Models with Hidden Structure

with Applications in Biology and Genomics

Lecture Notes
Master 2 'Mathématiques pour les Sciences du Vivant'

S. Robin
INRA / AgroParisTech

January 29, 2021

Contents

1 Introduction

The purpose of statistical modeling is often to retrieve some hidden process that is at work behind what is observed. The purpose of this lecture is to present a series of statistical models involving hidden, or latent, variables with application to (molecular) biology. In this field the hidden process often refers to some unobserved classification, so the hidden process is supposed to have a discrete state-space. Still, most of the techniques presented hereafter can be generalized to continuous state-space models.

Such models are part of so-called incomplete data models, the inference of which requires some specific developments. Most of the techniques that will be presented consists in variations around the expectation-maximization (EM) algorithm first proposed by Dempster *et al.* (1977). In the last decades, this family of algorithms has been re-considered and casted into a larger framework based on graphical models and variational techniques (see Jaakkola (2001) for an introduction or Wainwright and Jordan (2008) for a very complete review).

In such algorithms, the critical step is often the determination of the conditional distribution of the hidden variables given the observed ones, or at least the calculation of some of its moments. The three parts of these notes refer respectively to the cases where, in a frequentist framework,

Section 2: the calculation of the required conditional moments of the hidden variables is straightforward,

Section 3: this calculation is not straightforward but still possible,

Section 4: this calculation is not possible and approximations are needed.

The last section generalizes Section 4 to the Bayesian framework.

Section 5: the calculation of the joint conditional distribution of the parameters and the hidden variables is not possible and variational approximations can be derived.

Acknowledgements. The author is extremely grateful to Pierre Barbillon, Maud Delattre and Sarah Ouadah for their careful reading, and their helpful comments, remarks and advises.

Notations. All along these notes, we will use the following notations for the variables:

$Y =$ observed variables;

$Z =$ unobserved (hidden, latent) variables;

$\theta =$ parameters;

$x =$ covariates (if needed).

As for the distributions, we will denote

$f(\cdot)$ **or** $p(\cdot) =$ probability distribution function (pdf);

$f_\theta(\cdot) = f(\cdot;\theta)$ **or** $p_\theta(\cdot) = p(\cdot;\theta) =$ pdf with parameter θ;

$\mathbb{E}_\theta =$ expectation under p_θ.

The subscript θ may be replaced by the distribution itself (e.g. $\mathbb{E}_p$ or $\mathbb{E}_q$) or dropped when not necessary.

As for classical distributions, we will use the following notations:

$\mathcal{U}_{[a,b]} =$ uniform distribution over the interval $[a, b]$;

$\mathcal{N}(\mu, \sigma^2) =$ Gaussian distribution with mean μ and variance σ^2;

$\mathcal{M}(n, \pi) =$ multinomial distribution with n draws and vector of probabilities $\pi = (\pi_1, \dots, \pi_K)$, $(\sum_k \pi_k = 1)$;

$\mathcal{P}(\lambda) =$ Poisson distribution with mean λ;

$\mathbf{B}(\alpha, \beta) =$ beta distribution;

$\mathcal{D}(\alpha) =$ Dirichlet distribution with parameter $\alpha = (\alpha_1, \dots, \alpha_K)$;

$\mathcal{NB}(\pi, r) =$ negative binomial distribution with probability π and number of successes r;

$\mathcal{G}\mathbf{am}(a, b) =$ gamma distribution with shape parameter a and rate b.

We will also use the abbreviations *rv* for 'random variable', *iid* for 'independent and identically distributed' and *wrt* for 'with respect to'. We will denote $[\![i, j]\!] = \{i, i+1, \dots, j-1, j\}$.

2 Independent latent variables: Mixture models

2.1 Examples

In this chapter, we consider one of the most simple latent variable model: the mixture model. In this model, observations are supposed to be independent, each arising from a given class that is unobserved. One of the main goal when using mixture models is to retrieve the class from which each observation arises. Such a problem is often referred to as 'unsupervised classification' as we do not dispose of any observation with known label.

We first present a series of biological examples in which a mixture model turns out to be useful.

2.1.1 Gene expression

Functional analysis of one gene. To better understand the function of a given gene, on may measure its expression level Y_i in a large set of conditions (e.g. different drug treatments, tissues, patients with different diseases, ...) $i \in [\![1, I]\!]$. One hope then to be able to define a typology, that is to construct a classification, of conditions in which the gene under study is, e.g., highly, weakly or not expressed. Figure 2.1 displays an example of such measurements.
A latent (i.e. unobserved) status Z_i is associated with each condition and the following model is proposed:

$$\begin{aligned} (Z_i)_i \text{ iid} &\sim \mathcal{M}(1;\pi), \qquad Z_i \in [\![1, K]\!], \\ (Y_i)_i \text{ indep.} \,|\, (Z_i) &\sim F(\gamma_{Z_i}) \end{aligned}$$

where $F(\gamma)$ stands for some parametric distribution with parameter γ. Depending on the technology, one may take $F(\gamma_k) = \mathcal{N}(\mu_k, \sigma_k^2)$ for continuous measurements (such as fluorescence measurements provided by microarrays), $F(\gamma_k) = \mathcal{P}(\gamma_k)$ or $F(\gamma_k) = \mathcal{NB}(\mu_k, \phi_k)$ for count data (such as read counts provided by deep sequencing technologies).

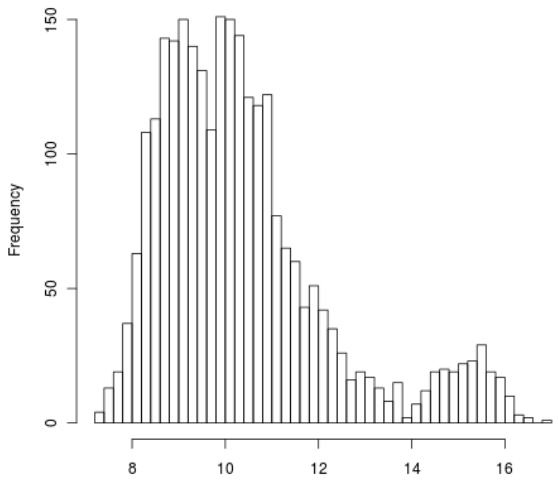

Figure 2.1: Histogram of the expression measurements of a specific nematode gene across $n = 2670$ conditions. From Martin-Magniette (pers. com.).

Genes involved in a given process. To exhibit which genes are involved in a given process (e.g. tumor growth or response to a stress), the expression levels of 'all' genes of the organism under study can be measured in a series of conditions (possibly with replicates) and a test of the null hypothesis $H_{0,i}$ = 'gene i is not involved in the process' is carried out.
For each gene, we get a test statistics T_i and a p-value Y_i. A latent status Z_i is associated with each gene and the same model as above can be used, taking e.g. $F(\gamma_k) = \mathrm{B}(\alpha_k, \beta_k)$, as $Y_i \in [0, 1]$ (Allison *et al.* (2002)). Figure 2.2 displays an example of the distribution of such p-values.
Note that, in this case, the distribution for the 'null' genes (i.e. genes for which $H_{0,i}$ is true) should be the uniform $\mathcal{U}_{[0,1]} = \mathrm{B}(1, 1)$. In this case, one of the emission distributions, say $F(\gamma_1)$, is known.

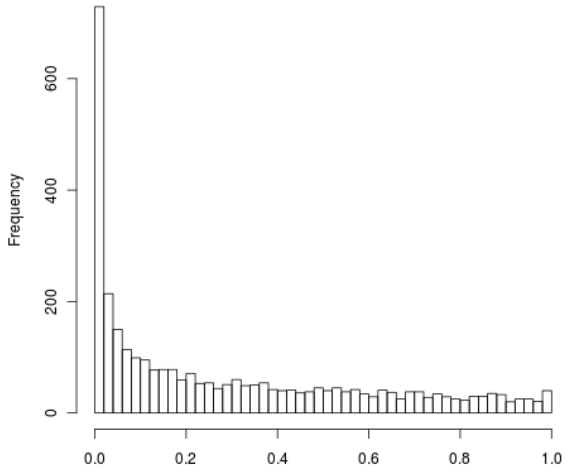

Figure 2.2: Distribution of the p-values associated with all human genes resulting from an analysis of variance model. Each p-value is associated to the null hypothesis stating that the considered gene has the same expression level in normal patients as in acute leukemia patients. From Hedenfalk *et al.* (2001).

2.1.2 Genetic structure of a population

Many efforts have been made in the last decade to better understand the genetic structure of populations. Most of them rely on the genotyping of large sets of individuals sampled in different places, environments or with different origins. In such experiments the genotype Y_{it} of a series of individuals $i \in [\![1, I]\!]$ at a series of locus $t \in [\![1, T]\!]$ is measured. One hopes to be able to distinguish sub-populations.

Model without 'admixture'. In this first simple model, each individual i is supposed to belong to one population, labeled Z_i (see Figure 2.3):

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

where $\pi = (\pi_1, \dots \pi_K)$ is the vector of the population proportions and γ_{kt} is the vector of the allelic frequencies at locus t in population k.

Note that the writing

$$(Y_{it})_{i,t}|Z_i \sim \mathcal{M}(1;\gamma_{Z_i t})$$

is equivalent to

$$(Y_{it})_{i,t}|(Z_i = k) \sim \mathcal{M}(1;\gamma_{kt}),$$

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

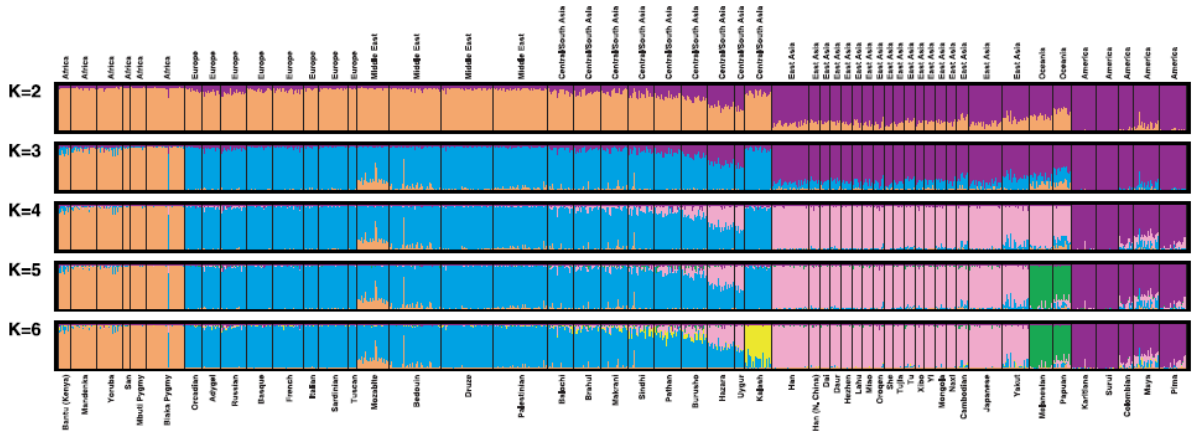

Figure 2.3: Population origine of series of human genomes with varying number of groups K. Each column corresponds to an individual. Colors represent the probability for the individual to belong to the color class given its genotype. From Rosenberg *et al.* (2002).

Model with admixture. In this second model, the genotype of each individual is supposed to come from a different population at each locus. Still, each individual has some preferential trends characterized by a latent variable Q_i (see Figure 2.4). This hidden variable can be interpreted as the position of individual i in the simplex of $\mathbb{R}^K$, the vertices of which correspond to fictitious individuals purely issued from each population:

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

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

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

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

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

Figure 2.4: Population origin of loci for 10 pairs of human chromosomes. The legend is similar to this of 2.3 except that the probability refers to the loci within the individual rather than to the whole individual. From Falush *et al.* (2003).

2.2 Model

The general model-based approach is based on the existence, for each individual i, of an unknown (or latent) label Z_i that can take a finite number of values among $[\![1, K]\!]$. The distribution of the observed variables Y_i depends on the value of this latent variable Z_i.

2.2.1 Model

The simple mixture model further assumes that the observations are independent with parametric distribution.

- The latent (Z_i) are iid with $P(Z_i = k) = \pi_k$;
- The observed (Y_i) are independent conditionally on the latent (Z_i);
- Conditionally on when $Z_i = k$, Y_i has a parametric distribution $F_k = F(\gamma_k)$ with probability distribution function (pdf) $f_k(\cdot) = f(\cdot; \gamma_k)$.

The purpose of the inference of a mixture model is to provide estimates of the parameters:

- π_k = proportion of the population k,
- γ_k = parameters of F_k.

All the parameters to be inferred are gathered into θ:

$$\theta := (\pi, \gamma) = ((\pi_k), (\gamma_k)).$$

In practice, mixture models are most often used with a classification purpose so the main aim is to infer the hidden status of each individual Z_i.

Remarks.

1. The elements $\pi_k = P(Z_i = k)$ of the distribution π are sometimes called *prior probabilities* although this denomination may be misleading in a non-Bayesian context. They are also often refereed to as the *proportions* of the mixture.
2. The distribution F_k is called the *emission* distribution in class k as it describes how observed data arising from class k are emitted. Respectively, f_k is called the emission pdf.

Definition 2.1 *An independent mixture model is defined as follows:*

$$\begin{array}{rrcl} (Z_i)_i \text{ iid}, & Z_i & \sim & \mathcal{M}(1; \pi), \\ (Y_i)_i \text{ indep. } |(Z_i), & Y_i|(Z_i = k) & \sim & F_k = F(\gamma_k), \end{array} \tag{2.1}$$

where $\pi = (\pi_1, \ldots, \pi_K)$. *We further denote* $f_k(\cdot) = f(\cdot; \gamma_k)$ *the pdf of distribution* $F(\gamma_k)$.

Remark. This model is equivalent to

$$\begin{aligned} p_\theta(Z) &= \prod_i \prod_k (\pi_k)^{Z_{ik}}, \\ p_\theta(Y|Z) &= \prod_i \prod_k f(Y_i, \gamma_k)^{Z_{ik}}, \end{aligned}$$

introducing the useful notation

$$Z_{ik} = \mathbb{I}\{Z_i = k\}.$$

Marginal distribution. The marginal distribution of the observation Y_i is the mixture distribution

$$g(y) = \sum_k \pi_k f(y; \gamma_k).$$

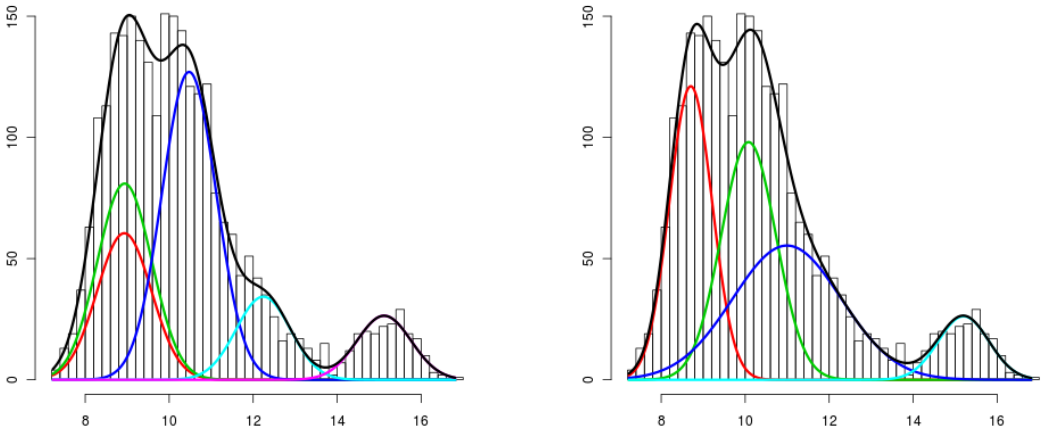

Figure 2.5: Expression level of a nematode gene across $n = 2670$ conditions: Gaussian mixture model with homogeneous variance (left) and heterogeneous variances (right). Same data as in Fig. 2.1.

Identifiability. Since the (Z_i) are not observed, the model is invariant for any permutation of the labels $[\![1, K]\!]$. Therefore, the mixture model with K classes has $K!$ equivalent definitions.

Number of parameters. The number of unknown parameters depends on both the dimension of the data and the number of groups. Due the constraint $\sum_k \pi_k = 1$, π involves only $K - 1$ independent parameters. As for the parameter γ, its dimension is typically proportional to the number of groups K. In the case where the F_k are univariate Poisson distributions with respective mean γ_k, γ has dimension K so the mixture model involves $2K - 1$ parameters. If the F_k are d-variate normal distributions (with respective mean vector μ_k and variance Σ_k), $(K - 1) + Kd + Kd(d + 1)/2 \simeq Kd^2/2$ parameters have to be estimated.

2.2.2 Dependency structure

All the dependencies involved in a mixture model can be encoded in the graphical model displayed in Figure 2.6. We refer to Appendix A.1 for a reminder on the definition of graphical models. From this figure we see that

- The (Z_i) are independent;
- the (Y_i) are independent conditionally to $Z = (Z_i)$;
- the couples $\{(Y_i, Z_i)\}_i$ are iid.

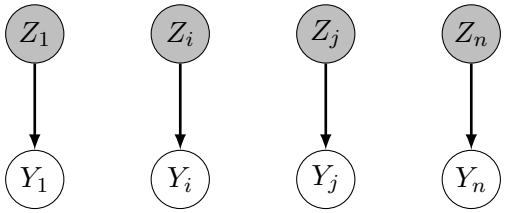

Figure 2.6: Graphical representation of a mixture model.

Remarks.

1. Note that the variables (Y_i, Y_j) are *not* independent conditionally on the event $Z_i = Z_j$.
2. Because the $\{(Y_i, Z_i)\}_i$ are independent, we have that
$$p_\theta(Z_i|Y) = p_\theta(Z_i|Y_i)$$
which means that the information about the classification of individual i is contained in the observation Y_i.

2.3 Inference

A general introduction to finite mixture models and their inference can be found in McLahan and Peel (2000). Several methods have been proposed for the inference of mixture models, but the most popular is definitely the maximum likelihood approach. The specificity of mixture models, shared with all latent variable models, is that the observed data $Y = (Y_i)$ can be seen as *incomplete*, as the latent variables $Z = (Z_i)$ are not observed. Such models are therefore often refereed to as *incomplete data models*.

2.3.1 Likelihoods

Definition 2.2 *The observed data log-likelihood (also called 'incomplete log-likelihood') is the marginal log-likelihood of the observed variables Y:*

$$\log p_\theta(Y).$$

The complete data log-likelihood is the joint log-likelihood of the observed Y and latent Z variables:

$$\log p_\theta(Y, Z).$$

Proposition 2.1 *For the mixture model* (2.1)*, the incomplete log-likelihood is*

$$\log p_\theta(Y) = \sum_i \log \left[\sum_k \pi_k f(Y_i; \gamma_k) \right],$$

and, denoting $Z_{ik} = \mathbb{I}\{Z_i = k\}$, the complete log-likelihood is

$$\log p_\theta(Y, Z) = \sum_i \sum_k Z_{ik} \left[\log \pi_k + \log f(Y_i; \gamma_k) \right].$$

Proof: The dependency structure described in Figure 2.6 ensures that

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

□

2.3.2 EM Algorithm

The EM algorithm was first proposed by Dempster *et al.* (1977) for a large class of incomplete data models, including mixture models. It is based on a decomposition of the incomplete data likelihood.

