Latent variable models in biology and ecology

Chapter 3: Hidden Markov Models

Sophie Donnet. INRAE

Master 2 MathSV. February 9, 2024

université PARIS-SACLAY

AgroParisTech
INSTITUT DES SCIENCES ET INDUSTRIES DU VIVANT ET DE L'ENVIRONNEMENT
PARIS INSTITUTE OF TECHNOLOGY FOR LIFE, FOOD AND ENVIRONMENTAL SCIENCES

Context

- Aim: Modelling linearly organized data $(y_t)_{t \geq 0}$
- **For example**:
  - Time series : observations are collected along time
  - Spatial data along a covariable gradient
  - Genomic applications where measurements are collected at places (*loci*) located along the genome.
- Introduce dependence between the $(Y_t)_{t \geq 0}$

Movement ecology i

Digital biotelemetry technologies are enabling the collection of bigger and more accurate data on the movements of free-ranging wildlife in space and time

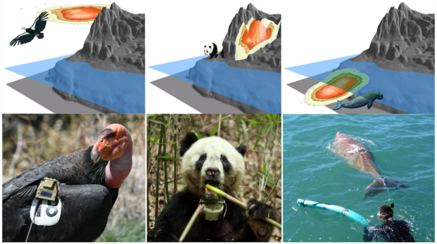


Figure 1: Examples of avian, terrestrial, and aquatic animal biotelemetry data sets and their spatial domains. Left: California condor with a GPS biologger attached to its patagium. Center: A giant panda telemetered with a GPS collar. Right: A dugong fitted with a tail mounted GPS biologger. [Tracey et al., 2014]

Movement ecology

- Y_t : characteristic of movement at time t.
  - Possibly multivariate: speed, depth, angular speed, etc...
- **Idea** : the value of this characteristic depends on the type of activity of the animal at time t: travelling, searching for food, sleeping...
- Let Z_t represent the behavior state at time t:

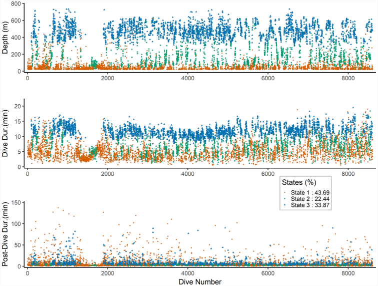
$$Y_t | Z_t = k \sim \mathcal{F}(\cdot, \gamma_k)$$

- Time dependeance in Z_t : Markov property

$$P(Z_t = z | Z_{t-1} = z_{t-1}, \ldots, Z_1 = z_1, Z_0 = z_0) = P(Z_t = z | Z_{t-1} = z_{t-1})$$

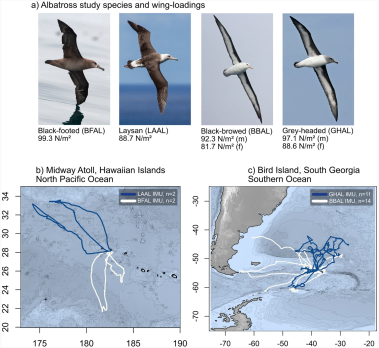
Narwhals [Ngô et al., 2019]

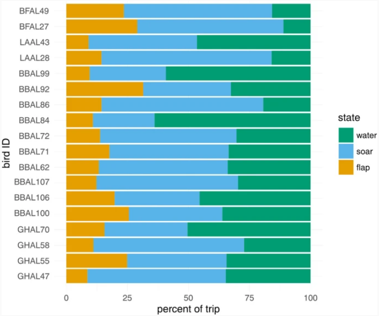
Understanding narwhal diving behaviour using Hidden Markov Models



Albatross [Conners et al., 2021]

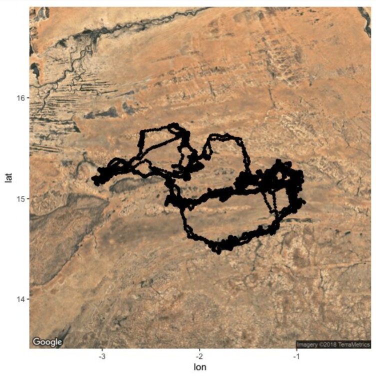
Hidden Markov models identify major movement modes in accelerometer and magnetometer data from four albatros species

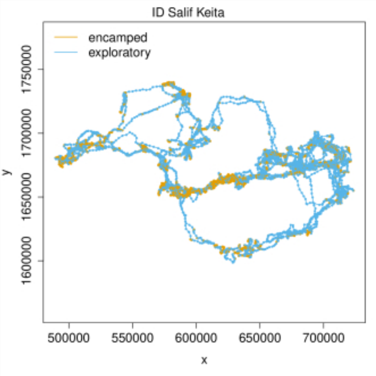




Other animals [McClintock and Michelot, 2018]

R-package for HMM inference. Trajectories of elephants, fur seal...





HMM for Human genetic

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

HMM model for population genetics

For each individual i and locus t, Z_{it} unknown population origins.

