Plan

Hidden semi-Markov models [1]

Vlad BARBU

Laboratoire de Mathématiques Appliquées de Compiègne
Université de Technologie de Compiègne

vlad.barbu@dma.utc.fr

[1] joint work with Nikolaos LIMNIOS, Université de Technologie de Compiègne

Plan

Plan

Plan

Plan

Plan

Plan

Hidden model

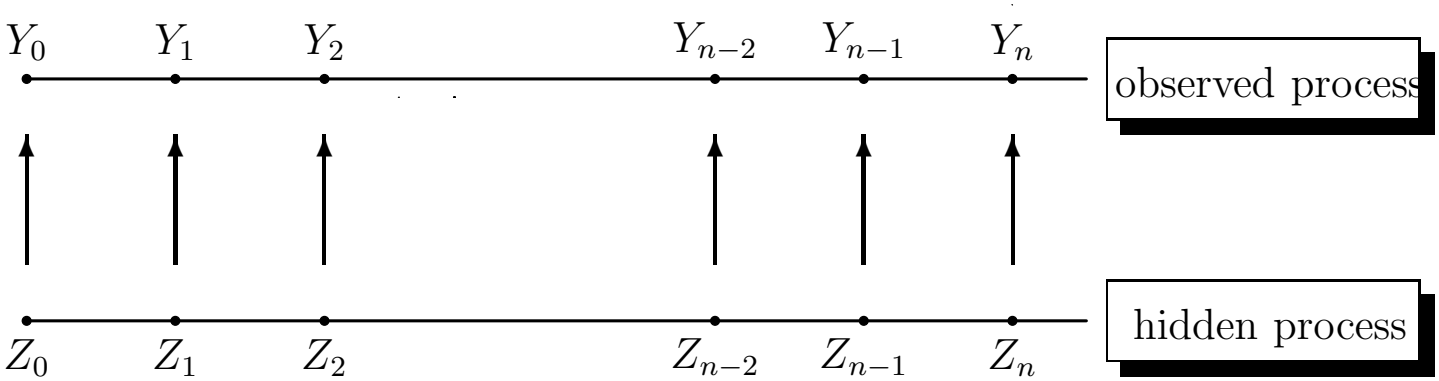

Figure 1: Hidden model

$(Y_n)_{n\in\mathbb{N}}$: received signal / GPS position / DNA sequence
$(Z_n)_{n\in\mathbb{N}}$: emitted signal / real position / related to DNA structure

Hidden Markov models (Baum and Petrie, 1966)