Proposition 2.2

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

Proof: It suffices to develop

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

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

Remarks.

1. The decomposition given in Proposition 2.2 is convenient as it makes a connexion between the observed likelihood $\log p_\theta(Y)$, which is often intractable, and the complete likelihood $\log p_\theta(Y,Z)$, which is generally more manageable.
2. The second term corresponds to the entropy of the latent variables Z given the observed Y: $H[p(Z|Y)] := -\mathbb{E}[\log p(Z|Y)|Y]$.

Proposition 2.2 suggests the following algorithm to get the MLE of θ

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

Algorithm 2.1 *Repeat until convergence:*

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

Maximization step: *update the estimate of θ as*

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

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

Proposition 2.3 (Dempster *et al.* (1977)) *The log-likelihood of the observed data $\log p_\theta(Y)$ increases at each step:*

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

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

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

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

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

and the proof is completed. □

E step. As mentioned in the introduction, the E step is straightforward for independent mixture models.

Proposition 2.4 *In a mixture model* (2.1)*, the hidden states Z_i are independent conditional on the observations:*

$$p_\theta(Z|Y) = \prod_i p_\theta(Z_i|Y_i)$$

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

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

Proof: The first result is a direct consequence of the second remark p.11. The second results follows the Bayes formula:

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

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

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

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

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

Remark. The conditional probability τ_{ik} is sometimes referred to as the *posterior probability* for observation i to belong to class k (as opposed to the *prior probability* π_k). Again this phrase is misleading in a non-Bayesian context and 'conditional probability' should be preferred.

M step. The actual estimation of the parameter θ is achieved at the M step. As indicated in Algo. 2.1, this step consists in the maximization of the the conditional expectation of the complete likelihood that appears in the decomposition of Proposition 2.2. We use Proposition 2.1 to get an explicit formula for this quantity

$$\begin{aligned}\mathbb{E}_\theta[\log p_\theta(Y,Z)|Y] &= \mathbb{E}_\theta\left[\sum_i\sum_k Z_{ik}[\log\pi_k + \log f(Y_k;\gamma_k)]|Y\right] \\ &= \sum_i\sum_k \mathbb{E}_\theta(Z_{ik}|Y_i)[\log\pi_k + \log f(Y_k;\gamma_k)] \\ &= \sum_i\sum_k \tau_{ik}[\log\pi_k + \log f(Y_k;\gamma_k)].\end{aligned}$$

This quantity has to be maximized with respect to $\theta = (\pi, \gamma)$, the τ_{ik} being fixed. The solution of this optimization problem has no general form as it strongly depends on the model at hand, especially on the complexity of the emission distributions. However, some general formula can be derived in the case of the exponential family, as we will see in Section 2.3.4. As for the proportions, one straightforwardly get $\pi_k^h = n^{-1}\sum_i \tau_{ik}^h$.

Entropy. The entropy term that appears in Proposition 2.2 and in the second remark p.12 can be calculated using the conditional independence of the Z_i given the data Y:

$$\begin{aligned}H[p_\theta(Z|Y)] &= \sum_i H[p_\theta(Z_i|Y_i)] \\ &= -\sum_i \mathbb{E}_\theta[\log P(Z_i = k|Y_i)|Y_i] \\ &= -\sum_i\sum_k \tau_{ik}\log\tau_{ik}.\end{aligned} \tag{2.2}$$

2.3.3 Variational interpretation

A more general view on the EM algorithm and its extensions is given by the following property. We first recall the definition of the Kullback-Leibler divergence between distribution q and p:

$$KL[q(Z)||p(Z)] = \mathbb{E}_q\{\log[q(Z)/p(Z)]\}$$

and recall that it is always positive[1] and is null iff $q = p$.

The following proposition gives a lower bound of the log-likelihood.

Proposition 2.5 *For any distribution q for Z, we have*

$$\log p(Y) \geq \mathbb{E}_q[\log p(Y,Z)] + H[q(Z)].$$

Proof: We write that

$$\begin{aligned}\log p(Y) &\geq \log p(Y) - KL[q(Z)||p(Z|Y)] \\ &= \log p(Y) - \mathbb{E}_q[\log q(Z) - \log p(Y,Z) + \log p(Y)] \\ &= \log p(Y) - \mathbb{E}_q[\log q(Z)] + \mathbb{E}_q[\log p(Y,Z)] - \log p(Y)\end{aligned}$$

as $\mathbb{E}_q[\log p(Y)] = \log p(Y)$ since q is a distribution for Z and the result follows. □

[1] as $\mathbb{E}_q \log(q/p) = -\mathbb{E}_q \log(p/q) \leq -\log \mathbb{E}_q(p/q) = -\log\int p = -\log(1) = 0$

Remark. The decomposition given by Proposition 2.2 is similar to the lower bound of Proposition 2.5 with an equality when taking $q(Z) = p(Z|Y)$. Furthermore, the E step of the EM algorithm can be viewed as the solution of the variational problem:

$$q^*(Z) = \arg\min_q KL[q(Z)||p(Z|Y)],$$

which, in absence of restriction for q is $q^*(Z) = p(Z|Y)$. From this point of view, the EM algorithm alternates the minimization of $KL[q(Z)||p(Z|Y)]$ wrt q (E step) and the maximization of the lower bound $\mathbb{E}_q[\log p_\theta(Y,Z)] + H[q(Z)]$ wrt θ (M step).

2.3.4 Case of the exponential family

Definition 2.3 *The distribution p belongs to exponential family with* canonical parameter θ *if*

$$p_\theta(y) = \exp[\theta^\intercal t(y) - a(y) - b(\theta)]$$

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

We recall two general properties that show connections between maximum likelihood estimates and moment estimates for this class of distribution. The proofs of both are given in Appendix A.2.

Proposition 2.6 $b'(\theta) = \mathbb{E}_\theta[t(Y)]$.

Proposition 2.7 *For an iid sample $(Y_1, \dots Y_n)$, the MLE $\widehat{\theta}$ of θ satisfies*

$$b'(\widehat{\theta}) = n^{-1}\sum_i t(Y_i) =: \bar{t}(Y).$$

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

Proposition 2.8 *If all emission distributions f_k belong to the exponential family with respective sufficient statistics t_k and normalizing functions a_k and b_k, the maximization in the M step results in the weighted moment estimates based on the expectation of the sufficient statistics, i.e. γ_k^{h+1} satisfies:*

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

where $U \sim f(\cdot, \gamma_k^{h+1})$, $\tau_{ik}^{h+1} = \mathbb{E}_{\theta^{h+1}}(Z_{ik}|Y_i)$, $N_k^{h+1} = \sum_i \tau_{ik}^{h+1}$ and $T_k^{h+1} = \sum_i \tau_{ik}^{h+1} t_k(Y_i)$.

Proof: The complete-likelihood is

$$\log p(Y,Z) = \sum_i \sum_k Z_{ik}[\log \pi_k + \log f_k(Y_i)] = \sum_i \sum_k Z_{ik}[\log \pi_k + \gamma_k^\intercal t_k(Y_i) - a_k(Y_i) - b_k(\gamma_k)]$$

so its conditional expectation is

$$\begin{aligned}\mathbb{E}[\log p(Y,Z)|Y] &= \mathbb{E}\left[\sum_i \sum_k Z_{ik}\left[\log \pi_k - b_k(\gamma_k)\right]|Y\right] + \mathbb{E}\left[\sum_i \sum_k Z_{ik}\left[\gamma_k^\intercal t_k(Y_i) - a_k(Y_i)\right]|Y\right] \\ &= \sum_k N_k[\log \pi_k - b_k(\gamma_k)] + \sum_k \gamma_k^\intercal T_k - \sum_i \tau_{ik} a_k(Y_i).\end{aligned}$$

The derivative with respect to γ_k is null iff $b_k'(\gamma_k) = T_k/N_k$ and the result follows from the general properties of the exponential family given in Propositions 2.6 and 2.7. □

Note that T_k^{h+1}/N_k^{h+1} is an empirical weighted moment of the Y_i so the estimate of γ_k resulting from Proposition 2.7 is a moment-type estimate. Also note that, depending on the form of $\mathbb{E}_{\gamma_k}[t_k(U)]$ as a function of γ_k, this estimate can have a close form or not. The popular cases listed below are cases where $\mathbb{E}_{\gamma_k}[t_k(U)]$ has a simple form.

Some popular models. As a consequence of this, we derive the estimates for a series of models:

- Multinomial mixture: $F_k = \mathcal{M}(1; \gamma_k)$, denoting $Y_{ia} = \mathbb{I}\{Y_i = a\}$:

$$\widehat{\gamma}_{ka} = N_k^{-1} \sum_i \tau_{ik} Y_{ia}.$$

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

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

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

$$\widehat{\gamma}_k = N_k^{-1} \sum_i \tau_{ik} Y_i.$$

2.3.5 Asymptotic variance and Fisher information

We remind that the asymptotic variance of the maximum likelihood estimate

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

is provided by the Fisher information matrix I by

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

where

$$S_\theta(Y) = \partial_\theta \log p_\theta(Y) \qquad \text{and} \qquad I_\theta = \mathbb{E}[S_\theta(Y) S_\theta(Y)^\intercal] = -\mathbb{E}\left[\partial^2_{\theta^2} \log p_\theta(Y)\right].$$

Louis (1982) provides a convenient way to compute the Hessian matrix

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

which only uses by-products of the EM algorithm.

Proposition 2.9 (Louis (1982))

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

The proof is given in Appendix A.3.1. This formula has two main interests:

- the first two terms involve the complete likelihood and can, most of the times, be easily computed;
- the last term is null when evaluated at $\widehat{\theta} = \arg\max \log p_\theta(Y)$ since $p'_\theta(Y)|_{\widehat{\theta}} = 0$.

Case of a Poisson mixture. Consider a mixture model (2.1) where $F(\gamma_k) = \mathcal{P}(\gamma_k)$. The complete log-likelihood is

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

where $\pi_K = 1 - \sum_{k<K} \pi_k$. The first derivatives are

$$\partial_{\pi_k} \log p_\theta(Y, Z) = \frac{\sum_i Z_{ik}}{\pi_k} - \frac{\sum_i Z_{iK}}{\pi_K} \qquad \text{and} \qquad \partial_{\gamma_k} \log p_\theta(Y, Z) = -\sum_i Z_{ik} + \frac{\sum_i Z_{ik} Y_i}{\gamma_k}$$

and the second derivatives:

$$\partial^2_{\pi_k^2} \log p_\theta(Y, Z) = -\frac{\sum_i Z_{ik}}{\pi_k^2} + \frac{\sum_i Z_{iK}}{\pi_K^2}, \qquad \partial^2_{\pi_k, \pi_\ell} \log p_\theta(Y, Z) = \frac{\sum_i Z_{iK}}{\pi_K^2}$$

and

$$\partial^2_{\gamma_k^2} \log p_\theta(Y, Z) = -\frac{\sum_i Z_{ik} Y_i}{\gamma_k^2}, \qquad \partial^2_{\gamma_k, \gamma_\ell} \log p_\theta(Y, Z) = 0.$$

The first term of Prop. 2.9 requires the calculation of the following moments, denoting here $\mathbb{E}^Y(\cdot) = \mathbb{E}(\cdot|Y)$:

$$\mathbb{E}^Y\left(\textstyle\sum_i Z_{ik}\right) \;=\; \textstyle\sum_i \tau_{ik} =: N_k, \qquad \mathbb{E}^Y\left(\textstyle\sum_i Z_{ik} Y_i\right) \;=\; \textstyle\sum_i \tau_{ik} Y_i =: S_k.$$

The second term requires these of

$$\begin{array}{rcl}
\mathbb{E}^Y\left[\left(\sum_i Z_{ik}\right)\left(\sum_i Z_{i\ell}\right)\right] &=& \mathbb{E}^Y\left(\sum_i Z_{ik} Z_{j\ell} + \sum_{i\neq j} Z_{ik} Z_{j\ell}\right) \\
&=& \sum_i \mathbb{E}^Y(Z_{ik} Z_{j\ell}) + \sum_{i\neq j} \mathbb{E}^Y(Z_{ik}) \mathbb{E}^Y(Z_{j\ell}) \\
&=& \sum_i \delta_{k\ell} \tau_{ik} + \sum_{i\neq j} \tau_{ik} \tau_{j\ell} \quad = \quad \delta_{k\ell} N_k + N_k N_\ell - \sum_i \tau_{ik} \tau_{i\ell}, \\
\mathbb{E}\left[\left(\sum_i Z_{ik} Y_i\right)\left(\sum_i Z_{i\ell}\right)\right] &=& \delta_{k\ell} S_k + S_k N_\ell - \sum_i Y_i \tau_{ik} \tau_{i\ell}, \\
\mathbb{E}^Y\left[\left(\sum_i Z_{ik} Y_i\right)\left(\sum_i Z_{i\ell} Y_i\right)\right] &=& \delta_{k\ell} Q_k + S_k S_\ell - \sum_i Y_i^2 \tau_{ik} \tau_{i\ell},
\end{array}$$

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

2.3.6 How many states?

The number of hidden states K is not known general. Because a model with $K-1$ classes is nested in a model with K classes, the likelihood increases as well and is therefore not a relevant criterion to estimate K. Also note that the dimension of the parameter θ increases with K.

The most common strategy to estimate K is based on penalized likelihood criteria. We define $\widehat{\theta}_K$ as the maximum likelihood estimate of θ for a model with K components:

$$\widehat{\theta}_K = \arg\max_{\theta \in \Theta_K} \log p_\theta(Y)$$

where Θ_K stands for the parameter space for a mixture model with K components. A penalized likelihood estimate of K is defined as

$$\widehat{K} = \arg\max_K \left(\log p_{\widehat{\theta}_K}(Y) - \text{pen}(K)\right).$$

BIC. The most commonly used criterion is the Bayesian information criterion (BIC, Schwarz (1978)), which is originally defined in a Bayesian setting with three levels of hierarchy:

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

The model selection problem is then rephrased in terms of conditional distribution of K given the observations:

$$p(K|Y) = \frac{p(Y,K)}{p(Y)} = \frac{p(K)\int p(Y|\theta,K)p(\theta|K)\,\mathrm{d}\theta}{p(Y)}.$$

Ideally, one would choose

$$\begin{aligned}\widehat{K} &= \arg\max_K\ p(K|Y) \ = \ \arg\max_K\ (\log p(K) + \log p(Y|K)) \\ &= \arg\max_K\ \left(\log p(K) + \log \int p(Y|\theta,K)p(\theta|K)\,\mathrm{d}\theta\right).\end{aligned}$$

The main difficulty is raised by the evaluation of the last integral for which a Laplace approximation is used to show the following proposition.

Proposition 2.10 *Under regularity conditions,*

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

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

A detailed proof of this result can be found in Lebarbier and Mary-Huard (2006), together with precise comparative study between BIC and another popular model selection criterion: AIC (Akaike (1974)).

As the term $\log p(K)$ remains fix when n grows large, it is neglected to define the BIC selection criterion, which is defined as follows.

Definition 2.4

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

Integrated Complete Likelihood (ICL). Using Proposition 2.2, the BIC criterion can be rewritten as

$$\log p_{\widehat{\theta}_K}(Y) - \frac{d_K}{2}\log n = \mathbb{E}_{\widehat{\theta}_K}\left[\log p_{\widehat{\theta}_K}(Y,Z)|Y\right] - \mathbb{E}_{\widehat{\theta}_K}\left[\log p_{\widehat{\theta}_K}(Z|Y)|Y\right] - \frac{d_K}{2}\log n.$$

Remind that $-\mathbb{E}_{\widehat{\theta}_K}\left[\log p_{\widehat{\theta}_K}(Z|Y)|Y\right]$ is an entropy term, which is small when observation are classified with reasonable confidence. Biernacki *et al.* (2000) propose to account for the classification uncertainty in the selection of K, by adding this term to the penalty, which results in the following criterion.

Definition 2.5

$$\widehat{K}_{ICL} = \arg\max_K \left(\mathbb{E}_{\widehat{\theta}_K}\left[\log p_{\widehat{\theta}_K}(Y,Z)|Y\right] - \frac{d_K}{2}\log n\right).$$

This criterion can be derived in a similar way as BIC, based on the conditional probability $p(K|Y,Z)$. In the original paper Biernacki *et al.* (2000), the authors use an estimate $\widehat{Z}$ of Z to compute the criterion. The form given in Definition 2.5 corresponds to the one given in McLahan and Peel (2000).

Illustration. The following figures illustrate the behavior of this two criteria for the nematode gene expression data (Ex. 2.1.1). We see that BIC favors the fit to the marginal distribution of the observation, whereas ICL favors the separation between the components.

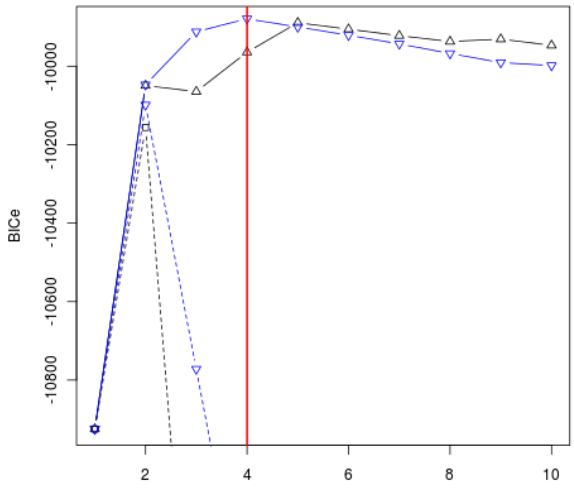

Figure 2.7: BIC and ICL criteria for the choice of a Gaussian mixture model for the nematode gene expression data as a function of the number of groups K. $\triangle$: equal variances, $\triangledown$: heterogeneous variances. Solid lines: BIC, dashed lines: ICL. (same data as in Fig. 2.1)

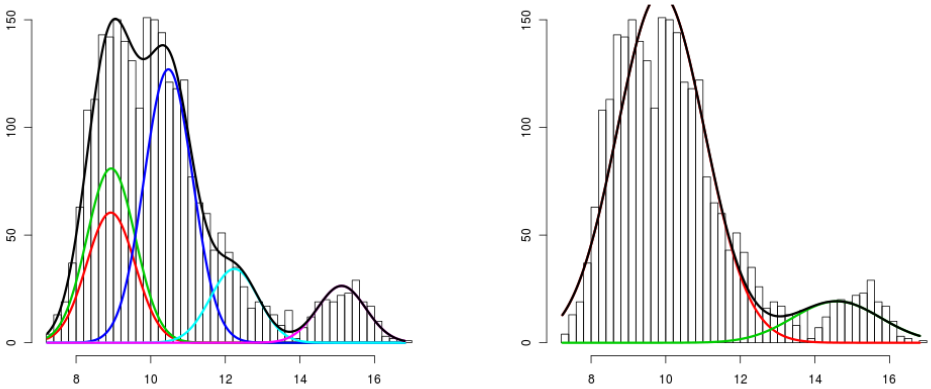

Figure 2.8: Comparison of the homogeneous variance mixture models selected with BIC (left) and ICL (right) for the nematode gene expression data. (same data as in Fig. 2.1)

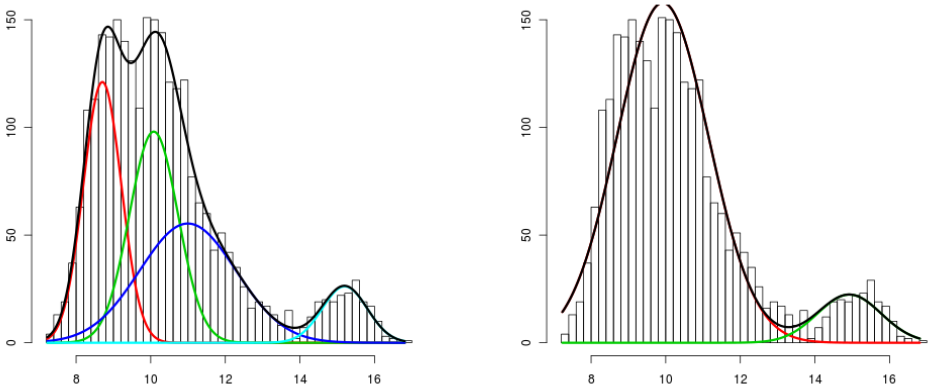

Figure 2.9: Comparison of the heterogeneous variance mixture models selected with BIC (left) and ICL (right) for the nematode gene expression data. (same data as in Fig. 2.1)

2.4 Classification

As mentioned in the introduction, classification is often the main aim when using a mixture model. Maximum likelihood inference provides estimates of the parameters. The EM algorithm also gives access, as a by product, to some information about the conditional distribution of the hidden classes Z conditional to the observed data Y but no formal classification of the observation into groups.

