Statistical and computational issues for hidden semi-Markov models

Vlad Stefan BARBU

LMRS, Université de Rouen-Normandie, France
barbu@univ-rouen.fr

19ième journée du réseau AIGM, December 6, 2018, Toulouse

Plan

A general framework

The hidden semi-Markov model

Nonparametric estimation and EM algorithm

HSMM with backward recurrence time dependence and SEM

Viterbi algorithm

Conclusions and perspectives

This talk is mainly based on works of

- V. S. Barbu, University of Rouen-Normandy
- N. Limnios, University of Technology of Compiègne
- S. Malefaki, University of Patras
- C.-E. Pertsinidou, Aristotle University of Thessaloniki
- S. Trevezas, University of Athens
- E. Votsi, University of Maine

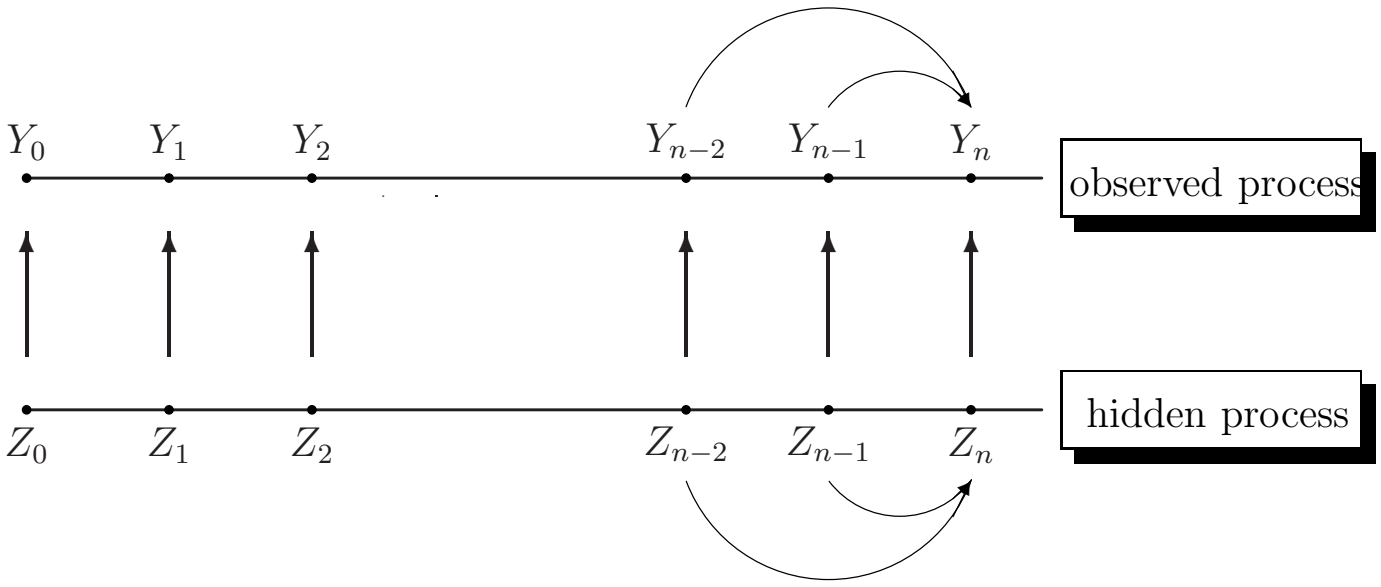
$\Rightarrow$ V. S. Barbu, N. Limnios (2006, 2008), S. Trevezas, N. Limnios (2009), S. Malefaki, S. Trevezas, N. Limnios (2010), C.-E. Pertsinidou, N. Limnios (2015), I. Votsi, N. Limnios, G. Tsaklidis, E. Papadimitriou (2014)

A general framework : hidden models

Assume :

- The process of interest $(Z_n)_{n\in\mathbb{N}^*}$ is not observed directly (it is hidden).
- Instead, one observes, say, $(Y_n)_{n\in\mathbb{N}^*}$, a function of $(Z_n)_{n\in\mathbb{N}^*}$

Typical case of hidden model



Remark : a very general problem

$(Y_n)_{n\in\mathbb{N}}$: received signal / DNA sequence / GPS position / failure occurrence / rainfall / wind speed
$(Z_n)_{n\in\mathbb{N}}$: emitted signal / gene / real position / failure intensity / weather regime / weather regime

In the literature (statistics, signal processing, speech recognition, bioinformatics, reliability, energy studies, etc.) :

$\approx$ Hidden Markov Model (Baum and Petrie, 1966 ; ...), i.e. : $(Z_n)_{n\in\mathbb{N}}$ is a Markov chain, while $(Y_n)_{n\in\mathbb{N}}$ conditionally independent, conditionally Markov chain, etc.

Some references on MLE asymptotics for H(S)MM

- L. E. Baum, T. Petrie (1966)
  Both processes with finite state space
- B. G. Leroux (1992)
  $(Y_n)_{n\in\mathbb{N}}$ with a general state space
- P. J. Bickel, Y. Ritov, T. Ryden (1998)
  $(Y_n)_{n\in\mathbb{N}}$ with a general state space
- V. S. Barbu, N. Limnios (2006, 2008)
  $(Z_n)_{n\in\mathbb{N}}$ SMC, both processes with finite state space
- S. Trevezas, N. Limnios (2009)
  $(Z_n)_{n\in\mathbb{N}}$ SMC, $(Y_n)_{n\in\mathbb{N}}$ with a general state space ;
  backward recurrence time dependence

DNA applications

- $(Y_n)_{n\in\mathbb{N}}$: DNA sequence
  $(Z_n)_{n\in\mathbb{N}}$: CpG islands

- CpG refers to a C nucleotide immediately followed by a G.
- The 'p' in 'CpG' refers to the phosphate group linking the two bases.
- Detection of regions of genomic sequences that are rich in the CpG pattern is important because such regions are resistant to methylation and tend to be associated with genes which are frequently switched on.
- Regions rich in the CpG pattern are known as CpG islands.

