \overline{desc}(V, \mathcal{G}) \mid pa(V, \mathcal{G}).$$

Consequently: $p(V \mid \overline{desc}(V, \mathcal{G}), pa(V, \mathcal{G})) = p(V \mid pa(V, \mathcal{G}))$.

Moralization of a DAG. The moral version of a directed graph $\mathcal{G}$ is obtained by marrying (i.e. setting an edge between) all the pairs of parents and then by removing the directions on the edges. The moral version of the DAG of Figure A.1 is given in Figure A.3.

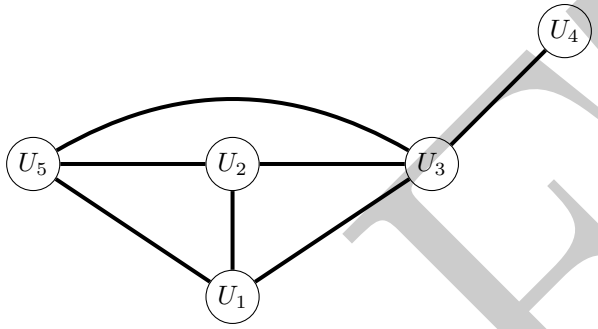


Figure A.3: Moral graph of the DAG from Figure A.1.

In an undirected graph, an undirected path between two nodes is a set of edges (ignoring the directions) linking one node to the other. In Figure A.3, (U_5, U_3), (U_5, U_2, U_3) and (U_5, U_1, U_3) are all paths relating U_5 to U_3.

Proposition A.13. *Let $p(U_1, \dots U_n)$ be a joint distribution factorized with respect to the directed graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ (with $\mathcal{V} = \{U_1, \dots U_n\}$).*
Let I, J and K be three subsets of $\{1, \dots N\}$ and define the corresponding subsets of random variables as $U_I = \{U_i\}_{i \in I}$, $U_J = \{U_j\}_{j \in J}$ and $U_K = \{U_k\}_{k \in K}$.
Then, in the moral graph deduced from $\mathcal{G}$, if all the paths from I to J pass through K then $U_K I$ is independent from U_J conditionally on U_K:

$$U_I \perp\!\!\!\perp U_J \mid U_K.$$

As a consequence: $p(U_I \mid U_J, U_K) = p(U_I \mid U_K)$.

Illustration of Proposition A.13 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:

$$\mathbb{P}(U_5, U_2, U_1, U_4 \mid U_3) = \mathbb{P}(U_5, U_2, U_1 \mid U_3)\mathbb{P}(U_4 \mid U_3)$$

2. $\mathbb{P}(U_1 \mid U_2, U_3, U_4, U_5) = \mathbb{P}(U_1 \mid U_2, U_3, U_5)$

A.3.3 Using the DAG to set independence properties in the HMM

Proposition A.14. *For the HMM whose DAG is provided in Figure A.4, the following independence properties hold:*

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

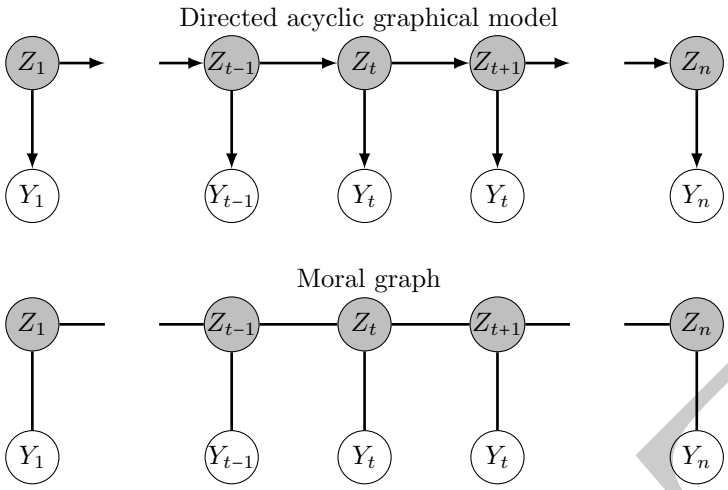


Figure A.4: Graphical models associated with an HMM. Top: Directed (DAG) graph, Bottom: Moralized (undirected) graph.

Proof of Proposition A.14

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

Proposition A.15. *Conditionally on the observed data $\mathbf{Y} = Y_{1:n}$, (Z_t) is still a Markov chain. In addition,*
$$\mathbb{P}(Z_{t+1} \mid Z_{1:t}, Y_{1:n}) = \mathbb{P}(Z_{t+1} \mid Z_t, Y_{t+1:n})\,.$$

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

Proof of Proposition A.15 (Version 1)

1. We have to prove that conditionnally on the observations, (Z_t) is still a Markov Chain, i.e.
$$\mathbb{P}(Z_{t+1} \mid Z_{1:t}, Y_{1:n}) = \mathbb{P}(Z_{t+1} \mid Z_t, Y_{1:n})\,.$$
Using that $\mathbb{P}(Z_{t+1} \mid Z_{1:t}, Y_{1:n}) = \mathbb{P}\left(\underbrace{Z_{t+1}}_{I} \mid \underbrace{Z_{1:t-1}}_{J}, \underbrace{Z_t, Y_{1:n}}_{K}\right)$, 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 $\mathbb{P}(Z_{t+1} \mid Z_{1:t}, Y_{1:n}) = \mathbb{P}(Z_{t+1} \mid Z_t, Y_{t+1:n})$. We have:

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

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

Proof of Proposition A.15 (Version 2)

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

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

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

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

Finally:

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

A.4 Derivation of the Bayesian Information Criterion (BIC)

This section presents key mathematical results essential for the derivation of the BIC.

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

Lemma A.2 (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*

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

Proof of Proposition A.2

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

$$\int_{\mathbb{R}^D} \exp\left(-\frac{1}{2} \|u\|_A^2 \right) \mathrm{d}u = |A|^{-1/2} (2\pi)^{D/2} \tag{A.9}$$

(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.9), taking $A = -nL''(u^*)$ (which is positive definite because, u^* being a maximum, $L''(u^*)$ is negative definite) and observing that $|-nL''(u^*)| = n^D \left|-L''(u^*)\right|$.