Soft classification. The classification of observations into groups is not always necessary (or relevant) and a soft classification is provided by the $\tau_{ik} = P(Z_i = k|Y)$. This probability gives a measure of the confidence with which an observation could be classified into a given group. The uncertainty of the classification can be summarized by the conditional entropy of Z_i, sometimes referred to as the *classification uncertainty* for observation i:

$$H[p_\theta(Z_i|Y)] = H[p_\theta(Z_i|Y_i)] = -\sum_k \tau_{ik} \log \tau_{ik}.$$

Note that the entropy of the whole conditional distribution of Z given Y is simply the sum of all the individual's uncertainties (see (2.2) p.14).

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

Definition 2.6 *The MAP classification rule is given by:*

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

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

$$\widehat{Z}_i = \arg\max_k \tau_{ik}$$

to the whole set of label Z. In the case of mixture, the two are equivalent:

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

since the Z_i are independent conditionally on Y.

3 Dependent hidden variables: Hidden Markov models and others

We now consider unsupervised classification problems in which the data are linearly organized. Such situations are faced in many applications such as time series analysis or signal processing, where observations are collected along time. Many genomic applications also fit this framework as measurements are collected at places (*loci*) located along the genome.

In such applications, it does not seem natural anymore to assume that the hidden status are independent, but rather to assume that they depend on each other. The Markovian dependency structure is one of the most simple to be considered and a strong attention has been paid to it for several decades now, resulting in hidden Markov models (HMMs).

3.1 Examples

3.1.1 Copy number variation

Genome rearrangements such as losses or amplifications of large genomic regions are associated with many disorders, including cancers or mental retardation. Microarrays can be used for many purposes and allow to observe such events. Comparative genomic hybridization (CGH) arrays provide, for a series of probes $t = 1, \dots n$ located along the genome, a fluorescence measurement Y_t that varies according to the relative number of copies of DNA at the position between a test (e.g. tumor) and a reference (normal) sample.

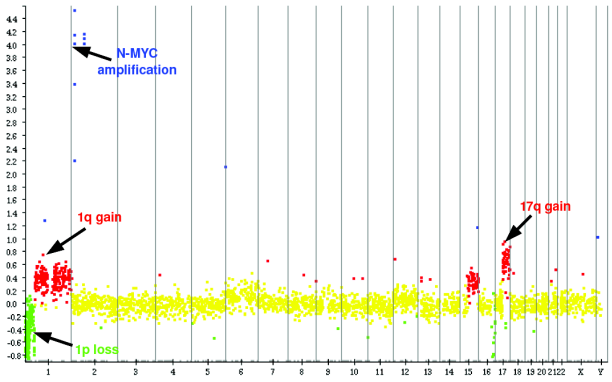

Figure 3.1: CGHarray signal. Vertical blocks refer to the 24 chromosomes (22 + X + Y). The color code results from the analysis: yellow = normal (= 2 copies), green = loss (< 2 copies), red = amplification (> 2 copies), blue = massive amplification (≫ 2 copies). The annotation results from the cytogenetic data presented in Fig. 3.2. From Hupé (2008).

Figure 3.1 provides an example of such data and Figure 3.2 provides the corresponding cytogenetic picture that gives a more precise picture (but at a much higher cost) of the alterations and translocations that actually occurred in the sample.

Model. A natural way to analyze such data is assume that the measure Y_t have a mixture distribution,

$$Y_t | Z_t = k \sim F(\gamma_k)$$

depending of the hidden status Z_t of the probe, e.g. $Z_t \in \{\text{loss}, \text{normal}, \text{gain}\}$ (colored in red, yellow and green respectively in Figure 3.1). As the probes are linearly organized along the

genome, it is not realistic to assume that the Z_t are independent (neighbor probes are likely to share the same status), so we rather suppose that $Z = (Z_t)$ is an homogeneous Markov chain

$$Z_t \sim MC(\nu, \pi)$$

where ν stands for the initial distribution over $[\![1, K]\!]$ and π for the $K \times K$ transition matrix:

$$\nu_k = P(Z_1 = k), \qquad \pi_{k,\ell} = P(Z_{t+1} = \ell | Z_t = k).$$

An HMM model including several refinements was proposed by Fridlyand *et al.* (2004) for the analysis of CGH arrays.

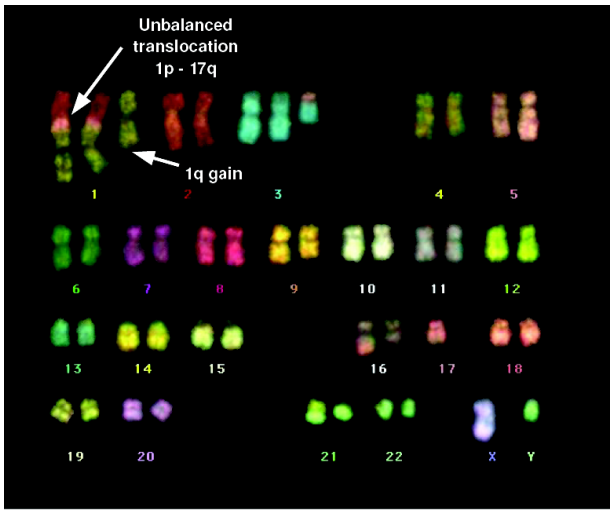

Figure 3.2: Karyotype of tumor cell with annotated translocations to which sole CGH experiments do not give access (same data as Figure 3.1) From Hupé (2008).

3.1.2 Genetic structure of a population with admixture

In Section 2.1.2, one was interested in classifying regions of an individual's chromosome based on its genotype at a series of loci $t = 1, \ldots, T$. In the Chapter 2, the unknown population origins at each loci Z_t were supposed to be independent from one loci to the next. The dependency between neighbor loci can be accounted for using the following model:

$$\begin{aligned} (Z_i) \quad &\text{iid} \quad Z_i = (Z_{i1}, \ldots, Z_{iT}), \\ (Z_{it})_t \quad &\sim \quad MC(\nu, \pi), \\ (Y_{it})_{it} \text{ indep.} \,|\, (Z_{it}) \quad &\sim \quad F(\gamma_{Z_{it}}), \end{aligned}$$

with multinomial emission distribution $F(\gamma_k) = \mathcal{M}(1; \gamma_k)$.

3.1.3 Sequence evolution

Phylogeny aims at reconstructing the evolutionary history of species, based on the observation of their genome. For a set I species, the observed data consist in the sequences $Y_i = (Y_{i1}, \ldots, Y_{iT})$ at T positions (after sequence alignment), where each Y_{it} is one of the $K =$ four nucleotides `a`, `c`, `g` or `t`. Note that all observed sequences are contemporary, i.e. observable today, and that we miss ancestral sequences from which they derive. Maximum likelihood approaches have been introduced in this field by Felsenstein (1981), who first showed that conditional independences can be used to compute the likelihoods efficiently.

We further assume that a phylogeny is available, providing the topology of the evolutionary tree. The inference of this tree will not be discussed in this course.

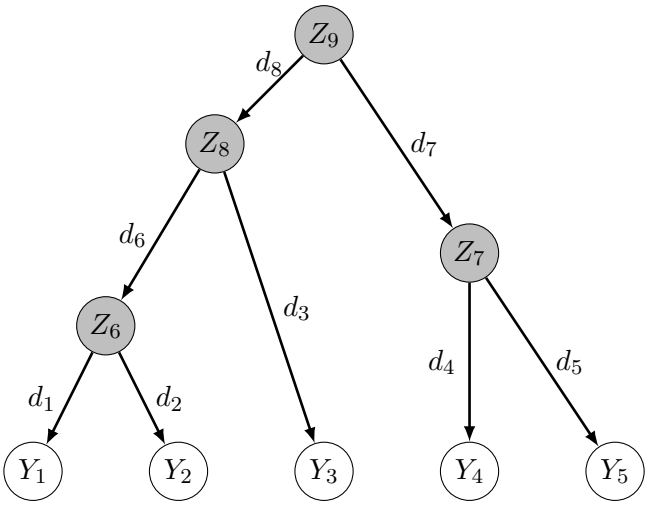

Figure 3.3: Sequence evolution model. Y_i sequences are contemporary (and observed), Z_j sequences are ancestral sequences (not observed). Branch lengths d_i refer to the evolutionary time from one sequence to the next.

The simplest model assumes that each nucleotide at each locus evolves independently from each other, according to a continuous time Markov process with infinitesimal mutation rates $\rho_{k\ell} = \rho_{k\to\ell}$, with $k, \ell \in \{\mathtt{a}, \mathtt{c}, \mathtt{g}, \mathtt{t}\}$.

The aim is then to estimate the mutation rates and the branch lengths of the phylogenetic tree. The parameters of the models is then made of the branch lengths $d = (d_i)$ and the mutation rates $\rho = (\rho_{k\ell})$:

$$\theta = (d, \rho)$$

This problem has some connexions with unsupervised classification as it involves hidden variables, which are the sequences of the species located at the interior nodes of the phylogenetic tree. These missing variables correspond to ancestral sequences that existed in the past Z_i. Recovering these ancestral sequences amounts at predicting their nucleotides at each locus t, which is an unsupervised classification problem.

3.2 Hidden Markov model

3.2.1 Model

Definition 3.1 *The general hidden Markov chain model is defined as follows:*

$$\begin{array}{rrcl} & (Z_t)_t & \sim & MC(\nu, \pi), \\ (Y_t)_t \text{ indep. } |(Z_t), & Y_i|(Z_i = k) & \sim & F_k = F(\gamma_k), \end{array} \tag{3.1}$$

The Markov chain $MC(\nu, \pi)$ is defined over the state space $[\![1, K]\!]$, K being the number of hidden states.

The parameters of this model are gathered into

$$\theta = (\nu, \pi, \gamma).$$

Marginal distribution. We denote ν_t the distribution of the hidden state at time t:

$$\nu_t = (\nu_{t1}, \ldots, \nu_{tK}), \qquad \nu_{tk} = P(Z_t = k).$$

As (Z_t) is an homogeneous Markov chain, we have

$$\nu_t = \nu^\intercal \pi^{t-1}$$

and the marginal distribution of observation Y_t is a mixture with proportion ν_t:

$$Y_t \sim \sum_k \nu_{tk} f(\cdot; \gamma_k).$$

(Z_t) is often further assumed to be a stationary Markov chain, meaning that

$$\nu = \nu^\intercal \pi.$$

In this case $\nu_t = \nu$ at all t, so the marginal distribution of Y_t remains the same mixture along time:

$$Y_t \sim \sum_k \nu_k f(\cdot; \gamma_k).$$

3.2.2 Dependency structure

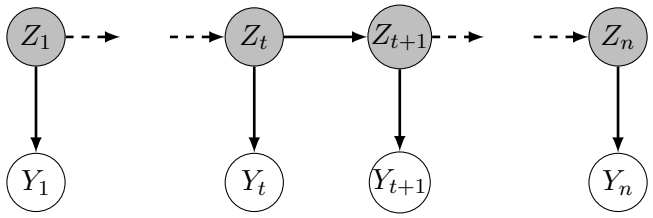

Figure 3.4: Graphical representation of an hidden Markov model.

Graphical model. We can derive the following various properties from the topology of the graph. Denoting $Z_s^t = (Z_s, \ldots Z_t)$ (for $s \leq t$ and the same for Y_s^t)n we have that

(a) all paths from Y_1^t to Z_{t+1} go through Z_1^t, meaning that Z_{t+1} is independent from Y_1^t conditionally on Z_1^t;

(b) 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);

(c) all paths from Y_1^t to Y_1^{t+1} go through Z_{t+1} meaning that Y_1^{t+1} is independent from Y_1^t to conditionally on Z_{t+1} (and the same holds with Z_t).

As a consequence, the conditional distribution of the hidden states (Z_t) conditional on the observed data $Y = Y_1^n$ is still a Markov chain. Indeed,

$$\begin{aligned} p(Z_{t+1}|Z_1^t, Y_1^n) &= p(Z_{t+1}|Z_1^t, Y_{t+1}^n) \quad (a) \\ &= p(Z_{t+1}|Z_t, Y_{t+1}^n) \quad (b) \\ &= p(Z_{t+1}|Z_t, Y_1^n) \quad (c) \end{aligned}$$

Similar arguments will be used in the 'Forward' and 'Backward' recursions given in Proposition 3.1 p.25.

3.3 Inference

For a general presentation of hidden Markov models and their inference, one may refer to Cappé *et al.* (2005). In this section, we only consider EM-like algorithms.

3.3.1 Likelihoods

The marginal (or 'observed') log-likelihood is

$$\begin{aligned}\log p_\theta(Y) &= \log\left[\sum_Z p_\theta(Z)p_\theta(Y|Z)\right] \\ &= \log\left[\sum_Z \left(\prod_k \nu_k^{Z_{1k}} \prod_{t\geq 2}\prod_{k,\ell} \pi_{k\ell}^{Z_{t-1,k}Z_{t,\ell}}\right)\left(\prod_{t,k} f(Y_t;\gamma_k)^{Z_{tk}}\right)\right].\end{aligned}$$

The complete log-likelihood corresponds to one term of the sum, that is

$$\begin{aligned}\log p_\theta(Y,Z) &= \log\left[p_\theta(Z)p_\theta(Y|Z)\right] \\ &= \sum_k Z_{1k}\log\nu_k + \sum_{t\geq 2}\sum_{k,\ell} Z_{t-1,k}Z_{t,\ell}\log\pi_{k\ell} + \sum_{t,k} Z_{tk}\log f(Y_t;\gamma_k).\end{aligned}$$

Remark. Note that the form of the complete likelihood derives from the graphical model given in Figure 3.4 where Z_1 is the root (with no parent), each Z_t ($t \geq 2$) has parent Z_{t-1} and each Y_t has parent Z_t. Using Definition A.1 (p.58) of directed graphical models, we get

$$p_\theta(Y,Z) = p_\theta(Z_1)\left(\prod_{t\geq 2} p_\theta(Z_t|Z_{t-1})\right)\left(\prod_t p_\theta(Y_t|Z_t)\right).$$

3.3.2 EM: Forward-Backward algorithm

The likelihood decomposition given in Proposition 2.2 p.12 and the resulting EM Algorithm 2.1 p.12, still hold. We remind that the M step consist in the maximization of

$$\mathbb{E}[\log p_\theta(Y,Z)|Y] = \sum_k \tau_{1k}\log\nu_k + \sum_{t\geq 2}\sum_{k,\ell}\eta_{tk\ell}\log\pi_{k\ell} + \sum_{t,k}\tau_{tk}\log f(Y_i;\gamma_k)$$

where

$$\tau_{tk} = \mathbb{E}[Z_{tk}|Y] = P(Z_t = k|Y), \qquad \eta_{tk\ell} = \mathbb{E}[Z_{t-1,k}Z_{t,\ell}|Y] = P(Z_{t-1}=k, Z_t=\ell|Y).$$

Remark. Due to the dependency structure, τ_{tk} is not equal to $P(Z_t = k|Y_t)$, as opposed to the mixture model. More generally, the conditional distribution $p(Z|Y)$ does not factorize over t any more.

Proposition 3.1 *The conditional probabilities τ_{tk} and $\eta_{tk\ell}$ can be computed via the two following recursions.*

Forward: *denoting $F_{tk} = P_\theta(Z_t = k|Y_1^t)$, with $Y_1^t = (Y_1,\ldots,Y_t)$, compute*

$$\begin{aligned}F_{1\ell} &\propto \nu_\ell f_\ell(Y_1), \\ F_{t\ell} &\propto f_\ell(Y_t)\sum_k F_{t-1,k}\pi_{k\ell}\end{aligned}$$

such that, for all t : $\sum_k F_{t\ell} = 1$.

Backward: *starting with $\tau_{nk} = F_{nk}$; compute*

$$G_{t+1,\ell} = \sum_k \pi_{k\ell} F_{tk}, \qquad \eta_{tk\ell} = \pi_{k\ell} \frac{\tau_{t+1,\ell}}{G_{t+1,\ell}} F_{tk}, \qquad \tau_{tk} = \sum_\ell \eta_{tk\ell}.$$

Proof:

Forward: The first step relies on Bayes formula:

$$F_{1\ell} = P(Z_1 = \ell | Y_1) = p(Y_1 | Z_1 = \ell) P(Z_1 = \ell) / p(Y_1) \propto \nu_\ell f_\ell(Y_1)$$

and the recursion follows as

$$\begin{aligned}
F_{t\ell} &= P(Z_t = \ell | Y_1^t) = \sum_k P(Z_{t-1} = k, Z_t = \ell | Y_1^t) \\
&= \sum_k \frac{p(Z_t = \ell, Z_{t-1} = k, Y_1^t)}{p(Y_1^t)} \\
&= \sum_k \frac{p(Y_1^{t-1}) P(Z_{t-1} = k | Y_1^{t-1}) P(Z_t = \ell | Z_{t-1} = k) p(Y_t | Z_t = \ell)}{p(Y_1^t)} \\
&\quad \text{(using conditional independences, from the past to present } t) \\
&= \frac{p(Y_1^{t-1})}{p(Y_1^t)} f_\ell(Y_t) \sum_k \pi_{k,\ell} F_{t-1,k}.
\end{aligned}$$

Note that the normalizing coefficient is $p(Y_1^t)/p(Y_1^{t-1}) = p(Y_t | Y_1^{t-1})$.