An application : Gene/CpG island detection

Problem 1 - CpG islands' detection

$$CpG \text{ islands} \begin{cases} \text{high frequency of the pair CG} \\ \text{high frequency of the bases C and G} \\ \text{regions with an essential role in the genome} \end{cases}$$

- $Y : \underbrace{TAGTGGAATG}\,\underbrace{CGACG}\cdots -$ DNA sequence
- $Z : 0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,1\,1\,1\,1\cdots -$ indicators for CpG islands

$$Z : \begin{cases} 1 \leftarrow \text{ island} \\ 0 \leftarrow \text{ non-islands} \end{cases}$$

Problem 2 - gene detection

- $Y : \underbrace{TAGTGGAATG}\,\underbrace{CGACG}\cdots -$ DNA sequence
- $Z : 0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,1\,1\,1\,1\cdots -$ indicators for coding regions

$$Z : \begin{cases} 1 \leftarrow \text{ coding (exon)} \\ 0 \leftarrow \text{ non-coding (intron)} \end{cases}$$

Semi-Markov framework - discrete time

All the stochastic processes are defined on a complete probability space $(\Omega, \mathcal{F}, \mathbb{P})$.

- $Z = (Z_k)_{k\in\mathbb{N}}$, chain with state space $E = \{1, 2, \ldots, s\}$, $s < \infty$, or $E = \mathbb{N}^*$
- $S = (S_n)_{n\in\mathbb{N}}$, jump times
  $J = (J_n)_{n\in\mathbb{N}}$, visited states
  $X = (X_n)_{n\in\mathbb{N}}$, sojourn times of Z

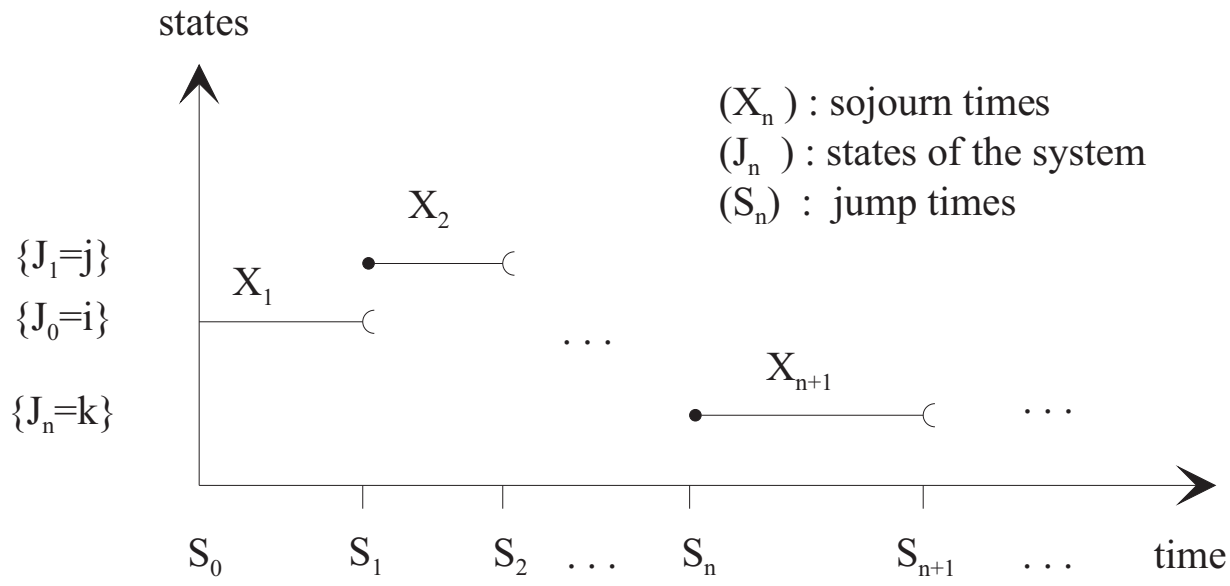


FIGURE – A trajectory of a semi-Markov chain

If (J,S) satisfies the relation

$$\begin{aligned} & \mathbb{P}(J_{n+1}=j, S_{n+1}-S_n=k|J_0,\cdots,J_n;S_1,\cdots,S_n) \\ = \quad & \mathbb{P}(J_{n+1}=j, S_{n+1}-S_n=k|J_n), j\in E, k\in\mathbb{N} \end{aligned}$$

- (J,S) Markov renewal chain
- $Z=(Z_k)_{k\in\mathbb{N}}-$ semi-Markov chain associated to (J,S)

$$Z_k := J_{N(k)} \quad \Leftrightarrow \quad J_n = Z_{S_n}$$

with $N(k) := \max\{n\in\mathbb{N} \mid S_n \leq k\},\ k\in\mathbb{N}$

Remark : $J=(J_n)_{n\in\mathbb{N}}$ is a Markov chain, called the embedded Markov chain.

Notation/definitions

the initial distribution $\alpha(i) := \mathbb{P}(J_0 = i)$
the homogeneous semi-Markov kernel $\mathbf{q} = \big(q_{ij}(\cdot)\big)_{i,j\in E}$

$$q_{ij}(k) := \begin{cases} \mathbb{P}(J_{n+1} = j, X_{n+1} = k \mid J_n = i), & k \in \mathbb{N}^* \\ 0, & k = 0 \end{cases}$$

the conditional sojourn time distribution $\mathbf{f} = \big(f_{ij}(\cdot)\big)_{i,j\in E}$

$$f_{ij}(k) := \mathbb{P}(X_{n+1} = k \mid J_n = i, J_{n+1} = j),\ f_{ij}(0) := 0$$

the transition matrix of the MC $(J_n)_{n\in\mathbb{N}}$, $\mathbf{p} = (p_{ij})_{i,j\in E}$

$$p_{ij} := \mathbb{P}(J_{n+1} = j \mid J_n = i),\ p_{ii} := 0$$

Remark : $q_{ij}(k) = p_{ij}\, f_{ij}(k)$

Semi-Markov framework - continuous time

If $(J,S) = (J_n, S_n)_{n\in\mathbb{N}}$ satisfies the relation

$$\begin{aligned} & \mathbb{P}(J_{n+1} = j, S_{n+1} - S_n \leq k | J_0, \cdots, J_n; S_1, \cdots, S_n) \\ = \quad & \mathbb{P}(J_{n+1} = j, S_{n+1} - S_n \leq k | J_n), j \in E, k \in \mathbb{R}_+ \end{aligned}$$