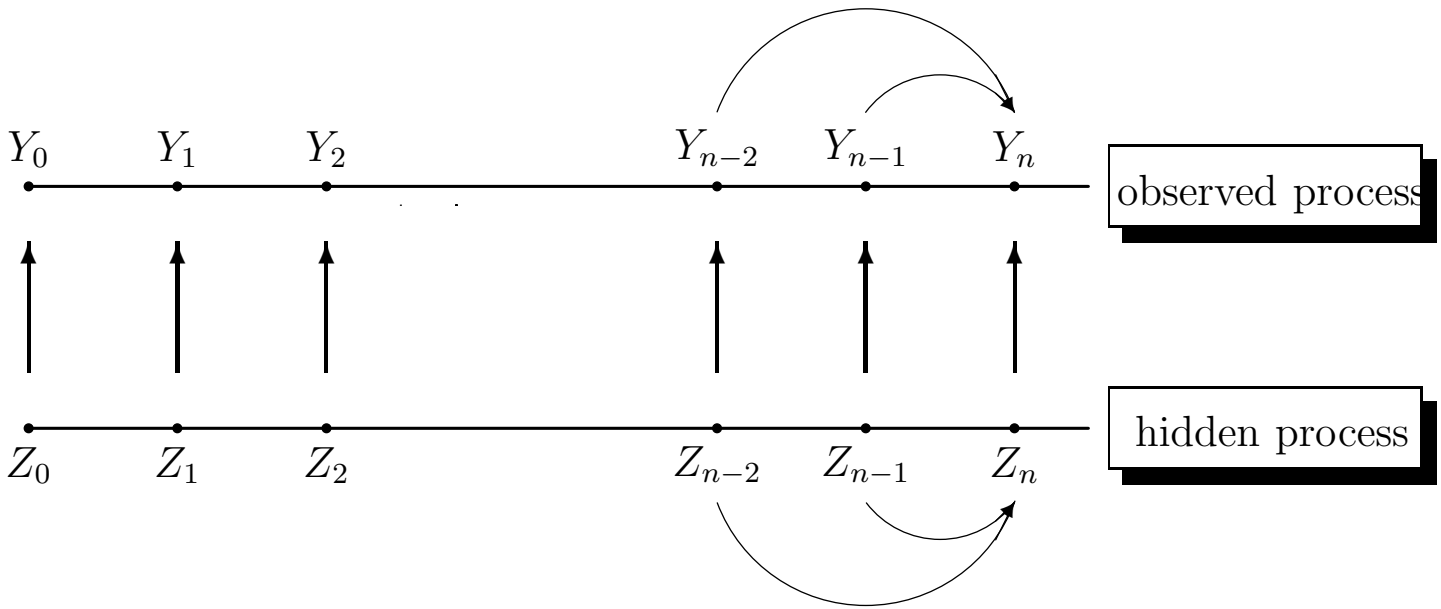

Figure 2: Hidden Markov model

$(Z_n, Y_n)_{n\in\mathbb{N}}$ – $(Z_n)_{n\in\mathbb{N}}$ Markov process (of order k)

– $(Y_n)_{n\in\mathbb{N}}$ depends on $(Z_n)_{n\in\mathbb{N}}$, conditionally independent (or Markov)

- applications : Churchill(1989, 1992), Durbin et al.(1998), Rabiner(1989), Ephraim and Merhav(2002)
- statistical properties : Leroux (1992), Bickel et al. (1998)
- associated algorithmic studies

Problem : constraints on the sojourn times (Markov)

$\Downarrow$

Hidden semi-Markov models : $(Z_n)_{n\in\mathbb{N}}$ semi-Markov process
Ferguson (1980), Levinson(1986), Guédon (2003)

- applications : Churchill(1989, 1992), Durbin et al.(1998), Rabiner(1989), Ephraim and Merhav(2002)
- statistical properties : Leroux (1992), Bickel et al. (1998)
- associated algorithmic studies

Problem : constraints on the sojourn times (Markov)

$\Downarrow$

Hidden semi-Markov models : $(Z_n)_{n\in\mathbb{N}}$ semi-Markov process
Ferguson (1980), Levinson(1986), Guédon (2003)

- applications : Churchill(1989, 1992), Durbin et al.(1998), Rabiner(1989), Ephraim and Merhav(2002)
- statistical properties : Leroux (1992), Bickel et al. (1998)
- associated algorithmic studies

Problem : constraints on the sojourn times (Markov)

$\Downarrow$

Hidden semi-Markov models : $(Z_n)_{n\in\mathbb{N}}$ semi-Markov process
Ferguson (1980), Levinson(1986), Guédon (2003)

Plan

Discrete-time semi-Markov model

- $Z = (Z_k)_{k\in\mathbb{N}}-$ with state space $E = \{1, 2, \ldots, s\}$
- $J = (J_n)_{n\in\mathbb{N}}$, visited states, $E-$ valued MC
  $S = (S_n)_{n\in\mathbb{N}}$, *jump times*, $\mathbb{N}-$ valued chain
  $X = (X_n)_{n\in\mathbb{N}}$, *times between two successive jumps*