Derivation of the BIC. We only provide here an idea of the derivation of Equation (2.26)

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

where $\widehat{\theta}_m = \arg\max_\theta p(\boldsymbol{y} \mid \theta, M = m)$ is the maximum likelihood estimate of θ under model m. We consider the independent and identically distributed setting, that is: we assume that the data are $\boldsymbol{Y} = (Y_1, \dots Y_n)$, where the Y_i are all independent conditional on θ and m, with distribution $p(y_i \mid \theta, m)$.

Proof of Equation (2.26)

We propose a sketch of proof of Equation (2.26) in the i.i.d. case. We define the normalized (conditional) log-likelihood of the observe dataset $\boldsymbol{y} = (y_1, \dots y_n)$ as

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

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

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

where the latter equality holds as long as neither $\log p(\theta \mid m)$ nor $\log p(m)$ depend on n. Then, using Lemma A.2, we have that:

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

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

the logarithm gives

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

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 4.1.1), or the multivariate Poisson log-normal model (Section 5.3).*

Appendix B

Proofs

Contents

B.1 Proof of Proposition 3.6

Proof of Proposition 3.6

Complete score function The complete score function gives the derivative of the complete log-likelihood with respect to β, σ^2 and $\left(\gamma_j^2\right)_{1\leq j\leq r}$.

From Equation (3.5), we have

$$\begin{aligned} S_\beta(\boldsymbol{y},\boldsymbol{Z}) &= \nabla_\beta \log p_\theta(\boldsymbol{y},\boldsymbol{Z}) \\ &= \frac{1}{\sigma^2}\boldsymbol{X}^\top\left(\boldsymbol{y}-\boldsymbol{X}\beta-\boldsymbol{U}\boldsymbol{Z}\right). \qquad \text{(B.1)} \\ S_{\sigma^2}(\boldsymbol{y},\boldsymbol{Z}) &= \nabla_{\sigma^2} \log p_\theta(\boldsymbol{y},\boldsymbol{Z}) \\ &= -\frac{n}{2\sigma^2}+\frac{1}{2\sigma^4}\|\boldsymbol{y}-\boldsymbol{X}\beta-\boldsymbol{U}\boldsymbol{Z}\|^2 \qquad \text{(B.2)} \\ \forall 1\leq j\leq r,\quad S_{\gamma_j^2}(\boldsymbol{y},\boldsymbol{Z}) &= \nabla_{\gamma_j^2} \log p_\theta(\boldsymbol{y},\boldsymbol{Z}) \\ &= -\frac{m_j}{2\gamma_j^2}+\frac{1}{2\gamma_j^4}\|\boldsymbol{Z}_j\|^2. \qquad \text{(B.3)} \end{aligned}$$

Hessian matrix We now focus on the hessian matrix, which we will consider block-wise:

$$\mathbf{H}_\theta\left(\log p_\theta(\boldsymbol{y})\right) = \begin{pmatrix} \mathbf{H}_\beta(\boldsymbol{y}) & \mathbf{H}_{\beta,\sigma^2}(\boldsymbol{y}) & \mathbf{H}_{\beta,\gamma_{1:r}^2}(\boldsymbol{y}) \\ \mathbf{H}_{\beta,\sigma^2}^\top(\boldsymbol{y}) & \mathbf{H}_{\sigma^2}(\boldsymbol{y}) & \mathbf{H}_{\sigma^2,\gamma_{1:r}^2}(\boldsymbol{y}) \\ \mathbf{H}_{\beta,\gamma_{1:r}^2}^\top(\boldsymbol{y}) & \mathbf{H}_{\sigma^2,\gamma_{1:r}^2}^\top(\boldsymbol{y}) & \mathbf{H}_{\gamma_{1:r}^2}(\boldsymbol{y}). \end{pmatrix}$$

As we are only interested in computing confidence intervals for β^*, let's first prove that:

$$\mathbb{E}_{\widehat{\theta}}\left[\mathbf{H}_{\beta,\sigma^2}(\boldsymbol{Y})\right] = \mathbb{E}_{\widehat{\theta}}\left[\mathbf{H}_{\beta,\gamma_{1:r}^2}(\boldsymbol{Y})\right] = 0\,.$$

This would be of great interest, as, in this case:

$$I_n(\widehat{\theta})a = \mathbb{E}_{\widehat{\theta}}[\mathbf{H}_\theta\left(p_\theta(\boldsymbol{Y})\right)] = \begin{pmatrix} \mathbb{E}_{\widehat{\theta}}[\mathbf{H}_\beta(\boldsymbol{Y})] & 0 & 0 \\ 0 & \mathbb{E}_{\widehat{\theta}}[\mathbf{H}_{\sigma^2}(\boldsymbol{Y})] & \mathbb{E}_{\widehat{\theta}}\left[\mathbf{H}_{\sigma^2,\gamma^2_{1:r}}(\boldsymbol{Y})\right] \\ 0 & \mathbb{E}_{\widehat{\theta}}\left[\mathbf{H}^\top_{\sigma^2,\gamma^2_{1:r}}(\boldsymbol{Y})\right] & \mathbb{E}_{\widehat{\theta}}\left[\mathbf{H}_{\gamma^2_{1:r}}(\boldsymbol{Y})\right] \end{pmatrix}$$
$$a = \begin{pmatrix} \mathbb{E}_{\widehat{\theta}}[\mathbf{H}_\beta(\boldsymbol{Y})] & 0 \\ 0 & \boldsymbol{M}_{\widehat{\sigma}^2,\widehat{\gamma}^2_{1:r}}(\boldsymbol{Y}) \end{pmatrix},$$

where $\boldsymbol{M}_{\widehat{\sigma}^2,\widehat{\gamma}^2_{1:r}}$ is a matrix containing all terms not related to the variance of β^*. Therefore, $\widehat{\mathbb{V}}\left[\widehat{\theta}\right] = I_n(\widehat{\theta})^{-1}$ would also be block diagonal, and our variance of interest will be given by $\mathbb{E}_{\widehat{\theta}}[\mathbf{H}_\beta(\boldsymbol{Y})]^{-1}$

Let's prove that is is indeed the case.

