Multiple Markov models

42. Finnish summer school on probability and statistics

May, 27-31, 2024, Lammi

Jüri Lember

University of Tartu

Hidden Markov Model (HMM)

Y – Markov Chain with finite state space $\mathcal{Y}$
Y is sometimes called as the regime.

To each state $l \in \mathcal{Y}$ corresponds an emission distribution on a measurable space (not necessarily discrete) $\mathcal{X}$ with densities $p(\cdot|l)$ w.r.t. some reference measure λ.

HMM:

To any realization $y_1, y_2, \ldots$ of Y corresponds a sequence of independent random variables $X_1, X_2, \ldots$, where $X_t \sim p(\cdot|y_t)$.

A realization $x_1, \ldots, x_n$ of $X_1, \ldots, X_n$ – observations; the corresponding realization $y_1, \ldots, y_n$ of $Y_1, \ldots, Y_n$ is typically hidden.

Key properties of HMM

- (X, Y) is a two-dimensional Markov chain;
- (hidden process) Y is a Markov chain;
- The distribution of X_t depends on Y_t, only;
- Observations $X_1, X_2, \ldots$ are conditionally independent (given realization of Y).

Sometimes too restrictive, hence the generalizations.

Markov switching model and HMM-DN

Markov switching model: $Y_1, \ldots, Y_n$ is a (homogenuous) Markov chain, but the conditional distribution of X_t depends also on X_{t-1}, i.e.

$$X_t | Y_t = k, X_{t-1} = x_{t-1} \quad \text{has density} \quad p(\cdot | k, x_{t-1}).$$

Given $Y_1, \ldots, Y_n$ the observations $X_1, \ldots, X_n$ are not (conditionally) independent any more; HMM is a special case.

HMM with dependent noise (HMM-DN): $Y_1, \ldots, Y_n$ is a (homogenuous) Markov chain, but the conditional distribution of X_t depends also on X_{t-1} and Y_{t-1}, i.e.

$$X_t | Y_t = k, X_{t-1} = x_{t-1}, Y_{t-1} = y_{t-1} \quad \text{has density} \quad p(\cdot | k, x_{t-1}, y_{t-1}).$$

Given $Y_1, \ldots, Y_n$ the observations $X_1, \ldots, X_n$ are not (conditionally) independent any more and X_t depends also on Y_{t-1}; Markov switching model is a special case.

Pairwise Markov model (PMM)

Pairwise Markov model: $(X_1, Y_1), (X_2, Y_2), \ldots$ is a (2-dim) Markov chain with state space $\mathcal{Z} \subseteq \mathcal{X} \times \mathcal{Y}$.

Formal definition of PMM. $\mathcal{X}$ Polish (separable completely metrizable) with Borel σ-field $\mathcal{B}(\mathcal{X})$ w.r.t. topology τ in $\mathcal{X}$ and $\mathcal{Y}$ is finite.

Equip $\mathcal{X} \times \mathcal{Y}$ with product topology $\tau \times 2^{\mathcal{Y}}$, then $\mathcal{B}(\mathcal{X} \times \mathcal{Y}) = \mathcal{B}(\mathcal{X}) \otimes 2^{\mathcal{Y}}$ (the smallest σ-field containing $A \times B$, where $A \in \mathcal{B}(\mathcal{X})$ and $B \in 2^{\mathcal{Y}}$).

Let μ be σ-finite measure on $\mathcal{B}(\mathcal{X})$, c be a counting measure on $2^{\mathcal{Y}}$, and $\mu \times c$ the product measure on $\mathcal{B}(\mathcal{X}) \otimes 2^{\mathcal{Y}}$.

Let $\mathcal{Z} \subseteq \mathcal{X} \times \mathcal{Y}$ be a measurable set – state space – thus $\mathcal{B}(\mathcal{Z})$ is the Borel σ-algebra on $\mathcal{Z}$ and $\mu \times c$ is the (restricted) measure on $(\mathcal{Z}, \mathcal{B}(\mathcal{Z}))$.

Any set $C \subset \mathcal{B}(\mathcal{Z}))$ has the form $C = \cup_j A_j \times \{j\}$, where $A_j \in \mathcal{B}(\mathcal{X})$ (easy to verify).

Formal definition of PMM. We have defined a measure space $(\mathcal{Z}, \mathcal{B}(\mathcal{Z}), \mu \times c)$. Let

$$p\colon \mathcal{Z}^2 \to [0, \infty), \quad (z', z) \mapsto p(z'|z)$$

be a measurable non-negative function such that for each $z \in \mathcal{Z}$ the function $z' \mapsto p(z'|z)$ is a probability density function with respect to product measure $\mu \times c$.

This function defines a Markov/transition kernel

$$P(z, C) := \int_C p(z'|z)\mu \times c(dz), \quad \forall C \in \mathcal{B}(\mathcal{Z}), \forall z \in \mathcal{Z}.$$

This means (easy to verify):

1) For every $C \in \mathcal{B}(Z)$, $P(\cdot, C)$ is a non-negative measurable function on $\mathcal{Z}$;
2) For every $z \in \mathcal{Z}$, $P(z, \cdot)$ is a probability measure on $(\mathcal{Z}, \mathcal{B}(\mathcal{Z}))$.

Therefore $p(z'|z)$ will be referred to as transition kernel density.

Formal definition of PMM. We have a transition kernel density $p(z'|z)$ on $\mathcal{Z}^2$, which defines a transition kernel $P(z, C)$.

A Pairwise Markov model (PMM) $Z = \{Z_t\}_{t\geq 1} = \{(X_t, Y_t)\}_{t\geq 1}$ is a homogeneous Markov chain on $\mathcal{Z}$ with transition kernel density $p(z'|z)$.

This means: for any $t \geq 1$, any $C \in \mathcal{B}(\mathcal{Z})$ and any $z \in \mathcal{Z}$.

$$P(Z_{t+1} \in C | Z_t = z) = \int_C p(z'|z)\mu \times c(dz).$$

Since $C = \cup_j A_j \times \{j\}$, the probability above reads

$$\begin{aligned} P(Z_{t+1} \in C | X_t = x, Y_t = i) &= \sum_{j\in\mathcal{Y}} P(X_{t+1} \in A_j, Y_{t+1} = j | X_t = x, Y_t = i) \\ &= \sum_{j\in\mathcal{Y}} \int_{A_j} p(x', j|x, i)\mu(dx'). \end{aligned}$$

Formal definition of PMM. We defined PMM as a (bivariate) homogeneous Markov chain $Z = (X, Y)$ with transition density $p(z'|z)$ and corresponding transition kernel $P(z, C)$.

We additionally assume that the distribution of Z_1 has density with respect to $\mu \times c$. Then (easy to verify), for every n, the vector $Z_{1:n} := (Z_1, \dots, Z_n)$ has a density with respect to $(\mu \times c)^n$. We shall denote use (generic) p for all (joint and conditional) densities (even when they are probabilities). Thus, $Z_{1:n}$ has density

$$p(z_{1:n}) = p(z_1)p(z_2|z_1)\cdots p(z_n|z_{n-1}) = p(x_1, y_1)p(x_2, y_2|x_1, y_1)\cdots p(x_n, y_n|x_{n-1}, y_{n-1}),$$

Z_t has density $p(z_t) = p(x_t, y_t)$, $Y_{1:n}$ has density (which actually is probability) $p(y_{1:n})$ etc.

When $\mathcal{X}$ is countable, then take $\tau = 2^{\mathcal{X}}$, $\mu = c$. Then Z is a homogeneous Markov chain with countable state space $\mathcal{Z}$ and transition matrix $p(z'|z)$ – a discrete PMM.

Classification of PMM's by factorization of kernel

Recall $p(x_t, y_t|x_{t-1}, y_{t-1})$ is the density of Markov kernel (w.r.t. $\mu \times \mathrm{c}$) .

- $p(x_t, y_t|x_{t-1}, y_{t-1}) = p(y_t|y_{t-1}, x_{t-1})p(x_t|y_t, x_{t-1}, y_{t-1})$ – general PMM
- $p(x_t, y_t|x_{t-1}, y_{t-1}) = p(y_t|y_{t-1})p(x_t|y_t, x_{t-1}, y_{t-1})$ – HMM-DN
- $p(x_t, y_t|x_{t-1}, y_{t-1}) = p(y_t|y_{t-1})p(x_t|y_t, x_{t-1})$ – Markov switching model
- $p(x_t, y_t|x_{t-1}, y_{t-1}) = p(y_t|y_{t-1})p(x_t|y_t)$ – HMM
- $p(x_t, y_t|x_{t-1}, y_{t-1}) = p(y_t)p(x_t|y_t)$ – mixture model

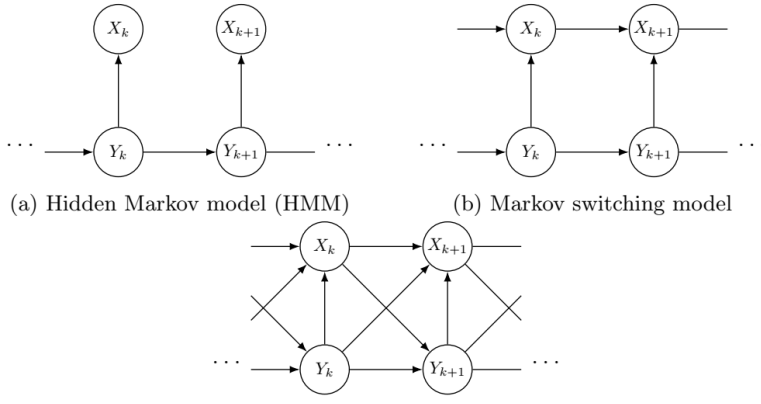


Figure 1: Directed dependence graphs of different types of PMM's

Of course, some other special cases possible

- $p(x_t, y_t|x_{t-1}, y_{t-1}) = p(y_t|y_{t-1}, x_{t-1})p(x_t|y_t, x_{t-1}, y_{t-1})$ – general PMM
- $p(x_t, y_t|x_{t-1}, y_{t-1}) = p(y_t|y_{t-1}, x_{t-1})p(x_t|y_t, y_{t-1})$ – PMM-IN
- $p(x_t, y_t|x_{t-1}, y_{t-1}) = p(y_t|y_{t-1})p(x_t|y_t, y_{t-1})$ – HMM-2
- $p(x_t, y_t|x_{t-1}, y_{t-1}) = p(y_t|y_{t-1}, x_{t-1})p(x_t|y_t)$ – another special case of PMM-IN
- $p(x_t, y_t|x_{t-1}, y_{t-1}) = p(y_t|y_{t-1}, x_{t-1})p(x_t|y_t, x_{t-1})$

The number of different classes can be (at least formally) reduced by blocking:

$$U_1 = (Y_1, Y_2), \quad U_t := (Y_t, Y_{t+1}), \quad V_1 = (X_1, X_2), \quad V_t = X_{t+1}, \quad t = 2, 3, \ldots.$$

Easy to see that $p(v_t, u_t | v_{t-1}, u_{t-1}) = p(x_{t+1}, y_{t+1} | x_t, y_t), \quad \forall t = 1, 2, \ldots.$

- (X, Y) HMM-2, then (U, V) is HMM:

$$p(v_t, u_t | v_{t-1}, u_{t-1}) = p(y_{t+1}|y_t) p(x_{t+1} | \overbrace{y_{t+1}, y_t}^{u_t}) = p(u_t|u_{t-1}) p(v_t|u_t)$$

- (X, Y) HMM-DN, then (U, V) is Markov switching:

$$p(v_t, u_t | v_{t-1}, u_{t-1}) = p(y_{t+1}|y_t) p(x_{t+1} | \overbrace{y_{t+1}, y_t}^{u_t}, x_{t-1}) = p(u_t|u_{t-1}) p(v_t|u_t, v_{t-1})$$

- (X, Y) PMM-IN, then (U, V) is a special case:

$$p(v_t, u_t | v_{t-1}, u_{t-1}) = p(y_{t+1}|y_t, x_t) p(x_{t+1} | \overbrace{y_{t+1}, y_t}^{u_t}) = p(u_t|u_{t-1}, v_{t-1}) p(v_t|u_t)$$

Key properties of PMM

- In general neither Y nor X is a Markov chain.

  The simplest (counter)example is HMM, where the observations $X_1, X_2, \ldots$ are not Markov.

- Conditionally on (a realization of) Y, X is a (inhomogenuous) Markov chain;

- Conditionally on (a realization of) X, Y is a (inhomogenuous) Markov chain;

The last two properties justify the name (PMM) introduced by W. Pieczynski.

Proof that given $X_{1:n}$, the random variables $Y_{1:n}$ have Markov property

$$p(y_{t+1}|y_{1:t}, x_{1:n}) = \frac{p(y_{t+1}, x_{t+1:n}|z_{1:t})p(z_{1:t})}{p(x_{t+1:n}|z_{1:t})p(z_{1:t})} = \frac{p(y_{t+1}, x_{t+1:n}|z_t)}{p(x_{t+1:n}|z_t)} = p(y_{t+1}|y_t, x_{t:n}),$$

where the second equality holds by pairwise Markov property. By the same property,

$$p(y_{t+1}|y_t, x_{t:n}) = \frac{p(y_{t+1}, x_{t+1:n}|z_t)}{p(x_{t+1:n}|z_t)} = \frac{p(y_{t+1}, x_{t+1:n}|x_{1:n}, y_t)}{p(x_{t+1:n}|x_{1:n}, y_t)} = p(y_{t+1}|y_t, x_{1:n}).$$

Hence, for any state sequence $y_1, \ldots, y_{t+1}$,

$$\begin{aligned} P(Y_{t+1} = y_{t+1}|Y_{1:t} = y_{1:t}, X_{1:n} = x_{1:n}) &= P(Y_{t+1} = y_{t+1}|Y_t = y_t, X_{1:n} = x_{1:n}) \\ &= \underbrace{P(Y_{t+1} = y_{t+1}|Y_t = y_t, X_{t:n} = x_{t:n})}_{\text{independent of } x_{1:t-1}}. \end{aligned}$$

The proof that given $Y_{1:n}$ the random variables $X_{1:n}$ have Markov property is the same.

NB! Holds without assumption that (X, Y) is homogeneous.

When Y is a Markov chain?

A sufficient condition (recall HMM-DN):

$$p(y_t|y_{t-1}, x_{t-1}) = p(y_t|y_{t-1}) \quad \forall y_t, y_{t-1} \text{ and for } \mu - \text{a.e. } x_{t-1}.$$

Indeed, then by law of total probability

$$\begin{aligned} p(y_t|y_{1:t-1}) &= \int p(y_t|x_{t-1}, y_{1:t-1}) p(x_{t-1}|y_{1:t-1}) \mu(dx_{t-1}) \\ &= \int p(y_t|x_{t-1}, y_{t-1}) p(x_{t-1}|y_{1:t-1}) \mu(dx_{t-1}) = p(y_t|y_{t-1}). \end{aligned}$$

The second equality holds because (X, Y) is Markov.