Backward: The initialization is given by the last step of the forward recursion:

$$\tau_{nk} = P(Z_n = k | Y) = P(Z_1 = k | Y_1^n) = F_{nk}$$

and the recursion follows as

$$\begin{aligned}
\tau_{tk} &= P(Z_t = k | Y_1^n) = \sum_\ell \underbrace{P(Z_t = k, Z_{t+1} = \ell | Y_1^n)}_{\eta_{tk\ell}} \\
&= \sum_\ell \frac{P(Z_t = k, Z_{t+1} = \ell, Y_1^n)}{p(Y_1^n)} \\
&= \sum_\ell \frac{p(Y_1^t) P(Z_t = k | Y_1^t) P(Z_{t+1} = \ell | Z_t = k) p(Y_{t+1}^n | Z_{t+1} = \ell)}{p(Y_1^n)} \\
&= \sum_\ell P(Z_t = k | Y_1^t) P(Z_{t+1} = \ell | Z_t = k) \frac{p(Y_1^t) p(Y_{t+1}^n | Z_{t+1} = \ell)}{p(Y_1^n)} \\
&= F_{tk} \sum_\ell \pi_{k\ell} \frac{p(Y_1^t) p(Y_{t+1}^n | Z_{t+1} = \ell)}{p(Y_1^n)}
\end{aligned}$$

and

$$\begin{aligned}
\frac{p(Y_1^t) p(Y_{t+1}^n | Z_{t+1} = \ell)}{p(Y_1^n)} &= \frac{p(Y_1^t) p(Y_{t+1}^n | Z_{t+1} = \ell)}{p(Y_1^n)} \frac{p(Y_1^t | Z_{t+1} = \ell)}{p(Y_1^t | Z_{t+1} = \ell)} \\
&= \frac{p(Y_1^t) p(Y_1^n | Z_{t+1} = \ell)}{p(Y_1^n) p(Y_1^t | Z_{t+1} = \ell)} = \frac{P(Z_{t+1} = \ell | Y_1^n)}{P(Z_{t+1} = \ell | Y_1^t)} \\
&\quad \text{(inverting the conditioning: } P(A|B)/P(A) = P(B|A)/P(B)) \\
&= \frac{\tau_{t+1,\ell}}{P(Z_{t+1} = \ell | Y_1^t)}
\end{aligned}$$

where $P(Z_{t+1} = \ell | Y_1^t) = \sum_k P(Z_{t+1} = \ell, Z_t = k | Y_1^t) = \sum_k F_{tk} \pi_{k\ell} =: G_{t+1,\ell}$.

□

Remarks.

1. The computational complexity of this double recursion is $O(Kn^2)$.
2. The normalization constant $p(Y_t|Y_1^{t-1})$ of the forward step can be stored to compute the log-likelihood as
$$\log p(Y) = \log p(Y_1) + \sum_{t\geq 2} \log p(Y_t|Y_1^{t-1}).$$
3. The Forward formula states that $F_{t\ell} = P(Z_t = \ell|Y_1^t)$ only depends on Y_t and on $F_{t-1,k}$, which means that, conditional on Y_1^t and Z_{t-1}, Z_t is independent from Z_1^{t-2}. $(Z_t|Y_1^t)$ is therefore an (heterogeneous) Markov chain. The same formula further provides the transition probabilities of this Markov chain:
$$P_\theta(Z_t = \ell|Y_1^t, Z_{t-1} = k) = \frac{\pi_{k\ell} f_\ell(Y_t)}{\sum_j \pi_{kj} f_j(Y_t)}.$$
Conditional on Y_1^t, the transitions $\pi_{k\ell}$ are biased according to the likelihood of the data under the arrival state $f_\ell(Y_t)$.

3.4 Classification

A classification at each position t can be defined based on the MAP rule (see Definition 2.6, p.20), applied to the marginal distribution of each label given the data:

$$\widehat{Z}_t = \arg\max_k P(Z_t = k|Y) = \arg\max_k \tau_{tk}.$$

3.4.1 Joint MAP: Viterbi algorithm

As mentioned in Section 2.4, when the labels are not independent, the marginal MAP does not retrieve the joint MAP. In many application, one is interested in the joint MAP, as it corresponds to the most probable hidden path given the observations:

$$\widehat{Z} = \arg\max_z P(Z = z|Y).$$

Proposition 3.2 *The most probable hidden path given the data is given by the following forward-backward recursion:*

Forward: $V_{1k} = \nu_k f_k(Y_1)$ *and for* $t \geq 2$*:*

$$\begin{aligned} V_{t\ell} &= \max_k V_{t-1,k}\pi_{k\ell} f_\ell(Y_t), \\ S_{t-1}(\ell) &= \arg\max_k V_{t-1,k}\pi_{k\ell} f_\ell(Y_t). \end{aligned}$$

Backward: $\widehat{Z}_n = \arg\max_k V_{nk}$ *and for* $t < n$*:*

$$\widehat{Z}_t = S_t(\widehat{Z}_{t+1}).$$

Proof: First note that

$$\arg\max_z P(Z = z|Y) = \arg\max_z p(Z = z, Y)$$

The forward recursion consists in a succession of optimal choices as for the hidden label at the preceding times, so that

$$V_{tk} = \max_{z_1^{t-1}} \; p(Z_1^{t-1} = z_1^{t-1}, z_t = k, Y_1^t)$$

and, finally,

$$\max_k V_{nk} = \max_z \; p(Z = z, Y).$$

The backward recursion traces back the succession of the optimal choices and retrieves the optimal path. □
The rational (for $n = 4$) behind this algorithm is that, for a function of the form[2]

$$F(z_1^4) = f_1(z_1) + f_2(z_1, z_2) + f_3(z_2, z_3) + f_4(z_3, z_4),$$

we have the decomposition

$$\begin{aligned}
\max_{z_1^4} F(z_1^4) &= \max_{z_4}\left[\max_{z_3}\left(\max_{z_2}\left\{\max_{z_1}\left[f_1(z_1) + f_2(z_1, z_2)\right] + f_3(z_2, z_3)\right\} + f_4(z_3, z_4)\right)\right] \\
&= \max_{z_4}\left[\max_{z_3}\left(\max_{z_2}\left\{F_1^2(z_2) + f_3(z_2, z_3)\right\} + f_4(z_3, z_4)\right)\right] \\
&\qquad \text{where} \quad F_1^2(z_2) = \max_{z_1} f_1(z_1) + f_2(z_1, z_2) \\
&= \max_{z_4}\left[\max_{z_3}\left(F_1^3(z_3) + f_4(z_3, z_4)\right)\right] \\
&\qquad \text{where} \quad F_1^3(z_3) = \max_{z_2} F_1^2(z_2) + f_3(z_2, z_3) \\
&= \max_{z_4}\left[F_1^4(z_4)\right] \\
&\qquad \text{where} \quad F_1^4(z_4) = \max_{z_3} F_1^3(z_3) + f_4(z_3, z_4)
\end{aligned}$$

so both the maximal value of F and the optimal solution $\widehat{z}_1^4$ are obtained by storing the $F_1^t(z_t)$ and the $\widehat{z}_{t-1}(z_t) = \arg\max_{z_{t-1}} F_1^{t-1}(z_{t-1}) + f(z_{t-1}, z_t)$.

Remark. The calculation of the Viterbi path sometimes raises numerical issues due the addition of a large number of small terms. It is therefore high recommended to make all calculation in a log scale, that is

$$\begin{aligned}
\log V_{t\ell} &= \max_k \; \left(\log V_{t-1,k} + \log \pi_{k\ell} + \log f_\ell(Y_1)\right), \\
S_{t-1}(\ell) &= \arg\max_k \; \left(\log V_{t-1,k} + \log \pi_{k\ell} + \log f_\ell(Y_1)\right).
\end{aligned}$$

Illustration. An example of the difference between the marginal and the joint MAP classification rule is given in Figure 3.5 where an isolated point is classified as green by the marginal MAP rule, whereas it is classified as red (because of tis neighbors) by the Viterbi algorithm.

[2]That is to take $f_1(Z_1) = \log\left(\nu_{z_1} f_{z_1}(Y_1)\right)$, $f_t(z_{t-1}, z_t)) = \log\left(\pi_{z_{t-1},z_t} f_{z_t}(Y_t)\right)$ and $F(z_1^4) = \log p(z_1^4, Y_1^4)$

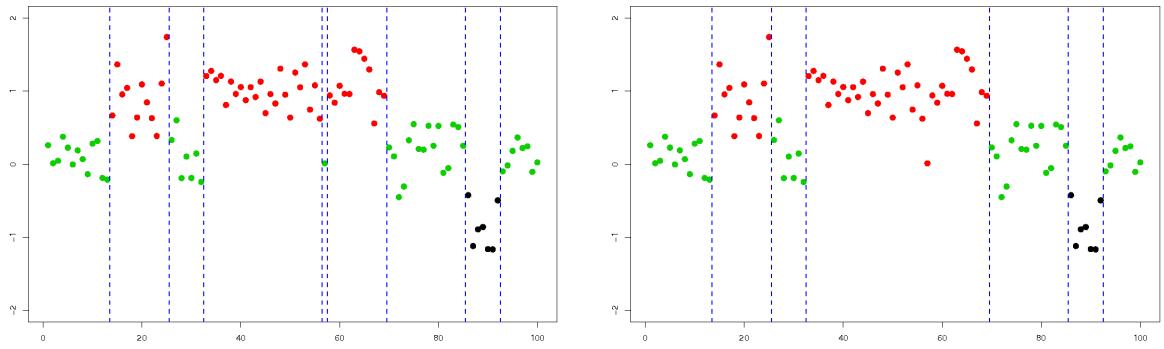

Figure 3.5: Comparison of the marginal MAP (left) and the joint MAP (Viterbi) classification rule on a simulated example.

3.4.2 Posterior entropy

As seen for mixture models, the entropy provides some insight about the certainty of the classification. When considering the whole hidden path, this amounts at computing (dropping the index θ for the sake of clarity)

$$H[p(Z|Y)] = -\mathbb{E}[\log p(Z|Y)|Y]$$

Because Z is an (heterogeneous) Markov conditionally on Y, we have that

$$H[p(Z|Y)] = -\mathbb{E}\left[\log p(Z_1|Y) + \sum_{t=2}^{n} \log p(Z_t|Z_{t-1}, Y)|Y\right]$$

The expectation of each term of the sum have to be taken wrt to Z_1 and to (Z_{t-1}, Z_t), respectively, so we have

$$\mathbb{E}[\log p(Z_1|Y)] = \sum_k P(Z_1 = k|Y) \log P(Z_1 = k|Y) = \sum_k \tau_{1k} \log \tau_{1k}$$

and, using $p(Z_t|Z_{t-1}, Y) = p(Z_t, Z_{t-1}|Y)/p(Z_{t-1}|Y)$,

$$\begin{aligned}\mathbb{E}[\log p(Z_t|Z_{t-1}Y)|Y] &= \sum_{k,\ell} P(Z_{t-1} = k, Z_t = \ell|Y) \log P(Z_t = \ell|Z_{t-1} = k, Y) \\ &= \sum_{k,\ell} \eta_{tk\ell}(\log \eta_{tk\ell} - \log \tau_{t-1,k}).\end{aligned}$$

So the conditional entropy can be computed as a by product of the backward step:

$$H[p(Z|Y)] = -\sum_k \tau_{1k} \log \tau_{1k} - \sum_{t=2}^{n} \sum_{k,\ell} \eta_{tk\ell}(\log \eta_{tk\ell} - \log \tau_{t-1,k}).$$

3.5 Some extensions

3.5.1 Connexion with the Kalman filter

The so-called Kalman filter is widely used in signal processing to retrieve an original signal (Z_t) from a noisy signal (Y_t). The model is the following

$$Y_t = Z_t\beta + F_t, \qquad Z_t = Z_{t-1}\pi + E_t, \qquad Z_1 \sim \mathcal{N}(0, 1)$$

where $E = (E_t)$ and $F = (F_t)$ are independent Gaussian white noises with respective variances $\mathbb{V}(E_t) = 1-\pi^2$ (without loss of generality) and $\mathbb{V}(F_t) = \sigma^2$. Note that the process Z is stationary with zero mean and unit variance. The parameters of this model are π and $\gamma = (\beta, \sigma^2)$.

The complete log-likelihood is then

$$\begin{aligned}\log p_\theta(Y, Z) &= \log p_\theta(Z) + \log p_\theta(Y|Z) \\ &= \log p_\theta(Z_1) + \sum_{t\geq 2} \log p_\theta(Z_t|Z_{t-1}) + \sum_t \log p_\theta(Y_t|Z_t)\end{aligned}$$

which only involves linear and quadratic functions of the Gaussian rv's Z_t and Y_t. So for the E step we only need the conditional mean and variance of the Z_t's, which can be derived using standard results on Gaussian vectors. The parameter estimation at the M step results in (weighted) linear regression estimates (see Ghahramani and Hinton (1996)).

3.5.2 Maximum likelihood inference for sequence evolution

We now consider the sequence evolution problem introduced in 3.1.3.

Transition probabilities. According to the mutation model defined above, if sequence $Y' = (Y'_t)$ derives from sequence $Y = (Y_t)$ after a duration d, then the transition probability from one to the other is given by

$$P(Y'_t = \ell | Y_t = k) = [\exp(d\rho)]_{k\ell} =: \pi_{k\ell}(d)$$

where $[\exp(s\rho)]$ stands for the matrix exponential of $s\rho$. Remind that, the nucleotides at different positions t are supposed to evolve independently.

For identifiability reasons, such mutation models are supposed to be *time reversible*, meaning that the rate matrix satisfies

$$\nu_k \rho_{k\ell} = \nu_\ell \rho_{\ell k}$$

where ν denotes the stationary distribution of the Markov chain. This assumption means that the evolutionary process behaves similarly forward and backward in time.

Complete likelihood. If the case of the phylogenetic tree from Figure 3.3, the complete likelihood can be written using the Definition A.1 of directed graphical models as

$$p_\theta(Y, Z) = p_\theta(Z_{2I-1}) \times \prod_{j=I+1}^{2I-2} p_\theta(Z_j | Z_{\mathrm{par}(j)}) \times \prod_{i=1}^{I} p_\theta(Y_i | Z_{\mathrm{par}(i)})$$

where $\mathrm{par}(i)$ denotes the parent node of node i. An important property of such model is that the corresponding graph is a tree, so each variable has only one parent. Denoting d_i the duration elapsed between sequences $Y_{\mathrm{par}(i)}$ and Y_i, we have $p_\theta(Y_{it} = \ell \,|\, Y_{\mathrm{par}(i),t} = k) = \pi_{k,\ell}(d_i)$ so we have, thank to site independence, for $x = (x_t)$ and $y = (y_t)$:

$$p_\theta(Y_i = y \,|\, Y_{\mathrm{par}(i)} = x) = \prod_t \pi_{x_t, y_t}(d_i) =: \pi_{x,y}(d_i).$$

E step. As for the E step, the conditional distribution of the unobserved sequences can be computed via the *upward-downward* recursions (see Durand *et al.* (2004) for a general presentation and Lartillot (2014) for genomic applications). As for the HMM, these two recursions

relies on conditional independences that are depicted in the graphical model given in Figure 3.3. Define:

$$\begin{aligned} L_{jz} &= p_\theta(Y_{\mathrm{sub}(j)}|Z_j = z), \\ F_{jz} &= p_\theta(Z_j = z \,|\, Y_{\mathrm{sub}(j)}) \quad \propto \quad \nu_z L_{jz}, \end{aligned}$$

where $\mathrm{sub}(j)$ denotes the set of indices of the observed sequences located downward Z_j in the tree (excluding Z_j itself). We now use the fact that the graphical model is a binary tree so each inside node j has only two offsprings, arbitrarily called 'left(j)' and 'right(j)'. We can then get the following recursion:

$$\begin{aligned} L_{jz} &= \left[\sum_x p_\theta(Z_{\mathrm{left}(j)} = x \,|\, Z_j = z) L_{\mathrm{left}(j),x}\right] \left[\sum_y p_\theta(Z_{\mathrm{right}(j)} = y \,|\, Z_j = z) L_{\mathrm{right}(j),y}\right] \\ &= \left[\sum_x \pi_{z,x}(d_{\mathrm{left}(j)}) L_{\mathrm{left}(j),x}\right] \left[\sum_y \pi_{z,y}(d_{\mathrm{right}(j)}) L_{\mathrm{right}(j),y}\right]. \end{aligned}$$

The recursion is initialized with observed nodes i as

$$L_{iy} = \mathbb{I}\{Y_i = y\}, \quad i \in [\![1, I]\!].$$

For Figure 3.3, we get

$$\begin{aligned} L_{6z} &= \left[\sum_x \pi_{zx}(d_1) L_{1x}\right] \left[\sum_y \pi_{zy}(d_2) L_{2y}\right], \\ L_{7z} &= \left[\sum_x \pi_{zx}(d_4) L_{4x}\right] \left[\sum_y \pi_{zy}(d_5) L_{5y}\right], \\ L_{8z} &= \left[\sum_x \pi_{zx}(d_6) L_{6x}\right] \left[\sum_y \pi_{zy}(d_3) L_{3y}\right], \\ L_{9z} &= \left[\sum_x \pi_{zx}(d_8) L_{8x}\right] \left[\sum_y \pi_{zy}(d_7) L_{7y}\right] \end{aligned}$$

The observed likelihood is a by-product of the recursion as

$$p_\theta(Y) = \sum_k \nu_k L_{9k}, \qquad \text{where } \nu_k = \prod_t \nu_{k_t}.$$

A *downward* (backward-like) recursion can be derived in the same way as the backward recursion for HMM, denoting

$$\begin{aligned} \tau_{jx} &= p_\theta(Z_j = x|Y) \\ \eta_{jxy} &= p_\theta(Z_{\mathrm{par}(j)} = x, Z_j = y|Y), \end{aligned}$$

we first note that $\tau_{2I-1,z} = F_{2I-1,z}$. We then have

$$\begin{aligned} \tau_{jy} &= \sum_x \eta_{jxy} \\ &= \frac{1}{p_\theta(Y)} \sum_x p_\theta(Z_{\mathrm{par}(j)} = x, Z_j = y, Y) \\ &= \frac{1}{p_\theta(Y)} \sum_x p_\theta(Y_{\mathrm{sub}(j)}) p_\theta(Z_j = y|Y_{\mathrm{sub}(j)}) p_\theta(Z_{\mathrm{par}(j)} = x|Z_j = y) p_\theta(Y_{\overline{\mathrm{sub}(j)}}|Z_{\mathrm{par}(j)} = x) \\ &= \frac{p_\theta(Y_{\mathrm{sub}(j)})}{p_\theta(Y)} F_{jy} \sum_x p_\theta(Z_{\mathrm{par}(j)} = x|Z_j = y) p_\theta(Y_{\overline{\mathrm{sub}(j)}}|Z_{\mathrm{par}(j)} = x) \end{aligned}$$

reminding that, thanks to reversibility

$$p_\theta(Z_{\mathrm{par}(j)} = x | Z_j = y) = \nu_x \pi_{xy}(d_j)/\nu_y.$$

We now only have to compute

$$\begin{aligned}
\frac{p_\theta(Y_{\overline{\mathrm{sub}(j)}}|Z_{\mathrm{par}(j)} = x)p_\theta(Y_{\mathrm{sub}(j)})}{p_\theta(Y)} &= \frac{p_\theta(Y_{\overline{\mathrm{sub}(j)}}|Z_{\mathrm{par}(j)} = x)p_\theta(Y_{\mathrm{sub}(j)}|Z_{\mathrm{par}(j)} = x)p_\theta(Y_{\mathrm{sub}(j)})}{p_\theta(Y)p_\theta(Y_{\overline{\mathrm{sub}(j)}}|Z_{\mathrm{par}(j)} = x)} \\
&= \frac{p_\theta(Y|Z_{\mathrm{par}(j)} = x)p_\theta(Y_{\mathrm{sub}(j)})}{p_\theta(Y)p_\theta(Y_{\overline{\mathrm{sub}(j)}}|Z_{\mathrm{par}(j)} = x)} \\
&= \frac{p_\theta(Z_{\mathrm{par}(j)} = x|Y)}{p_\theta(Z_{\mathrm{par}(j)} = x|Y_{\overline{\mathrm{sub}(j)}})} \\
&= \tau_{\mathrm{par}(j)x} \Big/ p_\theta(Z_{\mathrm{par}(j)} = x|Y_{\overline{\mathrm{sub}(j)}})
\end{aligned}$$

where

$$\begin{aligned}
p_\theta(Z_{\mathrm{par}(j)} = x|Y_{\overline{\mathrm{sub}(j)}}) &= \sum_y p_\theta(Z_{\mathrm{par}(j)} = x, Z_j = y|Y_{\overline{\mathrm{sub}(j)}}) \\
&= \sum_y F_{jy} p_\theta(Z_{\mathrm{par}(j)} = x|Z_j = y).
\end{aligned}$$

Thanks to site independence, all summations over x, y or z can be made independently from each other, so the computation complexity is proportional to the number of sequences I, their common length n and the squared number of possible states K, that is the four nucleotides.