Proof that $\mathbf{H}_{\beta,\sigma^2}(\boldsymbol{y}) = 0$ Thanks to Equation (2.21), restricted to β and $\gamma^2_{1:r}$ we have:

$$\mathbf{H}_{\beta,\sigma^2}(\boldsymbol{y}) = \mathbb{E}_{\widehat{\theta}}[\mathbf{J}_{\sigma^2}\left(S_\beta(\boldsymbol{y},\boldsymbol{Z})\right) \mid \boldsymbol{Y} = \boldsymbol{y}] + \mathbb{C}\mathrm{ov}_{\widehat{\theta}}\left(S_\beta, S_{\sigma^2} \mid \boldsymbol{Y} = \boldsymbol{y}\right). \tag{B.4}$$

Considering the first term, we have:

$$\begin{aligned}\mathbb{E}_{\widehat{\theta}}[\mathbf{J}_{\sigma^2} S_\beta(\boldsymbol{y},\boldsymbol{Z}) \mid \boldsymbol{Y} = \boldsymbol{y}] &= -\frac{1}{\sigma^4}\boldsymbol{X}^\top\left(\boldsymbol{y} - \boldsymbol{X}\beta - \boldsymbol{U}\mathbb{E}_{\widehat{\theta}}[\boldsymbol{Z} \mid \boldsymbol{Y} = \boldsymbol{y}]\right) \\ &= -\frac{1}{\sigma^4}\boldsymbol{X}^\top\left(\boldsymbol{y} - \boldsymbol{X}\beta - \boldsymbol{U}\widehat{m}\right). \end{aligned} \tag{B.5}$$

Moreover, let's denote

$$W := \boldsymbol{y} - \boldsymbol{X}\beta - \boldsymbol{U}\boldsymbol{Z}.$$

We have that:

$$m^{(W)} := \mathbb{E}_{\widehat{\theta}}[W \mid \boldsymbol{Y} = \boldsymbol{y}] = \boldsymbol{y} - \boldsymbol{X}\beta - \boldsymbol{U}\widehat{m}, \tag{B.6}$$
$$V^{(W)} := \mathbb{V}_{\widehat{\theta}}[W \mid \boldsymbol{Y} = \boldsymbol{y}] = \boldsymbol{U}\widehat{O}\boldsymbol{U}^\top. \tag{B.7}$$

Therefore:

$$\begin{aligned}\mathbb{C}\mathrm{ov}_{\widehat{\theta}}\left(S_\beta, S^2_\sigma \mid \boldsymbol{Y} = \boldsymbol{y}\right) &= \frac{1}{2\sigma^6}\boldsymbol{X}^\top\mathbb{C}\mathrm{ov}_{\widehat{\theta}}\left(W, \|W\|^2\right) \\ &= \frac{1}{2\sigma^6}\boldsymbol{X}^\top\left(\mathbb{E}_{\widehat{\theta}}\left[\|W\|^2 \times W \mid \boldsymbol{Y} = \boldsymbol{y}\right] - \mathbb{E}_{\widehat{\theta}}\left[\|W\|^2 \mid \boldsymbol{Y} = \boldsymbol{y}\right] \times \mathbb{E}_{\widehat{\theta}}[W \mid \boldsymbol{Y} = \boldsymbol{y}]\right) \end{aligned} \tag{B.8}$$

The third expectation is given by (B.6), let's consider the second one:

$$\begin{aligned}\mathbb{E}_{\widehat{\theta}}\left[\|W\|^2 \mid \boldsymbol{Y} = \boldsymbol{y}\right] &= \mathbb{E}_{\widehat{\theta}}\left[W^\top W \mid \boldsymbol{Y} = \boldsymbol{y}\right] \\ &= (\boldsymbol{y} - \boldsymbol{X}\beta)^\top(\boldsymbol{y} - \boldsymbol{X}\beta) - 2(\boldsymbol{y} - \boldsymbol{X}\beta)^\top\boldsymbol{U}\mathbb{E}_{\widehat{\theta}}[\boldsymbol{Z} \mid \boldsymbol{Y} = \boldsymbol{y}] + \mathbb{E}_{\widehat{\theta}}\left[\boldsymbol{Z}^\top\boldsymbol{U}^\top\boldsymbol{U}\boldsymbol{Z} \mid \boldsymbol{Y} = \boldsymbol{y}\right] \\ &= (\boldsymbol{y} - \boldsymbol{X}\beta)^\top(\boldsymbol{y} - \boldsymbol{X}\beta) - 2(\boldsymbol{y} - \boldsymbol{X}\beta)^\top\boldsymbol{U}\widehat{m} + \mathrm{tr}\left(\boldsymbol{U}\mathbb{E}_{\widehat{\theta}}\left[\boldsymbol{Z}\boldsymbol{Z}^\top \mid \boldsymbol{Y} = \boldsymbol{y}\right]\boldsymbol{U}^\top\right) \\ &= (\boldsymbol{y} - \boldsymbol{X}\beta)^\top(\boldsymbol{y} - \boldsymbol{X}\beta) - 2(\boldsymbol{y} - \boldsymbol{X}\beta)^\top\boldsymbol{U}\widehat{m} + \mathrm{tr}\left(\boldsymbol{U}\widehat{m}\widehat{m}^\top\boldsymbol{U}^\top\right) + \mathrm{tr}\left(\boldsymbol{U}\widehat{O}\boldsymbol{U}^\top\right) \\ &= \|\boldsymbol{y} - \boldsymbol{X}\beta - \boldsymbol{U}\widehat{m}\|^2 + \mathrm{tr}\left(\boldsymbol{U}\widehat{O}\boldsymbol{U}^\top\right)\end{aligned}$$

Then, we have that:

$$\mathbb{E}_{\widehat{\theta}}\left[\|W\|^2 \mid \boldsymbol{Y} = \boldsymbol{y}\right] \times \mathbb{E}_{\widehat{\theta}}[W \mid \boldsymbol{Y} = \boldsymbol{y}] = \left(\|\boldsymbol{y} - \boldsymbol{X}\beta - \boldsymbol{U}\widehat{m}\|^2 + \mathrm{tr}\left(\boldsymbol{U}\widehat{O}\boldsymbol{U}^\top\right)\right)(\boldsymbol{y} - \boldsymbol{X}\beta - \boldsymbol{U}\widehat{m}) \tag{B.9}$$

Consider then the expectation:

$$\mathbb{E}_{\widehat{\theta}}\left[\|W\|^2 \times W \mid \boldsymbol{Y} = \boldsymbol{y}\right] = \sum_{i=1}^n \mathbb{E}_{\widehat{\theta}}\left[W_i^2 \times W \mid \boldsymbol{Y} = \boldsymbol{y}\right].$$