- (J,S) Markov renewal chain
- $Z = (Z_k)_{k\in\mathbb{R}_+}$ – semi-Markov chain associated to (J,S)

$$Z_k := J_{N(k)} \quad \Leftrightarrow \quad J_n = Z_{S_n}$$

with $N(k) := \max\{n \in \mathbb{N} \mid S_n \leq k\},\ k \in \mathbb{R}_+$

- $Q_{ij}(k) := \mathbb{P}(J_{n+1} = j, S_{n+1} - S_n \leq k | J_n = i)$ – the semi-Markov kernel

Markov processes as a particular case of SM processes (1)

1. A Markov chain (MC) $(Y_n)_{n\in\mathbb{N}}$ with transition matrix $(\widetilde{p}_{ij})_{i,j\in E}$, $\widetilde{p}_{ii} \neq 1$ for all states $i \in E$ is a semi-Markov chain, whose semi-Markov kernel $q_{ij}(k)$, transition matrix $(p_{ij})_{i,j\in E}$ of the embedded Markov chain $(J_n)_{n\in\mathbb{N}}$, and conditional sojourn time distribution $f_{ij}(k)$, $i,j \in E, k \in \mathbb{N}$, are given by :

$$
\begin{aligned}
q_{ij}(k) &= \begin{cases} \widetilde{p}_{ij}\,(\widetilde{p}_{ii})^{k-1}, & \text{if } i \neq j \text{ and } k \in \mathbb{N}^*, \\ 0, & \text{otherwise,} \end{cases} \\
p_{ij} &= \begin{cases} \frac{\widetilde{p}_{ij}}{1-\widetilde{p}_{ii}}, & \text{if } i \neq j, \\ 0, & \text{otherwise,} \end{cases} \\
f_{ij}(k) &= \begin{cases} (1-\widetilde{p}_{ii})\,(\widetilde{p}_{ii})^{k-1}, & \text{if } p_{ij} \neq 0, \\ 0, & \text{otherwise.} \end{cases}
\end{aligned}
$$

Markov processes as a particular case of SM processes (2)

2. A Markov process $(Y_n)_{n\in\mathbb{R}_+}$ with state space E and infinitesimal generator matrix $A = (a_{ij})_{i,j\in E}$ is a semi-Markov process with :

$$
\begin{aligned}
Q_{ij}(k) &= \begin{cases} \frac{a_{ij}}{a_i}(1-\exp(-a_i k)), & \text{if } i \neq j \text{ and } k \in \mathbb{R}_+, \\ 0, & \text{otherwise,} \end{cases} \\
p_{ij} &= \begin{cases} \frac{a_{ij}}{a_i}, & \text{if } i \neq j, \\ 0, & \text{otherwise,} \end{cases} \\
F_{ij}(k) &= \begin{cases} 1-\exp(-a_i k), & \text{if } p_{ij} \neq 0, \\ 0, & \text{otherwise,} \end{cases}
\end{aligned}
$$

where $a_i := -a_{ii}, i \in E$.

The hidden model SM-M0

$(Z, Y) = (Z_n, Y_n)_{n\in\mathbb{N}}-$ hidden chain $SM - M0$ if :

H1. $Z = (Z_n)_{n\in\mathbb{N}}-$ semi-Markov chain

H2. $Y = (Y_n)_{n\in\mathbb{N}}-$ conditionally independent random variables, with state space $A = \{1, \ldots, d\}$:
$\mathbb{P}(Y_n = a \mid Y_{n-1} = \cdot, \ldots, Y_0 = \cdot, Z_n = i, Z_{n-1} = \cdot, \ldots, Z_0 = \cdot) =$
$\mathbb{P}(Y_n = a \mid Z_n = i) =: R_{i;a}$

The hidden model SM-Mk

$(Z, Y) = (Z_n, Y_n)_{n\in\mathbb{N}}-$ hidden chain $SM - MK$ $(k \geq 1)$ if :

H1 $Z = (Z_n)_{n\in\mathbb{N}}-$ semi-Markov chain

H2' $Y = (Y_n)_{n\in\mathbb{N}}-$ conditionally Markov chain of order $k, k \geq 1$

$$\begin{aligned} & \mathbb{P}(Y_{n+1} = a_k \mid Y_{n-k+1}^{n} = a_0^{k-1}, Y_0^{n-k} = \cdot, Z_{n+1} = i, Z_n = \cdot, \ldots, Z_0 = \cdot) \\ = \quad & \mathbb{P}(Y_{n+1} = a_k \mid Y_{n-k+1}^{n} = a_0^{k-1}, Z_{n+1} = i) =: R_{i;a_0,\ldots,a_k} \end{aligned}$$

The hidden semi-Markov models

- quite “few” references in the literature
- first used for speech recognition : Ferguson (1980), Levinson (1986), etc. ; models known under various names : variable duration HMMs, explicit duration HMMs, etc.
- Guédon (2003, 2007), Barbu and Limnios (2008), Trevezas and Limnios (2009), Malefaki, Trevezas and Limnios (2010), Yu (2010)
- various applications : bioinformatics (Burge and Karlin, 1997) - GENSCAN (gene prediction), plant growth (Guédon, 2003), meteorology (Sansom and Thomson, 2001), internet traffic (Yu and Kobayashi, 2003), finance (Bulla and Bulla, 2006), seismology (Votsi, Limnios and Tsaklidis, 2011)
- `R` packages : `hsmm` (Bulla, Bulla and Nenadic, 2008), `mhsmm` (O’Connell and Hojsgaard, 2009)

The problem

- $Z = (Z_k)_{k \in \mathbb{N}}-$ semi-Markov chain
- Z is not directly observed
- The observations are described by a chain $Y = (Y_n)_{n \in \mathbb{N}}$, conditioned by Z
- Starting from a sample path $y = y_0^M = (y_0, \ldots, y_M)$, we want to estimate :
  - the characteristics of the SM chain
  - the conditional distribution of Y
- We consider $(Z, Y)-$ hidden chain $SM - M0$

We consider $U = (U_n)_{n\in\mathbb{N}}$ - the backward recurrence times of the semi-Markov chain $(Z_n)_{n\in\mathbb{N}}$:

$$U_n := n - S_{N(n)}.$$

Result : The chain $(Z_n, U_n)_{n\in\mathbb{N}}$ is a Markov chain with state space $E \times \mathbb{N}$ and transition matrix $\widetilde{p} = (p_{(i,t_1)(j,t_2)})_{i,j;t_1,t_2}$,

$$p_{(i,t_1)(j,t_2)} = \begin{cases} q_{ij}(t_1+1)/\overline{H_i}(t_1) & \text{if } i \neq j \text{ and } t_2 = 0, \\ \overline{H_i}(t_1+1)/\overline{H_i}(t_1) & \text{if } i = j \text{ and } t_2 - t_1 = 1, \\ 0 & \text{otherwise.} \end{cases}$$

Hypotheses **H1**

1. Regularity conditions for the SM chain (irreducible, ...).
2. The conditional sojourn time distributions $f_{ij}(\cdot), i, j \in E$, have finite support.