- In Chapter 1 : $(Z_{it})_t$ are independant
- Here, one may assume that the popupulation origins at locus t depends of the one at locus $t-1$.
- Dependency between neighbor loci

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

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

About Markov Chains

Z_t is a Markov Chain on the finite state space $[1, K]$: $Z_t \sim \mathrm{MC}(\nu, \pi)$ where

- $\nu = (\nu_1, \ldots, \nu_K)$ with $\nu_k = P(Z_1 = k)$ (initial distribution)
- π is the $K \times K$ transition matrix:

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

A few properties

- Let $\nu_t = (\nu_{t1}, \ldots, \nu_{tK})$ be the distribution of the hidden state at time t: $\nu_{tk} = P(Z_t = k)$. Then, (Z_t) being an homogeneous Markov chain, we have

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

- If (Z_t) is a stationary Markov chain i.e. $\nu = \nu^\intercal \pi$. then

$$\nu_t = \nu, \forall t.$$

Model

Definition

The general hidden Markov chain model is defined as follows:

$$\begin{aligned} (Z_t)_t &\sim \mathrm{MC}(\nu, \pi), \\ (Y_t)_t \text{ indep. } |(Z_t), \qquad Y_t|(Z_t = k) &\sim F_k = F(\gamma_k), \end{aligned} \tag{1}$$

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

Parameters: $\theta = (\nu, \pi, \gamma)$

About the emission distribution

- Must be adapted to the data one wants to modelize
- If $Y_t \in \mathbb{R}^d$: multivariate gaussian

$$Y_t | Z_t = k \sim \mathcal{M}_d(\mu_k, \Sigma_k)$$

- If Y_t is a speed : gamma distribution.
- If Y_t is a speed which can be null : gamma distribution and Dirac mass.
- If Y_t is an angular speed : adapted distribution!

Marginal distribution of Y_t

$$Y_t \sim \sum_{k=1}^{K} \nu_{tk} f(\cdot; \gamma_k).$$

Indeed:

$$p(Y_t) \quad = \quad \sum_{k=1}^{K} p(Y_t | Z_t = k) P(Z_t = k) = \sum_{k=1}^{K} f(Y_t; \gamma_k) \nu_{tk}$$

If (Z_t) is stationnary i.e. $\nu_{tk} = \nu_k$ then: $Y_t \sim \sum_{k=1}^{K} \nu_k f(\cdot; \gamma_k)$.

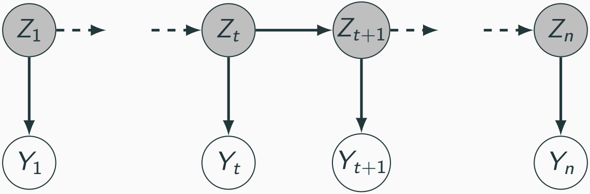
Dependencies structure

We do not have the same independancy properties as in the mixture model.

Useful notations:

- $Z_s^t = (Z_s, \dots Z_t)$ (for $s \leq t$)
- $Y_s^t = (Y_s, \dots Y_t)$

DAG of HMM



Factorisation

$$\mathbb{P}(Y_1, \ldots, Y_n, Z_1, \ldots, Z_n) = \prod_{t=1}^{n} \mathbb{P}(Y_t | Z_t)\, \mathbb{P}(Z_1) \prod_{t=1}^{n-1} \mathbb{P}(Z_{t+1} | Z_t)$$

About directed acyclic graphes (DAG)

See book by Lauritzen or see here for an introduction to DAG.

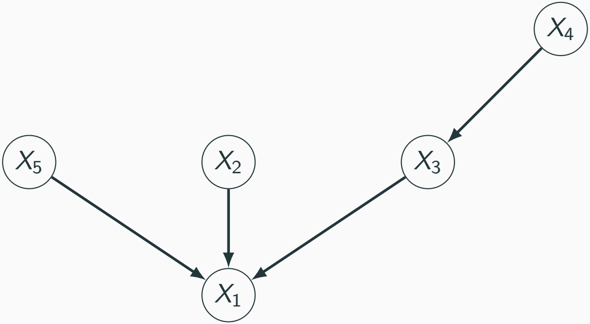
Factorized distributions

Let $\mathcal{V} = (V_1, \ldots, V_N)$ be a set of dependant random variables with joint distribution $\mathbb{P}$ and let $\mathcal{G} = (\mathcal{V}, E)$ be a directed acyclic graphe. $\mathbb{P}$ is said to be factorized with respect to $\mathcal{G}$ if

$$\mathbb{P}(V_1, \ldots, V_N) = \prod_{i=1}^{N} \mathbb{P}(V_i | Pa(V_i, \mathcal{G}))$$

where $Pa(V_i, \mathcal{G})$ denotes the parents of node V_i in $\mathcal{G}$.

Other example

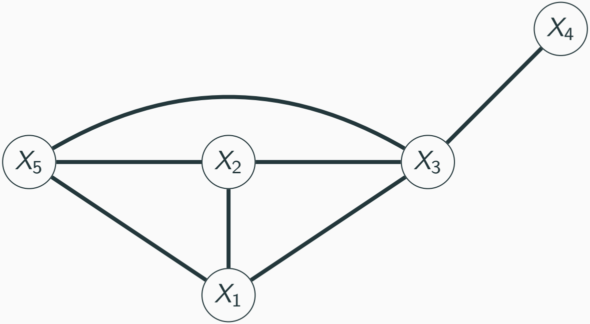


$$\mathbb{P}(X_1, X_2, X_3, X_4, X_5) = \mathbb{P}(X_1|X_2, X_3, X_5)\mathbb{P}(X_2)\mathbb{P}(X_5)\mathbb{P}(X_3|X_4)\mathbb{P}(X_4)$$

Moralization of a DAG

Moral graphe

The moral version of a graphe $\mathcal{G}$ is obtained by marrying the parents and by removing the directions on the edges.



Moralization of a DAG

Independancy properties

Let I, J and K, 3 subsets of $\mathcal{V}$.

1. In the moral graph deduced from $\mathcal{G}$, if all the paths from I to J pass through K then
$$(X_i)_{i\in I} \perp\!\!\!\perp (X_j)_{j\in J}|(X_k)_{k\in K}.$$
2. In a DAG, conditionnally to its parents, a variable is independant from its non-descendant.