Thanks to Proposition A.8, we can compute the j-th element ($1 \leq j \leq n$) of this vector:

$$\begin{aligned}\mathbb{E}_{\widehat{\theta}}\Big[\|W\|^2 \times W \mid \boldsymbol{Y} = \boldsymbol{y}\Big]_j &= \sum_{i=1}^{n} \mathbb{E}_{\widehat{\theta}}\big[W_i^2 W_j \mid \boldsymbol{Y} = \boldsymbol{y}\big] \\ &= \sum_{i=1}^{n} \left(\left(m_i^{(W)}\right)^2 + V_{ii}^{(W)}\right) \times m_j^{(W)} + 2\sum_{i=1}^{n} V_{ji}^{(W)} m_i^{(W)}.\end{aligned}$$

Then, stacking all these elements, we have:

$$\begin{aligned}\mathbb{E}_{\widehat{\theta}}\Big[\|W\|^2 \times W \mid \boldsymbol{Y} = \boldsymbol{y}\Big] &= \left(\left\|m^{(W)}\right\|^2 + \operatorname{tr}\left(V^{(W)}\right)\right) m^{(W)} + 2V^{(W)} m^{(W)} \\ &= \left(\|\boldsymbol{y} - \boldsymbol{X}\beta - \boldsymbol{U}\widehat{m}\|^2 + \operatorname{tr}\left(\boldsymbol{U}\widehat{O}\boldsymbol{U}^\top\right)\right)(\boldsymbol{y} - \boldsymbol{X}\beta - \boldsymbol{U}\widehat{m}) + 2\boldsymbol{U}\widehat{O}\boldsymbol{U}^\top(\boldsymbol{y} - \boldsymbol{X}\beta - \boldsymbol{U}\widehat{m}) \,.\end{aligned} \tag{B.10}$$

Plugging (B.9) and (B.10) in (B.8), we conclude:

$$\mathbb{C}\mathrm{ov}_{\widehat{\theta}}\left(S_\beta, S_\sigma^2 \mid \boldsymbol{Y} = \boldsymbol{y}\right) = \frac{1}{\sigma^6}\boldsymbol{X}^\top\boldsymbol{U}\widehat{O}\boldsymbol{U}^\top(\boldsymbol{y} - \boldsymbol{X}\beta - \boldsymbol{U}\widehat{m}) \,. \tag{B.11}$$

Combining (B.5) and (B.11) in (B.4), we have:

$$\mathbf{H}_{\beta,\sigma^2}(\boldsymbol{y}) = \boldsymbol{M}(\boldsymbol{y} - \boldsymbol{X}\beta - \boldsymbol{U}\widehat{m}) \,,$$

where $\boldsymbol{M}$ is a matrix constant with respect to $\boldsymbol{y}$. Now, consider the expression of $\hat{m}$ which, at convergence, depends linearly on $\left(\boldsymbol{y} - \boldsymbol{X}\hat{\beta}\right)$. Therefore, the previous equation can be factorized with respect to $\left(\boldsymbol{y} - \boldsymbol{X}\hat{\beta}\right)$ and we have:

$$\mathbb{E}_{\widehat{\theta}}\big[\mathbf{H}_{\widehat{\beta},\widehat{\sigma}^2}(\boldsymbol{Y})\big] = \widehat{\boldsymbol{M}}\mathbb{E}_{\widehat{\theta}}\big[\boldsymbol{Y} - \boldsymbol{X}\hat{\beta}\big] = 0 \,,$$

where $\widehat{\boldsymbol{M}}$ is again some matrix not depending on $\boldsymbol{Y}$, and the expectation is 0 by hypothesis of Model 3.5.

Proof that $\mathbf{H}_{\beta,\gamma_{1:r}^2}(\boldsymbol{y}) = 0$ Thanks to Equation (2.21), restricted to β and $\gamma_{1:r}^2$ we have:

$$\mathbf{H}_{\beta,\gamma_{1:r}^2}(\boldsymbol{y}) = \mathbb{E}_{\widehat{\theta}}\Big[\mathbf{J}_{\gamma_{1:r}^2}\left(S_\beta(\boldsymbol{y},\boldsymbol{Z})\right) \mid \boldsymbol{Y} = \boldsymbol{y}\Big] + \mathbb{C}\mathrm{ov}_{\widehat{\theta}}\Big(S_\beta(\boldsymbol{y},\boldsymbol{Z}), S_{\gamma_{1:r}^2}(\boldsymbol{y},\boldsymbol{Z}) \mid \boldsymbol{Y} = \boldsymbol{y}\Big).$$

For the first term, performing deriving (B.1) with respect to $\gamma_{1:r}^2$, we have:

$$\mathbb{E}_{\widehat{\theta}}\Big[\mathbf{J}_{\gamma_{1:r}^2} S_\beta(\boldsymbol{y},\boldsymbol{Z}) \mid \boldsymbol{Y} = \boldsymbol{y}\Big] = 0 \,,$$

Moreover, we have, for $1 \leq j \leq r$:

$$\begin{aligned}\mathbb{C}\mathrm{ov}_{\widehat{\theta}}\left(S_\beta, S_{\gamma_j^2} \mid \boldsymbol{Y} = \boldsymbol{y}\right) &= \frac{1}{2\widehat{\sigma}^2\widehat{\gamma}_j^2}\boldsymbol{X}^\top\boldsymbol{U}\,\mathbb{C}\mathrm{ov}_{\widehat{\theta}}\left(\boldsymbol{Z}, \|\boldsymbol{Z}_j\|^2\right) \\ &= \frac{1}{2\widehat{\sigma}^2\widehat{\gamma}_j^2}\boldsymbol{X}^\top\boldsymbol{U}\left(\mathbb{E}_{\widehat{\theta}}\Big[\|\boldsymbol{Z}_j\|^2 \times \boldsymbol{Z} \mid \boldsymbol{Y} = \boldsymbol{y}\Big] - \mathbb{E}_{\widehat{\theta}}\Big[\|\boldsymbol{Z}_j\|^2 \mid \boldsymbol{Y} = \boldsymbol{y}\Big] \times \mathbb{E}_{\widehat{\theta}}[\boldsymbol{Z} \mid \boldsymbol{Y} = \boldsymbol{y}]\right)\end{aligned} \tag{B.12}$$