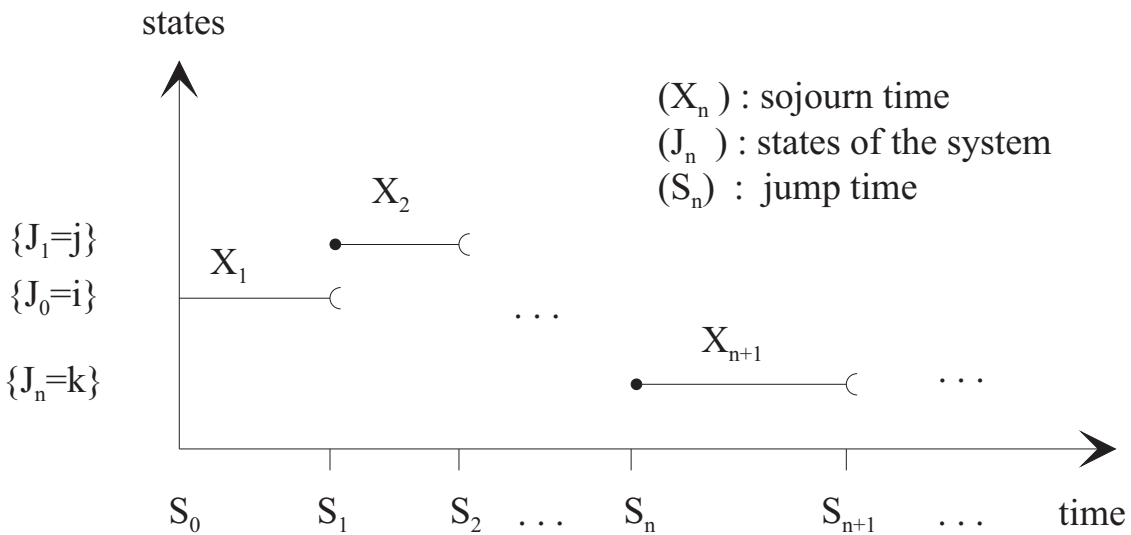

FIG.: A sample path of a semi-Markov chain

Discrete-time semi-Markov model

- $Z = (Z_k)_{k\in\mathbb{N}}-$ with state space $E = \{1, 2, \ldots, s\}$
- $J = (J_n)_{n\in\mathbb{N}}$, visited states, $E-$ valued MC
  $S = (S_n)_{n\in\mathbb{N}}$, *jump times*, $\mathbb{N}-$ valued chain
  $X = (X_n)_{n\in\mathbb{N}}$, *times between two successive jumps*

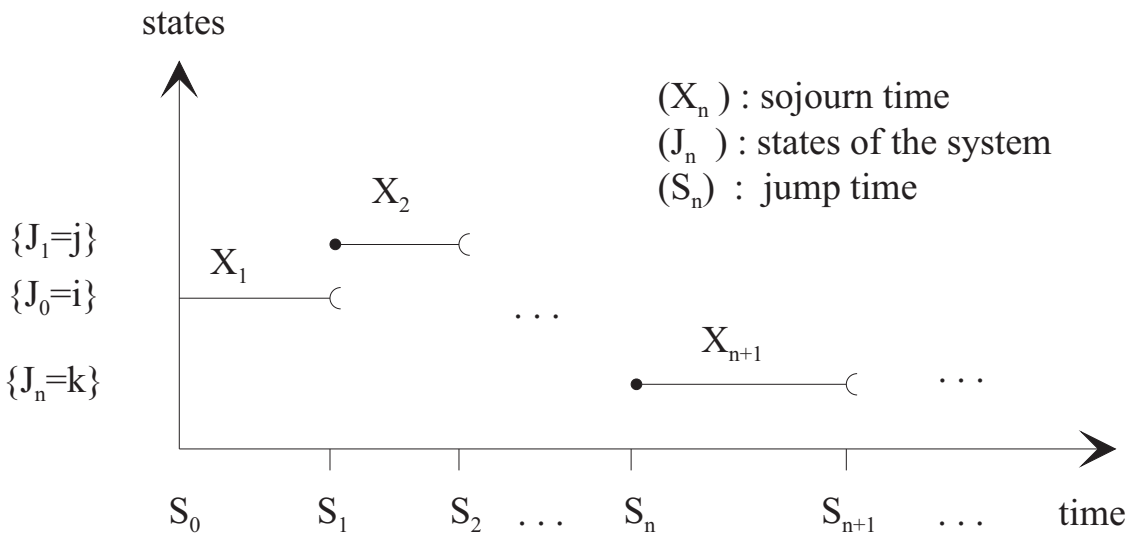

FIG.: A sample path of a semi-Markov chain

If (J, S) verifies :