Consequence of 1.

$$P(X_I|X_J,X_K) = \frac{P(X_I,X_J|X_K)}{P(X_J|X_K)} = \frac{P(X_I|X_K)P(X_J|X_K)}{P(X_J|X_K)} = P(X_I|X_K)$$

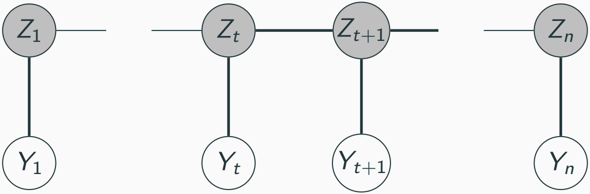
Example

1. $I = \{5, 2, 1\}, J = \{4\}, K = \{3\}$

$$P(X_5, X_2, X_1, X_4 | X_3) = P(X_5, X_2, X_1 | X_3) P(X_4 | X_3)$$

2. $P(X_1 | X_2, X_3, X_4, X_5) = P(X_1 | X_2, X_3, X_5)$

Application for HMM



1. $p(Z_{t+1}|Y_1^t, Z_1^t) = p(Z_{t+1}|Z_t)$
2. $p(Z_{t+1}|Z_1^t) = p(Z_{t+1}|Z_t)$

$$I = \{Y_{t+1}\}, K = \{Z_{t+1}\}, J = \{Z_1^{t-1}, , Y_1^t\}$$

3. $p(Y_{t+1}|Y_1^t, Z_1^{t+1}) = p(Y_{t+1}|Z_{t+1})$

Application for HMM

Consequences

(*a*) all paths from Y_1^t to Z_{t+1} go through $Z_1^t \Rightarrow Z_{t+1}$ is independent from Y_1^t conditionally on Z_1^t

$$p(Z_{t+1}|Y_1^t, Z_t) = p(Z_{t+1}|Z_t)$$

(*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^{t+1} go through Z_{t+1} meaning that Y^{t+1} is independent from Y_1^t to conditionally on Z_{t+1}

$$p(Y_{t+1}|Y_1^t, Z_{t+1}) = p(Y_{t+1}|Z_{t+1})$$

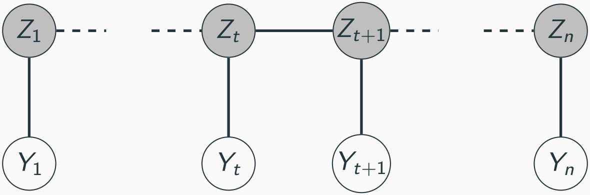
Property

Proposition

(Z_t) conditional on the observed data $\mathbf{Y} = Y_1^n$ is still a Markov chain.
<u>And</u>

$$p(Z_{t+1}|Z_1^t, Y_1^n) = p(Z_{t+1}|Z_t, Y_{t+1}^n)$$

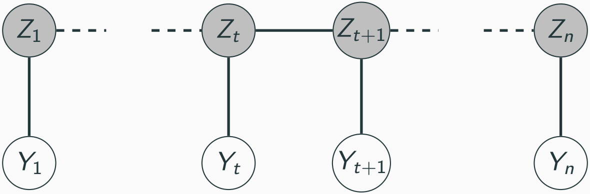
Proof (i)



Use DAG properties (Exercice) or :

$$
\begin{aligned}
p(Z_{t+1}|Z_1^t, Y_1^n)1 &= p(Z_{t+1}|Z_1^t, Y_1^t, Y_{t+1}^n) = \frac{p(Z_{t+1}, Z_1^t, Y_1^t, Y_{t+1}^n)}{p(Z_1^t, Y_1^t, Y_{t+1}^n)} \\
&= \frac{p(Y_{t+1}^n | Y_1^t, Z_{t+1}, Z_1^t) p(Y_1^t, Z_{t+1}, Z_1^t)}{p(Y_{t+1}^n | Z_1^t, Y_1^t) p(Z_1^t, Y_1^t)} \\
&= \frac{p(Y_{t+1}^n | Z_{t+1}) p(Y_1^t | Z_{t+1}, Z_1^t) p(Z_{t+1}|Z_1^t) p(Z_1^t)}{p(Y_{t+1}^n | Z_1^t, Y_1^t) p(Y_1^t | Z_1^t) p(Z_1^t)}
\end{aligned}
$$

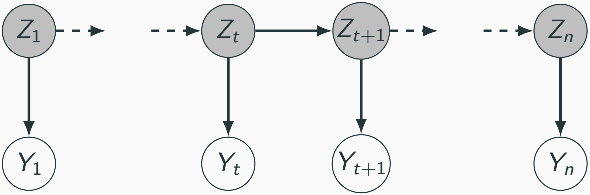
Proof (ii)



But $p(Y_1^t|Z_{t+1}, Z_1^t) = p(Y_1^t|Z_1^t)$ So

$$p(Z_{t+1}|Z_1^t, Y_1^n) \quad = \quad \frac{p(Y_{t+1}^n|Z_{t+1})p(Y_1^t|Z_{t+1}, Z_1^t)p(Z_{t+1}|Z_t)}{p(Y_{t+1}^n|Z_1^t, Y_1^t)p(Y_1^t|Z_1^t)}$$

Proof (iii)



Moreover

$$
\begin{aligned}
p(Y_{t+1}^n | Z_1^t, Y_1^t) &= \sum_{k=1}^{K} p(Y_{t+1}^n | \cancel{Z_1^t, Y_1^t}, Z_{t+1} = k) p(Z_{t+1} = k | Z_{\cancel{1}}^t, \cancel{Y_1^t}) \\
&= p(Y_{t+1}^n | Z_t)
\end{aligned}
$$

Proof (iv)

Finally:

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

Complete log-likelihood

Notations: $\mathbf{Y} = Y_1^n, \quad \mathbf{Z} = Z_1^n$

$$
\begin{aligned}
\log p_\theta(\mathbf{Y}, \mathbf{Z}) &= \log\left[p_\theta(\mathbf{Z}) p_\theta(\mathbf{Y}|\mathbf{Z})\right] \\
&= \log p_\theta(Z_1) p_\theta(Y_1|Z_1) + \sum_{t=2}^{n} \left[\log p_\theta(Z_t|Z_{t-1}) + \log p_\theta(Y_t|Z_t)\right] \\
&= \sum_{k=1}^{K} Z_{1k} \log \nu_k + \sum_{t=2}^{n} \sum_{k,\ell=1}^{K} Z_{t-1,k} Z_{t,\ell} \log \pi_{k\ell} \\
&\quad + \sum_{t=1,k=1}^{n,K} Z_{tk} \log f(Y_t; \gamma_k).
\end{aligned}
$$

Marginal (or 'observed') log-likelihood

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

EM algorithm : reminder

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

Algorithm (EM)

Repeat until convergence:

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

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

E-step: compute $\mathbb{E}_{\theta^{(h)}}[\log p_\theta(Y, Z)|Y]$

Using Slide 33

$$\mathbb{E}[\log p_\theta(\mathbf{Y}, \mathbf{Z})|\mathbf{Y}] = \mathbb{E}\left[\sum_{k=1}^{K} Z_{1k} \log \nu_k + \sum_{t=2}^{n} \sum_{k,\ell=1}^{K} Z_{t-1,k} Z_{t,\ell} \log \pi_{k\ell} |\mathbf{Y}\right]$$

$$+\mathbb{E}\left[\sum_{t=1,k=1}^{n,K} Z_{tk} \log f(Y_t; \gamma_k)|\mathbf{Y}\right].$$

$$= \sum_{k=1}^{K} \tau_{1k} \log \nu_k + \sum_{t=2}^{n} \sum_{k,\ell=1}^{K} \eta_{tk\ell} \log \pi_{k\ell} + \sum_{t=1,k=1}^{n,K} \tau_{tk} \log f(Y_t; \gamma_k)$$

where

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

Remark

As opposed to the mixture model:

$$\tau_{tk} = P(Z_t = k|\mathbf{Y}) \neq P(Z_t = k|Y_t)$$

More generally, $p(\mathbf{Z}|\mathbf{Y})$ does not factorize over t any more.

Foward- backward formulae

Proposition

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

- *Forward (for $t = 1, \ldots, n$): Denoting $F_{tk} = P_\theta(Z_t = k | Y_1^t)$ compute*

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

  such that, for all t : $\sum_{k=1}^{K} F_{t\ell} = 1$.

- *Backward (for $t = n, \ldots, 1$)*

$$\begin{aligned} \tau_{nk} &= P(Z_n = k | \mathbf{Y}) = P_\theta(Z_n = k | Y_1^n) = F_{nk} \\ G_{t+1,\ell} &= \sum_{k=1}^{K} \pi_{k\ell} F_{tk}, \qquad \eta_{tk\ell} = \pi_{k\ell} \frac{\tau_{t+1,\ell}}{G_{t+1,\ell}} F_{tk}, \qquad \tau_{tk} = \sum_{\ell=1}^{K} \eta_{tk\ell}. \end{aligned}$$

Proof of the Forward formula i

For $t = 1$

$$\begin{aligned} F_{1\ell} &= P(Z_1 = \ell | Y_1) \\ &= p(Y_1 | Z_1 = \ell) P(Z_1 = \ell) / p(Y_1) \\ &\propto \nu_\ell f_\ell(Y_1) \quad (F1) \end{aligned}$$

by the Bayes Formula.

Proof of the Forward formula ii

For $t \geq 2$

$$
\begin{aligned}
F_{t\ell} &= P(Z_t = \ell | Y_1^t) \quad = \sum_{k=1}^{K} P(Z_{t-1} = k, Z_t = \ell | Y_1^t) \\
&= \sum_{k=1}^{K} \frac{p(Z_t = \ell, Z_{t-1} = k, Y_1^t)}{p(Y_1^t)} \\
&= \sum_{k=1}^{K} \frac{\overbrace{p(Y_1^{t-1})}^{\perp\!\!\!\perp k} \overbrace{P(Z_{t-1} = k | Y_1^{t-1})}^{F_{t-1k}} \overbrace{P(Z_t = \ell | Z_{t-1} = k)}^{\pi_{k,\ell}} \overbrace{p(Y_t | Z_t = \ell)}^{\perp\!\!\!\perp k \text{ and } = f_\ell(Y_t)}}{p(Y_1^t)} \\
&\quad \text{(using conditional independences, from the past to present } t\text{)} \\
&= \frac{p(Y_1^{t-1})}{p(Y_1^t)} f_\ell(Y_t) \sum_{k=1}^{K} \pi_{k\ell} F_{t-1,k} \\
F_{t\ell} &= P(Z_t = \ell | Y_1^t) \propto f_\ell(Y_t) \sum_{k=1}^{K} \pi_{k\ell} F_{t-1,k} \quad (F2)
\end{aligned}
$$

About the normalizing constant i

Note that

$$\sum_{\ell=1}^{K} F_{t\ell} = \sum_{\ell=1}^{K} P(Z_t = \ell | Y_1^t) = 1$$

So

$$\begin{aligned}
& \sum_{\ell=1}^{K} \frac{p(Y_1^{t-1})}{p(Y_1^t)} f_\ell(Y_t) \sum_{k=1}^{K} \pi_{k\ell} F_{t-1,k} = 1 \\
\Leftrightarrow \quad & \frac{p(Y_1^{t-1})}{p(Y_1^t)} \sum_{\ell=1}^{K} f_\ell(Y_t) \sum_{k=1}^{K} \pi_{k\ell} F_{t-1,k} = 1 \\
\Leftrightarrow \quad & \frac{p(Y_1^t)}{p(Y_1^{t-1})} = \sum_{\ell=1}^{K} f_\ell(Y_t) \sum_{k=1}^{K} \pi_{k\ell} F_{t-1,k}
\end{aligned}$$