Let's denote $\widehat{m}_j = \mathbb{E}_{\widehat{\theta}}[\boldsymbol{Z}_j \mid \boldsymbol{Y} = \boldsymbol{y}]$ and $\widehat{O}_j = \mathbb{V}_{\widehat{\theta}}[\boldsymbol{Z}_j \mid \boldsymbol{Y} = \boldsymbol{y}]$. We have, using again Propositions A.1 and A.8, and calculations similar to those of the previous paragraph to set (denoting $\mathcal{I}_j$ the set of indices of $\boldsymbol{Z}$ whose marginal variance is γ_j^2):

$$\begin{aligned}\mathbb{E}_{\widehat{\theta}}\Big[\|\boldsymbol{Z}_j\|^2 \mid \boldsymbol{Y} = \boldsymbol{y}\Big] \times \mathbb{E}_{\widehat{\theta}}[\boldsymbol{Z} \mid \boldsymbol{Y} = \boldsymbol{y}] &= \left(\|\widehat{m}_j\|^2 + \operatorname{tr}\left(\widehat{O}_j\right)\right)\widehat{m} \\ \mathbb{E}_{\widehat{\theta}}\Big[\|\boldsymbol{Z}_j\|^2 \boldsymbol{Z} \mid \boldsymbol{Y} = \boldsymbol{y}\Big] &= \left(\|\widehat{m}_j\|^2 + \operatorname{tr}\left(\widehat{O}_j\right)\right)\widehat{m} + 2\widehat{O}_{\mathcal{I}_j}\widehat{m} \,,\end{aligned}$$

where $\widehat{O}_{\mathcal{I}_j}$ is the $d_z \times d_z$ matrix where all rows corresponding to those of Γ that does not contain γ_j are 0 (the other being the ones of $\widehat{O}$). Therefore, we have:

$$\mathbb{C}\mathrm{ov}_{\widehat{\theta}}\left(S_\beta, S_{\gamma_j^2} \mid \boldsymbol{Y} = \boldsymbol{y}\right) = \frac{1}{\widehat{\sigma}^2\widehat{\gamma}_j^2}\boldsymbol{X}^\top\boldsymbol{U}\widehat{O}_{\mathcal{I}_j}\widehat{m}.$$

Again, using the fact that $\widehat{m}$ depends linearly on $\boldsymbol{y} - \boldsymbol{X}\hat{\beta}$, we have:

$$\mathbb{E}_{\widehat{\theta}}\Big[\mathbf{H}_{\widehat{\beta},\widehat{\gamma}^2_{1:r}}(\boldsymbol{Y})\Big] = \mathbb{E}_{\widehat{\theta}}\Big[\mathbb{C}\mathrm{ov}_{\widehat{\theta}}\Big(S_{\widehat{\beta}}(\boldsymbol{y},\boldsymbol{Z}), S_{\widehat{\gamma}^2_{1:r}}(\boldsymbol{y},\boldsymbol{Z}) \mid \boldsymbol{Y} = \boldsymbol{y}\Big)\Big] = \widehat{\boldsymbol{M}}\mathbb{E}_{\widehat{\theta}}\big[\boldsymbol{Y} - \boldsymbol{X}\hat{\beta}\big] = 0.$$

We then proved that the Fisher information $I_n(\widehat{\theta})$ is block diagonal.

Expression of $\mathbb{V}\big[\widehat{\beta}\big]$

$$\mathbf{H}_\beta(\boldsymbol{y}) = \mathbb{E}_{\widehat{\theta}}[\mathbf{J}_\beta\left(S_\beta(\boldsymbol{y},\boldsymbol{Z})\right) \mid \boldsymbol{Y} = \boldsymbol{y}] + \mathbb{V}_{\widehat{\theta}}[S_\beta(\boldsymbol{y},\boldsymbol{Z}) \mid \boldsymbol{Y} = \boldsymbol{y}]\,.$$

We have on the one hand:

$$\mathbb{E}_{\widehat{\theta}}[\mathbf{J}_\beta S_\beta(\boldsymbol{y},\boldsymbol{Z}) \mid \boldsymbol{Y} = \boldsymbol{y}] = -\mathbb{E}_{\widehat{\theta}}\Big[\frac{1}{\sigma^2}\boldsymbol{X}^\top\boldsymbol{X} \mid \boldsymbol{Y} = \boldsymbol{y}\Big] = -\frac{1}{\sigma^2}\boldsymbol{X}^\top\boldsymbol{X}\,.$$

On the other hand:

$$\begin{aligned}
\mathbb{V}_{\widehat{\theta}}[S_\beta(\boldsymbol{y},\boldsymbol{Z}) \mid \boldsymbol{Y} = \boldsymbol{y}] &= \mathbb{E}_{\widehat{\theta}}\big[S_\beta(\boldsymbol{y},\boldsymbol{Z})S_\beta(\boldsymbol{y},\boldsymbol{Z})^\top \mid \boldsymbol{Y} = \boldsymbol{y}\big] - \mathbb{E}_{\widehat{\theta}}[S_\beta(\boldsymbol{y},\boldsymbol{Z}) \mid \boldsymbol{Y} = \boldsymbol{y}]\mathbb{E}_{\widehat{\theta}}[S_\beta(\boldsymbol{y},\boldsymbol{Z}) \mid \boldsymbol{Y} = \boldsymbol{y}]^\top \\
&= \frac{1}{\sigma^4}\boldsymbol{X}^\top\Big\{\left(\boldsymbol{y} - \boldsymbol{X}\beta\right)\left(\boldsymbol{y} - \boldsymbol{X}\beta\right)^\top - \left(\boldsymbol{y} - \boldsymbol{X}\beta - \boldsymbol{U}\boldsymbol{Z}\right)\mathbb{E}_{\widehat{\theta}}[\boldsymbol{Z} \mid \boldsymbol{Y} = \boldsymbol{y}]^\top\boldsymbol{U}^\top \\
&\quad - \boldsymbol{U}\mathbb{E}_{\widehat{\theta}}[\boldsymbol{Z} \mid \boldsymbol{Y} = \boldsymbol{y}]\left(\boldsymbol{y} - \boldsymbol{X}\beta\right)^\top + \boldsymbol{U}\mathbb{E}_{\widehat{\theta}}\big[\boldsymbol{Z}\boldsymbol{Z}^\top \mid \boldsymbol{Y} = \boldsymbol{y}\big]\boldsymbol{U}^\top \\
&\quad - \left(\boldsymbol{y} - \boldsymbol{X}\beta\right)\left(\boldsymbol{y} - \boldsymbol{X}\beta\right)^\top + \left(\boldsymbol{y} - \boldsymbol{X}\beta\right)\mathbb{E}_{\widehat{\theta}}[\boldsymbol{Z} \mid \boldsymbol{Y} = \boldsymbol{y}]^\top\boldsymbol{U}^\top \\
&\quad + \boldsymbol{U}\mathbb{E}_{\widehat{\theta}}[\boldsymbol{Z} \mid \boldsymbol{Y} = \boldsymbol{y}]\left(\boldsymbol{y} - \boldsymbol{X}\beta\right)^\top - \boldsymbol{U}\mathbb{E}_{\widehat{\theta}}[\boldsymbol{Z} \mid \boldsymbol{Y} = \boldsymbol{y}]\mathbb{E}_{\widehat{\theta}}[\boldsymbol{Z} \mid \boldsymbol{Y} = \boldsymbol{y}]^\top\boldsymbol{U}^\top\Big\}\boldsymbol{X} \\
&= \frac{1}{\sigma^4}\boldsymbol{X}^\top\boldsymbol{U}\widehat{O}\boldsymbol{U}^\top\boldsymbol{X}
\end{aligned}$$