That explains the name – HMM-DM (Y is a Markov chain).

Is it also necessary ? Under some additional conditions like (stationarity and) reversibility – yes.

Recall that Z is reversible, if for any n $Z_{1:n} = (Z_1, \dots, Z_n)$ has the same distribution as $(Z_n, \dots, Z_1)$ (implies stationarity). Then $y_{t-1} = y_{t+1}$ implies

$$p(x_t, y_t, y_{t-1}) = p(x_t, y_t, y_{t+1}) \quad \text{and} \quad p(x_t|y_t, y_{t-1}) = p(x_t|y_t, y_{t+1}).$$

It can be shown (via Jensen) that under this condition, the Markov property implies that

$$p(x_t|y_t, y_{t+1}) = p(x_t|y_t), \quad \mu - \text{a.e. } x_t$$

and this is equivalent to

$$p(y_{t+1}|x_t, y_t) = p(y_{t+1}|y_t) \quad \mu - \text{a.e. } x_t,$$

which is HMM-DN property. (Pieczynski *et al*, 2003, 2011), (Kuljus, L. 2023).

Is it also necessary ? In general – no.

The idea for a counterexample: consider a stationary, homogeneous PMM (X, Y) that has the following properties:

1) (X, Y) is a HMM-DN;
2) The reversed chain is not a HMM-DN. Implies that (X, Y) is not reversible;
3) $|\mathcal{Y}| = 2$.

By 1) Y is a MC. A two states MC with no zeros in transition matrix is always reversible (an easy exercise). Hence after reversing time, Y is the same MC, but by 2), the model is not HMM-DN.

PMM examples

Finite $\mathcal{X}$: HMM-DN. Let $|\mathcal{X}| = k$, $|\mathcal{Y}| = l$, let $\mathcal{X} \times \mathcal{Y}$ be ordered as follows:

$$\mathcal{X} \times \mathcal{Y} = \{(\chi_1, \gamma_1), \ldots, (\chi_k, \gamma_1), (\chi_1, \gamma_2), \ldots, (\chi_k, \gamma_2), \ldots, (\chi_1, \gamma_l), \ldots, (\chi_k, \gamma_l)\}.$$

Then (X, Y) is an HMM-DN if and only if the $kl \times kl$ transition matrix factorizes:

$$\begin{pmatrix} p_{11} \cdot A_{11} & p_{12} \cdot A_{12} & \ldots & p_{1l} \cdot A_{1l} \\ p_{21} \cdot A_{21} & p_{22} \cdot A_{22} & \ldots & p_{2l} \cdot A_{2l} \\ \ldots & \ldots & \ldots & \ldots \\ p_{l1} \cdot A_{l1} & p_{l2} \cdot A_{l2} & \ldots & p_{ll} \cdot A_{ll} \end{pmatrix}, \tag{1}$$

where $p_{ij} = P(Y_t = \gamma_j | Y_{t-1} = \gamma_i)$ and A_{ij} are the following transition matrices:

$$A_{ij} = \begin{pmatrix} a_{11}^{(ij)} & \ldots & a_{1k}^{(ij)} \\ \ldots & \ldots & \ldots \\ a_{k1}^{(ij)} & \ldots & a_{kk}^{(ij)} \end{pmatrix}, \quad a_{uv}^{(ij)} = P(X_t = \chi_v | Y_t = \gamma_j, Y_{t-1} = \gamma_i, X_{t-1} = \chi_u).$$

Recall that for any PMM (X is conditionally Markov),

$$p(x_t|x_{1:t-1}, y_{1:n}) = p(x_t|x_{t-1}, y_{t-1:n}).$$

HMM-DN: $p(y_t|z_{t-1}) = p(y_t|y_{t-1})$. Hence, for HMM-DN

$$\begin{aligned} p(x_t|x_{t-1}, y_{t-1:n}) &= \frac{p(x_{t-1}, x_t, y_{t-1:n})}{p(x_{t-1}, y_{t-1:n})} = \frac{p(z_t|z_{t-1})p(y_{t+1:n}|z_t)}{p(y_t|z_{t-1})p(y_{t+1:n}|y_t, z_{t-1})} \\ &= \frac{p(z_t|z_{t-1})}{p(y_t|z_{t-1})} = p(x_t|y_t, y_{t-1}, x_{t-1}), \end{aligned}$$

because by HMM-DN property, $p(y_{t+1:n}|z_t) = p(y_{t+1:n}|y_t)$ and by the same property

$$p(y_{t+1:n}|y_t, z_{t-1}) = \sum_{x_t} p(y_{t+1:n}|x_t, y_t, z_{t-1})p(x_t|y_t, z_{t-1}) = p(y_{t+1:n}|y_t).$$

Finite $\mathcal{X}$: HMM-DN.

$$\begin{pmatrix} p_{11} \cdot A_{11} & p_{12} \cdot A_{12} & \dots & p_{1l} \cdot A_{1l} \\ p_{21} \cdot A_{21} & p_{22} \cdot A_{22} & \dots & p_{2l} \cdot A_{2l} \\ \dots & \dots & \dots & \dots \\ p_{l1} \cdot A_{l1} & p_{l2} \cdot A_{l2} & \dots & p_{ll} \cdot A_{ll} \end{pmatrix}, \tag{2}$$

We saw that

$$p(x_t|x_{1:t-1}, y_{1:n}) = p(x_t|y_t, y_{t-1}, x_{t-1}) = a_{x_t, x_{t-1}}^{(y_{t-1}, y_t)}.$$

To generate $z_1, \dots, z_n$:

1) Generate $y_{1:n}$ from (p_{ij})

2) Given $y_{1:n}$ generate $x_{1:n}$ as a in homogeneous MC with matrices A_{y_{t-1}, y_t}

Finite $\mathcal{X}$: Markov switching model. When (X, Y) is a Markov switching model, then

$$\begin{aligned} a_{uv}^{(ij)} &= P(X_t = \chi_v | Y_t = \gamma_j, Y_{t-1} = \gamma_i, X_{t-1} = \chi_u) \\ &= P(X_t = \chi_v | Y_t = \gamma_j, X_{t-1} = \chi_u) = a_{uv}^{(j)} \end{aligned}$$

so that the $kl \times kl$ transition matrix factorizes as follows:

$$\begin{pmatrix} p_{11} \cdot A_1 & p_{12} \cdot A_2 & \dots & p_{1l} \cdot A_l \\ p_{21} \cdot A_1 & p_{22} \cdot A_2 & \dots & p_{2l} \cdot A_l \\ \dots & \dots & \dots & \dots \\ p_{l1} \cdot A_1 & p_{l2} \cdot A_2 & \dots & p_{ll} \cdot A_l \end{pmatrix}, \tag{3}$$

When (X, Y) is a HMM, then

$$a_{uv}^{(j)} = P(X_t = \chi_v | Y_t = \gamma_j, X_{t-1} = \chi_u) = P(X_t = \chi_v | Y_t = \gamma_j) = a_v^{(j)}$$

so that all rows of A_j are equal.

Regime switching model. Consider inhomogenous MC like (here $\mathcal{X} = \{0, 1\}$, $\mathcal{Y} = \{A, B, C\}$)

$$\underbrace{0100110101010}_{A}\underbrace{1111110000111111000111110}_{B}\underbrace{0100000100000110 0}_{C}\underbrace{1111110000111111}_{B}$$

Transition matrices in different regimes (A-rapid change, B-long blocks, C-more 0-s)

$$P_A = \begin{pmatrix} 0.3 & 0.7 \\ 0.7 & 0.3 \end{pmatrix}, \quad P_B = \begin{pmatrix} 0.6 & 0.4 \\ 0.3 & 0.7 \end{pmatrix}, \quad P_C = \begin{pmatrix} 0.8 & 0.2 \\ 0.9 & 0.1 \end{pmatrix}$$

Having (the estimates) the expected (average) time in every regime μ_A, μ_B, μ_C as well as the regime transition probabilities, we model the regimes as a Markov chain Y with tr matrix (the elements on main diagonal are relatively large)

$$R = \begin{pmatrix} r_{AA} & r_{AB} & r_{AC} \\ r_{BA} & r_{BB} & r_{BC} \\ r_{CA} & r_{CB} & r_{CC} \end{pmatrix}, \quad \frac{1}{(1 - r_{AA})} = \mu_A, \quad \frac{1}{(1 - r_{BB})} = \mu_B, \quad \frac{1}{(1 - r_{CC})} = \mu_C.$$

Regime switching model. Incorporate (X, Y) into a homogeneous PMM with (6×6) transition matrix

$$\begin{pmatrix} r_{AA}P_A & r_{AB}P_{AB} & r_{AC}P_{AC} \\ r_{BA}P_{BA} & r_{BB}P_B & r_{BC}P_{BC} \\ r_{CA}P_{CA} & r_{CB}P_{CB} & r_{CC}P_C \end{pmatrix}.$$

Hence our PMM is a HMM-DN. So Y is a Markov chain with transition matrix R, given $y_{t-1} = y_t = A$, the transitions $x_{t-1} \to x_t$ evolve along P_A and so on.

The matrices P_A, P_B, P_C are intra-regime matrices, they are given.

How to choose inter-regime matrices $P_{AB}, P_{CB}, \cdots$? Since r_{AB}, r_{CB}, etc are typically very small, the choice of inter-regime matrices is overlooked in the literature, but it matters!

Regime switching model. There is only one choice ($P_{AB} = P_{CB} = P_B$ etc) to make this model a Markov switching model.

There is only one choice (all rows of P_{AB} being equal to stationary distribution of P_B etc) so that in the long run the proportion of pairs (i, j) under a regime, say, A is the same as for the stationary MC with transition matrix P_A. Important in applications, because in practice the matrices are estimated from training samples.

Classical example: CpG islands. Matrix of island (+) and non-island (-) (Durbin *et al.*)

$P_+ =$

$+$	A	C	G	T
A	.18	.274	.426	.12
C	.171	.368	**.274**	.188
G	.161	.339	.375	.125
T	.079	.355	.384	.182

$P_- =$

$-$	A	C	G	T
A	.3	.202	.285	.21
C	.322	.298	**.078**	.302
G	.248	.246	.298	.208
T	.177	.239	.292	.292

Semi-Markov model. Let $\mathcal{X} = \{A, B, C\}$ with the matrix of jump probabilities

$$\begin{pmatrix} 0 & p_{AB} & p_{AC} \\ p_{BA} & 0 & p_{BC} \\ p_{CA} & p_{CB} & 0 \end{pmatrix} \qquad p_{AB} = P(X_t = B | X_t \neq A, X_{t-1} = A), \quad \text{etc.}$$

To every $a \in \mathcal{X}$ corresponds duration distribution $q_a = (q_a(1), \ldots, q_a(k))$.

Semi-Markov: After jumping to a state, say B, the duration (sojourn time) $d \sim q_B$ is emitted. And then the process stays in the state B exactly d times, Then it jumps to the next state (not B), say A, new duration from q_A is emitted and so on.

When q_a is geometric distribution, then semi-Markov model becomes Markov.

Semi-Markov model. Enbed X into homogeneous PMM (X, Y), where $\mathcal{Y} = \{1, \dots, k\}$ shows the rest of the time chain is a particular state:

$$P\big((X_t, Y_t) = (a, l) | (X_{t-1}, Y_{t-1} = (a, k)\big) = \begin{cases} 1, & \text{when } b = a \; l = k - 1; \\ p_{ab} q_b(l), & \text{when } b \neq a, \; k = 1 \; ; \\ 0, & \text{else.} \end{cases}$$

A typical path (up to last block, nothing is hidden)

$$(A,5), (A,4), (A,3), (A,2), (A,1), (C,4), (C,3), (C,2), (C,1), (A,2), (A,1), (B,4), (B,3), \cdots$$

For instance,

$$P\big((A, k-1) | (A, k)\big) = 1, \quad P\big((C,4) | (A,1)\big) = p_{AC} \cdot q_C(4).$$

Since the probability

$$P(Y_t = l | Y_{t-1} = 1, X_{t-1} = A) = q_B(l) p_{AB} + q_C(l) p_{AC}$$

is not (in general) independent of A, (X, Y) not a HMM-DN. Neither Y nor X is a Markov chain.

Semi-Markov model and regime-switching model can be combined into one PMM in several ways:

1) the regime times are not geometrically distributed, i.e. regime evolves according to a semi-Markov model. Inside a regime, the states X evolve according to the corresponding Markov model.

2) Inside a regime, the states X evolve according to a semi-Markov model with duration distributions and jump matrices depending on the regime. The number of jumps in one regime is geometrically distributed.

Milano model. Let $\mathcal{X} = \{1, 2\}$ and $\mathcal{Y} = \{a, b\}$. Consider PMM (X, Y) with the following transition matrix

$$\mathbb{P} = \begin{array}{c} \\ (1,a) \\ (1,b) \\ (2,a) \\ (2,b) \end{array} \begin{array}{c} \begin{array}{cccc} (1,a) & (1,b) & (2,a) & (2,b) \end{array} \\ \begin{pmatrix} p\lambda_1 & p(1-\lambda_1) & p(1-\lambda_1) & 1+p\lambda_1-2p \\ p\lambda_2 & p(1-\lambda_2) & q-p\lambda_2 & 1+p\lambda_2-q-p \\ q\mu_1 & q(1-\mu_1) & p-q\mu_1 & 1+q\mu_1-p-q \\ q\mu_2 & q(1-\mu_2) & q(1-\mu_2) & 1+q\mu_2-2q \end{pmatrix} \end{array}$$

Introducing additional parameters θ_i and ρ_i, the matrix can be rewritten

$$\mathbb{P} = \begin{array}{c} \\ (1,a) \\ (1,b) \\ (2,a) \\ (2,b) \end{array} \begin{array}{c} \begin{array}{cccc} (1,a) & (1,b) & (2,a) & (2,b) \end{array} \\ \begin{pmatrix} p\lambda_1 & p(1-\lambda_1) & (1-p)\theta_1 & (1-p)(1-\theta_1) \\ p\lambda_2 & p(1-\lambda_2) & (1-p)\theta_2 & (1-p)(1-\theta_2) \\ q\mu_1 & q(1-\mu_1) & (1-q)\rho_1 & (1-q)(1-\rho_1) \\ q\mu_2 & q(1-\mu_2) & (1-q)\rho_2 & (1-q)(1-\rho_2) \end{pmatrix} \end{array},$$

Milano model. We recognize the HMM-DN form (1) and so Y is a Markov chain with transition matrix

$$\begin{pmatrix} p & 1-p \\ q & 1-q \end{pmatrix}$$

Reversing the roles X and Y, we see the same shape and so X is also Markov chain with the same transition matrix.