About the normalizing constant ii

$$\frac{p(Y_1^t)}{p(Y_1^{t-1})} = \frac{p(Y_1^{t-1}, Y_t)}{p(Y_1^{t-1})} = p(Y_t | Y_1^{t-1})$$

Useful formula

$$p(Y_t | Y_1^{t-1}) = \sum_{\ell=1}^{K} f_\ell(Y_t) \sum_{k=1}^{K} \pi_{k\ell} F_{t-1,k} \quad (2)$$

▸ Use of the formula to compute the marginal likelihood

Proof of the Backward formula i

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

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

and the recursion follows as: for $t \leq n-1$

$$\tau_{tk} = P(Z_t = k|Y_1^n) = \sum_{\ell=1}^{K} \underbrace{P(Z_t = k, Z_{t+1} = \ell|Y_1^n)}_{\eta_{tk\ell}} = \sum_{\ell=1}^{K} \eta_{tk\ell} \quad (B3)$$

$$\eta_{tk\ell} = \frac{\overbrace{P(Z_t = k, Z_{t+1} = \ell, Y_1^n)}^{(\bullet)}}{p(Y_1^n)}$$

Proof of the Backward formula ii

with

$$(\bullet) = P(Z_t = k, Z_{t+1} = \ell, Y_1^n) = P(Z_t = k, Z_{t+1} = \ell, Y_1^t, Y_{t+1}^n)$$

$$= p(Y_{t+1}^n | Z_{t+1} = \ell, \cancel{Z_t = k}, \cancel{Y_1^n}) \overbrace{p(Z_{t+1} = \ell | Z_t = k, \cancel{Y_1^t})}^{\pi_{k\ell}} \underbrace{P(Z_t = k | Y_1^t)}_{=F_{tk}} p(Y_1^t)$$

And so:

$$\eta_{tk\ell} = \frac{(\bullet)}{p(Y_1^n)} = \pi_{k\ell} \overbrace{\frac{p(Y_1^t) p(Y_{t+1}^n | Z_{t+1} = \ell)}{p(Y_1^n)}}^{(\bullet)} F_{tk} \qquad (\approx B2)$$

Proof of the Backward formula iii

and

$$
\begin{aligned}
(\bullet) &= \frac{p(Y_1^t)p(Y_{t+1}^n|Z_{t+1}=\ell)}{p(Y_1^n)} \\
&= \frac{p(Y_1^t)p(Y_{t+1}^n|Z_{t+1}=\ell)}{p(Y_1^n)} \frac{p(Y_1^t|Z_{t+1}=\ell)}{p(Y_1^t|Z_{t+1}=\ell)} \\
&= \frac{p(Y_1^t)p(Y_1^n|Z_{t+1}=\ell)}{p(Y_1^n)p(Y_1^t|Z_{t+1}=\ell)}
\end{aligned}
$$

Because $p(Y_1^n|Z_{t+1}=\ell) = p(Y_{t+1}^n|\cancel{Y_1^t}, Z_{t+1}=\ell)p(Y_1^t|Z_{t+1}=\ell)$

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

Proof of the Backward formula iv

$$\eta_{tk\ell} = \pi_{kl} \frac{\tau_{t+1,\ell}}{P(Z_{t+1} = \ell | Y_1^t)} F_{tk} \qquad (\approx B2)$$

Now

$$\begin{aligned}
P(Z_{t+1} = \ell | Y_1^t) &= \sum_{k=1}^{K} P(Z_{t+1} = \ell, Z_t = k | Y_1^t) \\
&= \sum_{k=1}^{K} P(Z_{t+1} = \ell | Z_t = k, \cancel{Y_1^t}) P(Z_t = k | Y_1^t) \\
&= \sum_{k=1}^{K} \pi_{k\ell} F_{tk} =: G_{t+1,\ell} \qquad (B1)
\end{aligned}$$

Remarks on the EM Forward Backward

1. The formula is a double recursion
2. Computational complexity : $O(nK^2)$.

M-step

Assume that τ_{tk} and $\eta_{tk\ell}$ have been calculated by the FB algorithm. Now we have to find

$$\arg\max_{(\nu,\pi,\gamma)} \mathbb{E}_{\theta^{(h)}}[\log p_\theta(Y,Z)|Y]$$

where

$$\begin{aligned}\mathbb{E}_{\theta^{(h)}}[\log p_\theta(Y,Z)|Y] &= \sum_{k=1}^{K} \tau_{1k} \log \nu_k + \sum_{t=2}^{n} \sum_{k,\ell=1}^{K} \eta_{tk\ell} \log \pi_{k\ell} \\ &+ \sum_{t=1,k=1}^{n,K} \tau_{tk} \log f(Y_t;\gamma_k)\end{aligned}$$

and

$$\sum_{k=1}^{K} \nu_k = 1 \quad \text{and} \quad \sum_{\ell=1}^{K} \pi_{k\ell} = 1, \qquad \forall k = 1,\dots,K$$

M-step i

Lagrange multipliers:

$$\begin{aligned}
&\sum_{k=1}^{K} \tau_{1k} \log \nu_k + \sum_{t=2}^{n} \sum_{k,\ell=1}^{K} \eta_{tk\ell} \log \pi_{k\ell} + \sum_{t=1,k=1}^{n,K} \tau_{tk} \log f(Y_t; \gamma_k) \\
- \quad &\lambda_0 \left(\sum_{k=1}^{K} \nu_k - 1 \right) - \sum_{k=1}^{K} \lambda_k \left(\sum_{\ell=1}^{K} \pi_{k\ell} - 1 \right)
\end{aligned}$$

implies:

$$\begin{aligned}
\frac{\tau_{1k}}{\nu_k} - \lambda_0 &= 0, \quad \forall k = 1, \dots, K \\
\frac{\sum_{t=2}^{n} \eta_{tk\ell}}{\pi_{k\ell}} - \lambda_k &= 0, \quad \forall k, \ell = 1, \dots, K
\end{aligned}$$