Therefore, as none of the terms depend on $\boldsymbol{y}$, the Fisher information matrix is simply equal to the sum, and we retrieve the result of the proposition:

$$\mathbb{V}\big[\hat{\beta}\big] = \sigma^2\left(\boldsymbol{X}^\top\boldsymbol{X} - \frac{1}{\sigma^2}\boldsymbol{X}^\top\boldsymbol{U}\widehat{O}\boldsymbol{U}^\top\boldsymbol{X}\right)^{-1}.$$

B.2 Proof of Proposition 4.11

Proof of Proposition 4.11

We have to prove each recursion formulas.

Upward recursion: The initialization is straightforward, than the recursion is the same as the recursion given in 4.54 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 (4.53), is evaluated thanks to (4.54) and the marginal distribution $p_\theta(Z_j)$ is evaluated with (4.51) and (4.50).

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 4.4: 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

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

Appendix C

Technical detail for Stochastic Block Models

Contents

In this section, we provide technical details related to the Stochastic Block Model (SBM) used for analyzing network datasets. The first part focuses on the derivation of the ICL criterion used to select the number of blocks in the SBM. The second part addresses the Stochastic Block Model for bipartite networks, with a focus on the VEM algorithm and the ICL criterion.

C.1 Derivation of the model selection penalty term in ICL for 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. Equations (5.19) and (5.20) are recalled here:

$$\log\left(\int p(\boldsymbol{y}, \boldsymbol{z}, \theta, K)\, \mathrm{d}\theta\right) = \log p_{\widehat{\theta}_K}(\boldsymbol{y}, \boldsymbol{z}) - \mathrm{pen}(K) + O_n(1),$$

where

$$\mathrm{pen}(K) = \frac{1}{2}\left(\frac{K(K+1)}{2}\log\left(\frac{n(n-1)}{2}\right) + (K-1)\log n\right).$$

We provide a sketch of proof of these equations.

Proof of Equations 5.19 and 5.20

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 to get

Equation (2.26). Hence we get for the first term

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

because the $n(n-1)/2$ edges of the network $\boldsymbol{Y}$ are conditionally independent given $(\boldsymbol{Z}, \alpha, K)$ and because there are $K(K+1)/2$ independent parameters in α in the binary or Poisson version of Model 5.1. As for the second term, we have that

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

because the n latent variables Z_i underlying the network $\boldsymbol{Y}$ are conditionally independent given $(\boldsymbol{\omega}, K)$ and because there are $(K-1)$ independent parameters in $\boldsymbol{\omega}$. Gathering Equations (C.1) and (C.2) gives (5.19) and (5.20) using frequentist notations.

The ICL is obtained by integrating the latent variation in Equation (5.19) with respect to $p_{\widehat{\theta}_m}(\boldsymbol{Z} \mid \boldsymbol{Y})$. However, this distribution is not explicit in the SBM and so is replaced by its variational approximation $\hat{q}(\boldsymbol{Z})$, leading to the variational ICL ($\mathrm{ICL}_{\mathsf{V}}$, see Section (5.1.5)).

C.2 Mathematical details for the inference of bipartite SBM

The extension of the Stochastic Block Model (SBM) to bipartite networks is briefly introduced in Section 5.2.7. In this section, we provide the technical details, beginning with a reminder of the model definition, followed by an outline of the VEM algorithm, and concluding with the derivation of the ICL criterion for model selection.

C.2.1 Model, parameters and complete likelihood

We first recall the model:

Model C.1. ***(bipartite SBM)*** *Let $(Y_{ij})_{i=1,\dots,n,j=1,\dots,p}$ be the incidence matrix of a bipartite network involving $n \times p$ nodes.*

$$\begin{aligned} Z_i &\overset{iid}{\sim} \mathrm{Cat}\left(\boldsymbol{\omega}^{(1)}\right) \quad \forall i = 1, \dots, n \\ W_j &\overset{iid}{\sim} \mathrm{Cat}\left(\boldsymbol{\omega}^{(2)}\right) \quad \forall j = 1, \dots, p \end{aligned} \tag{C.3}$$

Then, conditionally to the latent variables $\boldsymbol{Z}$ and $\boldsymbol{W}$, the nodes are connected independently:

$$Y_{ij} \mid \{Z_i = k, W_j = \ell\} \overset{ind}{\sim} \mathcal{F}(\alpha_{k\ell}) \qquad \forall (i,j) \in \{1, \dots, n\} \times \{1, \dots, p\} \tag{C.4}$$

The vector of proportions $\boldsymbol{\omega}^{(1)} = (\omega_1^{(1)}, \dots, \omega_K^{(1)}) \in [0,1]^K$ and $\boldsymbol{\omega}^{(1)}$ are such that their sum is equal to 1.

Parameters The parameters of the model are $\theta = (\boldsymbol{\omega}^{(1)}, \boldsymbol{\omega}^{(2)}, (\alpha_{k\ell})_{k,\ell=1,\dots,K})$, $\boldsymbol{\omega}^{(1)}$ being the vector of the row block proportions while $\boldsymbol{\omega}^{(2)}$ is the vector of the column block proportions. The $\alpha_{k\ell}$ are the interblock connection parameters. We do not talk about intra-block connection parameters proportions since the row and colum blocks are defined on different types of nodes (plants/pollinators for intance, or trees/fungis).