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

- (J, S) Markov renewal chain (MRC)
- $Z = (Z_k)_{k \in \mathbb{N}}$ – the associated semi-Markov chain
- $P = \left(P_{ij}(\cdot)\right)_{i,j \in E}$ – the transition matrix of Z

$$P_{ij}(k) := \mathbb{P}(Z_k = j \mid Z_0 = i), \ i,j \in E, \ k \in \mathbb{N}.$$

If (J, S) verifies :

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

- (J, S) Markov renewal chain (MRC)
- $Z = (Z_k)_{k \in \mathbb{N}}$ – the associated semi-Markov chain
- $P = \left(P_{ij}(\cdot)\right)_{i,j \in E}$ – the transition matrix of Z

$$P_{ij}(k) := \mathbb{P}(Z_k = j \mid Z_0 = i), \ i, j \in E, \ k \in \mathbb{N}.$$

If (J, S) verifies :

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

- (J, S) Markov renewal chain (MRC)
- $Z = (Z_k)_{k \in \mathbb{N}}$ – the associated semi-Markov chain
- $P = \left(P_{ij}(\cdot)\right)_{i,j \in E}$ – the transition matrix of Z

$$P_{ij}(k) := \mathbb{P}(Z_k = j \mid Z_0 = i), \ i, j \in E, \ k \in \mathbb{N}.$$

The S-M chain Z is characterized by :

1. the initial distribution $\alpha(i) := \mathbb{P}(J_0 = i)$,
2. • the S-M kernel $q = \left(q_{ij}(\cdot)\right)_{i,j\in E}$

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

or

- the sojourn time distributions $f = \left(f_{ij}(\cdot)\right)_{i,j\in E}$

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

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

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

$$q_{ij}(k) = p_{ij} f_{ij}(k)$$

The S-M chain Z is characterized by :

1. the initial distribution $\alpha(i) := \mathbb{P}(J_0 = i)$,
2. • the S-M kernel $q = \big(q_{ij}(\cdot)\big)_{i,j\in E}$

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

or

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

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

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

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

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

The S-M chain Z is characterized by :

1. the initial distribution $\alpha(i) := \mathbb{P}(J_0 = i)$,
2. • the S-M kernel $q = \left(q_{ij}(\cdot)\right)_{i,j\in E}$

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

or

- the sojourn time distributions $f = \left(f_{ij}(\cdot)\right)_{i,j\in E}$

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

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

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

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

The S-M chain Z is characterized by :

1. the initial distribution $\alpha(i) := \mathbb{P}(J_0 = i)$,
2. • the S-M kernel $q = \left(q_{ij}(\cdot)\right)_{i,j\in E}$

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

or

- the sojourn time distributions $f = \left(f_{ij}(\cdot)\right)_{i,j\in E}$

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

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

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

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

The S-M chain Z is characterized by :

1. the initial distribution $\alpha(i) := \mathbb{P}(J_0 = i)$,
2. • the S-M kernel $q = \left(q_{ij}(\cdot)\right)_{i,j\in E}$

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

or

- the sojourn time distributions $f = \left(f_{ij}(\cdot)\right)_{i,j\in E}$

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

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

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

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

The hidden $SM1 - M0$ model

$(Z, Y) = (Z_n, Y_n)_{n\in\mathbb{N}}-$ hidden $SM1 - M0$ chain if :

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

H2. $Y = (Y_n)_{n\in\mathbb{N}}-$ conditionally independent r.v., 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 $SM1 - M0$ model

$(Z, Y) = (Z_n, Y_n)_{n\in\mathbb{N}}-$ hidden $SM1 - M0$ chain if :

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

H2. $Y = (Y_n)_{n\in\mathbb{N}}-$ conditionally independent r.v., 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 $SM1 - Mk$ model

$(Z, Y) = (Z_n, Y_n)_{n\in\mathbb{N}}-$ hidden $SM1 - Mk$ chain $(k \geq 1)$ if :

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

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

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

The hidden $SM1 - Mk$ model

$(Z, Y) = (Z_n, Y_n)_{n\in\mathbb{N}}-$ hidden $SM1 - Mk$ chain $(k \geq 1)$ if :

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

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

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

Plan

The problem

- $(Z, Y)-$ hidden $SM1 - M0$ 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 underlying semi-Markov chain
  - the conditional distribution of Y.