Hence both marginals are Markov chain (in our case they both have the same transition matrix),

Milano model. In this model, both marginals are Markov chains, the parameters λ_i and μ_i determine the dependence between X and Y.

When $\lambda_1 = \mu_1 = p$, $\lambda_2 = \mu_2 = q$ and initial distribution factorizes $\pi(i, a) = \pi_X(i)\pi_Y(a)$, then X and Y are independent.

When $\lambda_1 = \mu_2 = 1$ and $\pi(1, a) = \pi(2, b) = 0.5$, then X and Y are maximally (positively) dependent: $X_t = 1$ iff $Y_t = a$.

When, in addition, $q = 1 - p$, $\lambda_2 = \mu_1 = 0$ and $\pi(1, b) = \pi(2, a) = 0.5$, then then X and Y are maximally (negatively) dependent: $X_t = 1$ iff $Y_t = b$.

Generalizations possible: X and Y have different transition matrices, number of states are greater than two.

Linear Markov switching model. Here $\mathcal{X} = \mathbb{R}$ (not discrete PMM), Y is a Markov chain with state space $\mathcal{Y}$;

$$\alpha : \mathcal{Y} \to \mathbb{R}, \quad \xi_1(j), \xi_2(j), \ldots \text{ are iid } \forall j \in \mathcal{Y}$$

Thus $\xi_1(Y_1), \xi_2(Y_2), \ldots$ or, equivalently, $(\xi_1, Y_1), (\xi_2, Y_2), \ldots$ is a HMM.

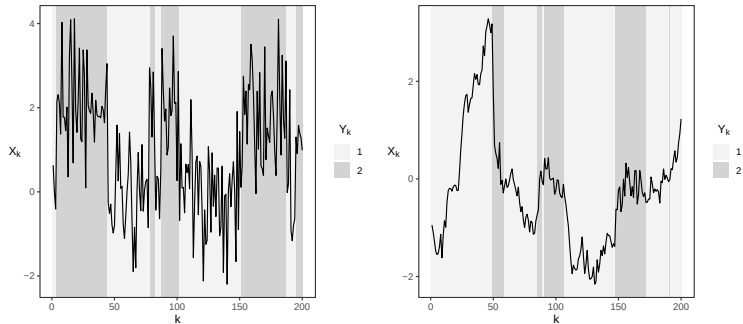
The model: $X_t = \alpha(Y_t)X_{t-1} + \xi_t(Y_t)$

Simple, but very flexible (Markov switching) model:

- When $\mathcal{Y} = \{1\}$ (nothing depends on Y), the AR(1);
- When $\alpha(j) \equiv \alpha$, (α independent of Y), then AR(1) with HMM-noise
- When $\alpha(j) \equiv 0$, then HMM;
- When $\xi_t(j) = \xi_t$ (noise independent of $\mathcal{Y}$), then Markov modulated AR(1)

Here Y is MC with transition matrix

$$\begin{array}{c} \\ 1 \\ 2 \end{array} \begin{array}{c} \begin{array}{cc} 1 & 2 \end{array} \\ \begin{pmatrix} 0.95 & 0.05 \\ 0.05 & 0,95 \end{pmatrix} \end{array} \quad \xi_t(i) \sim \mathcal{N}(\mu_i, 1), \quad \mu_1 = 0, \mu_2 = 2, \quad \alpha(1) = 1.01, \quad \alpha(2) = 0.5$$



Left: HMM $Y_t = \xi_t(Y_t)$ Right: Switching $Y_t = \alpha(Y_t)X_{t-1} + \xi_t(Y_t)$.

PMM tools

The popularity of HMM's is largely due to the simple algorithms:

- Forward algorithm for calculating $\alpha_t(j) = p(x_1, \ldots, x_t; y_t = j)$;
- Backward algorithm for calculating $\beta_t(j) = p(x_{t+1}, \ldots, x_n | y_t = j)$ Thus

$$\alpha_t(j)\beta_t(j) = p(x_1, \ldots, x_n; y_t = j), \quad p(y_t = j | x_1, \ldots, x_n) = \frac{\alpha_t(j)\beta_t(j)}{\sum_i \alpha_t(i)\beta_t(i)};$$

- Viterbi algorithm for finding

$$\arg\max_{y_1, \ldots y_n} p(y_1, \ldots, y_n | x_1, \ldots, x_n);$$

- EM (Baum-Welch) algorithm for estimating parameters (relies on forward-backward algorithms).

All these algorithms rely on a single property:

Conditionally on $x_1, \ldots, x_n, Y_1, \ldots, Y_n$ is a (inhomogenuous) Markov chain

We already know that every PMM has this property, no matter whether Y in a Markov chain or not!

This implies that with some obvious modifications, all these algorithms hold for PMM as well!

The ”obvious modification” example

$$\beta_t(j) = p(x_{t+1}, \ldots, x_n | y_t = j, x_t)$$

Forward-backward algorithms for PMM's

$$\alpha(y_t, x_{1:t}) = \sum_{y_{t-1}} \alpha(y_{t-1}, x_{1:t-1}) p(z_t | x_{t-1}, y_{t-1})$$
$$\beta(x_{t:n} | z_{t-1}) = \sum_{y_t} \beta(x_{t+1:n} | x_t, y_t) p(x_t, y_t | z_{t-1}).$$

Then

$$p(y_t | x_{1:n}) = \frac{\alpha(y_t, x_{1:t}) \beta(x_{t+1:n} | x_t, y_t)}{\sum_{y'_t} \alpha(y'_t, x_{1:t}) \beta(x_{t+1:n} | x_t, y'_t)}$$ smoothing probabilities,

$$p(y_t, y_{t+1} | x_{1:n}) = \frac{\alpha(y_t, x_{1:t}) p(z_{t+1} | z_t) \beta(x_{t+2:n} | z_{t+1})}{\sum_{y'_t, y'_{t+1}} \alpha(x_{1:t}, y'_t) p(y'_{t+1}, x_{t+1} | x_t, y'_t) \beta(x_{t+2:n} | x_{t+1}, y'_{t+1})}.$$

To prevent the numerical underflow, either scaled versions of α, β variables or log-sum tricks are used.

To recapitulate:

- PMM is a large and powerful/flexible class of dynamic models, HMM is just a narrow subclass;
- Yet all basic HMM-tools – dynamic programming algorithms – can be (at least in principle) used for PMM's as well (after obvious modifications).

Segmentation/decoding/denoising

Let (X, Y) be any two dimensional process (not necessary PMM) with Y_t taking values on $\mathcal{Y}$.

The first n elements

$$x_{1:n} := (x_1, \ldots, x_n)$$

of a realization of $X_{1:n} := (X_1, \ldots, X_n)$ are observed. The corresponding $y_{1;n} := (y_1, \ldots, y_n)$ (realization of $Y_{1:n}$) are not observed (Y is hidden).

The problem of segmentation: To find out / prognose / estimate the hidden state sequence $y_{1\ldots n}$. Formally, we are looking for a function – classifier –

$$g = (g_1, \ldots, g_n) : \mathcal{X}^n \to \mathcal{Y}^n,$$

that maps the observed sequence into state sequence. What is the best g ?

Framework of Statistical learning

Given observation $x_{1;n}$ define the conditional risk $R(y_{1:n}|x_{1:n})$ that measures the goodness of alignment/path $y_{1:n}$. Then the best classifier is the one that minimizes conditional (and also unconditional) risk over all possible classifiers:

$$g(x_{1:n}) := \arg \min_{y_{1:n} \in \mathcal{Y}^n} R(y_{1:n}|x_{1:n}).$$

$R(\cdot|x_{1:n})$ depends on the task, often defined via loss function

$$L : \mathcal{Y}^n \times \mathcal{Y}^n \to [0, \infty],$$

where $L(a_{1;n}, y_{1:n})$ is loss, when the actual state sequence is $a_{1:n}$ and the prognose is $y_{1:n}$. The conditional expectation

$$R(y_{1:n}|x_{1:n}) := E[L(Y_{1:n}, y_{1:n})|X_{1:n} = x_{1:n}]$$

is the conditional risk of $y_{1:n}$.

Standard classifiers: Viterbi path

When

$$L(a_{1;n}, y_{1:n}) = \begin{cases} 1, & \text{when } a_{1:n} \neq y_{1:n}; \\ 0, & \text{when } a_{1:n} = y_{1:n}. \end{cases}$$

then

$$R(y_{1:n}|x_{1:n}) = 1 - p(y_{1:n}|x_{1:n})$$

and, therefore, $g(x_{1:n})$ is Viterbi path or maximum a posteriori (MAP) path:

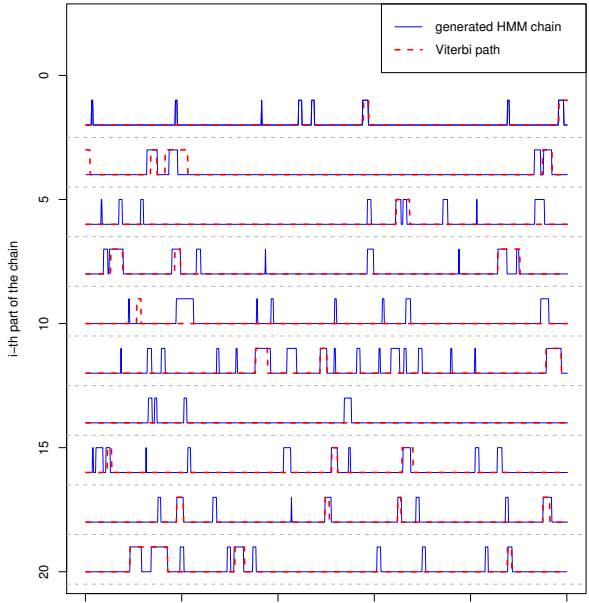
$$g(x_{1:n}) = \arg\max_{y_{1:n}} p(y_{1:n}|x_{1:n}).$$

Inherits its name from celebrated Viterbi algorithm.

For many models very models too conservative. For instance, the HMM with following transition matrix

$$\begin{pmatrix} 0.9 & 0.1 \\ 0.01 & 0.99 \end{pmatrix}$$

The typical path contains ”islands”.


omparison of Viterbi alignment and the actual HMM chain, splitted into parts with lengt
el parameters: p_11 = 0.90, p_22 = 0.99, emssion distributions are discrete and non−sy
Note: read the graph like normal text.
generated HMM chain
Viterbi path
i−th part of the chain
0
5
10
15
20

Standard classifiers: PMAP paths

When

$$L(a_{1:n}, y_{1:n}) = \sum_{t=1}^{n} I(a_t \neq y_t) \quad \text{(or a more general pointwise loss} \quad \ell(a_t, y_t))$$

is Hamming distance (measures the entry-wise difference between paths).

Then

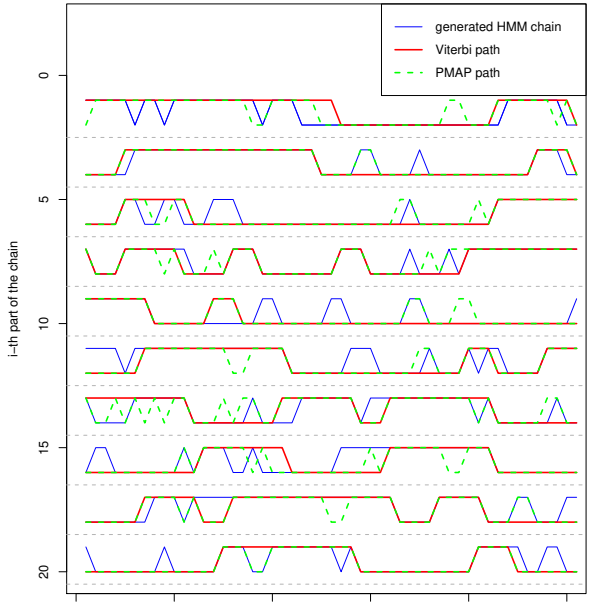
$$R(y_{1:n}|x_{1:n}) = \sum_{t=1}^{n} (1 - p_t(y_t|x_{1:n}))$$

and therefore $g(x_{1;n}) = w_{1:n}$ is pointwise maximum a posteriori (PMAP) path:

$$g_t(x_{1;n}) = \arg\max_{y \in \mathcal{Y}} p_t(y|x_{1:n}), \quad p_t(y|x_{1:n}) := P(Y_t = y|X_{1:n} = x_{1:n})$$ smoothing pr. -ies.

Minimizes expected numbers of misclassification errors.

Problem: In the presence of forbidden transitions (zeros in transition matrix, for example) PMAP-path might might be inadmissible, i.e. with zero probability.


ın of Viterbi alignment, PMAP alignment and the actual HMM chain, splitted into parts w
odel parameters: p_11 = 0.80, p_22 = 0.80, emssion distributions are discrete and symn
Note: read the graph like normal text.
generated HMM chain
Viterbi path
PMAP path
i–th part of the chain
0
5
10
15
20

Logarithmic versions of standard risks

$$\overline{R}_\infty(y_{1:n}|x_{1:n}) := -\ln p(y_{1:n}|x_{1:n}) = -\ln P(Y_{1:n} = y_{1:n}|X_{1:n} = x_{1:n})$$

$$\overline{R}_1(y_{1:n}|x_{1:n}) := -\sum_{t=1}^{n} \ln p_t(y_t|x_{1:n}) = -\sum_{t=1}^{n} \ln P(Y_t = y_t|X_{1:n} = x_{1:n}).$$

Clearly Viterbi alignment minimizes

$$\overline{R}_\infty(\cdot|x_{1:n})$$

and PMAP alignment minimizes

$$\overline{R}_1(\cdot|x_{1:n}).$$

If the aim is to minimize the number of errors, then one should use PMAP-alignment. Unfortunately, it can be with very low or zero likelihood (inadmissible).

The simplest solution: restricted optimization:

$$\min_{y_{1:n}:p(y_{1:n}|x_{1:n})>0} R_1(y_{1:n}|x_{1:n}) \quad \Leftrightarrow \quad \max_{y_{1:n}:p(y_{1:n}|x_{1:n})>0} \sum_{t=1}^{n} p_t(y_t|x_{1;n}). \tag{4}$$

In the presence of restrictions, (4) is not necessarily the solution of the following problem:

$$\min_{y_{1:n}:p(y_{1:n}|x_{1:n})>0} \overline{R}_1(y_{1:n}|x_{1:n}) \quad \Leftrightarrow \quad \max_{y_{1:n}:p(y_{1:n}|x_{1:n})>0} \prod_{t=1}^{n} p_t(y_t|x_{1:n}). \tag{5}$$

The solution of (5): posterior Viterbi decoding (PVD).