The associated hidden Markov model

$$\text{Hidden semi-Markov model} \quad \Leftrightarrow \quad \text{Hidden Markov model}$$

$$(Z_n, Y_n)_{n\in\mathbb{N}} \quad \Leftrightarrow \quad ((Z_n, U_n), Y_n)_{n\in\mathbb{N}}$$

Remark : We consider the natural (minimal) parameter space of the hidden Markov model $((Z,U),Y)$: (i) we remove all the identical zero parameters ; (ii) we remove all the dependent parameters.

The parameter space has the shape $\Theta := \Theta_1 \times \Theta_2$:

- Θ_1 - the parameter space of (Z,U)
- Θ_2 - the parameter space of Y

with a canonical element

$$\theta = \left((p_{(i,t_1)(j,t_2)})_{i,j,t_1,t_2}, (R_{ia})_{i,a} \right) .$$

Asymptotic results

Theorem

For a sample of observations $\{Y_0 = y_0, \ldots, Y_M = y_M\}$, *we have*

$$\widehat{\theta}(M) \xrightarrow[M\to\infty]{a.s.} \theta^0.$$

$\rightarrow$ Convergence of $(\widehat{R}_{ia}(M))_{i,a}$, $(\widehat{q}_{ij}(k,M))_{i,j,k}$

For a sample $(Y_0, \ldots, Y_M)$, let us consider the likelihood function for an observation of the chain $((Z,U),Y)$

$$p_\theta(Y_0^n) = \sum_{z_0^M, u_0^M} \pi_{z_0,u_0} \prod_{k=1}^{M} p_{(z_{k-1},u_{k-1})(z_k,u_k)} \prod_{k=0}^{M} R_{z_k;Y_k}, \qquad (1)$$

where $(\pi_{i,u})_{i,u}$ is the stationary distribution of (Z,U), $\pi_{i,u} = \frac{1-H_i(u)}{\mu_{ii}}$.

Let us define

$$\sigma_{Y_0^n}(\theta^0) := -\mathbb{E}_{\theta^0}\left(\frac{\partial^2 \log p(Y_0^n)}{\partial\theta_u \partial\theta_v}\Big|_{\theta=\theta^0}\right)_{u,v}$$

and

$$\sigma(\theta^0) := -\mathbb{E}_{\theta^0}\left(\frac{\partial^2 \log \mathbb{P}_\theta(Y_0 \mid Y_{-1}, Y_{-2}, \ldots)}{\partial\theta_u \partial\theta_v}\Big|_{\theta=\theta^0}\right)_{u,v}.$$

Hypothesis

H2 *There exists an $n \in \mathbb{N}$ such that the matrix $\sigma_{Y_0^n}(\theta^0)$ is invertible.*

Using Douc(2005), we obtain : $\sigma(\theta^0)$ invertible iff **H2** fulfilled.

Theorem

The vector

$$\sqrt{M}\left[\widehat{\theta}(M)-\theta^0\right]=\sqrt{M}\Bigg[\left((\widehat{p}_{(i,t_1)(j,t_2)}(M))_{i,j,t_1,t_2},(\widehat{R}_{ia}(M))_{i,a}\right)$$

$$-\quad\left((p^0_{(i,t_1)(j,t_2)})_{i,j,t_1,t_2},(R^0_{ia})_{i,a}\right)\Bigg]$$

is asymptotically normal, as $M\to\infty$, *with zero mean and covariance matrix* $\sigma(\theta^0)^{-1}$.

$\to$ Asymptotic normality of $(\widehat{R}_{ia}(M))_{i,a}$ and $(\widehat{q}_{ij}(k,M))_{i,j,k}$

Remark : The same type of result holds true when the emission probabilities of Y_n depend also on the backward recurrence times (Trevezas and Limnios, 2009).

How to find the estimators ?

Question : how can we practically obtain the estimators, since Z_k and U_k are not directly observable ?

Algorithmic approaches :

1. The popular algorithm EM (Expectation - Maximization)
2. Several works on EM algorithm for HSMM : Ferguson (1980), Guédon (2003), Yu et Kobayashi (2003), Barbu and Limnios (2006)
3. Known problems of the EM algorithm : dependence on the starting point, slow convergence, convergence to non-optimal solutions, choice of the stopping rule, etc.

Possible alternative : use a stochastic version of the EM algorithm

1. Several stochastic EM versions, based on the same idea : replace the computation of the conditional expectation during the E step by a simulation step.
2. SEM (Stochastic EM) - Celeux and Diebolt (1985)
   MCEM (Monte Carlo EM) - Wei and Tanner (1990)
   SAEM (Stochastic Approximation EM) - Delyon, Lavielle and Moulines (1999)
3. A stochastic EM version for HSMM : Malefaki, Trevezas and Limnios (2010)

EM algorithm

Dempster et al. (1977), Baum et al. (1970)

Consider :

- $(Z, Y)-$ a hidden chain $SM - M1$
- $\mathbf{y}_0^M = \{Y_0 = y_0, \ldots, Y_M = y_M\}-$ sample of Y
- $\mathbf{z}_0^M = \{Z_0 = z_0, \ldots, z_M = z_M\}-$ sample of Z
- $f_M(\mathbf{y}_0^M, \mathbf{z}_0^M \mid \theta)-$ the likelihood function

The idea :

Maximize $\mathbb{E}_{\theta^{(m)}}[\log(f_M(Y_0^M, Z_0^M \mid \theta)) \mid \mathbf{y}_0^M] =: Q(\theta \mid \theta^{(m)})$ instead of $\log(f_M(Y_0^M, Z_0^M \mid \theta))$.