The problem

- $(Z, Y)-$ hidden $SM1 - M0$ 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 underlying semi-Markov chain
  - the conditional distribution of Y.

Definition

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

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

Theorem

The chain $(Z_n, U_n)_{n\in\mathbb{N}}$ is a Markov chain with state space $E \times \mathbb{N}$.

$\rightarrow$ The transition matrix : $\widetilde{p} = (p_{(i,t_1)(j,t_2)})_{i,j\in E, t_1,t_2\in\mathbb{N}}$

Definition

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

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

Theorem

The chain $(Z_n, U_n)_{n\in\mathbb{N}}$ is a Markov chain with state space $E \times \mathbb{N}$.

$\rightarrow$ The transition matrix : $\widetilde{p} = (p_{(i,t_1)(j,t_2)})_{i,j\in E, t_1,t_2\in\mathbb{N}}$

Definition

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

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

Theorem

The chain $(Z_n, U_n)_{n\in\mathbb{N}}$ is a Markov chain with state space $E \times \mathbb{N}$.

$\rightarrow$ The transition matrix : $\widetilde{p} = (p_{(i,t_1)(j,t_2)})_{i,j\in E, t_1,t_2\in\mathbb{N}}$

Definition

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

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

Theorem

The chain $(Z_n, U_n)_{n\in\mathbb{N}}$ is a Markov chain with state space $E \times \mathbb{N}$.

$\rightarrow$ The transition matrix : $\widetilde{p} = \left(p_{(i,t_1)(j,t_2)}\right)_{i,j\in E, t_1,t_2\in\mathbb{N}}$

The associated hidden Markov model

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

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

$$\theta = ((q_{ij}(k)), (R_{ia})) \quad \Leftrightarrow \quad \theta = \left((p_{(i,t_1)(j,t_2)}), (R_{ia})\right)$$

Recall

$$\mathbb{P}(Y_n = a \mid Y_{n-1} = \cdot, \ldots, Y_0 = \cdot, Z_n = i, Z_{n-1} = \cdot, \ldots, Z_0 = \cdot) = \mathbb{P}(Y_n = a \mid Z_n = i)$$

$$\theta^0 = \left((p^0_{(i,t_1)(j,t_2)})_{\substack{i,j\in E \\ t_1,t_2\in D}}, (R^0_{ia})_{i\in E, a\in A} \right) \text{ – true value of } \theta$$

The associated hidden Markov model

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

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

$$\theta = ((q_{ij}(k)), (R_{ia})) \quad \Leftrightarrow \quad \theta = \left((p_{(i,t_1)(j,t_2)}), (R_{ia})\right)$$

Recall

$$\mathbb{P}(Y_n = a \mid Y_{n-1} = \cdot, \ldots, Y_0 = \cdot, Z_n = i, Z_{n-1} = \cdot, \ldots, Z_0 = \cdot) = \mathbb{P}(Y_n = a \mid Z_n = i)$$

$$\theta^0 = \left((p^0_{(i,t_1)(j,t_2)})_{\substack{i,j\in E \\ t_1,t_2\in D}}, (R^0_{ia})_{i\in E, a\in A} \right) - \text{true value of } \theta$$

The associated hidden Markov model

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

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

$$\theta = ((q_{ij}(k)), (R_{ia})) \quad \Leftrightarrow \quad \theta = \left((p_{(i,t_1)(j,t_2)}), (R_{ia})\right)$$

Recall

$$\mathbb{P}(Y_n = a \mid Y_{n-1} = \cdot, \ldots, Y_0 = \cdot, Z_n = i, Z_{n-1} = \cdot, \ldots, Z_0 = \cdot) = \mathbb{P}(Y_n = a \mid Z_n = i)$$

$\theta^0 = \left((p^0_{(i,t_1)(j,t_2)})_{\substack{i,j\in E \\ t_1,t_2\in D}}, (R^0_{ia})_{i\in E, a\in A} \right)$ – true value of θ

Assumptions

A1 The S-M chain $(Z_n)_{n\in\mathbb{N}}$ is irreducible.
A2 The conditional sojourn time distributions $f_{ij}(\cdot), i,j \in E$ have finite support.