Block-loss

A remedy against inadmissibility of PMAP-path is to use block-loss

$$L(a_{1:n}, y_{1:n}) = \sum_{t=0}^{n-k} I(a_{t+1:t+k}, y_{t+1:t+k}).$$

Minimizing $R(y_{1:n}|x_{1:n})$ corresponding to that loss is equivalent to maximizing

$$p(y_{1:k}|x_{1:n}) + p(y_{2:k+1}|x_{1:n}) \cdots p(y_{n-k:n}|x_{1:n}).$$

The case $k = 1$ corresponds to PMAP risk.

The case $k = 2$ corresponds to maximizing the expected number of correct pairs.

Unfortunately does not help – the obtained path might still have zero probability!
Indeed, the sum

$$p(y_{1:2}|x_{1:n}) + p(y_{2:3}|x_{1:n}) + \cdots + p(y_{n-1:n}|x_{1:n})$$

might be maximal even when one of the addends is zero, say $p(y_{2:3}|x_{1:n}) = 0$.

Hybrid paths

The idea – use product (sum of logs) instead of sum: maximize for $k = 2$

$$\ln p(y_1|x_{1:n}) + \underbrace{\ln p(y_{1:2}|x_{1:n}) + \ln p(y_{2:3}|x_{1:n}) + \cdots + \ln p(y_{n-1:n}|x_{1:n})}_{\text{log of product of pairs}} + \ln p(y_n|x_{1:n})$$

For $k > 2$ maximize with $q(y_{s:t}) := \ln p(y_{s:t}|x_{1:n})$

$$\overbrace{q(y_1) + q(y_{1:2}) + \cdots q(y_{1:k-1}) + \underbrace{q(y_{1:k}) + \cdots + q(y_{n-k:n})}_{\text{log of block product}} + q(y_{n-k+1:n}) + \cdots + q(y_n)}^{-\overline{R}_k(y_{1:n}|x_{1:n})},$$

The additional factors in the beginning and in the end ensure that every t would be encountered equally many times.

For example: $k = 3$, minimizing $\overline{R}_3(y_{1:7}|x_{1:7})$ is maximizing $[p(y_{s:t}) := p(y_{s:t}|x_{1:n})]$:

$$\underbrace{p(y_1)p(y_{1:2})}_{B_{\text{begin}}} \cdot \underbrace{p(y_{1:3})p(y_{2:4})p(y_{3:5})p(y_{4:7})}_{\text{blocks with length 3}} \cdot \underbrace{p(y_{6:7})p(y_7)}_{B_{\text{end}}}$$

For PMMs it holds (easy to verify): for any $k > 1$ and $y_{1:n}$

$$\overline{R}_k(y_{1:n}|x_{1:n}) = \overline{R}_\infty(y_{1:n}|x_{1:n}) + \overline{R}_{k-1}(y_{1:n}|x_{1:n}) = (k-1)\overline{R}_\infty(y_{1:n}|x_{1:n}) + \overline{R}_1(y_{1:n}|x_{1:n}).$$

In other words

$$\overline{R}_k(y_{1:n}|x_{1:n}) = -\Big[\sum_{t=1}^{n} \ln p_t(y_t|x_{1:n}) + (k-1)\ln p(y_{1:n}|x_{1:n})\Big].$$

Generalize: take $C > 0$ (regularization constant) arbitrary and define

$$R_{C+1}(y_{1:n}|x_{1:n}) = -\Big[\sum_{t=1}^{n} \ln p_t(y_t|x_{1:n}) + C\ln p(y_{1:n}|x_{1:n})\Big].$$

Any minimizer of $R_{C+1}(\cdot|x_{1:n})$ is referred to as a hybrid path. The case $C = 0$ corresponds to PMAP.

Example of different paths

A HMM with $\mathcal{Y} = \{1, 2, 3\}$ and transition matrix

$$\begin{pmatrix} 0.99 & 0.01 & 0 \\ 0.3 & 0.3 & 0.4 \\ 0 & 0.02 & 0.98 \end{pmatrix}$$

transitions $1 \to 3$ and $3 \to 1$ are not allowed and every path having these transitions is inadmissible

True:
111111111123333333333333321111111111111111111111···111111111111111111111111111111111111
Viterbi:
111111111111111111111111111111111111111111111111···111111111111111111111111111111111111
PMAP:
111111133333333333333333333333333333333311111111···111333333333333333333333333333333311
PairMAP:
111111111133333333333333333333333333333311111111···111333333333333333333333333333333311
Hyb(C=1):
111112333333333333333333333333333333333333211···111111111111111111111111111111111111

PairMAP maximizes the **sum** of (cond) probabilities of pairs ($p(y_{s:t}) := p(y_{s:t}|x_{1:n})$):
$p(y_{1:2}) + p(y_{2:3}) + \cdots + p(y_{n-1;n})$,
Hyp(C=1) maximizes the **product**: $B_{\text{begin}} \cdot p(y_{1:2}) \cdot p(y_{2:3}) \cdots p(y_{n-1;n}) \cdot B_{\text{end}}$.

Properties of hybrid paths

By definition, a hybrid path is a solution of the following problem

$$\max_{y_{1:n}\in\mathcal{Y}^n}\Big[\sum_{t=1}^{n}\ln p_t(y_t|x_{1:n}) + C\ln p(y_{1:n}|x_{1:n})\Big].$$

Equivalently a hybrid path is a solution of constrained maximization problems:

$$\max_{y_{1:n}\in\mathcal{Y}^n}\prod_{t=1}^{n} p_t(y_t|x_{1:n}) \qquad \max_{y_{1:n}\in\mathcal{Y}^n} p(y_{1:n}|x_{1:n})$$

subject to $p(y_{1:n}|x_{1:n}) \geq B$ subject to $\prod_{t=1}^{n} p_t(y_t|x_{1:n}) \geq A.$

"Equivalence" means: to every $C > 0$ and to every minimizer of R_{C+1} (hybrid path) $u_{1:n}$ correspond constants $B = p(u_{1:n}|x_{1:n})$ and $A = \prod_{t=1}^{n} p_t(u_t|x_{1:n})$ such that u is a solution of restricted optimization problems.

By definition, a hybrid path is a solution of the following problem

$$\max_{y_{1:n}\in\mathcal{Y}^n}\Big[\sum_{t=1}^{n}\ln\, p_t(y_t|x_{1:n}) + C\ln\, p(y_{1:n}|x_{1:n})\Big].$$

Alternative problem

$$\max_{y_{1:n}\in\mathcal{Y}^n}\Big[\sum_{t=1}^{n} p_t(y_t|x_{1:n}) + C\ln\, p(y_{1:n}|x_{1:n})\Big]. \tag{6}$$

Just like previously, every solution of (6) is a solution of the following problems (for suitable D, E)

$$\max_{y_{1:n}\in\mathcal{Y}^n}\sum_{t=1}^{n} p_t(y_t|x_{1:n}) \qquad \max_{y_{1:n}\in\mathcal{Y}^n} p(y_{1:n}|x_{1:n})$$

subject to $p(y_{1:n}|x_{1:n}) \geq D$ subject to $\sum_{t=1}^{n} p_t(y_t|x_{1:n}) \geq E.$

No nice block-interpretation.

To every $x_{1:n}$ (observations) corresponds an integer k and finitely many constants $0 < C_1 < \cdots < C_k$ such that:

- the set of hybrid paths is same for every $C \in (C_i, C_{i+1})$;
- all hybrid paths corresponding to $C \in (C_i, C_{i+1})$ have the same conditional probability;
- to every C_i correspond at least two different solutions $u_{1:n}$ and $u'_{1:n}$ with different conditional probabilities $p(u_{1:n}|x_{1:n}) \neq p(u'_{1:n}|x_{1:n})$;
- When $C > C_k$, then hybrid path equals to Viterbi path. In case Viterbi path is not unique, the hybrid path u corresponding to $C > C_k$ is the Viterbi path with the biggest $\sum_{t=1}^{n} p_t(u_t|x_{1:n})$ – *primus inter pares*;
- The hybrid path $u_{1:n}$ corresponding to $C \in (0, C_1)$ has the biggest $\sum_{t=1}^{n} p_t(u_t|x_{1:n})$ over all admissible paths.

Finding hybrid paths for PMMs

Viterbi path(s) can be found by Viterbi algorithm – holds for PMM's (one pass)

PMAP paths need to calculate smoothing probabilities $p_t(y|x_{1:n})$ ($y \in \mathcal{Y}$) and that can be done by forward-backward algorithms that also hold for PMM's (two passes)

For hybrid paths: first apply forward-backward algorithms to calculate smoothing probabilities $p_t(y|x_{1:n})$ and the Viterbi (type) algorithms for finding hybrid paths (in total three passes).

To recapitulate: For PMM's the hybrid paths are easily found for any C.

Local Viterbi property

Assume PMM, let observations $x_{1:n}$ be give. Let $y_{1:n} \in \mathcal{Y}^n$ be a path. We say that $y_{1:n}$ is m-locally Viterbi, when

L1 For any $t \in \{2, \dots, m\}$,

$$p(y_{1:t}|x_{1:n}) = \max_{y'_{1:t-1} \in \mathcal{Y}^{t-1}} p(y'_{1:t-1}, y_t|x_{1:n});$$

L2 For any $t_1 < t_2 \in \{1, \dots, n\}$ such that $1 < t_2 - t_1 \leq m$,

$$p(y_{t_1+1:t_2}|y_{t_1}, x_{1:n}) = \max_{y'_{t_1+1:t_2-1} \in \mathcal{Y}^{t_2-t_1}} p(y'_{t_1+1:t_2-1}, y_{t_2}|y_{t_1}, x_{1:n});$$

L3 For any $t \in \{n - m + 1, \dots, n\}$,

$$p(y_{t+1:n}|y_t, x_{1:n}) = \max_{y'_{t+1:n} \in \mathcal{Y}^{n-t-2}} p(y'_{t+1:n}|y_t, x_{1:n}).$$

L2 means: for every $t_1 < t_2 \leq t_1 + m$

$$y_1, y_2, \ldots, y_{t_1-1}, \underbrace{y_{t_1}, y_{t_1+1}, \ldots y_{t_2-1} y_{t_2}}_{\text{Viterbi piece (segment)}}, y_{t_2+1}, \ldots y_n$$

the piece $y_{t_1:t_2}$ is MAP piece given it starts with y_{t_1} and ends with y_{t_2}.

In other words: knowing that $y_{t_1} = i$ and $y_{t_2} = j$ as well as the corresponding piece of observations $x_{t_1:t_2}$ (tho whole sequence $x_{1:n}$ is not needed), the rest of the piece $y_{t_1:t_2}$ can be found by Viterbi algorithm (MAP piece amongst those pieces that start with i and end with j).

L1 is the same for the beginning: $t \leq m$

$$\underbrace{y_1, \ldots, y_{t-1}, y_t}_{\text{Viterbi piece}}, y_{t+1}, \ldots, y_n$$

L3 is the same for the end: $n - t \leq m - 1$

$$y_1, \ldots, y_{t-1}, \underbrace{y_t, y_{t+1}, \ldots, y_n}_{\text{Viterbi piece}}$$

Hybrid paths are m-locally Viterbi for C big enough

Let (X, Y) be a PMM.

Theorem

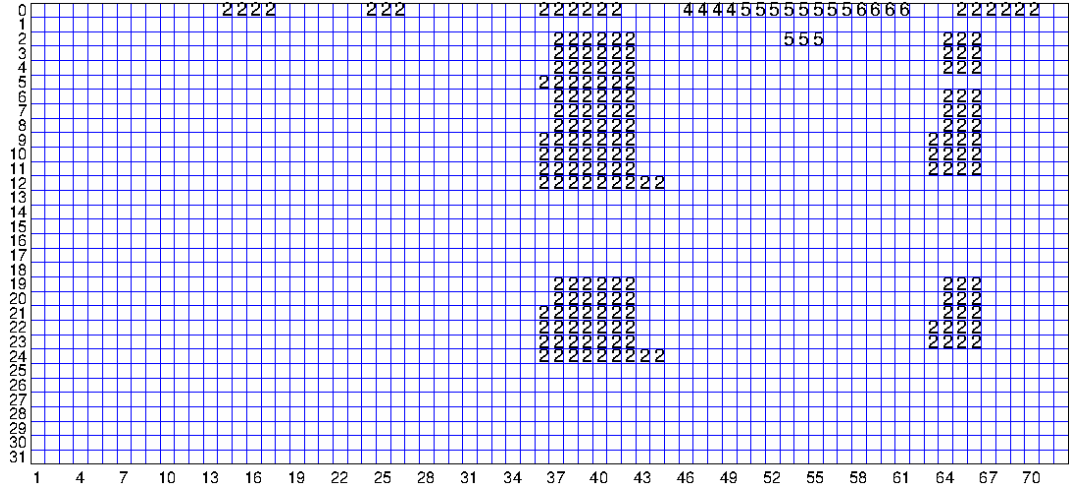
For any m there exists $C^m(x_{1:n})$ such that when $C > C^m$, then every hybrid path (corresponding to C) is m-locally Viterbi.

Hence when C grows, then hybrid paths approach to Viterbi path "locally" (that is expected).

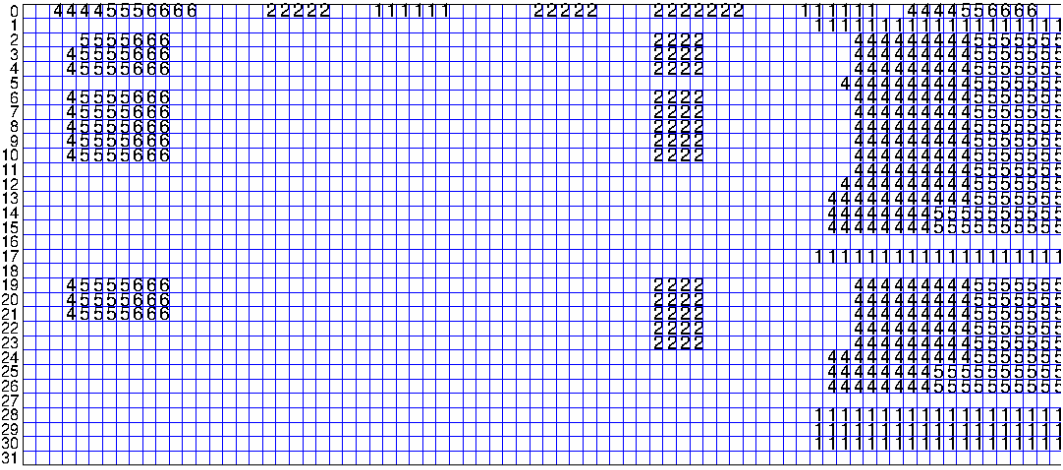
However, when all Viterbi pieces are unique (holds for many models) then any two m-locally Viterbi paths differ by pieces with lengths at least m: we call them m-different.