EM for SM-M0

We get the following auxiliary Q-function :

$$
\begin{aligned}
Q(\theta \mid \theta^{(m)}) \quad &:= \quad \mathbb{E}_{\theta^{(m)}} \left[\log(f_M(\mathbf{Z}_0^M, \mathbf{Y}_0^M \mid \theta)) \mid \mathbf{y} \right] \\
&= \quad \sum_{\substack{i,j,k \\ i \neq j}} \log q_{ij}(k) \sum_{n=0}^{M-k} \mathbb{P}_{\theta^{(m)}} \left(Z_{n+k} = j, \mathbf{Z}_n^{n+k-1} = \boldsymbol{i}, Z_{n-1} \neq i \mid \mathbf{y} \right) \\
&\quad + \sum_i \sum_\alpha \log R_{i;\alpha} \sum_{n=0}^{M} \mathbb{P}_{\theta^{(m)}} \left(Z_n = i, Y_n = \alpha \mid \mathbf{y} \right) \\
&\quad + \mathbb{E}_{\theta^{(m)}} \left(\log \bar{H}_{Z_M}(U_M) \mid \mathbf{y} \right). \qquad (2)
\end{aligned}
$$

Estimators for SM-M0

$$q_{ij}^{(m+1)}(k) = \frac{\sum_{n=0}^{M-k} \mathbb{P}_{\theta^{(m)}}(Z_{n+k} = j, \mathbf{Z}_n^{n+k-1} = \boldsymbol{i}, Z_{n-1} \neq i \mid \mathbf{y})}{\sum_{n=0}^{M-1} \mathbb{P}_{\theta^{(m)}}(Z_n = i, Z_{n-1} \neq i \mid \mathbf{y}) - \mathbb{P}_{\theta^{(m)}}(Z_M = i, Z_{M-1} = i \mid \mathbf{y})}, \tag{3}$$

$$R_{i;\alpha}^{(m+1)} = \frac{\sum_{n=0}^{M} \mathbb{1}_{\{Y_n = \alpha\}} \mathbb{P}_{\theta^{(m)}}(Z_n = i \mid \mathbf{y})}{\sum_{n=0}^{M} \mathbb{P}_{\theta^{(m)}}(Z_n = i \mid \mathbf{y})}. \tag{4}$$

Notation

For any $n \in \mathbb{N}, j \in E$:

$$
\begin{aligned}
F_n(j) &:= \mathbb{P}(Z_n = j, Z_{n-1} \neq j \mid \mathbf{y}_0^n) \\
L_n(j) &:= \mathbb{P}(Z_n = j \mid \mathbf{y}_0^M) = L_{1;n}(j) + L_{2;n}(j) \\
L_{1;n}(j) &:= \mathbb{P}(Z_n = j, Z_{n-1} \neq j \mid \mathbf{y}_0^M) \\
L_{2;n}(j) &:= \mathbb{P}(Z_n = j, Z_{n-1} = j \mid \mathbf{y}_0^M) \\
P_n &:= \mathbb{P}(\mathbf{y}_n \mid \mathbf{y}_0^{n-1})
\end{aligned}
$$

Estimation for SM-M1

Forward

- For $n = 0$:

$$F_0(j) \quad = \quad \mu(j, y_0) / \sum_{i \in E} \mu(i, y_0), \qquad P_0 = \sum_{i \in E} \mu(i, y_0),$$

with $\mu(i, a) := \mathbb{P}(Z_0 = i, Y_0 = a)$.

- For $n = 1, \ldots, M$:

$$P_n \quad = \quad \text{fnct}(P_k,\ F_k(i),\ q_{ij}(t)\ R_{i;y_k,y_{k+1}}, k = 0, \ldots, n-1, t \leq M)$$

$$F_n(j) \quad = \quad \sum_{t=1}^{n-1} \sum_{i \neq j} \frac{q_{ij}(t) F_{n-t}(i) R_{j;y_{n-1},y_n} \prod_{p=n-t+1}^{n-1} R_{i;y_{p-1},y_p}}{\prod_{p=n-t+1}^{n} P_p}$$

$$+ \sum_{i \neq j} \frac{q_{ij}(n) \mu(i, y_0) R_{j;y_{n-1},y_n} \prod_{p=1}^{n-1} R_{i;y_{p-1},y_p}}{\prod_{p=0}^{n} P_p}$$

Backward

- For $n = M$:

$$
\begin{aligned}
L_{1;M}(i) &= \mathbb{P}(Z_M = i, Z_{M-1} \neq i \mid \mathbf{y}_0^M) = F_M(i) \\
L_{2;M}(i) &= \sum_{u=2}^{M} \frac{[1 - \sum_{l=1}^{u-1} \sum_{j \in E} q_{ij}(l)] F_{M-u+1}(i) \prod_{p=M-u+2}^{M} R_{i;y_{p-1},y_p}}{\prod_{l=M-u+2}^{M} P_l} \\
&\quad + \frac{\mu(i, y_0)[1 - \sum_{l=1}^{M} \sum_{j \in E} q_{ij}(l)] \prod_{p=1}^{M} R_{i;y_{p-1},y_p}}{\prod_{l=0}^{M} P_l}
\end{aligned}
$$

- For $n = M - 1, \ldots, 1$:

$$
\begin{aligned}
L_{1;n}(i) &= \text{fnct}(L_{1;m}(i), m = M, \ldots, n+1, P_k,\ F_k(i),\ q_{ij}(t),\ R_{i;y_k,y_{k+1}}) \\
L_{2;n}(i) &= \text{fnct}(L_{1;m}(i), m = M, \ldots, n+1, P_k,\ F_k(i),\ q_{ij}(t),\ R_{i;y_k,y_{k+1}})
\end{aligned}
$$

- For $n = 0$:

$$
L_0(i) = \text{fnct}(L_{1;m}(i), m = M, \ldots, 1, P_k,\ F_k(i),\ q_{ij}(t),\ R_{i;y_k,y_{k+1}})
$$