- $D := \{0,\dots,\widetilde{n}\} \supseteq supp f_{ij}(\cdot)$
- $q = \left(q_{ij}(k)\right)_{i,j\in E,k\in D}$, $\widetilde{p} = \left(p_{(i,t_1)(j,t_2)}\right)_{i,j\in E,t_1,t_2\in D}$

Theorem

Under assumptions **A1** *and* **A2**, *given a sample of observations* $\{Y_0 = y_0,\dots,Y_M = y_M\}$, *we have*

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

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

Assumptions

A1 The S-M chain $(Z_n)_{n\in\mathbb{N}}$ is irreducible.
A2 The conditional sojourn time distributions $f_{ij}(\cdot), i,j \in E$ have finite support.

- $D := \{0,\dots,\widetilde{n}\} \supseteq \bigcup_{i,j} supp f_{ij}(\cdot)$
- $q = \left(q_{ij}(k)\right)_{i,j\in E,k\in D}$, $\widetilde{p} = \left(p_{(i,t_1)(j,t_2)}\right)_{i,j\in E,t_1,t_2\in D}$

Theorem

Under assumptions **A1** *and* **A2**, *given a sample of observations* $\{Y_0 = y_0,\dots,Y_M = y_M\}$, *we have*

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

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

Assumptions

A1 The S-M chain $(Z_n)_{n\in\mathbb{N}}$ is irreducible.
A2 The conditional sojourn time distributions $f_{ij}(\cdot), i,j \in E$ have finite support.

- $D := \{0,\dots,\widetilde{n}\} \supseteq suppf_{ij}(\cdot)$
- $q = \left(q_{ij}(k)\right)_{i,j\in E,k\in D}$, $\widetilde{p} = \left(p_{(i,t_1)(j,t_2)}\right)_{i,j\in E,t_1,t_2\in D}$

Theorem

Under assumptions **A1** *and* **A2**, *given a sample of observations* $\{Y_0 = y_0,\dots,Y_M = y_M\}$, *we have*

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

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

Assumptions

A1 The S-M chain $(Z_n)_{n\in\mathbb{N}}$ is irreducible.
A2 The conditional sojourn time distributions $f_{ij}(\cdot), i,j \in E$ have finite support.

- $D := \{0,\dots,\widetilde{n}\} \supseteq suppf_{ij}(\cdot)$
- $q = \left(q_{ij}(k)\right)_{i,j\in E,k\in D}$, $\widetilde{p} = \left(p_{(i,t_1)(j,t_2)}\right)_{i,j\in E,t_1,t_2\in D}$

Theorem

Under assumptions **A1** *and* **A2**, *given a sample of observations* $\{Y_0 = y_0,\dots,Y_M = y_M\}$, *we have*

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

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

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

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

Assumption

A3 There exists an integer $n \in \mathbb{N}$ such that the matrix $\sigma_{Y_0^n}(\theta^0)$ is nonsingular.

From Theorem 3 of Douc(2005) we know that $\sigma(\theta^0)$ is nonsingular if and only if **A3** is fulfilled.