Corollary

Let $u_{1:n}$ and $u'_{1:n}$ be two different hybrid paths corresponding to C and C', respectively. Assume u is m-locally Viterbi and $C' > C$. Also let u be unique hybrid path corresponding to C. Then u and u' are m-different.



Performance of different decoders in protein alignment study: 0-truth, 1-Viterbi, 2-PMAP, 3-PVD, 4 -constrained PMAP, 5-PairPMAP, 6 – 17 hybrid with incrreasing C; 19–30 hybrid (6) with increasing C.



Performance of different decoders in protein alignment study: 0-truth, 1-Viterbi, 2-PMAP, 3-PVD, 4 -constrained PMAP, 5-PairPMAP, 6 – 17 hybrid with incrreasing C; 19–30 hybrid (6) with increasing C.

The last (Viterbi) constant

There is a constant $C_k(x_{1:n})$ so that any hybrid path is Viterbi when $C > C_k$. What is the upper bound of C_k? Block-interpretation: hybrid with $k+1$ corresponds to maximizing

$$B_{\text{begin}} \cdot p(y_{1:k})p(y_{2:k+1}) \cdots p(y_{n-k+1:n}) \cdot B_{\text{end}}$$

suggesting that when $k \geq n$, then the hybrid is always Viterbi or $C_k \leq n$.

That intuition is wrong! In (Kuljus, L. 2023) the behavior of random variable $C_k(X_{1:1000})$ was studied by simulations, and given a (regime switching) model, the value ranges between 6 to 9193. So, depending on the observation, that constant can be about 10 times bigger than n. Even when typically the $C_k(X_{1:n})$ is not varying much, one can often construct a specific $x_{1:n}$ so that $C_k(x_{1:n})$ is very large.

To recapitulate: Constant $C_k(x_{1:n})$ depends on the model and can heavily depend on the observations. When C increases, then hybrid paths approach to Viterbi path in m-local sense. However, they do not approach "entrywise" – quite opposite is true, because as C grows the hybrid paths tend to differ from each other in larger and larger intervals.

Two-step segmentation

We have see that Viterbi path is often too conservative (especially for models with "islands"). How to make it more "jump"?

First step: find/estimate jumps or change points.

Second step: use restricted optimization.

We call t a change point or jump, if $y_t \neq y_{t-1}$. How to find/estimate them?

The first approach: Find, for every t, the jump-probability

$$\rho_t := p(y_{t-1} \neq y_t | x_{1:n})$$

(easy with forward-backward) and estimate t as a change point, if $\rho_t \geq \mathrm{treshold}$.

The second approach: Consider the loss function

$$L_B(a_{1:n}, y_{1:n}) = \sum_{t<s} \Big[AI(a_t = \cdots = a_s)(1 - I(y_t = \cdots = y_s)) + BI(y_t = \cdots = y_s)(1 - I(a_t = \cdots = a_s)). \Big]$$

Given $t < s$, the loss is A when there is no change points between t and t in $a_{1:n}$, but there is one in $y_{1:n}$; the loss is B when there is no change points between t and t in $y_{1:n}$, but there is one in $a_{1:n}$. Motivated by Binder loss function in cluster-analysis.

Find

$$R_B(y_{1:n}|x_{1:n}) = E[L_B(Y_{1:n}, y_{1:n})|X_{1:n} = x_{1:n}]$$

and find any $\hat{y}_{1:n}$ minimizing $R_B(y_{1:n}|x_{1:n})$. Only the change points in $\hat{y}_{1:n}$ matter.

For example when $y_{1:6} = 112223$ and $y'_{1:6} = 334443$, $R_B(y_{1:6}|x_{1:6}) = R_B(y'_{1:6}|x_{1:6})$.

When replacing $\sum_{t<s}$ with the sums over pairs $(t-1,t)$, i.e. considering loss

$$L(a_{1:n},y_{1:n})=\sum_{t=2}^{n}\Big[AI(a_{t-1}=a_t)(1-I(y_{t-1}=y_t))+BI(y_{t-1}=y_t)(1-I(a_{t-1}=a_t))\Big],$$

then minimizing the corresponding risk (expected loss) equals to determining the change points via ρ_t:

$$\hat{y}_{t-1}\neq\hat{y}_t \quad\Leftrightarrow\quad \rho_t\geq\frac{A}{A+B}$$

The miminizer of R_B-risk can be found via dynamic programming algorithm.

Second step: Given the change points (estimates), say a minimizer $\hat{y}_{1:n}$ of $R_B(y_{1:n}|x_{1:n})$, then the final solution is restricted Viterbi path:

$$\hat{v}_{1:n}:=\arg\max_{y_{1:n}:L_B(y_{1:n},\hat{y}_{1:n})=0}p(y_{1:n}|x_{1:n}).$$

The restriction $L_B(y_{1:n},\hat{y}_{1:n})=0$ guarantees that the change points of $\hat{v}$ equal to the change points of $\hat{y}$ (provided $A>0,B>0$). Restricted Viterbi path can be found by Viterbi-like algorithm.

Improving accuracy of Viterbi path

For any path $a_{1:n}$, the sum – accuracy –

$$\sum_{t=1}^{n} p_t(a_t|x_{1:n}) = E\big[\sum_{t=1}^{n} I(Y_t = a_t)\big|X_{1:n} = x_{1:n}\big]$$

is the expected number of correctly estimated states. The most accurate path is PMAP path. The smoothing probability

$$p_t(a_t|x_{1:n}) = P(Y_t = a_t|X_{1:n} = x_{1:n})$$

measures the accuracy of a_t.

For Viterbi path v, $p_t(v_t|x_{1:n})$ can be very slow. In fact, it is easy to construct a HMM satisfying the following property: for every $\epsilon > 0$, there exists a n and $x_{1:n}$ (having positive probability) such that $\min_t p_t(v_t|x_{1:n}) < \epsilon$ (Kuljus, L. 2015).

The idea is to improve Viterbi path to avoid points with low accuracy:

1) Batch-approach: Take $\epsilon_o > 0$ a threshold, find all times t where the accuracy of Viterbi path is $< \epsilon_o$, let the set be T, so

$$T = \{t \in \{1, \dots, n\} : p_t(v_t | x_{1:n}) < \epsilon_o\}.$$

For any $t \in T$, find PMAP-state

$$a_t = \arg\max_{y \in \mathcal{Y}} p_t(y | x_{1:n}).$$

Finally use restricted optimization: find max-probability path that passes a_t at every $t \in T$:

$$\hat{v} = \arg \max_{y_{1:n} : y_t = a_t \; t \in T} p(y_{1:n} | x_{1:n})$$

2) Iterative-approach: At first find the time t_1 with smallest Viterbi accuracy and the corresponding PMAP-point

$$t_1 := \arg\min_t p_t(v_t|x_{1:n}), \quad a_{t_1} = \arg\max_{y\in\mathcal{Y}} p_{t_1}(y|x_{1:n}).$$

Then find the constrained Viterbi path

$$v^{(1)} = \arg \max_{y_{1:n}: y_{t_1}=a_{t_1}} p(y_{1:n}|x_{1:n})$$

Then recalculate everything under condition $Y_{t=1} = a_{t_1}$. Hence the smoothing probabilities are now

$$p_t^{(1)}(\cdot|x_{1:n}) = P(Y_t \in \cdot | Y_{t_1} = a_{t_1}, X_{1:n} = x_{1:n}).$$

Then find

$$t_2 := \arg\min_t p_t^{(1)}(v_t^{(1)}|x_{1:n}), \quad a_{t_2} := \arg\max_{y\in\mathcal{Y}} p_{t_1}^{(1)}(y|x_{1:n})$$

and the new constrained path

$$v^{(2)} = \arg \max_{y_{1:n}: y_{t_2}=a_{t_2}} p(y_{1:n}|x_{1:n}, y_{t_1} = a_{t_1})$$

and so on. Works (in many ways) better than the batch-approach (Kuljus, L. 2016)

Asymptotics and segmentation

Given a classifier g and loss function L, we would often like to know the long-run or asymptotic behavior of the actual loss

$$\frac{1}{n}L\big(Y_{1:n}, g(X_{1:n})\big). \tag{7}$$

For example, when

$$L(Y_{1:n}, g(X_{1:n})) = \sum_{t=1}^{n} I(Y_t \neq g_t(X_{1:n}))$$

then the actual loss is the proportion of misclassified entries.

When there exists a constant R such that actual loss (7) converges to R, a.s. then it is easy to deduce the convergence

$$\frac{1}{n}E[L(Y_{1:n}, g(X_{1:n}))|X_{1:n}] = \frac{1}{n}R(g(X_{1:n})|X_{1:n}) \to R. \quad \text{a.s,}$$

so that the limit R is called asymptotic risk. Given classifer, it is a number that depends on the model only.

Asymptotics of Viterbi classifier

Consider Viterbi classifier, let us denote it as $v(x_{1:n})$. Any long-run (asymptotical) property of Viterbi classifier is very difficult to study because of the following property: adding new observation can change the whole path so far. Formally

$$v(x_{1;n}) \neq v(x_{1:n+1})_{1:n}.$$

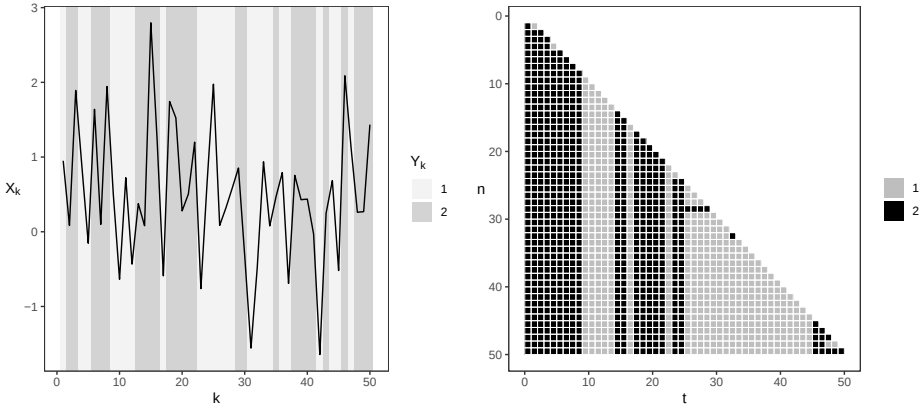
It might be that x_{n+1} changes the first element of Viterbi alignment, $v_1(x_{1:n})$ for arbitrary large n. Then, obviously, there is no asymptotics.

However, intuition suggests that typically $v_t(x_{1:n})$ stabilizes for n big enough. When it is so, then the Viterbi path can be extended to infinity.

Definition

Let $x_{1;\infty}$ be a realization of $X_{1:\infty}$. A sequence $v_{1:\infty} =: v(x_{1:\infty}) \in \mathcal{Y}^\infty$, is called the infinite Viterbi path (of $x_{1:\infty}$), if for any $t \geq 1$, there exists $m(t) \geq t$ (depending on $x_{1:\infty}$) so that

$$v(x_{1:n})_{1:t} = v_{1:t}, \quad \forall n \geq m(t).$$



Simulations from a HMM and corresponding Viterbi path with increasing n.

When infinite Viterbi path exists?

The existence of infinite Viterbi path is not trivial, there are several counterexamples (rather trivial HMM's), where it does not exists.

Example: Emissions:

$$p(\cdot|y_1 = 1) = p(\cdot|y_1 = 2) = I_{[0,1]}, \quad p(\cdot|y_1 = 3) = 4I_{[0,1/4]}, \quad p(\cdot|y_1 = 4) = 4I_{[3/4,1]}.$$

Transition matrix

$$\begin{array}{c} \\ 1 \\ 2 \\ 3 \\ 4 \end{array} \begin{array}{c} \begin{array}{cccc} 1 & 2 & 3 & 4 \end{array} \\ \begin{pmatrix} 3/4 & 0 & 1/4 & 0 \\ 0 & 3/4 & 0 & 1/4 \\ 1/2 & 1/2 & 0 & 0 \\ 1/2 & 1/2 & 0 & 0 \end{pmatrix} \end{array}$$

Initial distribution stationary $(4/10, 4/10, 1/10, 1/10)$

It is easy to verify that when ties are broken in co-lexicographic ordering $(1 \succ 2 \succ 3 \succ 4)$, then Viterbi path is the following

$$v(x_{1:n} = \begin{cases} 11 \cdots 13, & \text{if } x_n \in [0, 1/4]; \\ 22 \cdots 24, & \text{if } x_n \in [3/4, 1]; \\ 11 \cdots 11, & \text{else.} \end{cases}$$

Since

$$P(X_t \in [0, 1/4], \text{ i.o.}) = P(X_t \in [3/4, 1], \text{ i.o.}) = 1,$$

we see that a. e. $x_{1:\infty}$ passes both interval infinitely often and so $v_{1:\infty}$ does not exists a.s.

A sufficient condition for existence of infinite Viterbi path is the existence of infinitely many (r-order) nodes. Then the infinite path can be constructed piecewise

Let us consider a toy example of nodes

An easy but yet insightful special case

Consider an ergodic HMM with $\mathcal{Y} = \{1, \ldots, k\}$ and suppose there $\exists$ set A with the following property

$$P_1(A) > 0, \quad P_2(A) = \cdots = P_k(A) = 0.$$

To emit an observation from A, Y has to be in the state 1, a.s.

Suppose we have observations: ($a \in A$)

x_1	x_2	x_3	a	x_5	x_6	x_7	x_8	a	x_{10}	x_{11}	x_{12}	x_{13}	x_{14}	a	x_{17}	x_{18}

What can we say about the Viterbi (max likelihood) alignment?

An easy but yet insightful special case

Consider a HMM with $\mathcal{Y} = \{1, \dots, k\}$ and suppose there $\exists$ set A with the following property

$$P_1(A) > 0, \quad P_2(A) = \cdots = P_k(A) = 0.$$

To emit an observation from A, Y has to be in the state 1, a.s.

Suppose we have observations: ($a \in A$)

x_1	x_2	x_3	a	x_5	x_6	x_7	x_8	a	x_{10}	x_{11}	x_{12}	x_{13}	x_{14}	a	x_{16}	x_{17}
?	?	?	1	?	?	?	?	1	?	?	?	?	?	1	?	?

The a's correspond to the state 1. Then use the optimality principle (recall local optimality).