Maximization for SM-M1

Maximize $\mathbb{E}_{\theta^{(m)}}[\log(f_M(Y_0^M, Z_0^M \mid \theta)) \mid \mathbf{y}_0^M]$ with respect to θ :

$$
\begin{aligned}
f_{ij}^{(m+1)}(k) &= \frac{\sum_{n=0}^{M-k} \mathbb{P}_{\theta^{(m)}}(Z_{n+k}=j, Z_{n+k-1}=i, \ldots, Z_n=i, Z_{n-1}\neq i \mid \mathbf{y})}{\sum_{n=1}^{M} \mathbb{P}_{\theta^{(m)}}(Z_n=j, Z_{n-1}=i \mid \mathbf{y})}, \\
p_{ij}^{(m+1)} &= \frac{\sum_{n=1}^{M} \mathbb{P}_{\theta^{(m)}}(Z_n=j, Z_{n-1}=i \mid \mathbf{y})}{\sum_{n=0}^{M-1} \mathbb{P}_{\theta^{(m)}}(Z_n=i, Z_{n-1}\neq i \mid \mathbf{y})}, \\
q_{ij}^{(m+1)}(k) &= f_{ij}^{(m+1)}(k) p_{ij}^{(m+1)}, \\
R_{i;a,b}^{(m+1)} &= \frac{\sum_{n=1}^{M} L_n^{(m)}(i) \mathbf{1}_{\{Y_{n-1}=a, Y_n=b\}}}{\sum_{n=1}^{M} L_n^{(m)}(i) \mathbf{1}_{\{Y_{n-1}=a\}}}.
\end{aligned}
$$

$$
\theta^{(m+1)} = (p^{(m+1)}, f^{(m+1)}, R^{(m+1)})
$$

Application : CpG island detection

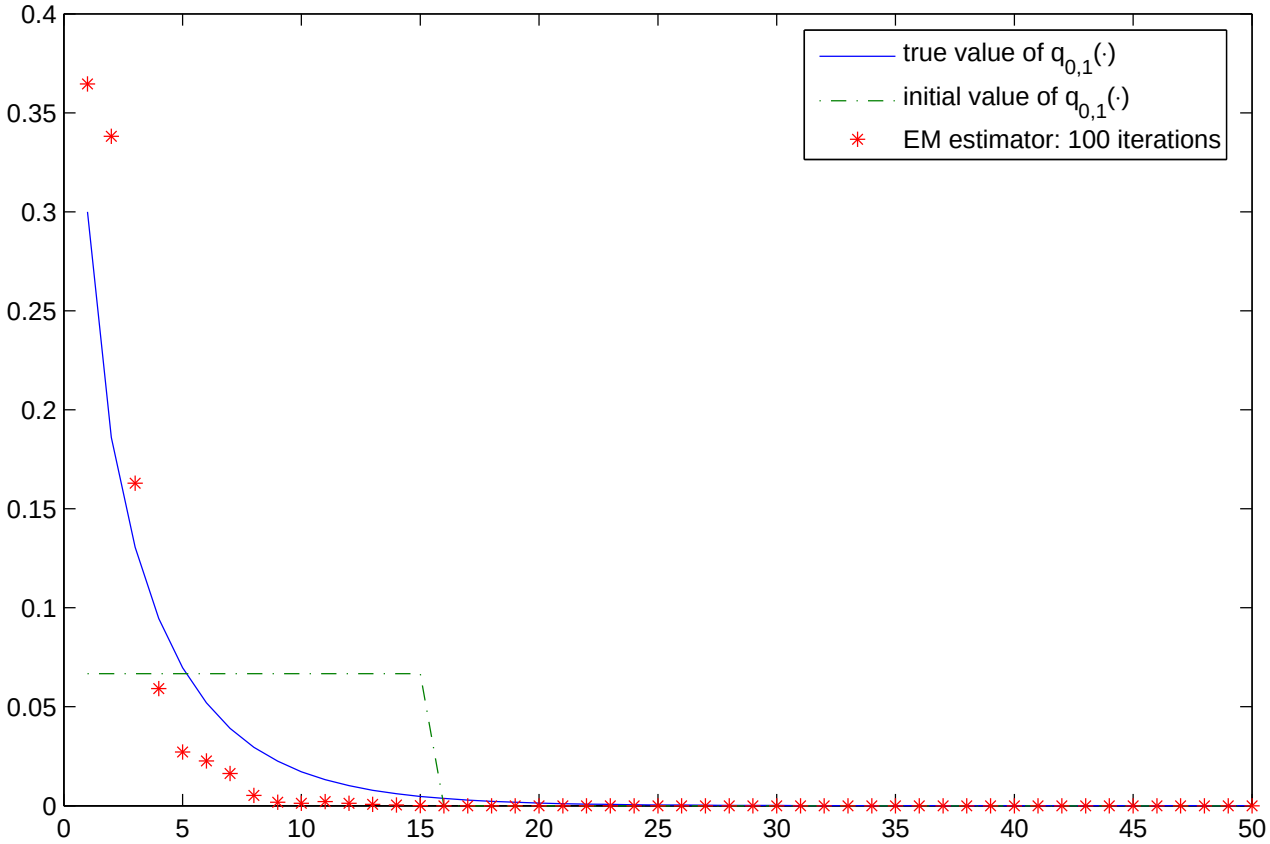
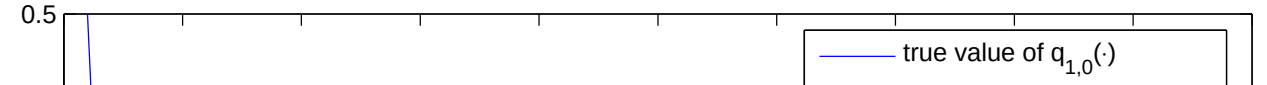