M-step ii

So:

$$\begin{aligned}\widehat{\nu}_k &= \frac{\tau_{1k}}{\lambda_0}, \quad \forall k = 1, \ldots, K \\ \widehat{\pi}_{k\ell} &= \frac{\sum_{t=2}^{n} \eta_{tk\ell}}{\lambda_k}, \quad \forall k, \ell = 1, \ldots, K\end{aligned}$$

Using the constraints we get:

- $\sum_{k=1}^{K} \widehat{\nu}_k = 1 = \sum_{k=1}^{K} \frac{\tau_{1k}}{\lambda_0}$. But $\sum_{k=1}^{K} \tau_{1k} = 1$ so $\lambda_0 = 1$.

M-step iii

- For all $k = 1, \ldots, K$,

$$1 = \sum_{\ell=1}^{K} \widehat{\pi}_{k\ell} = \sum_{\ell=1}^{K} \frac{\sum_{t=2}^{n} \eta_{tk\ell}}{\lambda_k} = \frac{1}{\lambda_k} \sum_{t=2}^{n} \underbrace{\sum_{\ell=1}^{K} \eta_{tk\ell}}_{=\tau_{tk}}$$

And so $\lambda_k = \sum_{t=2}^{n} \tau_{tk}$,

$$\widehat{\pi}_{k\ell} = \frac{\sum_{t=2}^{n} \eta_{tk\ell}}{\sum_{t=2}^{n} \tau_{tk}}, \quad \forall k, \ell = 1, \ldots, K$$

M-step:γ i

If $\mathcal{F}$ belongs to the exponential family

$$\log f_k(Y_t;\gamma_k) = \gamma_k^{\intercal} t_k(Y_t) - a_k(Y_t) - b_k(\gamma_k)$$

So:

$$\begin{aligned}
\frac{\partial}{\partial \gamma_k} \sum_{t=1,k=1}^{n,K} \tau_{tk} \log f(Y_t;\gamma_k) &= 0 \\
\frac{\partial}{\partial \gamma_k} \sum_{t=1}^{n} \tau_{tk} \left[\gamma_k^{\intercal} t_k(Y_t) - a_k(Y_t) - b_k(\gamma_k)\right] &= 0 \\
\sum_{t=1}^{n} \tau_{tk} \left[t_k(Y_t) - b_k'(\gamma_k)\right] &= 0
\end{aligned}$$

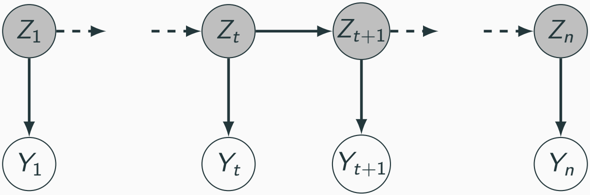
$$b_k'(\gamma_k) = \frac{\sum_{t=1}^{n} \tau_{tk} t_k(Y_t)}{\sum_{t=1}^{n} \tau_{tk}}$$

Remark i

EM fo HMM : Baum–Welch algorithm

Prediction of Z_{t+1} given $Y_1^{t+1}, Z_t = k$

About $P(Z_{t+1} = \ell | Y_1^{t+1}, Z_t = k)$



$$
\begin{aligned}
P(Z_{t+1} = \ell | Y_1^{t+1}, Z_t = k) &= P(Z_{t+1} = \ell | Y_{t+1}, Z_t = k) \\
&\propto P(Y_{t+1} | Z_{t+1} = \ell, \cancel{Z_t = k}) P(Z_{t+1} = \ell | Z_t = k) \\
&\propto f_\ell(Y_{t+1}) \pi_{k\ell} \\
&= \frac{\pi_{k\ell}}{\sum_{j=1}^{K} \pi_{kj} f_j(Y_{t+1})}
\end{aligned}
$$

Conditional on Y_1^{t+1}, the transitions $\pi_{k\ell}$ are biased according to the likelihood of the data under the arrival state $f_\ell(Y_{t+1})$.

BIC

$$BIC(K) = \log p_{\widehat{\theta_K}}(\mathbf{Y}) - \frac{d_K}{2} \log n$$

where

- n: indicates the length/size of the observation time-series
- d_K: number of free parameters:

$$d_K = \underbrace{K^2 - K}_{\boldsymbol{\pi}} + \sum_{k=1}^{K} \dim(\gamma_k) + \underbrace{(K-1)}_{\nu}$$

Computation of the marginal likelihood

$$\log p_\theta(\mathbf{Y}) = \log p_\theta(Y_1) + \sum_{t\geq 2} \log p_\theta(Y_t|Y_1^{t-1}).$$

Equation (2) which gives explicit formula of $p(Y_t|Y_1^{t-1})$

$$p(Y_t|Y_1^{t-1}) = \sum_{\ell=1}^{K} f_\ell(Y_t) \sum_{k=1}^{K} \pi_{k\ell} F_{t-1,k}$$

By product of the EM algorithm: can be computed by storing the adequate quantities in the forward step

From BIC to Integrated Complete Likelihood (ICL)

- BIC focus on the fit to the data.
- In classification problems, interesting to have a classification that separates well the observations.
- Entropy $\mathcal{H}[p_{\widehat{\theta_K}}(\mathbf{Z}|\mathbf{Y})]$ is small when the observations are classified with reasonable confidence.
- [Biernacki et al., 2000]: account for the classification uncertainty in the selection of K
- Penalize value of K with large entropy