An easy but yet insightful special case

Consider a HMM with $\mathcal{Y} = \{1, \dots, k\}$ and suppose there $\exists$ set A with the following property

$$P_1(A) > 0, \quad P_2(A) = \cdots = P_k(A) = 0.$$

To emit an observation from A, Y has to be in the state 1, a.s.

Suppose we have observations: ($a \in A$)

x_1	x_2	x_3	a	x_5	x_6	x_7	x_8	a	x_{10}	x_{11}	x_{12}	x_{13}	x_{14}	a	x_{16}	x_{17}
v_1	v_2	v_3	1	?	?	?	?	1	?	?	?	?	?	1	?	?

The observations to first a can be used to determine the first piece.

An easy but yet insightful special case

Consider a HMM with $\mathcal{Y} = \{1, \dots, k\}$ and suppose there $\exists$ set A with the following property

$$P_1(A) > 0, \quad P_2(A) = \cdots = P_k(A) = 0.$$

To emit an observation from A, Y has to be in the state 1, a.s.

Suppose we have observations: ($a \in A$)

x_1	x_2	x_3	a	x_5	x_6	x_7	x_8	a	x_{10}	x_{11}	x_{12}	x_{13}	x_{14}	a	x_{16}	x_{17}
v_1	v_2	v_3	1	v_5	v_6	v_7	v_8	1	?	?	?	?	?	1	?	?

The observations from first to second a can be used to determine the second piece.

An easy but yet insightful special case

Consider a HMM with $\mathcal{Y} = \{1, \dots, k\}$ and suppose there $\exists$ set A with the following property

$$P_1(A) > 0, \quad P_2(A) = \cdots = P_k(A) = 0.$$

To emit an observation from A, Y has to be in the state 1, a.s.

Suppose we have observations: ($a \in A$)

x_1	x_2	x_3	a	x_5	x_6	x_7	x_8	a	x_{10}	x_{11}	x_{12}	x_{13}	x_{14}	a	x_{16}	x_{17}
v_1	v_2	v_3	1	v_4	v_5	v_6	v_7	1	v_{10}	v_{11}	v_{12}	v_{13}	v_{14}	1	$?$	$?$

The observations from second to third a can be used to determine the second piece.

An easy but yet insightful special case

Consider a HMM with $\mathcal{Y} = \{1, \dots, k\}$ and suppose there $\exists$ set A with the following property

$$P_1(A) > 0, \quad P_2(A) = \cdots = P_k(A) = 0.$$

To emit an observation from A, Y has to be in the state 1, a.s.

Suppose we have observations: ($a \in A$)

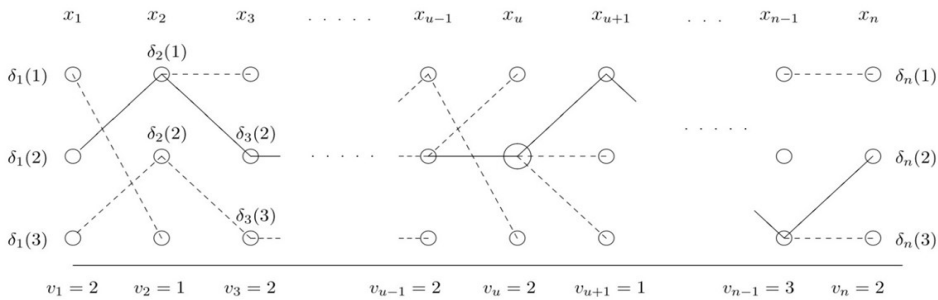
x_1	x_2	x_3	a	x_5	x_6	x_7	x_8	a	x_{10}	x_{11}	x_{12}	x_{13}	x_{14}	a	x_{16}	x_{17}
v_1	v_2	v_3	1	v_4	v_5	v_6	v_7	1	v_{10}	v_{11}	v_{12}	v_{13}	v_{14}	1	v_{16}	v_{17}

Finally the last piece. So, the whole alignment can be constructed piecewise. The process X is ergodic: every realization of the process has infinitely many a's. Hence, the piecewise alignment can be extended to infinity – we have an infinite (piecewise) alignment!

Node

How to generalize the concept of a? The answer lies in the Viterbi aligorithm – the dynamic programming algorithm to find the (max-likelihood) alignment.

$$\delta_t(l) := \max_{y_{1:t-1}} p(y_{1:t-1}, y_t = l; x_{1:t}), \quad t = 1, \ldots, n.$$



Node of order r

Restriction of the concept of node – r-order node:

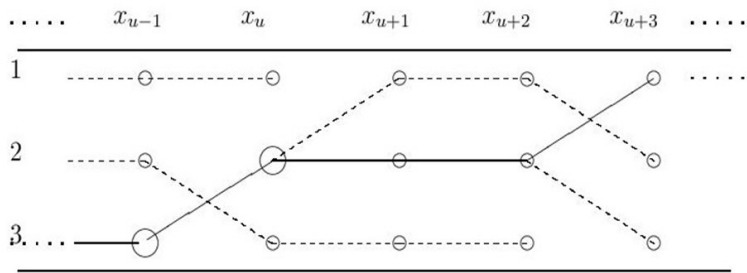


Figure 1: x_u is a 2^{nd} order 2-node, x_{u-1} is a 3^{rd}-order 3-node. Any alignment $v(x_{1:n})$ has $v(x_{1:n})_u = 2$.

Barrier

In general, to understand that x_u is an r-order node, one has to look at the observations

$$x_1, x_2, \ldots, x_u, x_{u+1}, \ldots, x_{u+r}.$$

On the other hand, a was a node independently of the previous observations. Could we have something like that as well?

A barrier is a block of observations that contains a (r-order) node independently of the observations before (and after) it.

$$x_1, x_2, x_3, x_4, x_5, \underbrace{x_6, x_7, x_8, x_9, x_{10}, x_{11}}_{\text{a barrier of length 6}}, x_{12}, x_{13}, \ldots$$

In previous example: a – barrier of length 1.

When $x_{1:\infty}$ has infinite many barriers, then it has infinitely many nodes and infinite Viterbi path exists.

Definition of r-order node and barrier

Let for any $x_{1:k} \in \mathcal{X}^k$ and any pair $i, j \in \mathcal{Y}$

$$p_{ij}(x_{1:k}) := \max_{y_{1:k}: y_1 = i, y_2 = j} p(x_{2:k}, y_{2:k} | x_1, y_1).$$

When $p_{ij}(x_{1:k}) > 0$, then it is possible to start from (x_1, i) and end in the state j having observations $x_{2:k}$ at the same time.

Let $x_{1:m}$ be a vector of observations and $i \in \mathcal{Y}$. The time $t \leq m$ is called an i-node of order $r = m - t$, when the following holds

$$\delta_t(i) p_{ij}(x_{t:m}) \geq \delta_t(s) p_{sj}(x_{t:m}) \quad \forall j, s \in \mathcal{Y}.$$

Time t is called a strong i-node of order $r = m - t$, when the inequality is strict for any j and $s \neq i$ for which the left side is positive.

Given $i \in \mathcal{Y}$, $b_{1:M}$ is called a (strong) i-barrier of order r and length M, if for any $x_{1:\infty} \in \mathcal{X}^\infty$ and $m \geq M$ satisfying $x_{m-M+1:n} = b_{1:M}$, $m-r$ is a (strong) i-node of order r.

A simple example of a barrier of length 3

Let (x_{t-1}, x_t, x_{t+1}) satisfy

$$p(x_t, i|x_{t-1}, u)p(x_{t+1}, j|x_t, i) \geq p(x_t, s|x_{t-1}, u)p(x_{t+1}, j|x_t, s), \quad \forall u, j, s \in \mathcal{Y}$$

Then

$$\begin{aligned}\delta_t(i)p(x_{t+1}, j|x_t, i) &= \max_u \delta_{t-1}(u)p(x_t, i|x_{t-1}, u)p(x_{t+1}, j|x_t, i) \\ &\geq \max_u \delta_{t-1}(u)p(x_t, s|x_{t-1}, u)p(x_{t+1}, j|x_t, s) \\ &= \delta_t(s)p(x_{t+1}, j|x_t, s),\end{aligned}$$

so t is i-node of order 1.

$P(u \to i \to j) \geq P(u \to s \to j)$ for any u, j, s.

Viterbi process

When our PMM is such that a.e. realization of $X_{1:\infty}$ has infinitely many barriers, then a.e. realization $x_{1:\infty}$ has infinite Viterbi path $v(x_{1:\infty})$. Then it is easy to verify (measurability) that there exists a stochastic process $V = V_{1:\infty}$ such that $V_{1:\infty} = v(X_{1:\infty})$, a.s. The process V is called Viterbi process.

When a.e. realization of $X_{1:\infty}$ has infinitely many barriers, then Viterbi process exists. In particular, it suffices to construct a set $\mathcal{X}^* \subset \mathcal{X}^M$ such that:

- any vector of $\mathcal{X}^*$ is a barrier;
- a.e. realization of $X_{1:\infty}$ contains infinitely many elements of $\mathcal{X}^*$ (guaranteed typically by ergodic/Harris recurrence argument).

Sufficient conditions for existence of $\mathcal{X}^*$: for HMM's in (Koloydenko, L., 2010) and for (Harris recurrent) PMM's in (Sova, L. 2020). Constructive proofs.

Harris recurrence of general state space Markov chain

Markov chain Z is called φ-irreducible for some σ-finite measure φ on $(\mathcal{Z})$, if $\varphi(A) > 0$ implies

$$\sum_{t=2}^{\infty} P_z(Z_t \in A) > 0 \quad \forall z \in \mathcal{Z}.$$

If Z is φ-irreducible, then there exists a maximal irreducibility measure ψ in the sense that for any other irreducibility measure φ' the measure ψ dominates φ', $\psi \succ \varphi'$. The symbol ψ will be reserved to denote the maximal irreducibility measure of Z.

Chain Z is called Harris recurrent when it is ψ-irreducible and $\psi(A) > 0$ implies

$$P_z(Z_t \in A \text{ i.o.}) = 1$$

for all $z \in \mathcal{Z}$. Note that if Z is Harris recurrent, then Z returns infinitely often a.s. to any set $A \in (\mathcal{Z})$ satisfying $\psi(A) > 0$.

A point $z \in \mathcal{Z}$ is called reachable, when for every open neighborhood O of z,

$$\sum_{t=2}^{\infty} P_z(Z_t \in O | Z_1 = z') > 0 \quad \forall z' \in \mathcal{Z}.$$

For φ-irreducible Z, the point z is reachable iff it belongs to support of φ. In our case $\mathcal{Z} \subset \mathcal{X} \times \mathcal{Y}$ and so $(x, i) \in \mathcal{Z}$ is reachable if for every open neighborhood O of x,

$$\sum_{t=2}^{\infty} P_z(X_t \in O, Y_t = i | Z_1 = z') > 0 \quad \forall z' \in \mathcal{Z}.$$

Let $\mathcal{X}^* \subset \mathcal{X}^M$. The fact that a.e. realization of $X_{1:\infty}$ contain infinitely many elements of $\mathcal{X}^*$ can be stated $P(X \in \mathcal{X}^* \text{ i.o.}) = 1$.

Formally

$$\{X \in \mathcal{X}^* \text{ i.o.}\} = \bigcap_{t=1}^{\infty} \bigcup_{l=t}^{\infty} \{X_{l:l+M-1} \in \mathcal{X}^*\}.$$

Existence of a reachable point + Harris recurrence is needed for $P(X \in \mathcal{X}^* \text{ i.o.}) = 1$.

A key condition for existence of $\mathcal{X}^*$

Recall

$$p_{ij}(x_{1:k}) := \max_{y_{1:k}: y_1 = i, y_2 = j} p(x_{2:k}, y_{2:k} | x_1, y_1).$$

Let $\mathcal{Y}^+(x_{1:k}) := \{(i,j) : p_{ij}(x_{1:k}) > 0\}$ and

$$\mathcal{Y}^+(x_{1:k})_{\text{input}} := \{i \in \mathcal{Y} : \exists j(i) \text{such that } p_{ij}(x_{1:k}) > 0\}$$
$$\mathcal{Y}^+(x_{1:k})_{\text{output}} := \{j \in \mathcal{Y} : \exists i(j) \text{such that } p_{ij}(x_{1:k}) > 0\}.$$

Observe: when $i \in \mathcal{Y}^+_{\text{input}}$ and $j \in \mathcal{Y}^+_{\text{output}}$, then not necessarily $(i,j) \in \mathcal{Y}^+$! In other words

$$\mathcal{Y}^+ \subset \mathcal{Y}^+_{\text{input}} \times \mathcal{Y}^+_{\text{output}}$$

and the inclusion can be strict.

The full theorem for the existence of $\mathcal{X}^*$ for PMM's is too technical to present here. A key condition (very close to be necessary) is the following:

A: There exists an open set $E \subset \mathcal{X}^k$ ($k \geq 2$) such that $\mathcal{Y}^+ = \mathcal{Y}^+(x_{1:k})$ is the same for every $x_{1:k} \in E$ and

$$\mathcal{Y}^+ = \mathcal{Y}^+_{\text{input}} \times \mathcal{Y}^+_{\text{output}}.$$

Furthermore there exists a reachable point in $\{x_1 : x_{1:k} \in E\} \times \mathcal{Y}^+_{\text{input}}$.

For HMM's, recall $p(\cdot|y = i)$ be the emission densities with respect to μ, $G_i := \{x : p(\cdot|y = i) > 0\}$. The condition $\mathbf{A}$ is implied by (easy to check):

Cluster condition: Markov chain Y with transition matrix (p_{ij}) is irreducible and there exists a cluster $C \subset \mathcal{Y}$ such that:

- $\mu[(\cap_i G_i) \setminus (\cup_{j \notin C} G_j)] > 0$;
- The sub-stochastic matrix $(p_{ij})_{i,j \in C}$ is irreducible and aperiodic.

Recall the example of HMM with no infinite Viterbi alignment. Emissions:

$$p(\cdot|y_1 = 1) = p(\cdot|y_1 = 2) = I_{[0,1]}, \quad p(\cdot|y_1 = 3) = 4I_{[0,1/4]}, \quad p(\cdot|y_1 = 4) = 4I_{[3/4,1]}.$$

Transition matrix

$$\begin{array}{c} \\ 1 \\ 2 \\ 3 \\ 4 \end{array} \begin{array}{c} \begin{array}{cccc} 1 & 2 & 3 & 4 \end{array} \\ \begin{pmatrix} 3/4 & 0 & 1/4 & 0 \\ 0 & 3/4 & 0 & 1/4 \\ 1/2 & 1/2 & 0 & 0 \\ 1/2 & 1/2 & 0 & 0 \end{pmatrix} \end{array} \quad G_1 = G_2 = [0, 1], G_3 = [0, 1/4], G_4 = [3/4, 1]$$

Possible clusters (satisfying the first condition): $\{1, 2\}$, $\{1, 2, 3\}$, $\{1, 2, 4\}$, but none of them satisfies another (matrix) condition – cluster condition fails.