FIGURE – Estimator of the hidden semi-Markov kernel



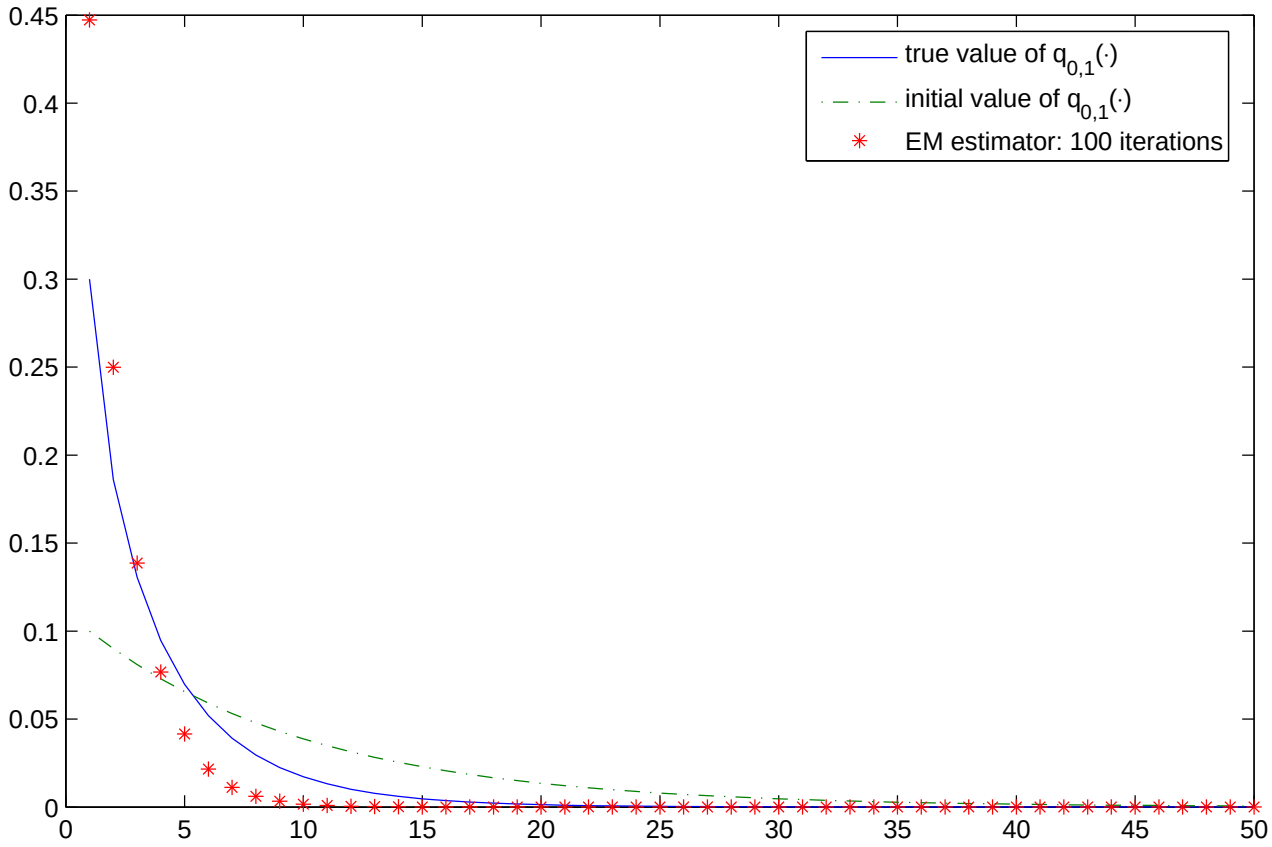
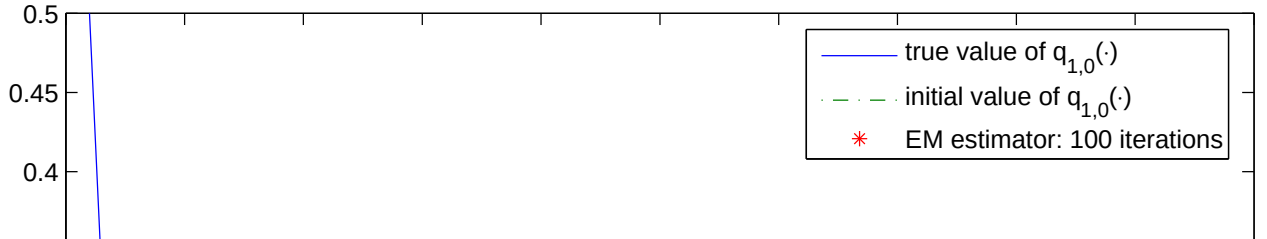


FIGURE – Estimator of the hidden semi-Markov kernel



General hidden semi-Markov models

- based on S. Trevezas, N. Limnios, 2009 (backward recurrence time HSMM) and S. Malefaki, S. Trevezas, N. Limnios, 2010 (SEM)

Let
$(Z, Y) := (Z_n, Y_n)_{n\in\mathbb{N}}$ defined on $(\Omega, \mathcal{A}, \mathbb{P}_\theta)$, where

- Z : an unobservable semi-Markov chain (SMC) with finite state space $E = \{1, 2, \ldots, s\}$
- Y : an observable process with $(Y_n)_{n\in\mathbb{N}} \mid Z$ forming a sequence of conditionally independent r.v. with values in a σ-finite measured space $(\mathcal{Y}, \mathcal{B}(\mathcal{Y}), \nu)$

Backward recurrence time dependence

- In the usual setting of a general HSMM,

$$\mathbb{P}_\theta(Y_0^n \in A_0 \times \ldots \times A_n \mid Z)$$

$$= \prod_{m=0}^{n} \mathbb{P}_\theta(Y_m \in A_m \mid Z_m), \tag{5}$$

- while of SM-M0-B type we have

$$= \prod_{m=0}^{n} \mathbb{P}_\theta(Y_m \in A_m \mid Z^m_{S_{N(m)}}). \tag{6}$$

Y_m does not depend only locally on Z_m, but also on the time period $m - S_{N(m)}$ that Z has stayed at this state.

Backward recurrence time dependence $SM - Mk - B$ type

- In the usual setting of a HSMM,

$$\mathbb{P}_\theta(Y_0^n = y_0^n \mid Z)$$

$$= \prod_{m=0}^{n} \mathbb{P}_\theta(Y_m = y_m \mid Z_m), \tag{7}$$

- while of SM-Mk-B type we have

$$= \prod_{m=0}^{n} \mathbb{P}_\theta(Y_m = y_m \mid Y_{m-k}^{m-1}, Z_{S_{N(m)}}^{m}). \tag{8}$$

Y_m does not depend only locally on Z_m, but also on the previous k-observations Y_{m-k}^{m-1} and on the time period $m - S_{N(m)}$ that Z has stayed at this state.