Complete log-Likelihood Denoting $\boldsymbol{y} = \{y_{ij}\}_{(i,j)=1\dots,n,j=1,\dots,p}$, $\boldsymbol{Z} = \{Z_i\}_{1 \le i \le n}$ and $\boldsymbol{W} = \{W_j\}_{1 \le j \le p}$, the complete likelihood of model (C.3) and (C.4) is

$$\begin{aligned} \log p_\theta(\boldsymbol{y}, \boldsymbol{Z}, \boldsymbol{W}) &= \log p_\theta(\boldsymbol{Z}) + \log p_\theta(\boldsymbol{W}) + \log p_\theta(\boldsymbol{y} \mid \boldsymbol{Z}, \boldsymbol{W}) \\ &= \sum_{i=1}^{n}\sum_{k=1}^{K} Z_{ik} \log \omega_k^{(1)} + \sum_{j=1}^{p}\sum_{\ell=1}^{L} W_{j\ell} \log \omega_k^{(2)} + \sum_{i,j=1}^{n,p}\sum_{k,\ell=1}^{K,L} Z_{ik} W_{j\ell} \log f(y_{ij}; \alpha_{k\ell})\,. \end{aligned} \tag{C.5}$$

where, as before $Z_{ik} = \mathbb{I}_{Z_i = k}$ and $W_{j\ell} = \mathbb{I}_{W_j = \ell}$,

C.2.2 VEM algorithm for the bipartite SBM

In the bipartite SBM, two sets of latent variables have been defined, namely $\boldsymbol{Z}$ (row clustering) and $\boldsymbol{W}$ (column clustering). We define $\mathcal{Q}$ such that $\forall q \in \mathcal{Q}$:

$$q(\boldsymbol{Z}, \boldsymbol{W}) = \prod_{i=1}^{n} q_i(Z_i) \prod_{j=1}^{p} q_j(W_j) = \prod_{i=1}^{n} \prod_{k=1}^{K} \tau_{ik}^{Z_{ik}} \prod_{j=1}^{p} \prod_{\ell=1}^{L} \rho_{j\ell}^{W_{i\ell}}, \tag{C.6}$$

where $\tau_{ik} = \mathbb{E}_q[Z_{ik}]$ and $\rho_{j\ell} = \mathbb{E}_q[W_{j\ell}]$.

VE step for the bipartite SBM

Proposition C.1. *In the bipartite SBM with approximate conditional distribution q chosen in $\mathcal{Q}$ defined before, the solution of the VE step are obtained by alternatively computing:*

$$\tau_{ik} := \mathbb{E}_q[Z_{ik}] \quad \propto \quad \omega_k^{(1)} \prod_{j=1}^{p} \prod_{\ell=1}^{L} f(y_{ij}; \alpha_{k\ell})^{\rho_{j\ell}}, \tag{C.7}$$

$$\rho_{j\ell} := \mathbb{E}_q[W_{j\ell}] \quad \propto \quad \omega_\ell^{(2)} \prod_{i=1}^{n} \prod_{k=1}^{K} f(y_{ij}; \alpha_{k\ell})^{\tau_{ik}}. \tag{C.8}$$

Proof of Proposition C.1

The objective function is given by:

$$\begin{aligned}\mathsf{ELBO}(q, \theta, \boldsymbol{y}) = &\sum_{i=1}^{n} \sum_{k=1}^{K} \tau_{ik} \log \omega_k^{(1)} + \sum_{j=1}^{p} \sum_{\ell=1}^{L} \rho_{j\ell} \log \omega_\ell^{(2)} + \sum_{i,j=1}^{n,p} \sum_{k,\ell=1}^{K,L} \tau_{ik} \rho_{j\ell} \log f(y_{ij}; \alpha_{k\ell}) \\ &- \sum_{i=1}^{n} \sum_{k=1}^{K} \tau_{ik} \log \tau_{ik} - \sum_{j=1}^{p} \sum_{\ell=1}^{L} \rho_{j\ell} \log \rho_{j\ell}.\end{aligned}$$

To maximize the function, we derive with respect to the τ_{ik}'s and $\rho_{j\ell}$'s, subject to the constraint $\sum_{k=1}^{K} \tau_{ik} = 1$, for all $i \in \{1, \ldots, n\}$ and $\sum_{\ell=1}^{L} \rho_{j\ell} = 1$, for all $j \in \{1, \ldots, p\}$. The derivative with respect to τ_{ik} is zero iff

$$\log \omega_k^{(1)} + \sum_{j=1}^{p} \sum_{\ell=1}^{L} \rho_{j\ell} \log f(y_{ij}; \alpha_{k\ell}) - \log \tau_{ik} - 1 - \lambda_i^{(1)} = 0,$$

where $\lambda_i^{(1)}$ is the Lagrange multiplier for the constraint $\sum_{k=1}^{K} \tau_{ik} = 1$. Similarly, the derivative with respect to $\rho_{j\ell}$ is zero iff

$$\log \omega_\ell^{(2)} + \sum_{i=1}^{n} \sum_{k=1}^{K} \tau_{ik} \log f(y_{ij}; \alpha_{k\ell}) - \log \rho_{j\ell} - 1 - \lambda_j^{(2)} = 0,$$

where $\lambda_j^{(2)}$ is the Lagrange multiplier for the constraint $\sum_{\ell=1}^{L} \rho_{j\ell} = 1$, which proves the proposition.

M step for the bipartite SBM

Proposition C.2. *In the bipartite SBM with approximate conditional distribution q chosen in $\mathcal{Q}$ and such that the emission distribution belongs to the exponential family:*

$$f(y; \gamma) = \exp[\alpha^\top S(y) - a(y) - b(\alpha)]$$

where $t(y)$ is the vector of the sufficient statistics, the solution of the M step of the VEM is:

$$\widehat{\omega_k^{(1)}} = \frac{\sum_{i=1}^{n} \tau_{ik}}{n}, \tag{C.9}$$

$$\widehat{\omega_k^{(2)}} = \frac{\sum_{j=1}^{p} \rho_{j\ell}}{n}, \tag{C.10}$$

$$\nabla b(\widehat{\alpha}_{k\ell}) = \frac{\sum_{i=1}^{n} \sum_{j=1}^{p} \tau_{ik} \rho_{j\ell} \, S(y_{ij})}{\sum_{i=1}^{n} \sum_{j=1}^{p} \tau_{ik} \rho_{j\ell}}. \tag{C.11}$$

In particular, if the emission distribution is $\mathcal{B}ern(p_{k\ell})$ for binary interaction or $\mathcal{P}(\mu_{k\ell})$ for weigthed interaction then

$$\widehat{p}_{k\ell} = \widehat{\mu}_{k\ell} = \frac{\sum_{i=1}^{n}\sum_{j=1}^{p} \tau_{ik}\rho_{j\ell}\, y_{ij}}{\sum_{i=1}^{n}\sum_{j=1}^{p} \tau_{ik}\rho_{j\ell}}\,.$$

Proof of Proposition C.2

The proof is the same as for SBM case (Proposition 5.4).