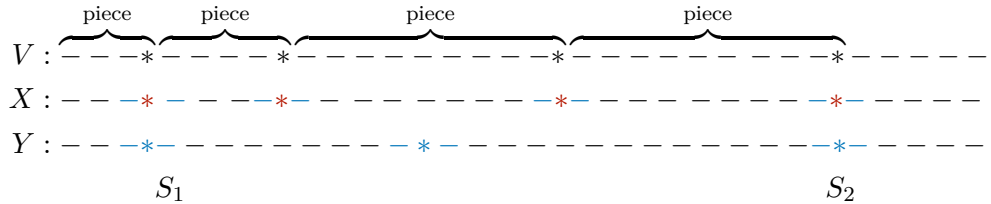
Regenerativity

A finite state (irreducible) Markov chain Y is regenerative with regenerative times $S_1, S_2, \ldots$ being the hitting times of a fixed state y_o:

$$S_1 := \min\{t \geq 1 : Y_t = y_o\}, \quad S_{k+1} := \min\{t \geq S_k + 1 : Y_t = y_o\}.$$

Clearly a HMM (X, Y) is regenerative (with respect to the same times) as well.

HMM: Under cluster condition (+ something else) the barrier set can be constructed as product set $\mathcal{X}^* = B_1 \times \cdots \times B_M$, it is not difficult to see that there exists regeneration times that correspond to a node inside barrier:



By piecewise construction (V, X, Y) a regenerative process, $E(S_{k+1} - S_k) < \infty$.

Due to the regenerativity, the cycles $\eta_1, \eta_2, \ldots$ are iid, where

$$\eta_k := (Y_{S_k+1:S_{k+1}}, X_{S_k+1:S_{k+1}}, V_{S_k+1:S_{k+1}}), \quad k = 1, 2, \ldots.$$

and so several limit theorems like the following SLLN (also CLT) can be proven:

$$\frac{1}{n}\sum_{t=1}^{n} f(V_t, X_t, Y_t) \to \frac{1}{E(S_2 - S_1)} E\Big[\sum_{t=S_1+1}^{S_2} f(V_t, X_t, Y_t)\Big] =: c(f), \quad \text{a.s.}$$

With some additional work, then the similar convergences for actual Viterbi path $V^n = v(X_1, \ldots, X_n)$ hold

$$\frac{1}{n}\sum_{t=1}^{n} f(V_t^n, X_t, Y_t) \to c(f), \quad \text{a.s..}$$

The convergences above were our initial goal. For example, with

$$f(V_t^n, X_t, Y_t) = I(V_t^n \neq Y_t),$$

we obtain that the limit proportion of misclassification errors (asymptotic risk) exists.

Regenerativity with PMM's

Technical complications arise.

- When $\mathcal{X}$ is not countable, then (X, Y) is a Markov chain with general state space. In general, there are no atoms in the state space and so the hitting times cannot be used as the regeneration times any more (HMM is a special case where atom exists). The solution is to create the pseudo-atoms by (Nummelin) splitting. With splitting, the regenerative times $S_1, S_2, \ldots$ for PMM (X, Y) inside the barriers – so that the reg. times are nodes — can be obtained.

- Since conditionally on Y the observations are not independent, $S_1, S_2, \ldots$ are not necessarily the regeneration times for Viterbi process V, because

$$V_{S_1+1:S_2} \text{ and } V_{S_2+1:S_3} \text{ both depend on } X_{S_2} \text{ (etc).}$$

Piecewise coding for PMM's

$$\ldots \overbrace{X_{S_1}, X_{S_1+1}, \ldots, X_{S_2}}^{V_{S_1+1:S_2}}, \underbrace{X_{S_2+1}, \ldots \overbrace{X_{S_3}}^{V_{S_3+1:S_4}}}_{V_{S_2+1:S_3}}, X_{S_3+1}, \ldots, X_{S_4} \ldots$$

$$\underbrace{\ldots, X_{S_1}}_{\text{independent}}, \underbrace{X_{S_1+1}, \ldots, X_{S_2}}_{\text{independent}}, \underbrace{X_{S_2+1}, \ldots X_{S_3}}_{\text{independent}}, \underbrace{X_{S_3+1}, \ldots, X_{S_4}}_{\text{independent}} \ldots$$

In particular, the cycles $\eta_1, \eta_2, \ldots$ of (X, Y, V) are not necessarily independent, but they are 1-dependent, i.e

$$(\eta_1, \ldots \eta_{k-1}) \text{ and } (\eta_{k+1}, \eta_{k+1}, \ldots) \text{ are independent for every } k.$$

For 1-dependent cycles the limit theorems (SAS, CLT) holds as well, so the existence of asymptotic risks can be proved!

Some special cases when cycles are independent: HMM, $\mathcal{X}$ is discrete, linear Markov switching model, ...

Exponential smoothing

Recall smoothing probabilities: $p_t(\cdot|x_{1:n}) := P(Y_t = \cdot|X_{1:n} = x_{1:n})$. When $t = n$, then $p_n(\cdot|x_{1:n})$ is often referred to as filtering probabilities.

Clearly $p_t(\cdot|x_{1:n})$ depends on the observation x_t, on its neighbors and so on. However, the intuition suggests: the bigger $|s - t|$, the less x_s influences $p_t(\cdot|x_{1:n})$. This intuition is postulated by exponential smoothing: for any $1 \le s \le t \le n$:

$$\|p_t(\cdot|X_{1:n}) - p_t(\cdot|X_{s:n})\|_{TV} \le C_s \alpha^{t-s},$$

where C_s is $\sigma(X_{s:\infty})$-measurable r.v. and $\alpha \in (0, 1)$ depends on the model.

A more general inequality: $1 \le l \le s \le t \le n \le \infty$ (r.h.s. independent of n and l)

$$\|P(Y_{t:\infty} \in \cdot|X_{l:n}) - P(Y_{t:\infty} \in \cdot|X_{s:n})\|_{TV} \le C_s \alpha^{t-s}. \tag{8}$$

General idea behind such kind of exponential forgetting (geometric ergodicity) results is bounding Dobrshin coefficient with help of Doeblin condition.

For a transition matrix $Q(i,j)$, Dobrshin coefficient is the maximum total variation distance over all row pairs divided by 2:

$$\delta(Q) = \frac{1}{2} \sup_{i,i'} \|Q(i,\cdot) - Q(i',\cdot)\|_{TV} \leq 1.$$

Sub-multiplicative: when Q_1 and Q_2 are two transition matrices, then

$$\delta(Q_1 Q_2) \leq \delta(Q_1)\delta(Q_2).$$

When π and π' are two probability row vectors (initial distributions), then

$$\|\pi Q - \pi' Q\|_{TV} \leq \delta(Q)\|\pi - \pi'\|_{TV} \leq 2\delta(Q).$$

So, when $Q_1, Q_2, \ldots, Q_{n-1}$ are transition matrices such that $\delta(Q_t) \leq (1-\epsilon), \forall t$

$$\left\| \pi \prod_{t=1}^{n-1} Q_t - \pi' \prod_{t=1}^{n-1} Q_t \right\|_{TV} \leq 2(1-\epsilon)^{n-1}.$$

So, when $Y_1, \ldots, Y_n$ is an inhomogeneous MC with transition matrices

$$Q_t(i,j) = P(Y_{t+1} = j | Y_t = i)$$

and $Y_1 \sim \pi$; $Y_1', \ldots, Y_n'$ is an inhomogeneous MC with the same transition matrices but $Y_1' \sim \pi'$, then

$$\|P(Y_n \in \cdot) - P(Y_n' \in \cdot)\|_{TV} \leq 2(1-\epsilon)^{n-1}.$$

A transition matrix Q is said to satisfy Doeblin condition, if there is a probability row vector (measure) ν and $\epsilon > 0$ such that

$$Q(i,j) \geq \epsilon \nu(j), \quad \forall i, j.$$

When Q satisfies that condition, then $\delta(Q) \leq (1-\epsilon)$. In particular with finite state space, it holds when all entries of Q are strictly positive.

We have to work with conditional chain $Y|X$ and somehow bound the Dobrushin coefficient. It turns out that above-mentioned key assumption **A** is needed:

A: There exists an open set $E \subset \mathcal{X}^k$ ($k \geq 2$) such that $\mathcal{Y}^+ = \mathcal{Y}^+(x_{1:k})$ is the same for every $x_{1:k} \in E$ and

$$\mathcal{Y}^+ = \mathcal{Y}^+_{\text{input}} \times \mathcal{Y}^+_{\text{output}}.$$

Another (technical) condition:

B: Chain Z is ψ-irreducible, with

$$\psi(\{x_1 : x_{1:k} \in E\} \times \mathcal{Y}^+_{\text{input}}) > 0 \quad \text{and} \quad \mu^{k-1}(\{x_{2:k} | x_{1:k} \in E\}) > 0.$$

B together with Harris recurrence implies that $P(X \in E, \quad \text{i.o}) = 1$.

A general exponential forgetting theorem

Assume **A**, **B** hold and let Z be Harris recurrent.

1. Then for all $s \geq l \geq 1$

$$\lim_{t\to\infty} \sup_{n\geq t} \|P(Y_{t:\infty} \in \cdot | X_{l:n}) - P(Y_{t:\infty} \in \cdot | X_{s:n})\|_{\mathrm{TV}} = 0, \quad \text{a.s.} \tag{9}$$

2. If Z is positive (i.e. stationary distribution exists), then there exists a constant $\alpha \in (0,1)$ such that the following holds: for every $s \geq 1$ there exist a $\sigma(X_{s:\infty})$-measurable random variable $C_s < \infty$ such that for all $t \geq s \geq l \geq 1$

$$\sup_{n\geq t} \|P(Y_{t:\infty} \in \cdot | X_{l:n}) - P(Y_{t:\infty} \in \cdot | X_{s:n})\|_{\mathrm{TV}} \leq C_s \alpha^{t-s} = \frac{C_s}{\alpha^s}\,\alpha^t, \quad \text{a.s.} \tag{10}$$

Part 1: converges to 0; part 2: converges to zero exponentially fast.

Applying Levy martingale convergence, we can replace $X_{l:n}$ and $X_{s:n}$ in (9) and (10) by $X_{l:\infty}$ and $X_{s:\infty}$ and leave $\sup$ out.

One can also consider arbitrary different initial distributions if Z_1, say π and π' (mutually equivalent) and restate the theorem as follows.

Assume **A**, **B** hold and let Z be Harris recurrent.

1. Then for all $s \geq 1$

$$\lim_{t\to\infty} \sup_{n\geq t} \|P_\pi(Y_{t:\infty} \in \cdot | X_{s:n}) - P_{\pi'}(Y_{t:\infty} \in \cdot | X_{s:n})\|_{\mathrm{TV}} = 0, \quad \text{a.s.}$$

2. If Z is positive (i.e. stationary distribution exists), then there exists a constant $\alpha \in (0,1)$ such that the following holds: for every $s \geq 1$ there exist a $\sigma(X_{s:\infty})$-measurable random variable $C_s < \infty$ such that for all $t \geq s \geq 1$

$$\sup_{n\geq t} \|P_\pi(Y_{t:\infty} \in \cdot | X_{s:n}) - P_{\pi'}(Y_{t:\infty} \in \cdot | X_{s:n})\|_{\mathrm{TV}} \leq C_s \alpha^{t-s} = \frac{C_s}{\alpha^s}\,\alpha^t, \quad \text{a.s.}$$

For $t = n$, such type of results are known as filter stability.

Again, n can be replaced by ∞.

Meaning of A

Condition **A** appears again and again. It generalizes the concept of "irreducibility" and "aperiodicity" for inhomogeneous chains.

A finite state homogeneous MC with state space $\mathcal{Y}$ and transition matrix P is irreducible and aperiodic iff it is primitive – there exists m such that $P^m(i,j) > 0$ for every i, j. It means that whenever Y is a MC with transition matrix P, then

$$P(Y_{m+1} = j | Y_1 = i) > 0, \quad \forall i, j \in \mathcal{Y}.$$

Suppose now Y is homogeneous MC, let $\mathcal{Y}_t$ be the finite state space of Y_t. A generalization of the condition above would be: for every t, there $\exists\, n > t$ such that

$$P(Y_n = j | Y_t = i) > 0, \quad \forall i \in \mathcal{Y}_t, j \in \mathcal{Y}_n. \tag{11}$$

Fix $n > t$ and define

$$\mathcal{Y}^+ = \{(i,j) : i \in \mathcal{Y}_t, j \in \mathcal{Y}_n, \quad P(Y_n = j | Y_t = i) > 0\},$$

then (11) means: $\mathcal{Y}^+ = \mathcal{Y}^+_{\text{input}} \times \mathcal{Y}^+_{\text{output}}$, where $\mathcal{Y}^+_{\text{input}} = \mathcal{Y}_t$ and $\mathcal{Y}^+_{\text{output}} = \mathcal{Y}_n$. Apply for conditional chain $Y|X$.

When A and B hold? A lot of different conditions in the literature (usually for general $\mathcal{Y}$ and/or for HMM). It turns out that **A** and **B** generalize most of them.

For HMM: Cluster condition $\Rightarrow$ **A** and **B**, but not vice versa.

An easily verified assumption for HMM's: Transition matrix of Y is irreducible and has at least one row consisting of non-zero entries. Implies **A** and **B** (Sova, L, 2021).

Application in segmentation: Recall PMAP classifier:

$$g_t(x_{1:n}) = \arg\max_{y \in \mathcal{Y}} p_t(y|x_{1:n}).$$

Exponential forgetting provides the tools to show that the proportion of misclassification errors convergences (to asymptotic risk):

$$\frac{1}{n}\sum_{t=1}^{n} I(Y_t \neq g_t(X_{I:n})) = 1 - \frac{1}{n}\sum_{t=1}^{n} \max_y \overbrace{p_t(y|X_{1:n})}^{\text{smoothing prob}} \to R, \quad \text{a.s.}$$

Triplet Markov models (TMM)

Recall $\mathcal{X}$ is arbitrary, $\mathcal{Y}$ finite. Let $\mathcal{U}$ be another finite set. A triplet Markov model (TMM) is a three-dimensional Markov chain $Z_t = \{X_t, Y_t, U_t\}$ taking values in $\mathcal{Z} \subset \mathcal{X} \times \mathcal{Y} \times \mathcal{U}$.

Of course, any TMM is a PMM (X, V), where $V = (Y, U)$. So, alternative definition: a PMM (X, V) is called a TMM, when V is two-dimensional.