3.6 Some applications of HMM in computational biology

3.6.1 Gene detection using tilling array data

Tilling arrays rely on the microarray technology. They are constituted of probes (almost) regularly spread along the genome of the species under study. As an example, about $n = 10^5$ probes are spread along each chromosome of the model plant *A. Thaliana*. When applied to transcriptome (i.e. to all the transcripts present in the cell), tilling arrays give access to a measure of the level of transcription at each probe location. As it does not rely on any prior annotation of the genome, this technology allows us to discover new genes or new regions that are actually transcribed. Figure 3.6 gives an example of the repartition of the probes along the genome and of the available annotation, which can be used to validate model-based predictions.

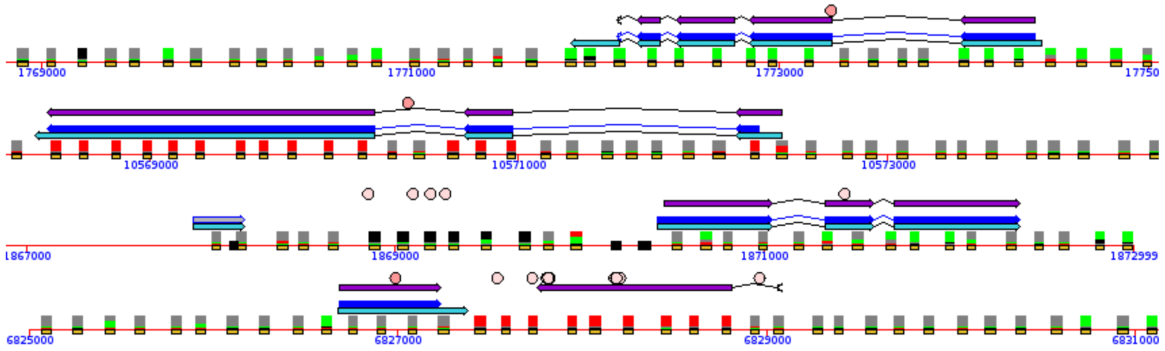

Figure 3.6: Distribution of the probes along a portion of the genome of *A. thaliana*. Blue numbers: position (in bp) along the genome. Blue, cyan and purple arrows: gene locations according to three different genome annotations, cyan being the consensus. Pink dots: observed expressed sequence tags (EST). Box color: gray: $\Pr\{Z_t = \circ | Y\} = \tau_{t\circ}$, black: $\tau_{t=}$, red: τ_{t+}, green: τ_{t-}. Source: Bérard *et al.* (2011).

The two-color version of this technology allows us to compare two samples collected under two different conditions, such as two different tissues (e.g. leaf and seed of a plant). Four categories of probes are then expected to be observed: non transcribed probes (i.e., transcribed in none of the two conditions, labeled '$\circ$'), probes that are transcribed equally under the two conditions ('='), probes that are more transcribed in the first condition than in the second ('+') and the probes displaying the opposite difference ('−'). Figure 3.7 displays and example of such data.

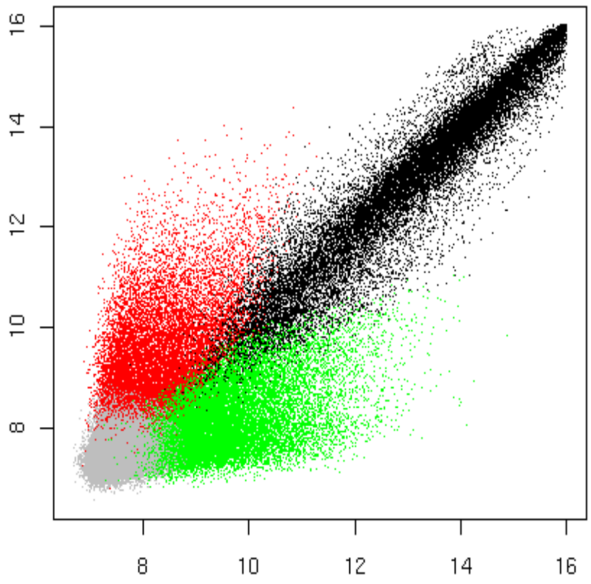

Figure 3.7: Four types of probes from a tilling array experiment on *A. thaliana*. x-axis $= Y_{t1}$, y-axis $= Y_{t2}$. Each dot represents a probe. The color code is the same as for the boxes in Fig. 3.6, but probes are classified according to the most probable hidden path.

Obviously, an HMM with four classes $\{\circ, =, +, -\}$ can be used to distinguish between these

four categories, accounting for the repartition of the probes along the genome. Denoting $Y_t = (Y_{t1}, Y_{t2})$ the signal observed at probe t under condition 1 and 2, respectively, we consider the model from Definition 3.1, where each emission distribution F_k is typically a bivariate normal distribution:

$$F(\gamma_k) = \mathcal{N}(\mu_k, \Sigma_k).$$

Note that the shape of the variance matrix Σ_k has to be carefully thought to get meaningful results (see Bérard *et al.* (2011)).

This modeling results in both the probability for each probe to belong to each of the four categories $\{\circ, =, +, -\}$ (displayed in Figure 3.7) and in their classification according to the most probable hidden path (displayed in Figure 3.6). In Figure 3.7, black, red and green probes are in good concordance with known genes. Interestingly, a series of 6 black probes around position 1,869,000 does not match with any known gene, but seems to be confirmed by observed ESTs.

3.6.2 Pair HMM for sequence alignment

Sequence alignment is a elementary tool for the comparison of genomic sequences. The problem is to find the 'best' possible alignment between two sequences $A = (A_1, \ldots, A_n)$ and $B = (B_1, \ldots, B_m)$. Suppose we consider the following sequences

$$A = (\text{gatctgaac}), \qquad B = (\text{gacgtta}).$$

A possible alignment of these two sequences would be

$$\begin{array}{lccccccccc} A: & \text{g} & \text{a} & \text{t} & \text{c} & - & \text{t} & \text{g} & \text{a} & \text{a} & \text{c} \\ B: & \text{g} & \text{a} & - & \text{c} & \text{g} & \text{t} & - & \text{t} & \text{a} & - \end{array} \tag{3.2}$$

where '–' stands for a deletion (or 'gap') in the sequence or, symmetrically, an insertion in the other sequence. A pair HMM model can be defined to find such an alignment.

Pair hidden Markov model. A pair HMM is an HMM resulting in paired observed sequences. For the purpose of sequence alignment, the model can be defined as follows (using the notations from Definition 3.1).

- $K = 3$ hidden states are considered: 0 = match, 1 = gap in sequence A (or insertion in sequence B), 2 = gap in sequence B (or insertion in sequence A).
- Due to the intrinsic symmetry of the problem, the transition matrix has the form

$$\pi = \begin{pmatrix} \pi_{00} & \pi_{01} & \pi_{01} \\ \pi_{10} & \pi_{11} & \pi_{12} \\ \pi_{10} & \pi_{12} & \pi_{11} \end{pmatrix},$$

  where the transition probabilities π_{00} and π_{10} are expected to be large, whereas all others are expected to be small.

- A specificity of the pair HMM for sequence alignment is that the emission distributions do not have the same domain. Namely, all emission distributions F_k ($k = 0, 1, 2$) are multinomial distributions: $F_k = \mathcal{M}(1, \gamma_k)$, but γ_0 is a distribution over $\{a, c, g, t\}^2$, γ_1 is a distribution over $' - ' \times \{a, c, g, t\}$ and γ_2 is a distribution over $\{a, c, g, t\} \times ' - '$.

Under this model, the alignment given in (3.2) can be rephrased as

Z_t	0	2	2	0	1	0	2	0	0	2
A_t	g	a	t	c	-	t	g	a	a	c
B_t	g	a	-	c	g	t	-	t	a	-

The hidden path Z is often as successions of moves along the two sequences as represented in Figure 3.8.

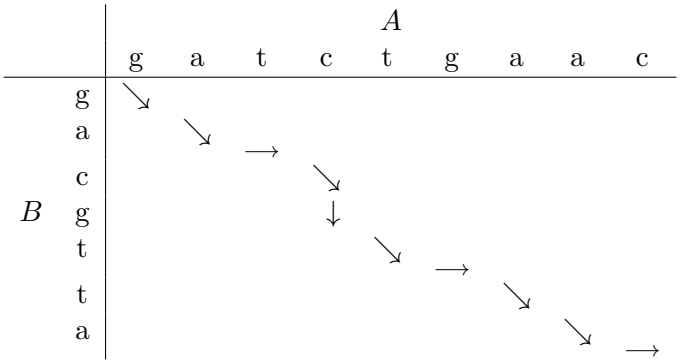

Figure 3.8: Hidden path corresponding to the alignment (3.2) of the two sequences A and B. Diagonal arrows: matches and mismatches, vertical: gap in sequence A, horizontal: gap in sequence B.

Sequence alignment. In many bioinformatics applications, the inference of the parameters π and γ is not considered and the parameters are set to arbitrary values constants, interpreted in terms of costs:

- $-\log \pi_{01}$ is interpreted as the cost for opening a gap and $-\log \pi_{11}$ as the cost for continuing a gap. π_{12} is sometimes set to 0 to avoid the alternation of gaps between the two sequences.
- The distributions γ_k often have a very simple form (both to account for symmetries and to ease interpretation). Typically:

$$\begin{array}{rcll} \gamma_0(a,b) & = & \gamma_0^+ & \text{if } a = b \\ \gamma_0(a,b) & = & \gamma_0^- & \text{if } a \neq b, \\ \gamma_1(-,b) & = & \gamma_2(a,-) & \forall (a,b) \end{array}$$

  where γ_0^+ is larger than γ_0^-. It is sometimes further assumed that $\gamma_1(-,b) = \gamma_0^-$. The quantity $-\log \gamma_0^-$ is then interpreted as the cost of a mismatch.

The main use of the HMM is then the determination of the alignment itself, that is the determination of the most probable hidden path. This path can be retrieved via the Viterbi algorithm, which is equivalent to the Smith & Waterman algorithm widely used in bioinformatics (see e.g. Waterman (1995)).

4 More complex dependency structures: Variational EM

Many models used in many applications display complex dependency structures due, e.g., to space-time organization or to interactions between entities. Biological networks constitute a typical example where the interactions between entities (e.g. genes, proteins, metabolites) are observed and where the goal is to better understand the underlying structure of the resulting network.

Again, latent variable models can be useful to describe such an underlying structure. However, due to the complex dependency structure the conditional distribution of the unobserved labels given the observations most often turns out to be intractable and approximations are required.

4.1 Examples

4.1.1 Stochastic Block-Model

Consider a set of individuals (e.g. humans or proteins) $i = 1, \dots, n$ between which the presence or absence of interaction is observed as

$$Y_{ij} = \mathbb{I}\{i \sim j\} = Y_{ji}$$

where $i \sim j$ means that i interacts with j. The resulting data is called an *interaction network* or, in social sciences, a *social network* as in Figure 4.1.

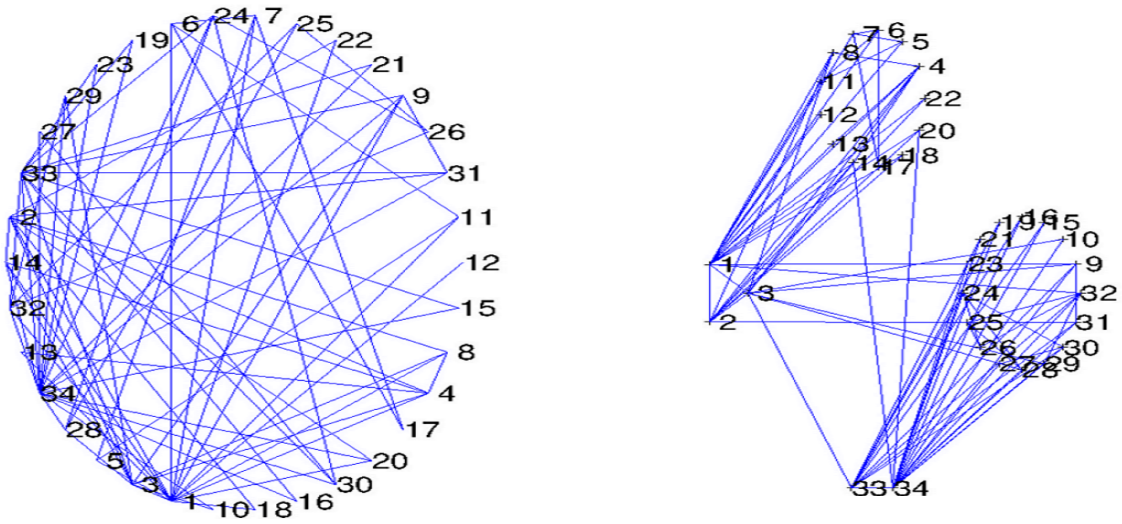

Figure 4.1: Example of a social network (Karate club from Zachary (1977)). Both pictures depict the same network, only the (arbitrary) position of the nodes is changed. Left: unclustered (= raw data), right: clustered in four groups.

One possible way to analyze such a network is to try to define groups of nodes sharing the same connectivity behavior. This can be encoded in the following mixture model known as the stochastic block-model (SBM: Nowicki and Snijders (2001)):

$$\begin{aligned} (Z_i)_i \text{ iid} &\sim \mathcal{M}(1;\pi), \qquad Z_i \in [\![1, K]\!], \\ (Y_{ij})_{i,j} \text{ indep.} \,|\, (Z_i) &\sim \mathcal{B}(\gamma_{Z_i Z_j}). \end{aligned} \tag{4.1}$$

Retrieving the latent labels Z allows to better understand the structure of the network, as in the right panel of Figure 4.1.

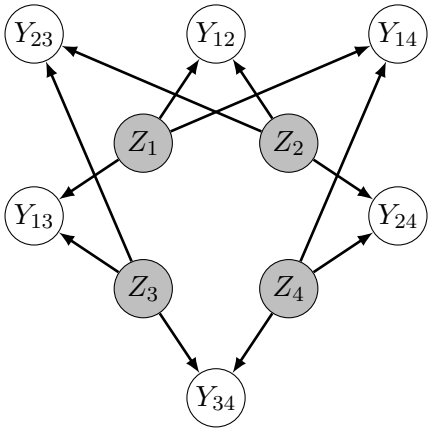

Figure 4.2: Graphical representation of the stochastic block-model.

4.1.2 Latent Block-Model

Consider a set of n genes ($i = 1 \dots n$) studied under p conditions ($j = 1 \dots p$). For each condition j, we measure the expression level Y_{ij} of gene i. We want to determine groups of genes that are preferentially expressed (or not expressed) in certain groups of conditions. Such a problem is often refereed to as *co-clustering* or *bi-clustering*.

A natural way to describe such a structure is to assume that unobserved labels U_i and V_j exist for genes and conditions, respectively. This results in the following model

$$\begin{aligned} (U_i)_i \text{ iid} &\sim \mathcal{M}(1;\pi), \\ (V_j)_j \text{ iid} &\sim \mathcal{M}(1;\nu), \\ (Y_{ij})_{i,j} \text{ indep.} \,|\, (U_i),(V_j) &\sim \mathcal{N}(\mu_{U_iV_j};\sigma^2_{U_iV_j}) \end{aligned}$$

or $\mathcal{N}(\mu_{U_i,V_j},\sigma^2)$ for an homoscedastic version.
In this model, the set of hidden variables is $Z = (U,V) = ((U_i),(V_j))$ and one main goal of the inference is to retrieve it through its conditional distribution

$$p_\theta(Z|Y) = p_\theta(U,V|Y).$$

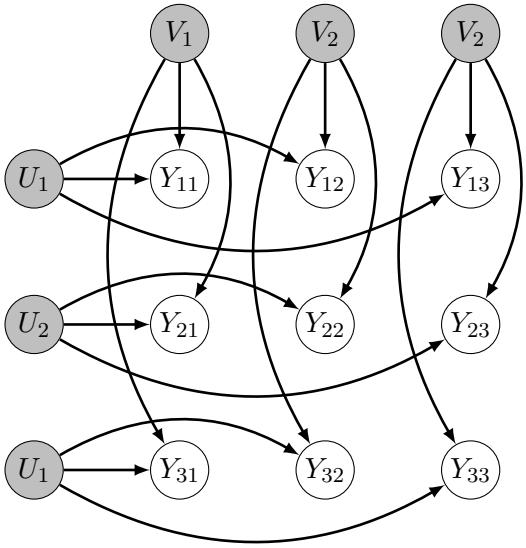

Figure 4.3: Graphical representation of the latent block-model.

4.1.3 Bayesian inference for a mixture model

In a Bayesian perspective, the aim of inference is to provide the conditional distribution of the parameters given the observations. In the case of the mixture models (with K components, K being fixed), the model (2.1) must be completed with prior distribution for the parameters π and γ, e.g. for a Poisson mixture model

$$\pi \sim \mathcal{D}(p), \tag{4.2}$$

$$(\gamma_k)_k \text{ iid} \sim \mathcal{G}\text{am}(a, b) \tag{4.3}$$

where $\mathcal{D}$ stands for the Dirichlet distribution and p, a and b are called the *hyper-parameters* of the model, as they control the distribution of the parameters.

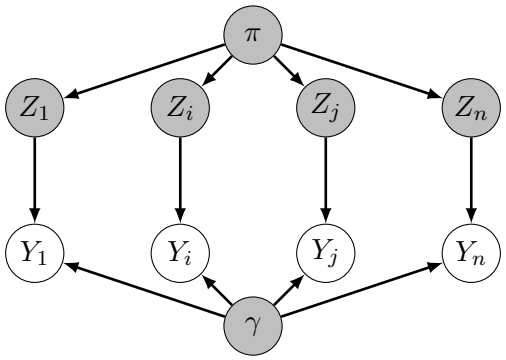

Figure 4.4: Graphical representation of the Bayesian mixture model.