C.2.3 ICL criterion for the bipartite SBM

Both the number of blocks of type 1 (K) and the number of blocks of type 2 (L) need usually to be estimated for the bipartite SBM. Using the same Bayesian reasoning as in Section 2.4 to define both the BIC and the ICL criteria (see also Appendix A.4), we need to approximate the integral

$$\int p(\boldsymbol{y}, \boldsymbol{z}, \boldsymbol{w}, \theta, K)\, \mathrm{d}\theta = \int p(\boldsymbol{y}, | \boldsymbol{z}, \boldsymbol{w}, \theta, K) p(\boldsymbol{z}, | \theta, K) p(\boldsymbol{w}, | \theta, K) p(\theta \mid K)\, \mathrm{d}\theta\,.$$

Again, assume that the prior distribution $p(\theta \mid K)$ is such that $\boldsymbol{\omega}^{(1)}$, $\boldsymbol{\omega}^{(2)}$, α are all independent conditionally on K and L, i.e.:

$$p(\boldsymbol{\omega}^{(1)}, \boldsymbol{\omega}^{(2)}, \alpha \mid K, L) = p(\boldsymbol{\omega}^{(1)} \mid K, L) p(\boldsymbol{\omega}^{(2)} \mid K, L) p(\alpha \mid K, L)\,.$$

Assume moreover that both dimensions of the network go to infinity at the same rate ($\lim_{n\to\infty} n/p = \mathrm{cst}$), we may prove for the binary and the Poisson bipartite SBM that

$$\log\left(\int p(\boldsymbol{y}, \boldsymbol{z}, \boldsymbol{w}, \theta, K, L)\, \mathrm{d}\theta\right) = \log_{\widehat{\theta}}(\boldsymbol{y}, \boldsymbol{z}, \boldsymbol{w}) - \mathrm{pen}(K, L) + O_n(1)\,, \tag{C.12}$$

where

$$\mathrm{pen}(K, L) = \frac{1}{2}\left\{KL\log(np) + (K-1)\log n + (L-1)\log p\right\}\,. \tag{C.13}$$

This penalty $\mathrm{pen}(K, L)$ is now composed of three terms corresponding to α, $\boldsymbol{\omega}^{(1)}$ and $\boldsymbol{\omega}^{(2)}$ and the respective number of variables (np edges Y_{ij} for α, n latent variables Z_i for $\boldsymbol{\omega}^{(1)}$ and p latent variables W_j for $\boldsymbol{\omega}^{(2)}$).

Proof of Equations (C.12) **and** (C.13)

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

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

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: $\lim_{n\to\infty} n/p = \mathrm{cst}$):

$$\begin{aligned}
\log\left(\int p(\boldsymbol{y} \mid \boldsymbol{z}, \boldsymbol{w}, \alpha, K) p(\alpha \mid K)\, \mathrm{d}\alpha\right) &= \log_{\widehat{\theta}}(\boldsymbol{y} \mid \boldsymbol{z}, \boldsymbol{w}) - \frac{KL}{2}\log(np) + O_n(1)\\
\log\left(\int p(\boldsymbol{z}, | \boldsymbol{\omega}^{(1)}, K) p(\boldsymbol{\omega}^{(1)} \mid K)\, \mathrm{d}\boldsymbol{\omega}\right) &= \log_{\widehat{\theta}}(\boldsymbol{z}) - \frac{K-1}{2}\log(n) + O_n(1)\\
\log\left(\int p(\boldsymbol{w}, | \boldsymbol{\omega}^{(2)}, K) p(\boldsymbol{\omega}^{(2)} \mid K)\, \mathrm{d}\boldsymbol{\omega}\right) &= \log_{\widehat{\theta}}(\boldsymbol{w}) - \frac{L-1}{2}\log(p) + O_n(1).
\end{aligned}$$

Gathering the three equations gives Equations (C.12) and (C.13).

Again, the ICL criterion is given by the dominant term of Equation (C.12) and we now have to deal with the unknown latent $\boldsymbol{Z}$ and $\boldsymbol{W}$. When factorized approximate distribution are used, the variational modes of $\boldsymbol{Z}$ and $\boldsymbol{W}$ can be defined as for the SBM model, defined as:

$$\widetilde{\boldsymbol{z}} = \left(\arg\max_{z_i} \widetilde{q}_i^{(1)}(z_i) \right)_{1 \leq i \leq n},$$
$$\widetilde{\boldsymbol{w}} = \left(\arg\max_{w_i} \widetilde{q}_i^{(2)}(w_i) \right)_{1 \leq i \leq n}.$$

We end up with the two variational versions of the ICL criterion:

$$\begin{aligned}
\mathrm{ICL}_{\mathsf{V},1}(K, L) &= \log p_{\widetilde{\theta}}(\boldsymbol{y}, \widetilde{\boldsymbol{z}}, \widetilde{\boldsymbol{w}}) - \mathrm{pen}(K, L), \\
\mathrm{ICL}_{\mathsf{V},2}(K, L) &= \mathbb{E}_{\widetilde{q}}\big[\log p_{\widetilde{\theta}}(\boldsymbol{y}, \boldsymbol{Z}, \boldsymbol{W})\big] - \mathrm{pen}(K, L) \\
&= \mathsf{ELBO}(\widetilde{q}, \widetilde{\theta}, \boldsymbol{y}) - \mathsf{Ent}_{\widetilde{q}^{(1)}}[\boldsymbol{Z}] - \mathsf{Ent}_{\widetilde{q}^{(2)}}[\boldsymbol{W}] - \mathrm{pen}(K, L)
\end{aligned}$$

where $\mathrm{pen}(K, L)$ is given in Equation (C.13).

Index