Rather than the dimension of the state space, in classifying a Markov chain as PMM or TMM, the roles of X, Y, U are important. In TMM, typically:

X stands for observed sequence

hidden Y-process – states — is of interest

hidden U -process is auxiliary of nuisance process that is necessary for modeling.

So, we are interested in process (X, Y) (observations and the latent variables of interest), which – as we know – need not be a Markov chain (PMM). If it happens to be so, then there is no need for TMM.

Examples of TMM

Probably the most common type of TMMs is PMM with independent noise:

$V = (Y, U)$ be a (finite-state) Markov chain – thus a PMM – and

$Z = (X, V) = (X, Y, U)$ is a HMM. So

$$p(x_t, y_t, u_t | x_{t-1}, y_{t-1}, u_{t-1}) = p(y_t, u_t | y_{t-1}, u_{t-1}) p(x_t | y_t, u_t).$$

Often X_t depends solely on Y_t, thus $p(x_t | y_t, u_t) = p(x_t | y_t)$.

The latter case models the situation, where the states Y_t are read/measured with noise.

A very simple (but not so silly as it might look) model: U is a Markov chain, (U, Y) is a HMM and Y_t is read with noise. Thus

$$p(x_t, y_t, u_t | x_{t-1}, y_{t-1}, u_{t-1}) = p(u_t | u_{t-1}) p(y_t | u_t) p(x_t | y_t).$$

Regime switching model with noise

Classical example is the regime switching model with noise – (Y, U) is a regime-switching model (HMM-DN), where U stands for the regime and Y are states.

Recall our earlier example, but now U stands for the regime, i.e. $\mathcal{Y} = \{0, 1\}$, $\mathcal{U} = \{A, B, C\}$

$$\underbrace{0100110101010}_{A}\underbrace{111111000011111100011110}_{B}\underbrace{0100000100000110 0}_{C}\underbrace{111111000011111}_{B}$$

Transition matrices in different regimes (A-rapid change, B-long blocks, C-more 0-s)

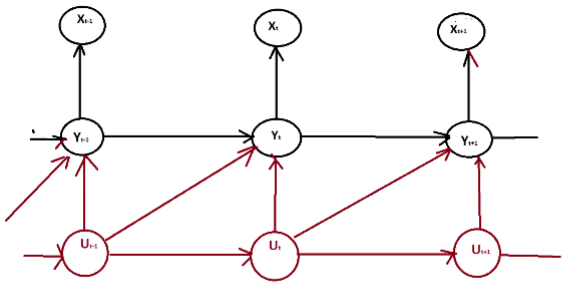
$$P_A = \begin{pmatrix} 0.3 & 0.7 \\ 0.7 & 0.3 \end{pmatrix}, \quad P_B = \begin{pmatrix} 0.6 & 0.4 \\ 0.3 & 0.7 \end{pmatrix}, \quad P_C = \begin{pmatrix} 0.8 & 0.2 \\ 0.9 & 0.1 \end{pmatrix}$$

The observations are conditionally (on $y_{1:n}$) independent and X_t on Y_t, only.

Often Gaussian noise, like $X_t = Y_t + \xi_t$, where $\xi_1, \xi_2, \ldots$ are iid $\mathcal{N}(0, \sigma)$-distributed variables.

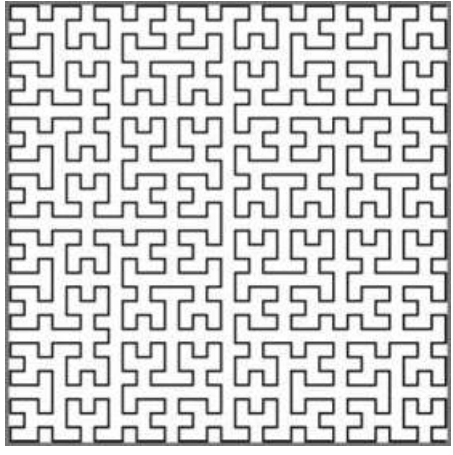
Regime switching model with noise

Dependence structure of regime switching model with independent noise



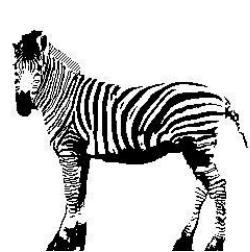
Regime switching model with noise

Historically (one of the) first TMM proposed for image segmentation by W. Pieczynski. In image segmentation, the 2D image is mapped into a 1D via Hilbert-Peano curve



Denoising zebra (Lachantin, Lapuyade-Lahorge, Pieczynski 2011)

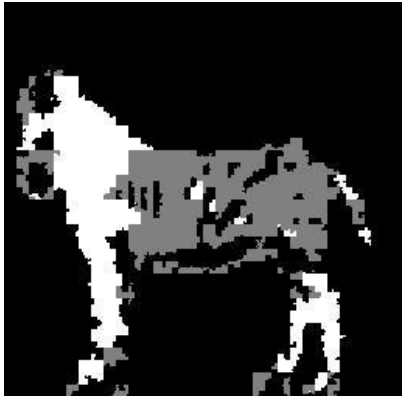
Denoising zebra



HMM reconstruction

TMM reconstruction

Denoising zebra. The U-variable models the frequency of color change.



Estimates of U

Another examples of PMM with independent noise:

Jumping noise HMM: U – MC modeling emission distribution, Y – MC modeling states in HMM, independent of U, X – observations, X_t depends on (U_t, Y_t):

$$p(x_t, y_t, u_t | x_{t-1}, y_{t-1}, u_{t-1}) = p(y_t|y_{t-1})p(u_t|, u_{t-1})p(x_t|y_t, u_t).$$

In other words (V, X) is a HMM, where $V = (U, Y)$ with U and Y being independent MC-s.

Hidden semi-Markov model: here $V = (U, Y)$ is a semi-Markov PMM (U counts remaining sojourn time) and (V, X) is a HMM, where X_t depends on Y_t.

Combining semi-Markov and regime-switching model

In regime switching model, the sojourn times of a particular regime is Geometric. It is possible to change the model so that these times have another distributions.

Suppose we have regimes A, B, C with corresponding sojourn time distributions q_A, q_B, q_C. Let $U = (U^r, U^s)$ be the semi-Markov PMM, where U^r models regimes, i.e. takes values in $\mathcal{U}^r = \{A, B, C\}$ and U^s models sojourn time. Thus (with $a, b \in \{A, B, C\}$)

$$P\big((U_t^r, U_t^s) = (a, l) | (U_{t-1}^r, U_{t-1}^s = (a, k)\big) = \begin{cases} 1, & \text{when } b = a \; l = k - 1; \\ p_{ab} q_b(l), & \text{when } b \neq a, \; k = 1\,; \\ 0, & \text{else.} \end{cases}$$

The process U^r is a semi-Markov chain.

Combining semi-Markov and regime-switching model

We now add the states Y and define a TMM $V = (Y, U^r, U^s)$ as follows

$$p(v_t|v_{t-1}) = p(u_t|u_{t-1})p(y_t|y_{t-1}, u_t^r, u_{t-1}^r),$$

where $p(u_t|u_{t-1})$ is defined as above. So (Y, U) is a HMM-DN or a regime switching model, but the transitions $Y_{t-1} \to Y_t$ depend on U_{t-1}^r and U_t^r.

Finally, independent noise to Y is added, so $Z = (X, Y, U)$ is a TMM with

$$p(z_t|z_{t-1}) = p(v_t|v_{t-1})p(x_t|y_t).$$

Observe that formally we have now 4D Markov process (X, Y, U^r, U^s) but rather than the dimension, the roles of marginal processes matters.

Segmentation with TMMs

Recall (X, Y, U) is a MC but (X, Y) is not a MC. In particular, given $X_{1:n}$ the process $Y_{1:n}$ is not necessarily an inhomogeneous MC. It means – no Viterbi algorithm for finding

$$\arg\max_{y_{1:n}} p(y_{1:n}|x_{1:n}). \tag{12}$$

Since $(Y, U)|X$ is Markov, so with Viterbi algorithm it is possible to find

$$\arg\max_{y_{1:n}, u_{1:n}} p(y_{1:n}, u_{1:n}|x_{1:n}). \tag{13}$$

However, if $(\hat{y}_{1:n}, \hat{u}_{1:n})$ is a solution of (13), then $\hat{y}_{1:n}$ is not necessarily a solution of (12).

Solving (12) is NP hard.

Since (V, X), where $V = (Y, U)$ is a PMM, then with forward-backward formulas, it is possible to find the smoothing probabilities

$$p_t(v_t|x_{1:n}) = p_t(u_t, y_t|x_{1:n}), \quad p_t(y_t|x_{1:n}) = \sum_{u_t \in \mathcal{U}} p_t(y_t, u_t|x_{1:n}).$$

summing above is possible, because $\mathcal{U}$ is small. Hence finding PMAP-path is possible. So far, the segmentation with TMMs has been done with PMAP, only.

When $k > 1$ is not a very big integer, then it is also possible to find the probabilities of k-blocks

$$p(y_{t:t+k-1}|x_{1:n}) = \sum_{u_{t:t+k-1} \in \mathcal{U}^k} p(y_{t:t+k-1}, u_{t:t+k-1}|x_{1:n}).$$

Since (X, V) is a PMM, also EM-training of parameters works.

Estimating Viterbi/hybrid path: blocks

Recall for PMM, the hybrid path with $C = k - 1$ maximizes

$$\sum_{t=1}^{n} \ln p_t(y_t|x_{1:n}) + (k-1) \ln p(y_{1:n}|x_{1:n}) \tag{14}$$

and at the same time, it also maximizes $p(y_{s:t}) := p(y_{s:t}|x_{1:n})$

$$p(y_1)p(y_{1:2})(y_{1;k-1}) \cdot p(y_{1:k})p(y_{2:k+1}) \cdots p(y_{n-k:n}) \cdot p(y_{n-k+1:n}) \cdots p(y_n). \tag{15}$$

For $k = 3, n = 7$:

$$\underbrace{p(y_1)p(y_{1:2})}_{B_{\text{begin}}} \cdot \underbrace{p(y_{1:3})p(y_{2:4})p(y_{3:5})p(y_{4:7})}_{\text{blocks with length 3}} \cdot \underbrace{p(y_{6:7})p(y_7)}_{B_{\text{end}}}$$

For TMMs, this equivalence does not hold any more (the proof relies on Markov property). Thus, with k not too big (say $k = 2, 3, 4$), it is possible to maximize (15), (k-block-probabilities+dynamic programming), but the solution is not a maximizer of (14).

Since k-block probabilites

$$p(y_{s:t}) := p(y_{s:t}|x_{1:n})$$

are possible to find, one can define

$$p_k(y_{1:n}) := p(y_{1:k}) \prod_{t=k+1}^{n} p(y_t|y_{t-k:t-1}) = p(y_{1:k}) \prod_{t=k+1}^{n} \frac{p(y_{t-k:t})}{p(y_{t-k:t-1})}.$$

If $Y|X$ would be a k-order Markov chain, then

$$p_k(y_{1;n}) = p(y_{1:n}|x_{1:n})$$

and maximizing p_k would be a right thing. In general, p_k-measure is k-order Markov approximation. It can be maximized, provided k is not too big. Works well in simulations.

Estimating Viterbi: segmentation EM

Consider Y as parameter, U as latent variable and apply EM.
In particular: given a path $y^{(i)} := y^{(i)}_{1:n}$ – the output of i-th step of iteration – update

$$\begin{aligned} y^{(i+1)} &= \arg\max_{y_{1:n}\in\mathcal{Y}^n} \sum_{u_{1:n}\in\mathcal{U}^n} \ln p(u_{1:n}, y_{1:n}|x_{1:n}) p(u_{1:n}|y^{(i)}, x_{1:n}) \\ &= \arg\max_{y_{1:n}\in\mathcal{Y}^n} \sum_{u_{1:n}\in\mathcal{U}^n} \ln p(y_{1:n}|u_{1:n}, x_{1:n}) p(u_{1:n}|y^{(i)}, x_{1:n}). \end{aligned}$$

Can be shown (as for any EM-procedure)

$$p(y^{(i+1)}|x_{1:n}) \geq p(y^{(i)}|x_{1:n}).$$

Algorithm is applicable, because $U, Y|X$ is Markov, so

$$\ln p(u_{1:n}, y_{1:n}|x_{1:n}) = \ln p(u_1, y_1|x_{1:n}) + \sum_{t=2}^{n} \ln p(u_t, y_t|u_{t-1}, y_{t-1}, x_{t-1:n}),$$

and so

$$\sum_{u_{1:n}\in\mathcal{U}^n} \ln p(u_{1:n}, y_{1:n}|x_{1:n}) p(u_{1:n}|y^{(i)}, x_{1:n}) = \sum_{u_1} \ln p(u_1, y_1|x_{1:n}) p(u_1|y^{(i)}, x_{1:n}) +$$
$$+ \sum_{t=2}^{n} \sum_{u_{t-1}, u_t} \ln p(u_t, y_t|u_{t-1}, y_{t-1}, x_{t-1:n}) p(u_{t-1}, u_t|y^{(i)}, x_{1:n}).$$

Simulations: depends heavily on initial value of iteration, gets stuck into a local maxima with few steps.

Another iterative methods: variational Bayes approach, (greedy) hill-climbing algorithms etc.

References

1. J. Lember, ”Local Viterbi property in decoding”, *Information and Inference*, 2023
2. K. Kuljus, J. Lember, ”Pairwise Markov Models and Hybrid Segmentation Approach”, *Methodology and Computing in Applied Probability*, 2023
3. J. Lember, J. Sova, ”Regenerativity of Viterbi process for pairwise Markov models”, *Journal of Theoretical Probability*, 2021
4. J. Lember, J. Sova, ”Exponential forgetting of smoothing distributions for pairwise Markov models”, *Electronic journal of Probability*, 2021
5. J. Lember, J. Sova, ”Existence of infinite Viterbi path for pairwise Markov models”, *Stochastic Processes and their Applications*, 2020
6. K. kuljus, J. Lember, ”On the accuracy of MAP inference in HMMs, *Methodology and Computing in Applied Probability*, 2015
7. K. Kuljus, J. Lember ”Asymptotic risks of Viterbi segmentation”, *Stochastic Processes and their Applications*, 2012
8. A. Koloydenko, J. Lember, ”A constructive proof of the existence of Viterbi processes”, *IEE Transactions on Information Theory*, 2010
9. W. Pieczynski, ”Pairwise Markov chains”, *IEEE Trans Pattern Anal Mach Intell*, 2003
10. P. Lachantin, J. Lapuyade-Lahorgue, Pieczynski, ”Unsupervised segmentation of randomly switching data with hidden non-Gaussian correlated noise”, *Signal processing*, 2011