4.2 Variational inference: VEM

We first consider frequentist inference. As seen in the preceding chapters, maximum likelihood inference is most often achieved with the EM algorithm, the E step of which relies on the calculation of the conditional distribution $p_\theta(Z|Y)$. As seen before, the feasibility of the calculation strongly relies on conditional independences allowing (or not) convenient factorizations.

Graph moralization. One important issue in the calculation of the conditional distribution $p_\theta(Z|Y)$ is due to the so-called *moralization* of the graphical model. Consider the simple example of Figure 4.5 where variables A and B are marginally independent and where the distribution of variable C depends on both A and B. In this setting, we have

$$p(A, B, C) = p(A)p(B)p(C|A, B)$$

so the joint conditional distribution of A and B given C is

$$p(A, B|C) = \frac{p(A, B, C)}{p(C)} = \frac{p(A)p(B)p(C|A, B)}{p(C)},$$

which cannot be factorized to separate A and B. Although they are marginally independent, A and B are conditionally dependent given C. This effect is called moralization as the parents get 'married' once their common offspring is observed. Note that the *moralized graph* is an undirected graphical model, as defined in Definition A.2.

$$p(A, B, C) = p(A)p(B)p(C|A, B) \qquad p(A, B|C) = p(A, B, C)/p(C)$$

Figure 4.5: Moralization of a graph. Left: Joint distribution. Right: Conditional distribution of the parents given the offspring.

Case of SBM. In the case of SBM, each couple of hidden variables (Z_i, Z_j) is dependent conditionally on the edge Y_{ij} they share, since

$$p(Z_i, Z_j|Y_{ij}) = \frac{p(Z_i, Z_j, Y_{ij})}{p(Y_{ij})} = \frac{p(Z_i)p(Z_j)p(Y_{ij}|Z_i, Z_j)}{\sum_{k,\ell} p(Y_{ij}|Z_i = k, Z_j = \ell)P(Z_i = k)P(Z_j = \ell)}$$

for which no factorization can be found. This results in the conditional dependency structure depicted in Figure 4.6, which shows that the conditional dependency graph between the hidden labels Z_i is a clique[3]. So, the calculation of the conditional distribution $p_\theta(Z|Y)$ requires the enumeration of the K possible configurations, which is impossible even for moderate sample sizes n.

[3]A clique is a complete graph, that is a graph in which all nodes are connected with each other.

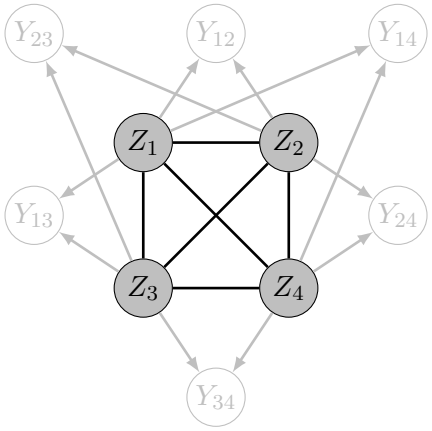

Figure 4.6: Graphical model of the conditional distribution $p(Z|Y)$ of the hidden variables Z_i given the observed ones Y_{ij} in the stochastic block-model.

4.2.1 Variational approximation

As the calculation of $p(Z|Y)$ is intractable, the regular EM algorithm cannot be used and we can only propose an approximate version of it, based on Proposition 2.5 p.14 which states that, for any distribution $q(Z)$,

$$\log p_\theta(Y) \geq \log p_\theta(Y) - KL[q(Z)||p_\theta(Z|Y)] = \mathbb{E}_q[\log p_\theta(Y,Z)] + H[q(Z)]. \tag{4.4}$$

According to this inequality, the smaller the KL divergence, the better the lower bound. However, as $p_\theta(Z|Y)$ can not be computed, the search has to be restricted to a limited class of distribution $\mathcal{Q}$, so the KL divergence can not be 0. Hence, the following algorithm aims at maximizing a lower bound of the likelihood, but not the likelihood itself.

Algorithm 4.1 *Repeat until convergence:*

Variational E step: *given the current estimate θ^h of θ, compute the best possible approximation $\widetilde{q}(Z)$ of $p_{\theta^h}(Z|Y)$ as*

$$\widetilde{q}(Z) = \arg\min_{q\in\mathcal{Q}} KL[q(Z)||p_{\theta^h}(Z|Y)]$$

where $\mathcal{Q}$ is a given class of distributions;

M step: *update the estimate of θ as*

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

Remark. The name 'variational' comes from the pseudo E step that amounts at minimizing a functional (the KL divergence) with respect to a function (the distribution $q(Z)$). In the case of SBM, the function is a distribution over a discrete space, so the optimization can be achieved with standard tools. In the Bayesian context, we will deal with continuous distributions so notions of calculus of variations will be needed.

4.2.2 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 $\widetilde{q}$ within a certain class of distributions $\mathcal{Q}$. The choice of $\mathcal{Q}$ is indeed 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).

The simplest class to be considered is the class of factorized distributions

$$\mathcal{Q}_{\text{fact}} = \left\{ q : q(Z) = \prod_i q_i(Z_i) \right\} \tag{4.5}$$

which result in a mean-field approximation.

Application to the SBM.

Proposition 4.1 *In the stochastic block-model* (4.1) *with approximate conditional distribution q chosen in $\mathcal{Q}_{fact}$, the solution of the VE step satisfies the fix-point relation*

$$\tau_{ik} \propto \pi_k \prod_{j \neq i} \prod_{\ell} f(Y_{ij}; \gamma_{k\ell})^{\tau_{j\ell}}$$

where $\tau_{ik} = \mathbb{E}_q(Z_{ik})$.

Proof: The complete likelihood of the SBM (4.1) is given by

$$\log p_\theta(Y, Z) \quad = \quad \sum_{i,k} Z_{ik} \log \pi_k + \sum_{i<j} \sum_{k,\ell} Z_{ik} Z_{j\ell} \log f(Y_{ij}; \gamma_{k\ell}).$$

Because of (4.4), it is equivalent to minimize the KL divergence $KL[q(Z)||p(Z|Y)]$ and to maximize the lower bound. Since q is chosen in $\mathcal{Q}_{\text{fact}}$, we have

$$q(Z) = \prod_i q_i(Z_i) = \prod_i \prod_k \tau_{ik}^{Z_{ik}} \qquad \text{where } \tau_{ik} = \mathbb{E}_q Z_{ik},$$

so

$$\begin{aligned} H[q(Z)] \quad &= \quad \sum_i H[q_i(Z_i)] = -\sum_i \sum_k \tau_{ik} \log \tau_{ik} \\ \text{and} \qquad \mathbb{E}_q[Z_{ik} Z_{j\ell}] \quad &= \quad \tau_{ik}\tau_{j\ell} \quad \text{for } k \neq \ell. \end{aligned}$$

We now have to maximize the lower bound $\mathbb{E}_q[\log p_\theta(Y, Z)] + H[q(Z)] =$

$$\sum_{i,k} \tau_{ik} \log \pi_k + \sum_{i<j} \sum_{k,\ell} \tau_{ik}\tau_{j\ell} \log f(Y_{ij}; \gamma_{k\ell}) - \sum_i \sum_k \tau_{ik} \log \tau_{ik}$$

with respect to the τ_{ik}'s, subject to $\sum_k \tau_{ik} = 1$ for all i. The derivative with respect to τ_{ik} is zero iff

$$\log \pi_k + \sum_{j \neq i} \sum_{\ell} \tau_{j\ell} \log f(Y_{ij}; \gamma_{k\ell}) - \log \tau_{ik} - 1 - \lambda_i = 0$$

(where λ_i is the Lagrange multiplier for the constraint $\sum_k \tau_{ik} - 1 = 0$), which proves the proposition. □

Remark. In the SBM, denoting $Z^i = \{Z_j, j \neq i\}$, we have

$$\begin{aligned} P(Z_i = k|Y, Z^i) &= \frac{P(Z_i = k, Y|Z^i)}{p(Y|Z^i)} = \frac{p(Y|Z_i = k)P(Z_i = k)}{p(Y|Z^i)} \\ &\propto \pi_k \prod_{j \neq i} \prod_{\ell} f(Y_{ij}; \gamma_{k\ell})^{Z_{j\ell}}. \end{aligned}$$

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 gives the name of the approximation: when considering one individual, the other elements of the field (i.e. the other individuals) are set to their respective means.

4.2.3 Properties of variational estimates

Not much is known about the general properties of variational estimates $\widehat{\theta}_V$. As they are not maximum likelihood estimates, they do not benefit from the general likelihood theory and, for example, their asymptotic variance is not this given by Proposition 2.9. From a general point of view, Gunawardana and Byrne (2005) showed that the VEM algorithm converges to an optimum that differs from the maximum likelihood estimates: $\widehat{\theta}_V \neq \widehat{\theta}_{ML}$. Mean field approximations have also been studied in statistical physics, e.g. Opper and Winther (2001) who showed that it is asymptotically exact for models with 'infinite range dependency'. Some more precise results have been recently obtained in the case of the SBM by Celisse *et al.* (2012), Bickel *et al.* (2013) or Mariadassou and Matias (2015), who proved the consistency of $\widehat{\theta}_V$ for SBM.

An intuition of this can obtained looking at the distribution of the degree of a node (i.e. its number of neighbors). Indeed, conditionally on $Z_i = k$, each edge Y_{ij} arising from i has Bernoulli distribution

$$(Y_{ij}|Z_i = k) \sim \mathcal{B}(\overline{\gamma}_k), \qquad \text{where } \overline{\gamma}_k = \sum_{\ell} \pi_\ell \gamma_{k\ell},$$

so the degree D_i of this node has a conditional binomial distribution

$$D_i = \sum_{j \neq i} Y_{ij}, \qquad (D_i|Z_i = k) \sim \mathcal{B}(n-1, \overline{\gamma}_k).$$

As a consequence, the degrees of all nodes from the same group k concentrate around their common mean $(n-1)\overline{\gamma}_k$ at an exponential rate given by Hoeffding's inequality:

$$P\left(\left|\frac{D_i}{n-1} - \overline{\gamma}_k\right| \geq t\right) \leq 2e^{2(n-1)t^2},$$

which makes the classification of the nodes asymptotically easy.

4.2.4 Alternative approximations and inference strategies

Composite likelihood. Another general approach to deal with complex dependency structures is to use *composite likelihoods*, which consist in the linear combination of likelihoods of sub-groups of variables

$$CL_\theta(Y) = \sum_C w_C \log p_\theta(Y_C), \qquad \text{where } Y_C = (Y_i)_{i \in C}.$$

that can be jointly optimized to get

$$\widehat{\theta}_{CL} = \arg\max_\theta CL_\theta(Y).$$

Varin *et al.* (2011) present a general introduction to these approaches, including general results on the asymptotic variance and distribution of maximum composite likelihood estimates $\widehat{\theta}_{CL}$. Some connexions between variational approximations and composite likelihoods are studied by Lyu (2011). Composite likelihood estimates are considered for SBM by Ambroise and Matias (2012).

Expectation propagation. Other lower bounds than the one given in Proposition 2.5 can be considered. Indeed, any divergence can be removed from the log-likelihood to get a lower bound. For example, a *message passing* version of EM can be obtained by simply inverting the roles of the distribution in KL divergence:

$$\log p(Y) \geq \log p(Y) - KL[p(Z|Y)||q(Z)].$$

A series of such alternatives are presented (and compared) in Minka (2005). To our knowledge, none of these alternatives give raise to tractable computations for SBM.

5 Bayesian inference: Variational Bayes approximations

5.1 A (very brief) reminder on Bayesian inference

In a Bayesian context, the parameter θ itself is supposed to be random and the aim of inference is to evaluate its conditional distribution given the observations (also called *posterior* distribution):

$$p(\theta|Y) = \frac{p(\theta)p(Y|\theta)}{p(Y)} \tag{5.1}$$

where

- $p(\theta)$ is the marginal distribution of the parameter (also called *prior* distribution, possibly depending on some given *hyperparameters*);
- $p(Y|\theta)$ is the *likelihood* function, which was denoted $p_\theta(Y)$ in the frequentist setting;
- $p(Y)$ is the marginal distribution of the data

$$p(Y) = \int p(\theta)p(Y|\theta)\,\mathrm{d}\theta$$

that may be difficult to compute in practice.

The evaluation of the posterior distribution (5.1) is the central task of Bayesian inference (see e.g. Marin and Robert (2007)). For latent variable models; this often moves to the evaluation of the joint conditional distribution

$$p(\theta, Z|Y) = p(\theta)p(Z|\theta)p(Y|\theta, Z)/p(Y).$$

Three main strategies exist to achieve such a tasks.

Exact derivation. In some very specific cases, such as the exponential family / conjugate prior framework (see Section 5.1.1 below and Appendix A.2), $p(\theta|Y)$ can be derived explicitly. Still this situation is limited to rather simple models. In many cases, neither $p(\theta|Y)$ nor $(\theta, Z|Y)$ can be derived in a close-form, nor computed in an exact manner.

Sampling. A huge literature is dedicated to stochastic algorithms, such as Monte-Carlo Markov chains (MCMC), that aim at sampling in these conditional distributions. This topic is out of the scope of this course.

Approximation. The rest of this chapter is dedicated to methods that aim at deriving approximate distributions $q(\theta) \approx p(\theta|Y)$ or $q(\theta, Z) \approx p(\theta, Z|Y)$ using variational techniques.

5.1.1 Case of the exponential family.

In the special case of the exponential family (see Definition 2.3 p.15), the posterior distribution can be obtained in a close form, provided that a *conjugate prior* distribution is used for the parameters.

Proposition 5.1 *If the likelihood of the observed variable belongs to the exponential family with canonical parameter θ*

$$p(Y|\theta) = \exp[\theta^\intercal t(Y) - a(Y) - b(\theta)]$$

and if the parameter θ has conjugate prior distribution with hyper-parameters ν and η:

$$p(\theta) = \exp[\theta^\intercal \nu - c(\nu, \eta) - \eta b(\theta)],$$

then its posterior distribution is the same as the prior distribution with parameters $\nu + t(Y)$ and $\eta + 1$:

$$p(\theta|Y) = \exp[\theta^\intercal(\nu + t(Y)) - c(\nu, \eta + 1) - (\eta + 1)b(\theta)].$$

The proof is given in Appendix A.2.

5.1.2 Latent variable models.

In the presence of hidden variable, the purpose of Bayesian inference is still to know the conditional distribution of the parameter but, in the perspective of classification we are also interested in the conditional distribution of the hidden variables. We are therefore interested in the joint conditional distribution

$$p(\theta, Z|Y).$$

Figures 4.4 and 5.1 present the graphical models for the Bayesian versions of both the mixture model (2.1) p.9 and the stochastic block-model (4.1). The moralization of these graphs leads to an intricate dependency structure between the parameter θ and the hidden variable Z so that, even for the simple independent mixture model, the joint conditional distribution can not be calculated in a close form to get an exact algorithm. The aim of variational Bayes inference is to provide an approximation q of the distribution of interest

$$q(\theta, Z) \approx p(\theta, Z|Y),$$

namely

$$\widetilde{q}(\theta, Z) = \arg\min_{q \in \mathcal{Q}} KL[q(\theta, Z)||p(\theta, Z|Y)]. \tag{5.2}$$

Lower bound of the marginal likelihood $p(Y)$. Note that solving the optimization problem (5.2) is equivalent to maximize the lower bound of $\log p(Y)$:

$$\begin{aligned} \log p(Y) &\geq \log p(Y) - KL[q(\theta, Z)||p(\theta, Z|Y)] \\ &= \mathbb{E}_q \left[\log p(\theta, Z, Y) - \log q(\theta, Z)\right]. \end{aligned} \tag{5.3}$$

Because this lower bound writes as an expectation according to the approximate distribution q, it is hopefully computable in many situations.

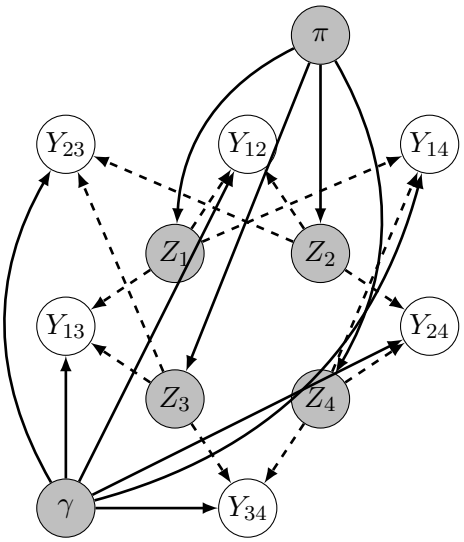

Figure 5.1: Graphical representation of the Bayesian stochastic block-model. Arrows from hidden (Z_i) to observed (Y_{ij}) variables are dashed to preserve clarity of the figure, but have the same status as solid arrows.

5.2 A first example: Bayesian logistic regression inference

Jaakkola and Jordan (2000) exemplify variational Bayes inference on a classical model that does not involve any latent variable: logistic regression. The model is the following: we consider n ($i = 1 \dots n$) individuals for which both a vector of covariates x_i and a binary response Y_i are observed. The aim is to study the effects of the covariates on the response, encoded in a vector of regression coefficients θ. Given the parameter θ the logistic regression model writes:

$$\{Y_i\}|\theta \text{ iid } \sim \mathcal{B}[g(x_i^\intercal \theta)], \qquad \text{where} \quad g(u) = 1 \big/ \left(1 + e^{-u}\right) ,$$

and g is the canonical link function (see Dobson (1990)). In a Bayesian framework, we also need to define a prior (i.e. marginal) distribution for the parameter θ, which can typically be taken as Gaussian:

$$\theta \sim \mathcal{N}(m, G^{-1})$$

where G is the (prior) precision matrix of θ, that is the inverse of its variance matrix.

Likelihood and posterior distribution. The likelihood of the observation is defined are their joint conditional distribution given the parameter and writes

$$p(Y|\theta) = \prod_i p(Y_i|\theta) = \prod_i g(x_i^\intercal \theta)^{Y_i} [1 - g(x_i^\intercal \theta)]^{1-Y_i}$$

so

$$\begin{aligned} \log p(Y|\theta) &= \sum_i Y_i \log g(x_i^\intercal \theta) + (1 - Y_i) \log[1 - g(x_i^\intercal \theta)] \\ &= \sum_i (Y_i - 1) x_i^\intercal \theta + \log g(x_i^\intercal \theta). \end{aligned}$$

because $1 - g(u) = e^{-u} g(u)$. Now, the posterior distribution of θ is proportional to $p(\theta, Y)$ where

$$\begin{aligned} \log p(\theta, Y) &= \log p(\theta) \log p(Y|\theta) \\ &= -\frac{1}{2}\|\theta - m\|_G^2 + \sum_i (Y_i - 1) x_i^\intercal \theta + \log g(x_i^\intercal \theta) + \frac{1}{2} \log |G| + \text{cst}. \end{aligned} \tag{5.4}$$

No classical (log-)distribution for θ can be recognized here. In particular, as (5.4) is not quadratic in θ, the (exact) posterior distribution of θ is not Gaussian.