Definition (ICL)

$$\widehat{K}_{ICL} = \arg\max_K \left(\log p_{\widehat{\theta_K}}(Y) - \mathcal{H}[p_{\widehat{\theta_K}}(\mathbf{Z}|\mathbf{Y})] - \frac{d_K}{2} \log n \right)$$

Computing ICL

Using Proposition from chap 2

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

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

$\mathbb{E}_{\widehat{\theta}_K}\left[\log p_{\widehat{\theta}_K}(\mathbf{Y},\mathbf{Z})|\mathbf{Y}\right]$: Forward Backward algorithm

About the conditional entropy i

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

Here

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

- $$\begin{aligned} \mathbb{E}[\log p(Z_1|\mathbf{Y})] &= \sum_k P(Z_1 = k|\mathbf{Y}) \log P(Z_1 = k|\mathbf{Y}) \\ &= \sum_k \tau_{1k} \log \tau_{1k} \end{aligned}$$

About the conditional entropy ii

- Using $p(Z_t|Z_{t-1},\mathbf{Y}) = p(Z_t, Z_{t-1}|\mathbf{Y})/p(Z_{t-1}|\mathbf{Y})$,

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

- Finally,

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

By product of the backward step of the E-step

MAP using the marginal

A classification at each position t can be defined based on the MAP rule applied to the marginal distribution of each label given the data:

$$\widehat{Z}_t = \underset{k=1,\ldots,K}{\arg\max}\, P(Z_t = k|\mathbf{Y}) = \underset{k=1,\ldots,K}{\arg\max}\, P(Z_t = k|Y_1^n) = \underset{k=1,\ldots,K}{\arg\max}\, \tau_{tk}.$$

Really easy.

Joint MAP

- Because of the conditional dependencies:

$$\underset{\mathbf{z}\in\{1,\dots,K\}^n}{\arg\max}\ P(\mathbf{Z}=\mathbf{z}|Y_1^n) \neq \left(\underset{k\in\{1,\dots,K\}}{\arg\max}\ P(Z_t=k|Y_1^n)\right)_{t=1,\dots,n}$$

- Most probable hidden path given the observations:

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

- Finding a MAP in $\{1,\dots,K\}^n$ much more difficult.

Joint MAP: Viterbi algorithm

Proposition

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* $t < n$*:*

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

Demonstration of Viterbi i

First note that

$$\arg\max_{\mathbf{z}} P(\mathbf{Z} = \mathbf{z} | \mathbf{Y}) = \arg\max_{\mathbf{z}} \frac{P(\mathbf{Z} = \mathbf{z}, \mathbf{Y})}{P(\mathbf{Y})} = \arg\max_{\mathbf{z}} p(\mathbf{Z} = \mathbf{z}, \mathbf{Y})$$

Forward recursion: Succession of optimal choices as for the hidden label at the preceding times, so that

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

and, finally,

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

Demonstration of Viterbi ii

$$
\begin{aligned}
V_{t\ell} &= \max_{z_1^{t-1}} \; p(Z_1^{t-1} = z_1^{t-1}, z_t = \ell, Y_1^t) \\
&= \max_k \max_{z_1^{t-2}} p(Z_1^{t-2} = z_1^{t-2}, Z_{t-1} = k, Z_t = \ell, Y_1^{t-1}, Y_t) \\
&= \max_k \max_{z_1^{t-2}} p(Y_t | \cancel{Z_1^{t-2} = z_1^{t-2}}, \cancel{Z_{t-1} = k}, Z_t = \ell, \cancel{Y_1^{t-1}}) \\
&\qquad\qquad p(Z_t = \ell | \cancel{Z_1^{t-2} = z_1^{t-2}}, Z_{t-1} = k, \cancel{Y_1^{t-1}}) \\
&\qquad\qquad p(Z_1^{t-2} = z_1^{t-2}, Z_{t-1} = k, Y_1^{t-1}) \\
&= \max_k \underbrace{\max_{z_1^{t-2}} p(Z_1^{t-2} = z_1^{t-2}, Z_{t-1} = k, Y_1^{t-1})}_{= V_{t-1\,k}} \\
&\qquad p(Y_t | Z_t = \ell) p(Z_t = \ell | Z_{t-1} = k) \\
&= \max_k V_{t-1\,k} \pi_{k\ell} f_\ell(Y_t)
\end{aligned}
$$

Demonstration of Viterbi iii

Backward recursion

- $$\begin{aligned} \widehat{Z}_n &= \arg\max_k V_{nk} = \arg\max_k \max_{z_1^{n-1}} p(Z_1^{n-1} = z_1^{n-1}, z_n = k, Y_1^{n}) \\ &= \arg\max_k \max_{z_1^{n-1}} p(Z_1^{n-1} = z_1^{n-1}, z_n = k, \mathbf{Y}) \\ &= \arg\max_k \max_{z_1^{n-1}} p(Z_1^{n-1} = z_1^{n-1}, z_n = k | \mathbf{Y}) \end{aligned}$$

Demonstration of Viterbi iv

- For $n-1$: $\widehat{Z}_{n-1} = S_{n-1}(\widehat{Z}_n)$. So:

$$
\begin{aligned}
\widehat{Z}_{n-1} &= S_{n-1}(\widehat{Z}_n) \\
&= \arg\max_k \; V_{n-1,k}\pi_{k\widehat{Z}_n} f_{\widehat{Z}_n}(Y_n) \\
&= \arg\max_k \max_{z_1^{n-2}} \; p(Z_1^{n-2} = z_1^{n-2}, Z_{n-1} = k, Y_1^{n-1})\pi_{k\widehat{Z}_n} f_{\widehat{Z}_n}(Y_n) \\
&= \arg\max_k \max_{z_1^{n-2}} p(Z_1^{n-2} = z_1^{n-2}, Z_{n-1} = k, Z_n = \widehat{Z}_n, Y_1^n) \\
&= \arg\max_k \max_{z_1^{n-2}} p(Z_1^{n-2} = z_1^{n-2}, Z_{n-1} = k, Z_n = \widehat{Z}_n | \mathbf{Y})
\end{aligned}
$$