SEM

The simplest stochastic version of EM algorithm is the Stochastic EM (SEM), which was proposed by Celeux and Diebolt (1985). The SEM algorithm replace the function $Q(\theta|\theta^{(m)})$ by

$$\hat{Q}(\theta|\theta^{(m)}) = \log f(\mathbf{Z}_0^M, \mathbf{Y}_0^M|\theta), \tag{9}$$

where $\mathbf{Z}_0^M$ is an observation drawn from $\mathbb{P}_{\theta^{(m)}}(\cdot|\mathbf{y}_0^M)$.

EM for SM-M1-B

$$
\begin{aligned}
Q(\theta \mid \theta^{(m)}) \quad = \quad & \sum_{i} \sum_{u=0}^{\widetilde{n}_i - 2} \log p_{(i,u)(i,u+1)} \sum_{n=0}^{M-1} \mathbb{P}_{\theta^{(m)}}(Z_n = i, U_n = u, Z_{n+1} = i \mid \mathbf{y}) \\
& + \sum_{\substack{i,j \\ i \neq j}} \sum_{u=0}^{\widetilde{n}_i - 1} \log p_{(i,u),(j,0)} \sum_{n=0}^{M-1} \mathbb{P}_{\theta^{(m)}}(Z_n = i, U_n = u, Z_{n+1} = j \mid \mathbf{y}) \\
& + \sum_{i} \sum_{u=0}^{\widetilde{n}_i - 1} \sum_{\alpha,b} \log R_{i,u;\alpha,b} \sum_{n=1}^{M} \mathbb{1}_{\{Y_{n-1}=\alpha, Y_n = b\}} \mathbb{P}_{\theta^{(m)}}(Z_n = i, U_n = u \mid \mathbf{y})
\end{aligned}
$$

Estimators for SM-M1-B

$$p^{(m+1)}_{(i,u)(j,0)} = \frac{\sum_{n=0}^{M-1} \mathbb{P}_{\theta^{(m)}}(Z_n = i, U_n = u, Z_{n+1} = j \mid \mathbf{y})}{\sum_{n=0}^{M-1} \mathbb{P}_{\theta^{(m)}}(Z_n = i, U_n = u \mid \mathbf{y})}, \tag{10}$$

$$p^{(m+1)}_{(i,u)(i,u+1)} = 1 - \sum_{j \neq i} p^{(m+1)}_{(i,u)(j,0)}, \tag{11}$$

$$R^{(m+1)}_{i,u;\alpha,b} = \frac{\sum_{n=1}^{M} \mathbb{1}_{\{Y_{n-1}=\alpha, Y_n=b\}} \mathbb{P}_{\theta^{(m)}}(Z_n = i, U_n = u \mid \mathbf{y})}{\sum_{n=1}^{M} \mathbb{P}_{\theta^{(m)}}(Z_n = i, U_n = u \mid \mathbf{y})}. \tag{12}$$

Complexity in time of the EM algorithms for HSMMs :

- This EM algorithm of S. Malefaki, S. Trevezas, N. Limnios (2010) : $O(M^2 s^2)$;
- The EM algorithm of V. S. Barbu, N. Limnios (2006, 2008) : $O(M^3 s^2)$;
- The EM algorithm of Y. Guédon (2003) : $O(M^2 s + M s^2)$.

Viterbi algorithm (1/2)

- Proposed in C.-E. Pertsinidou, N. Limnios (2015).
- Main idea : use the associated HMM $((Z_n, U_n), Y_n)_{n\in\mathbb{N}}$.
- Obtained complexity : $O(Ms^2)$, the same as the one of the Viterbi algorithm for HMMs. This is an important improvement as compared to existing Viterbi algorithms for HSMM.
- Specifications of the algorithm for : HSMMs of SM-M0 type and SM-M1 type ; with possible transitions to the same state.

Viterbi algorithm (2/2)

Algorithm 3.3. ***Step 1.*** *Initial conditions. For* $k = 0$,

$$d_0(i_0) = \log_2(\tilde{\alpha}(i_0)) + \log_2(R(i_0, y_0)), \quad b_0(i_0) = 0.$$

Step 2. *For* $k \geq 1$

$$d_k(i_k) = \max_{i_{k-1} \in E} [d_{k-1}(i_{k-1}) + \log_2(a_k(i_0^k))] + \log_2(R(i_k, y_k))$$

$$b_k(i_k) = \arg \max_{i_{k-1} \in E} [d_{k-1}(i_{k-1}) + \log_2(a_k(i_0^k))]$$

where

$$a_k(i_0^k) = \sum_{l_1=0}^{k-1} \sum_{l_2 \in \{0, l_1+1\}} \tilde{p}_{\{(i_{k-1}, l_1)(i_k, l_2)\}} \times 1_{\{i_{k-1} = \ldots = i_{k-l_1-1} \neq i_{k-l_1-2}\}}, \quad k \geq 1.$$

Step 3. *Termination If* $k = T - 1$ *(*T *is the length of the observation sequence),*

$$P = \max_{i_{T-1} \in E} [d_{T-1}(i_{T-1})], \quad P^* = 2^{max_{i_{T-1} \in E}[d_{T-1}(i_{T-1})]},$$

$$q_{T-1} = \arg \max_{i_{T-1} \in E} [d_{T-1}(i_{T-1})].$$

Step 4. *Sequence of states with backward-forward steps*

$$q_{k-1} = b_k(q_k), \quad k = T-1, \ldots, 1.$$

$$a_k\left(i_0^k\right) = \sum_{l_1=0}^{k-1} \sum_{l_2 \in \{0, l_1+1\}} \tilde{p}_{(i_{k-1}, l_1)(i_k, l_2)} \times 1_{\{i_{k-1} = \ldots = i_{k-l_1-1} \neq i_{k-l_1-2}\}}, \quad k \geq 1$$

Conclusions and perspectives

The hidden semi-Markov processes

- Flexibility/interest from a practical point of view
- Rich theoretical framework
- Can be adapted to various problems, in various application field ; methods can be “borrowed” from a field to another

Extensions

- modeling : choice between homogeneous $\leftrightarrow$ nonhomogeneous process
- statistical approach : choice between parametric $\leftrightarrow$ nonparametric estimation
- take into account the covariables, when extra data is available ; consider mixed hidden semi-Markov models
- study of algorithmic techniques to accelerate the estimation of parameters.