A lower bound. Jaakkola and Jordan (2000) observe that in (5.4) the first term is quadratic in θ, the second is linear so all difficulties come from the last term, which involves $\log g(u)$ that writes

$$\log g(u) = -\log(1 + e^{-u}) = \frac{u}{2} + \left(e^{u/2} + e^{-u/2}\right)$$

and that the function $f(u) = (e^{u/2} + e^{-u/2})$ is convex in u^2, so that it is always above its quadratic tangent:

$$\forall w, u, \qquad f(u) \geq f(w) + \lambda(w)(u^2 - w^2), \qquad \text{where} \quad \lambda(w) = \partial_{w^2} f(w).$$

Approximate Gaussian posterior. Setting u to $x_i^\intercal\theta$ for each i, we end-up with a quadratic lower bound of $\log p(\theta, Y)$ depending on the variational parameters $w = (w_i)$:

$$\begin{aligned}\log p(\theta, Y) &\geq -\frac{1}{2}\|\theta - m\|_G^2 + \sum_i \left[(Y_i - 1)x_i^\intercal\theta + \frac{x_i^\intercal\theta}{2} + \lambda(w_i)(\theta^\intercal x_i x_i^\intercal\theta - w_i^2)\right] \\ &\quad + \frac{1}{2}\left[\log|G| - d\log(2\pi)\right].\end{aligned}$$

The quadratic and linear terms in θ can be re-organized as

$$\begin{aligned}\log p(\theta, Y) &\geq -\frac{1}{2}\theta^\intercal\left[G + 2\sum_i \lambda(w_i)x_i x_i^\intercal\right]\theta + \left[Gm + \sum_i \left(Y_i - \frac{1}{2}\right)x_i\right]^\intercal\theta \\ &\quad + \sum_i \lambda(w_i)w_i^2 - \frac{1}{2}\left[\|m\|_G^2 - \log|G| + d\log(2\pi)\right]\end{aligned}$$

so that, taking

$$\widetilde{G}(w) = G + 2\sum_i \lambda(w_i)x_i x_i^\intercal \quad \widetilde{G}(w) = \widetilde{G}(w)^{-1}\left[Gm + \sum_i \left(Y_i - \frac{1}{2}\right)x_i\right]$$

we get the lower bound

$$\log p(\theta, Y) \geq \log q_w(\theta) + \sum_i \lambda(w_i)w_i^2 - \frac{1}{2}\left[\|m\|_G^2 - \log|G| - \|\widetilde{m}(w)\|_{\widetilde{G}(w)}^2 + \log|\widetilde{G}(w)|\right], \quad (5.5)$$

where q_w stands for the Gaussian distribution $\mathcal{N}(\widetilde{m}(w), \widetilde{G}(w))$, to be used as an approximate posterior.

Maximizing the lower bound. As (5.5) hold for any w, the lower bound can be maximized wrt the variational parameter w so to get the optimal variational approximation of this form. The complete procedure is described in the aforementioned article by Jaakkola and Jordan (2000). Note that, in this case, the optimization is still achieved wrt a finite dimensional parameter, i.e. w.

5.3 Variational Bayes EM inference

5.3.1 A (very brief) reminder on calculus of variations

In the following, the approximate distribution q will be defined as the minimizer of an integral. To find it, we need to define the equivalent of a derivative for functionals (see Frigyik *et al.* (2008) for an introduction). The following proposition gives a characterization of such a minimizer.

Proposition 5.2 *The function q that minimizes the functional*

$$\mathcal{F}(q) = \int L(x, q(x))\, dx.$$

satisfies the Euler-Lagrange differential equation:

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

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

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

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

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

which is, at $t=0$,

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

The fundamental lemma of calculus of variations states that

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

which completes the proof. □

5.3.2 Variational Bayes EM algorithm

We now introduce the variational Bayes EM (VBEM) algorithm as introduced by Ghahramani and Beal (2001) (and Beal and Ghahramani (2003)), which aims at retrieving the solution of (5.2), for

$$\mathcal{Q} = \{q(Z, \theta) = q_Z(Z) q_\theta(\theta)\}. \tag{5.6}$$

A tutorial on variational Bayes inference can be found in Fox and Roberts (2012).

Optimal q_Z and q_θ. We have to maximize $\mathcal{F}(q) = \mathcal{F}(q_Z q_\theta)$ with respect to both q_Z and q_θ that play completely symmetric roles in the optimization problem (5.2).

Proposition 5.3 *The minimizer $\widetilde{q}_Z$ of the functional*

$$\mathcal{F}(q_Z) = KL[q_Z(Z) q_\theta(\theta) || p(\theta, Z|Y)]$$

satisfies

$$\widetilde{q}_Z(Z) \propto \mathbb{E}_{q_\theta}\left[\log p(Y, Z, \theta)\right].$$

Proof: The optimization problem can be casted into the framework of Proposition 5.2, taking

$$\begin{aligned} L(Z, q_Z) &= q_Z(Z) \int q_\theta(\theta) \log \frac{p(Y, Z, \theta)}{q_Z(Z) q_\theta(\theta)}\, \mathrm{d}\theta \\ &= q_Z(Z) \int q_\theta(\theta) \log p(Y, Z, \theta)\, \mathrm{d}\theta \\ &\quad - q_Z(Z) \int q_\theta(\theta) \log q_\theta(\theta)\, \mathrm{d}\theta - q_Z(Z) \int q_\theta(\theta) \log q_Z(Z)\, \mathrm{d}\theta. \end{aligned}$$

The solution must satisfy

$$\begin{aligned} \partial_{q_Z(Z)} L(Z, q_Z(Z)) &= \int q_\theta(\theta) \log p(Y, Z, \theta)\, \mathrm{d}\theta - \int q_\theta(\theta) \log q_\theta(\theta)\, \mathrm{d}\theta \\ &\quad - [\log q_Z(Z) + 1] \int q_\theta(\theta)\, \mathrm{d}\theta \\ &= 0 \end{aligned}$$

that is

$$\log \widetilde{q}_Z(Z) = \int q_\theta(\theta) \log p(Y, Z, \theta) \, \mathrm{d}\theta + \text{cst}, \tag{5.7}$$

so

$$\widetilde{q}_Z(Z) \propto \exp \int q_\theta(\theta) \log p(Y, Z, \theta) \, \mathrm{d}\theta. \tag{5.8}$$

□

Similarly, it is easy to show that

$$\widetilde{q}_\theta(\theta) \propto \mathbb{E}_{q_Z} [\log p(Y, Z, \theta)]$$

so the two optimal distributions $\widetilde{q}_Z$ and $\widetilde{q}_\theta$ depend on each other.

Remarks.

1. The optimal solution is a mean-field approximation, since the approximate conditional distribution q_Z is the mean of the complete-data log-likelihood averaged according to q_θ.
2. The distribution q_Z must satisfy $\int q_Z(Z) \, \mathrm{d}Z = 1$ (idem for q_θ). Adding this constraint to the optimization problem is equivalent to take

$$L(Z, q_Z) = q_Z(Z) \left[\int q_\theta(\theta) \log \frac{p(Y, Z, \theta)}{q_Z(Z) q_\theta(\theta)} \, \mathrm{d}\theta + \lambda \right]$$

   so (5.7) still holds, and (5.8) is unchanged.

Conjugate exponential case. In the case of exponential emission distribution with conjugate prior, Beal and Ghahramani (2003) derived explicit formula for the optimal approximate distributions.

Proposition 5.4 *If the following conditions are fulfilled:*

(*i*) *the distribution* $p(Y, Z|\theta)$ *belongs to the exponential family:*

$$p(Y, Z) = \exp[\theta^\intercal t(Y, Z) - a(Y, Z) - b(\theta)],$$

(*ii*) *the prior distribution* $p(\theta)$ *is conjugate*

$$p(\theta) = \exp[\theta^\intercal \nu - c(\nu, \eta) - \eta b(\theta)],$$

the distribution $\widetilde{q}(\theta, Z) = \widetilde{q}_\theta(\theta)\widetilde{q}_Z(Z)$ *that minimizes* $KL[q(\theta, Z)||p(\theta, Z|Y)]$ *satisfies:*

$$\widetilde{q}_\theta(\theta) = \exp\{\theta^\intercal \tilde{\nu} - c(\tilde{\nu}, \tilde{\eta}) - \tilde{\eta} b(\theta)\}$$

where $\overline{t}(Y) = \int \widetilde{q}_Z(\theta) t(Y, Z)$, $\tilde{\nu} = \nu + \overline{t}(Y)$, $\tilde{\eta} = \eta + 1$ *and*

$$\widetilde{q}_Z(Z) \propto \exp\left\{ \overline{\theta}^\intercal t(Y, Z) - a(Y, Z) \right\}.$$

where $\overline{\theta} = \int \widetilde{q}_\theta(\theta) \theta \, d\theta$.

Proof: Under conditions (i) and (ii), the joint distribution of Z, Y and θ is

$$p(Y, Z, \theta) = \exp\{\theta^\intercal[\nu + \overline{t}(Y)] - a(Y, Z) - (\eta + 1)b(\theta)\}.$$

When applying (5.8) to q_θ, $\exp[-a(Y, Z)]$ is a constant so

$$\begin{aligned}\widetilde{q}_\theta(\theta) &\propto \exp \int \widetilde{q}_Z(Z) \log\left(\exp\{\theta^\intercal[\nu + t(Y, Z)]\} - (\eta + 1)b(\theta)\right) \, \mathrm{d}Z \\ &= \exp\left\{\theta^\intercal\left[\nu + \int \widetilde{q}_Z(Z)t(Z, Y) \, \mathrm{d}Z\right] - (\eta + 1)b(\theta)\right\}.\end{aligned}$$

When the same is applied to $\widetilde{q}_Z$, $\exp[-(\eta + 1)b(\theta)]$ is a constant so

$$\begin{aligned}\widetilde{q}_Z(Z) &\propto \exp \int \widetilde{q}_\theta(\theta) \log\left[\exp\{\theta^\intercal t(Y, Z) - a(Y, Z)\}\right] \, \mathrm{d}\theta \\ &= \exp\left\{\left[\int \widetilde{q}_\theta(\theta)\theta \, \mathrm{d}\theta\right]^\intercal t(Y, Z) - a(Y, Z)\right\}.\end{aligned}$$

□

This results leads to the following VBEM algorithm.

Algorithm 5.1 *The variational Bayes EM algorithm consists in alternative updates of $\widetilde{q}_\theta$ and $\widetilde{q}_Z$:*

E step: *update $\widetilde{q}_\theta$ as*

$$q_\theta^{h+1}(\theta) = \exp\{\theta^\intercal[\overline{t}^h(Y) + \nu] - c[\overline{t}^h(Y) + \nu, 1 + \eta] - (\eta + 1)b(\theta)\};$$

M step: *update $\widetilde{q}_Z$ as*

$$q_Z^{h+1}(Z) \propto \exp\left\{\left(\overline{\theta}^{h+1}\right)^\intercal t(Y, Z) - a(Y, Z)\right\}.$$

5.3.3 Example: Poisson mixture model

Consider a Poisson mixture model with the conjugate prior distributions given in (4.2). More precisely

$$\begin{aligned}\pi \sim \mathcal{D}(a): \qquad & p(\pi) = \Gamma\left(\sum_k a_k\right) \prod_k \left[\pi_k^{a_k - 1}/\Gamma(a_k)\right], \\ (\gamma_k) \text{ indep.} \sim \mathcal{G}\mathrm{am}(b_k, c_k): \qquad & p(\gamma_k) = \gamma_k^{b_k - 1} \exp(-c_k\gamma_k) c_k^{b_k}/\Gamma(b_k).\end{aligned}$$

so

$$\log p(\theta) = \sum_k (a_k - 1)\log \pi_k + (b_k - 1)\log \gamma_k - c_k\gamma_k := \theta^\intercal\nu - c(\nu, \eta) - \eta b(\theta) + \mathrm{cst}$$

where

$$\begin{aligned}\theta &= [\quad (\log \pi_k)_k \quad (\log \gamma_k)_k \quad (-\gamma_k)_k \quad], \\ \nu &= [\quad (a_k - 1)_k \quad (b_k - 1)_k \quad (c_k)_k \quad].\end{aligned}$$

The complete likelihood is

$$\log p(Y, Z|\theta) = \sum_{i,q} Z_{ik} \left[\log \pi_k + Y_i \log \gamma_k - \gamma_k - \log(Y_i!)\right] =: \theta^\intercal u(Y, Z) - a(Y, Z),$$

where

$$u(Y,Z) \quad = \quad [\quad (\sum_i Z_{ik})_k \quad (\sum_i Z_{ik}Y_i)_k \quad (\sum_i Z_{ik})_k \quad].$$

So we get the following update formula for $\widetilde{q}_\theta$

$$\widetilde{q}_\theta(\theta) \propto \exp\left\{\theta^\intercal[\nu + \overline{u}(Y)]\right\}$$

with

$$\overline{u}(Y) \quad = \quad [\quad (\sum_i \tau_{ik})_k \quad (\sum_i \tau_{ik}Y_i)_k \quad (\sum_i \tau_{ik})_k \quad],$$

and

$$\widetilde{\nu} \quad = \quad [\quad (\widetilde{a}_k - 1)_k \quad (\widetilde{b}_k - 1)_k \quad (\widetilde{c}_k)_k \quad].$$

where

$$\tilde{a}_k = a_k + \sum_i \tau_{ik}, \qquad \tilde{b}_k = b_k + \sum_i \tau_{ik}Y_i, \qquad \tilde{c}_k = c_k + \sum_i \tau_{ik};$$

and for $\widetilde{q}_Z$:

$$\widetilde{q}_Z(Z) \propto \exp\left\{\overline{\theta}^\intercal t(Y,Z) - a(Y,Z)\right\}$$

where (using Lemma A.1, p.59)

$$\overline{\theta} \quad = \quad [\quad (\psi_0(\tilde{a}_k) - \psi_0\left(\textstyle\sum_\ell \tilde{a}_\ell\right))_k \quad \left(\psi_0(\tilde{b}_k) - \log(\tilde{c}_k)\right)_k \quad \left(-\tilde{b}_k/\tilde{c}_k\right)_k \quad].$$

Note that we also have

$$\begin{aligned} \widetilde{q}_Z(Z) \quad &\propto \quad \exp\left\{\overline{\theta}^\intercal \left[\sum_i t(Y_i,Z_i)\right] - \sum_i a(Y_i,Z_i)\right\} \\ &\propto \quad \prod \widetilde{q}_{Z_i}(Z_i) \\ \text{where} \quad \widetilde{q}_{Z_i}(Z_i) \quad &\propto \quad \sum_k Z_{ik} \log \tau_{ik} \end{aligned}$$

where

$$\tau_{ik} \propto \exp\left\{\psi_0(\tilde{a}_k) - \psi_0\left(\sum_\ell \tilde{a}_\ell\right) + Y_i\left[\psi_0(\tilde{b}_k) - \log(\tilde{c}_k)\right] - \tilde{b}_k/\tilde{c}_k\right\},$$

subject to $\sum_k \tau_{ik} = 1$.

5.4 (Variational) Bayesian model selection or averaging

In many situations, several models can be considered to analyze a given dataset. A typical case is the series models with $K = 1, 2, ...$ hidden states. In this section we shall denote $(M_K)_{K\geq 1}$ the set of models at hand. In this situation two approaches can be considered:

Model selection, that is to find the 'best' model among the list or

Model averaging, that is to combine the predictions of all models, without choosing any specific one.

Both problems can be stated in a Bayesian way, considering the model as an additional parameter. This amounts to cast all models into a larger one defined by the following series of (conditional) distributions:

$$\begin{aligned} p(K) \quad &= \quad \text{prior distribution on the models;} \\ p(\theta|K) \quad &= \quad \text{conditional prior of the parameters given model } M_K; \\ p(Z|\theta,K) \quad &= \quad \text{conditional distribution of the latent variables given the parameters} \\ &\quad\quad \text{(and the model);} \\ p(Y|Z,\theta,K) \quad &= \quad \text{conditional distribution of the observed variables given all the rest.} \end{aligned}$$

For both model selection and model averaging, the critical quantity to evaluate is the posterior probability of each model, that is

$$p(K|Y) = \iint p(Z, \theta, K|Y) \, \mathrm{d}Z \, \mathrm{d}\theta. \tag{5.9}$$

Model selection: the best' model among the list can be defined as most probable conditionally on the date:

$$\widehat{K} = \arg\max_k P(K = k|Y),$$

which is the rational of the BIC criterion (Schwarz (1978)).

Model averaging: the posterior distribution of any function of interest $\Delta = h(\theta)$ can be obtained as

$$p(\Delta|Y) = \sum_K p(K|Y) p(\Delta|Y, K)$$

where $p(\Delta|Y, K)$ is the posterior distribution of Δ for model K (see Hoeting *et al.* (1999) for a general introduction).

Indeed, the calculation pf $p(K|Y)$ is often complex, not to say impossible, but a variational approximation of it can be derived using an easy-to-handle joint distribution $q(Z, \theta, K)$ which approximates $p(Z, \theta, K|Y)$. This is actually doable with no additional approximation that this made in the variational Bayes inference described in Section **??**, that is to assume that, the approximate conditional distribution $q(\theta, Z|L)$ belongs to same class $\mathcal{Q}$ as this considered for VBEM (for example this given by (5.6)) *for model* M_K. More specifically, we consider distributions q in the following class:

$$\overline{\mathcal{Q}} = \{q(Z, \theta, K) : \forall K, q(\theta, Z|K) \in \mathcal{Q}\}. \tag{5.10}$$

Volant *et al.* (2012) derive the general form of the variational Bayes approximation of $p(K|Y)$.

Proposition 5.5 *The joint distribution*

$$\widetilde{q} = \arg\min_{q \in \overline{\mathcal{Q}}} KL\left[q(Z, \theta, K)||p(Z, \theta, K|Y)\right]$$

satisfies

$$\widetilde{q}(\theta, Z|K) = \arg\min_{q \in \mathcal{Q}_K} KL\left[q(\theta, Z)||p(Z, \theta|Y, K)\right]$$

and, denoting $KL_K^* = KL\left[\widetilde{q}(\theta, Z|K)||p(Z, \theta|Y, K)\right]$,

$$\begin{aligned} \widetilde{q}(K) &\propto p(K|Y) e^{-KL_K^*} \\ &\propto p(K) \exp\left\{\mathbb{E}_{\widetilde{q}_{\theta,Z|K}}\left[\log p(Y, Z, \theta|K) - \log \widetilde{q}_{\theta,Z|K}(\theta, Z)\right]\right\} \end{aligned}$$

Proof: The KL divergence between $q(Z, \theta, K)$ and $p(Z, \theta, K|Y)$ is

$$\begin{aligned} KL &= \mathbb{E}_q\left[\log q(Z, \theta, K) - \log p(Z, \theta, K|Y)\right] \\ &= \mathbb{E}_q\left[\log q(Z, \theta|K) - \log p(Z, \theta|Y, K) + \log q(K) - \log p(K|Y)\right] \\ &= \sum_K q(K) \left\{\mathbb{E}_{q_{\theta,Z|K}}\left[\log q(Z, \theta|K) - \log p(Y, Z, \theta|Y, K)\right] + \log q(K) - \log p(K|Y)\right\} \\ &= \sum_K q(K) \left\{KL\left[q(\theta, Z|K)||p(Z, \theta|Y, K)\right] + \log q(K) - \log p(K|Y)\right\}. \end{aligned}$$

Now, we first minimize wrt $q(\theta, Z|K)$ to get

$$\min_{q \in \overline{\mathcal{Q}_K}} KL = \min_q \left(\min_{q_{\theta,Z|K} \in \mathcal{Q}_K} KL \right) = \min_q \sum_K q(K) \left\{ KL_K^* + \log q(K) - \log p(K|Y) \right\}.$$

The optimal distribution $\widetilde{q}(K)$ is obtained but setting the derivatives of this wrt to each $q(K)$ to zero (under the constraint that $\sum_K q(K) = 1$). We get

$$\partial_{q(K)} KL - \lambda \left(\sum_K q(K) - 1 \right) = KL_K^* + \log q(K) - \log p(K|Y) - \lambda$$

which is zero iff

$$q(K) \propto p(K|Y) e^{-KL_K^*}. \tag{5.11}$$

To get the second formulation, we observe that

$$\begin{aligned} p(K|Y) e^{-KL_K^*} &= p(K) \exp \left[\log p(Y|K) - KL_K^* \right] / p(Y) \\ &= p(K) \exp \left\{ \mathbb{E}_{\widetilde{q}_{\theta,Z|K}} \left[\log p(Y, Z, \theta|K) - \log \widetilde{q}_{\theta,z|K}(\theta, z) \right] \right\} \end{aligned}$$

where we recognize the lower bound (5.3) for model M_K to get the last equality. □

Remarks.