Demonstration of Viterbi v

- For $n-2$: $\widehat{Z}_{n-2} = S_{n-2}(\widehat{Z}_{n-1})$. So:

$$
\begin{aligned}
& \widehat{Z}_{n-2} = S_{n-2}(\widehat{Z}_{n-1}) \\
= \quad & \arg\max_k \; V_{n-2,k}\pi_{k\widehat{Z}_{n-1}} f_{\widehat{Z}_{n-1}}(Y_{n-1}) \\
= \quad & \arg\max_k \max_{z_1^{n-3}} \; p(Z_1^{n-3} = z_1^{n-3}, Z_{n-2} = k, Y_1^{n-2})\pi_{k\widehat{Z}_{n-1}} f_{\widehat{Z}_{n-1}}(Y_{n-1}) \\
= \quad & \arg\max_k \max_{z_1^{n-3}} p(Z_1^{n-3} = z_1^{n-3}, Z_{n-2} = k, Z_{n-1} = \widehat{Z}_{n-1}, Y_1^{n-1}) \\
= \quad & \arg\max_k \max_{z_1^{n-3}} p(Z_1^{n-3} = z_1^{n-3}, Z_{n-2} = k, Z_{n-1} = \widehat{Z}_{n-1}, Y_1^{n-1}) \\
& \qquad\qquad f_{\widehat{Z}_n}(Y_n)\pi_{\widehat{Z}_{n-1}\widehat{Z}_n} \\
= \quad & \arg\max_k \max_{z_1^{n-3}} p(Z_1^{n-3} = z_1^{n-3}, Z_{n-2} = k, Z_{n-1} = \widehat{Z}_{n-1}, Z_n = \widehat{Z}_n, Y_1^n) \\
= \quad & \arg\max_k \max_{z_1^{n-3}} p(Z_1^{n-3} = z_1^{n-3}, Z_{n-2} = k, Z_{n-1} = \widehat{Z}_{n-1}, Z_n = \widehat{Z}_n|\mathbf{Y})
\end{aligned}
$$

Demonstration of Viterbi vi

The backward recursion traces back the succession of the optimal choices and retrieves the optimal path.

A few more details to understand i

The rational (for $n = 4$) behind this algorithm is that, for a function of the form

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

For us it would be

$$f_1(Z_1) = \log\left(\nu_{z_1} f_{z_1}(Y_1)\right),$$

$$f_t(z_{t-1}, z_t)) = \log\left(\pi_{z_{t-1}, z_t} f_{z_t}(Y_t)\right)$$

and

$$F(z_1^4) = \log p(z_1^4, Y_1^4)$$

A few more details to understand ii

We have the decomposition

$$\max_{z_1^4} F(z_1^4) = \max_{z_4} \left[\max_{z_3} \left(\max_{z_2} \left\{ \max_{z_1} [f_1(z_1) + f_2(z_1, z_2)] + f_3(z_2, z_3) \right\} + f_4(z_3, z_4) \right) \right]$$

$$= \max_{z_4} \left[\max_{z_3} \left(\max_{z_2} \left\{ F_1^2(z_2) + f_3(z_2, z_3) \right\} + f_4(z_3, z_4) \right) \right]$$

$$\text{where} \quad F_1^2(z_2) = \max_{z_1} f_1(z_1) + f_2(z_1, z_2)$$

$$= \max_{z_4} \left[\max_{z_3} \left(F_1^3(z_3) + f_4(z_3, z_4) \right) \right]$$

$$\text{where} \quad F_1^3(z_3) = \max_{z_2} F_1^2(z_2) + f_3(z_2, z_3)$$

$$= \max_{z_4} \left[F_1^4(z_4) \right]$$

$$\text{where} \quad F_1^4(z_4) = \max_{z_3} F_1^3(z_3) + f_4(z_3, z_4)$$

A few more details to understand iii

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 on Viterbi computational details

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

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

Kalman Filter

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

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

- with
  - $E = (E_t)$ 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 $\gamma = (\beta, \sigma^2)$.

Estimation

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.

- E step: compute conditional mean and variance of the Z_t's, which can be derived using standard results on Gaussian vectors.
- M step results in (weighted) linear regression estimates (see [Ghahramani and Hinton, 1996])

Conclusion

- From Mixture models to HMM : more dependence in the latent variable
- More complexe but still explicit.
- R packages HiddenMarkov
- Next chapter : more complexe dependencies SBM

References

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

Conners, M. G., Michelot, T., Heywood, E. I., Orben, R. A., Phillips, R. A., Vyssotski, A. L., Shaffer, S. A., and Thorne, L. H. (2021).
Hidden Markov models identify major movement modes in accelerometer and magnetometer data from four albatross species.
Movement Ecology, 9(1):7.

Ghahramani, Z. and Hinton, G. E. (1996).
Parameter estimation for linear dynamical systems.

McClintock, B. T. and Michelot, T. (2018).
momentuhmm: R package for generalized hidden markov models of animal movement.
Methods in Ecology and Evolution, 9(6):1518–1530.

Ngô, M. C., Heide-Jørgensen, M. P., and Ditlevsen, S. (2019).
Understanding narwhal diving behaviour using hidden markov models with dependent state distributions and long range dependence.
PLOS Computational Biology, 15(3):1–21.

Tracey, J. A., Sheppard, J., Zhu, J., Wei, F., Swaisgood, R. R., and Fisher, R. N. (2014).
Movement-based estimation and visualization of space use in 3d for wildlife ecology and conservation.
PLOS ONE, 9(7):1–15.