1. Formula (5.11) is not directly computable as neither $p(K|Y)$ not $e^{-KL_K^*}$ are. Still, it has an intuitive interpretation in the sense that the approximate conditional probability of model M_K is proportional to both
   - the true posterior probability of the model $p(K|Y)$, which is the quantity we were primarily interested in, and
   - a correction term $e^{-KL_K^*}$ which can be seen as a penalization for a poor quality of the VB approximation in model M_K.
2. We stress that integrating the model as an additional missing variable does not require any additional approximation. The proximity between $\widetilde{q}(K)$ and $p(K|Y)$ is completely ruled by the quality of the VB approximation for each model.

Variational Bayes model selection. The VB approximation of $p(K|Y)$ can be used for model selection, taking

$$\widehat{M} = M_{\widehat{K}} \qquad \text{where} \quad \widehat{K} = \arg\max_K \widetilde{q}(K).$$

Variational Bayes model averaging. Similarly a variation approximation of the posterior expectation of any parameter of interest $\Delta = h(\theta, Z)$ is given by

$$\widetilde{\mathbb{E}}(\Delta) = \sum_K \widetilde{q}(K) \mathbb{E}_{\widetilde{q}_{\theta,Z|K}}(\Delta).$$

5.5 Sampling in the Posterior distribution

The aim of variational Bayes inference is to provide a distribution $\widetilde{q}_\theta(\theta)$ that approximates the true posterior distribution of the parameters $p(\theta|Y)$:

$$\widetilde{q}_\theta(\theta) \approx p(\theta|Y).$$

Bayesian parameter inference, typically requires the determination of credibility intervals

$$C_{1-\alpha} = [a; b], \quad \text{such that } \int_{C_{1-\alpha}} p(\theta|Y)\ \mathrm{d}\theta = 1 - \alpha.$$

More generally, we are interested in the evaluation of integral of the form

$$\int p(\theta|Y)h(\theta)\ \mathrm{d}\theta.$$

As $\widetilde{q}_\theta$ is only an approximation, the approximation of such integrals may be poor when replacing $p(\theta|Y)$ with $\widetilde{q}_\theta(\theta)$. Indeed, the variational approximation is known to provide accurate estimate for the posterior mode (see Minka (2005)) but to underestimate the posterior variance (see Bishop and Nasrabadi (2006)).

Importance sampling. Monte Carlo techniques allow to evaluate such integrals as

$$\int p(\theta|Y)h(\theta)\ \mathrm{d}\theta \approx \frac{1}{B}\sum_b h(\theta^b) \qquad \text{with } \{\theta^b\} \text{ iid } \sim p(\theta|Y).$$

In latent variable models sampling under $p(\theta|Y)$ is not possible, but the computation of $p(Y|\theta)$ is possible. The above sampling scheme can be modified, remarking that

$$\int p(\theta|Y)h(\theta)\ \mathrm{d}\theta \propto \int \frac{p(\theta)p(Y|\theta)}{q(\theta)} q(\theta)h(\theta)\ \mathrm{d}\theta = \int w_q(\theta)q(\theta)h(\theta)\ \mathrm{d}\theta,$$

denoting $w_q(\theta) = p(\theta)p(Y|\theta)/q(\theta)$, so that

$$\int p(\theta|Y)h(\theta)\ \mathrm{d}\theta \approx C^{-1}\sum_b w_q(\theta^b)h(\theta^b) \qquad \text{with } \{\theta^b\} \text{ iid } \sim q(\theta)$$

where the normalizing constant is evaluated by

$$C = \sum_b w_q(\theta^b) \qquad \text{with } \{\theta^b\} \text{ iid } \sim q(\theta).$$

The accuracy of such an estimate relies of the proximity between the proposal distribution $q(\theta)$ and the target distribution $p(\theta|Y)$. The variational posterior $\widetilde{q}_\theta$ can be used as a proxy.

References

Akaike, H. (1974). A new look at the statistical model identification. *Automatic Control, IEEE Transactions on.* **19 (6)** 716–723.

Allison, D. B., Gadbury, G., Heo, M., Fernandez, J., Lee, C.-K., Prolla, T. A. and Weindruch, R. A. (2002). Mixture model approach for the analysis of microarray gene expression data. *Computational Statistics & Data Analysis.* **39** 1–20.

Ambroise, C. and Matias, C. (2012). New consistent and asymptotically normal parameter estimates for random-graph mixture models. *Journal of the Royal Statistical Society: Series B.* **74 (1)** 3–35.

Beal, J., M. and Ghahramani, Z. (2003). The variational Bayesian EM algorithm for incomplete data: with application to scoring graphical model structures. *Bayes. Statist.* **7** 543–52.

Bérard, C., Martin-Magniette, M.-L., Brunaud, V., Aubourg, S., Robin, S. *et al.* (2011). Unsupervised classification for tiling arrays: ChIP-chip and transcriptome. *Statistical applications in genetics and molecular biology.* **10 (1)** 1–22.

Bickel, P., Choi, D., Chang, X. and Zhang, H. (2013). Asymptotic normality of maximum likelihood and its variational approximation for stochastic blockmodels. *The Annals of Statistics.* 1922–1943.

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

Bishop, C. M. and Nasrabadi, N. M. (2006). *Pattern recognition and machine learning.* volume 1. springer New York.

Cappé, O., Moulines, E. and Rydén, T. (2005). *Inference in Hidden Markov Models.* Springer.

Celisse, A., Daudin, J.-J. and Pierre, L. (2012). Consistency of maximum-likelihood and variational estimators in the stochastic block model. *Electron. J. Statis.* **6** 1847–99.

Dempster, A. P., Laird, N. M. and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society: Series B.* **39** 1–38.

Dobson, A. J. (1990). *An introduction to generalized linear models.* Chapman & Hall.

Durand, J.-B., Goncalves, P. and Guédon, Y. (2004). Computational methods for hidden Markov tree models-an application to wavelet trees. *IEEE Transactions on Signal Processing.* **52 (9)** 2551–2560.

Falush, D., Stephens, M. and Pritchard, J. K. (Aug, 2003). Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. *Genetics.* **164 (4)** 1567–1587.

Felsenstein, J. (1981). Evolutionary trees from DNA sequences: a maximum likelihood approach. *J. Mol. Evol.* **17 (6)** 368–376.

Fox, C. W. and Roberts, S. J. (2012). A tutorial on variational Bayesian inference. *Artificial Intelligence Review.* **38 (2)** 85–95.

Fridlyand, J., Snijders, A. M., Pinkel, D., Albertson, D. G. and Jain, A. N. (July, 2004). Hidden Markov models approach to the analysis of array CGH data. *Journal of Multivariate Analysis.* **90 (1)** 132–153.

Frigyik, B. A., Srivastava, S. and Gupta, M. R. (2008), An introduction to functional derivatives. Technical report, Dept. Electr. Eng., Univ. Washington, Seattle, WA.

Ghahramani, Z. and Beal, M. J. (2001). Propagation algorithms for variational Bayesian learning. In *Advances in neural information processing systems*, 507–513.

Ghahramani, Z. and Hinton, G. E. (1996), Parameter estimation for linear dynamical systems. Technical Report CRG-TR-96-2, University of Toronto, Dept. of Computer Science.

Gunawardana, A. and Byrne, W. (2005). Convergence theorems for generalized alternating minimization procedures. *J. Mach. Learn. Res.* **6** 2049–73.

Hedenfalk, I., Duggan, D., Chen, Y., Radmacher, M., Bittner, M., Simon, R., Meltzer, P., Gusterson, B., Esteller, M., Raffeld, M. *et al.* (2001). Gene-expression profiles in hereditary breast cancer. *New England Journal of Medicine.* **344 (8)** 539–548.

Hoeting, J. A., Madigan, D., Raftery, A. E. and Volinsky, C. T. (1999). Bayesian model averaging: a tutorial. *Statistical science.* **14 (4)** 382–401.

Hupé, P. (2008). *Biostatistical algorithms for omics data in oncology: Application to DNA copy number microarray experiments.* PhD thesis, AgroParisTech.

Jaakkola, T. (2001). *Advanced mean field methods: theory and practice.* chapter Tutorial on variational approximation methods, 129–160. MIT Press.

Jaakkola, T. S. and Jordan, M. I. (2000). Bayesian parameter estimation via variational methods. *Statistics and Computing.* **10 (1)** 25–37.

Lartillot, N. (2014). A Phylogenetic Kalman Filter for Ancestral Trait Reconstruction Using Molecular Data. *Bioinformatics.* **30 (4)** 488–496.

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

Louis, T. A. (1982). Finding the observed information matrix when using the EM algorithm. *Journal of the Royal Statistical Society. Series B.* 226–233.

Lyu, S. (2011). Unifying non-maximum likelihood learning objectives with minimum KL contraction. In *NIPS*, (J. Shawe-Taylor, R. S. Zemel, P. L. Bartlett, F. C. N. Pereira, and K. Q. Weinberger, ed.), 64–72.

Mariadassou, M. and Matias, C. (2015). Convergence of the groups posterior distribution in latent or stochastic block models. *Bernoulli.* **21 (1)** 537–573.

Marin, J.-M. and Robert, C. P. (2007). *Bayesian Core: A Practical Approach to Computational Bayesian Statistics.* Springer-Verlag: New-York.

McLahan, G. and Peel, D. (2000). *Finite Mixture Models*. Wiley.

Minka, T. (2005), Divergence measures and message passing. Technical Report MSR-TR-2005-173, Microsoft Research Ltd.

Nowicki, K. and Snijders, T. (2001). Estimation and prediction for stochastic blockstructures. *Journal of the American Statistical Association*. **96 (455)** 1077–1087.

Opper, M. and Winther, O. (2001). *Advanced mean field methods: Theory and practice*. chapter From Naive Mean Field Theory to the TAP Equations. The MIT Press.

Rosenberg, N. A., Pritchard, J. K., Weber, J. L., Cann, H. M., Kidd, K. K., Zhivotovsky, L. A. and Feldman, M. W. (Dec, 2002). Genetic structure of human populations. *Science*. **298 (5602)** 2381–2385.

Schwarz, G. (1978). Estimating the dimension of a model. *The annals of statistics*. **6 (2)** 461–464.

Varin, C., Reid, N. and Firth, D. (2011). An overview of composite likelihood methods. *Statistica Sinica*. **21** 5–42.

Volant, S., Martin Magniette, M.-L. and Robin, S. (2012). Variational bayes approach for model aggregation in unsupervised classification with Markovian dependency. *Comput. Statis. & Data Analysis*. **56 (8)** 2375 – 2387.

Wainwright, M. J. and Jordan, M. I. (2008). Graphical models, exponential families, and variational inference. *Found. Trends Mach. Learn.* **1 (1–2)** 1–305.

Waterman, M. S. (1995). *Introduction to Computational Biology*. Chapman & Hall.

Zachary, W. W. (1977). An information flow model for conflict and fission in small groups. *J. Anthrop. Res.* **33** 452–73.

A Some useful tools

A.1 Graphical models

Definition A.1 *A directed graphical model is a directed acyclic graph (DAG) $G = (V, E)$ where*

- *the set of vertices V is the set of all the random variables involved in the model:*

$$V := Y = \{Y_j\}$$

- *the joint distribution of all variables can be factorized as*

$$p_\theta(Y) = \prod_j p_\theta(Y_j | Y_{par(j)})$$

where $Y_{par(j)}$ stands for the (possibly empty) set of parents of Y_j in G:

$$Y_{par(j)} = \{Y_i : (i, j) \in E\}.$$

Definition A.2 *An undirected graphical model is an unoriented graph $G = (V, E)$ where*

- *the set of vertices V is the set of all the random variables involved in the model:*

$$V := Y = \{Y_j\}$$

- *the joint distribution of all variables can be factorized as*

$$p_\theta(Y) = \frac{1}{\kappa(\theta)} \prod_{c \in C(G)} f_\theta^c(Y_c)$$

where $C(G)$ stands for the set of all maximal cliques of G and Y_c stands for the set of vertices (variables) in clique c.

Remark. The functions f_θ^c from Definition A.2 are generally not pdf and the normalizing constant $\kappa(\theta)$ ensures that $p_\theta(Y)$ is a pdf.

A.2 Exponential family

A.2.1 Maximum likelihood inference

Proof of Proposition 2.6. Remind that the moment generating function of a rv V is defined as $m(z) = \mathbb{E}[e^{z^\intercal V}]$ and satisfies $m'(0) = \mathbb{E}(V)$. For the exponential family, consider the moment generating function of the sufficient statistics

$$m(z) := \mathbb{E}[e^{z^\intercal t(Y)}] = \int e^{z^\intercal t(y)} p_\theta(y) \, \mathrm{d}y = \int \exp[(z + \theta)^\intercal t(y) - a(y) - b(\theta)] \, \mathrm{d}y.$$

Because p_θ is a pdf, $e^{b(\theta)}$ is a normalizing constant, we have that

$$\int \exp[\theta^\intercal t(y) - a(y)] \, \mathrm{d}y = e^{b(\theta)}$$

so

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

and

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

The result follows from the fact that $m'(z) = b'(\theta + z)e^{b(z+\theta)-b(\theta)}$ so $m'(0) = b'(\theta)$. □

Proof of Proposition 2.7. Take the derivative of the log-likelihood

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

with respect to θ. □

A.2.2 Bayesian inference

Proof of Proposition 5.1. It suffices to note that, as $p(\theta)$ is a probability distribution, we have

$$\exp[c(\nu,\eta)] = \int \exp[\theta^\intercal \nu - \eta b(\theta)]\ \mathrm{d}\theta.$$

So

$$\begin{aligned}\int p(\theta)p(Y|\theta)\ \mathrm{d}\theta &= \int \exp[\theta^\intercal(\nu + t(Y)) - c(\nu,\eta) - (\eta+1)b(\theta) - a(Y)]\ \mathrm{d}\theta \\ &= \exp[-c(\nu,\eta) - a(Y)] \int \exp[\theta^\intercal(\nu + t(Y)) - (\eta+1)b(\theta)]\ \mathrm{d}\theta \\ &= \exp[-c(\nu,\eta) - a(Y)] \exp[c(\nu + t(Y), \eta + 1)].\end{aligned}$$

The term $\exp[-c(\nu,\eta) - a(Y)]$ then vanishes in the ratio (5.1). □

Lemme A.1 *If $\pi \sim \mathcal{D}(a)$ and $\gamma \sim \mathcal{G}am(b,c)$, then*

$$\mathbb{E}(\log \pi_k) = \psi_0(a_k) - \psi_0\left(\sum_\ell a_\ell\right), \quad \mathbb{E}(\gamma) = \frac{b}{c}, \quad \mathbb{E}(\log\gamma) = \psi_0(b) - \log(c)$$

where ψ_0 is the first derivative of the Γ function, known as the di-gamma function.

A.3 Latent variable models

A.3.1 Asymptotic variance

Proof of Proposition 2.9. First remind that, for a generic incomplete data model,

$$p_\theta(Y) = \int p_\theta(Y,Z)\ \mathrm{d}Z.$$

Similarly to $S_\theta(Y)$ and $S'_\theta(Y)$, we denote

$$S_\theta(Y,Z) = \partial_\theta \log p_\theta(Y,Z) \qquad S'_\theta(Y,Z) = \partial^2_{\theta^2} \log p_\theta(Y,Z)$$

and p'_θ (resp. p''_θ) the first (resp. second) derivative of p_θ with respect to θ. First, we have

$$\begin{aligned} S_\theta(Y) &= \frac{p'_\theta(Y)}{p_\theta(Y)} = \frac{\int p'_\theta(Y,Z)\ \mathrm{d}Z}{p_\theta(Y)} \qquad (A.1)\\ &= \frac{\int p_\theta(Y,Z) S_\theta(Y,Z)\ \mathrm{d}Z}{p_\theta(Y)} \qquad \text{(because } \partial_x f(x) = f(x)\partial_x \log f(x)\text{)} \\ &= \int p_\theta(Z|Y) S_\theta(Y,Z)\ \mathrm{d}Z = \mathbb{E}\left[S_\theta(Y,Z)|Y\right].\end{aligned}$$

Then we have

$$S'_\theta(Y) = \partial^2_{\theta^2} \log p_\theta(Y) = \frac{p_\theta(Y)p''_\theta(Y) - (p'_\theta(Y))(p'_\theta(Y))^\intercal}{p^2_\theta(Y)} \\ = \frac{\int \partial^2_{\theta^2} p_\theta(Y,Z)\,\mathrm{d}Z}{p_\theta(Y)} - \left(\frac{p'_\theta(Y)}{p_\theta(Y)}\right)\left(\frac{p'_\theta(Y)}{p_\theta(Y)}\right)^\intercal \tag{A.2}$$

Because of (A.1), we have that

$$\left(\frac{p'_\theta(Y)}{p_\theta(Y)}\right)\left(\frac{p'_\theta(Y)}{p_\theta(Y)}\right)^\intercal = \mathbb{E}\left[S_\theta(Y,Z)|Y\right]\mathbb{E}\left[S_\theta(Y,Z)|Y\right]^\intercal.$$

Furthermore, because

$$\partial^2_{x^2} f(x) = f(x)\partial^2_{x^2}\log f(x) + f(x)(\partial \log f(x))(\partial \log f(x))^\intercal,$$

we have that

$$\frac{\int \partial^2_{\theta^2} p_\theta(Y,Z)\,\mathrm{d}Z}{p_\theta(Y)} = \int \frac{p_\theta(Y,Z)}{p_\theta(Y)} S'_\theta(Y,Z)\,\mathrm{d}Z + \int \frac{p_\theta(Y,Z)}{p_\theta(Y)} S_\theta(Y,Z)S_\theta(Y,Z)^\intercal\,\mathrm{d}Z \\ = \mathbb{E}\left[S'_\theta(Y,Z)|Y\right] + \mathbb{E}\left[S_\theta(Y,Z)S_\theta(Y,Z)^\intercal|Y\right],$$

which completes the proof. □