Consider the partition of $\sigma^{-1}(\theta^0)$:

$$\sigma^{-1}(\theta^0) = \begin{matrix} \overbrace{}^{s^2 \times (\widetilde{n}+1)^2} & \overbrace{}^{sd} \\ \left[\begin{matrix} \sigma^{-1}(\theta^0)_{11} & \sigma^{-1}(\theta^0)_{12} \\ \sigma^{-1}(\theta^0)_{21} & \sigma^{-1}(\theta^0)_{22} \end{matrix} \right] & \begin{matrix} \} \\ \} \end{matrix} \begin{matrix} s^2 \times (\widetilde{n}+1)^2 \\ sd \end{matrix} \end{matrix}$$

Theorem

Under assumptions **A1**, **A2** *and* **A3**, *the MLE of* $(R^0_{ia})_{i\in E, a\in A}$ *is asymptotically normal :*

$$\sqrt{M}\left[(\widehat{R}_{ia}(M))_{i\in E,a\in A} - (R_{ia})_{i\in E,a\in A}\right] \xrightarrow[M\to\infty]{\mathcal{D}} \mathcal{N}\left(\mathbf{0}, \sigma^{-1}(\theta^0)_{22}\right).$$

Consider the partition of $\sigma^{-1}(\theta^0)$:

$$\sigma^{-1}(\theta^0) = \begin{bmatrix} \overbrace{\sigma^{-1}(\theta^0)_{11}}^{s^2 \times (\widetilde{n}+1)^2} & \overbrace{\sigma^{-1}(\theta^0)_{12}}^{sd} \\ \sigma^{-1}(\theta^0)_{21} & \sigma^{-1}(\theta^0)_{22} \end{bmatrix} \begin{matrix} \} \\ \} \end{matrix} \begin{matrix} s^2 \times (\widetilde{n}+1)^2 \\ sd \end{matrix}$$

Theorem

Under assumptions **A1**, **A2** *and* **A3**, *the MLE of* $(R^0_{ia})_{i\in E, a\in A}$ *is asymptotically normal :*

$$\sqrt{M}\left[\left(\widehat{R}_{ia}(M)\right)_{i\in E,a\in A} - (R_{ia})_{i\in E,a\in A}\right] \xrightarrow[M\to\infty]{\mathcal{D}} \mathcal{N}\left(\mathbf{0}, \sigma^{-1}(\theta^0)_{22}\right).$$

Theorem

Under assumptions **A1**, **A2** *and* **A3**, *the MLE of* $(q_{ij}^0(k))_{i,j\in E,k\in D}$ *is asymptotically normal :*

$$\sqrt{M}\left[(\widehat{q}_{ij}(k,M))_{i,j\in E,k\in D} - (q_{ij}^0(k))_{i,j\in E,k\in D}\right] \xrightarrow[M\to\infty]{\mathcal{D}} \mathcal{N}\left(\mathbf{0}, \sigma_q^2\right),$$

with

$$\sigma_q^2 = \mathit{Function}\left(\sigma^{-1}(\theta^0), (q_{ij}^0(k))_{i,j\in E,k\in D}\right).$$

Plan

Research of CpG islands

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

- Y : $\underbrace{TAGTGGAATG}\,\underbrace{CGACG}\cdots$ – DNA sequence
- $Z : 0000000000011111\cdots$ – indicators for CpG islands

$$Z : \left\{ \begin{array}{l} 1 \leftarrow \text{island} \\ 0 \leftarrow \text{non-island} \end{array} \right.$$

Consider : (Z, Y) hidden $SM1 - M0$ chain

Research of CpG islands

CpG 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
- $Z : 0000000000011111\cdots$ – indicators for CpG islands

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

Consider : (Z, Y) hidden $SM1 - M0$ chain

Research of CpG islands

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

- Y : $\underbrace{TAGTGGAATG}\,\underbrace{CGACG}\cdots$ – DNA sequence
- $Z : 0000000000011111\cdots$ – indicators for CpG islands

$$Z : \left\{ \begin{array}{ll} 1 \leftarrow & \text{island} \\ 0 \leftarrow & \text{non-island} \end{array} \right.$$

Consider : (Z, Y) hidden $SM1 - M0$ chain

Research of CpG islands

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

- 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

$$Z : \left\{ \begin{array}{ll} 1 \leftarrow & \text{island} \\ 0 \leftarrow & \text{non-island} \end{array} \right.$$

Consider : (Z, Y) hidden $SM1 - M0$ chain

Research of CpG islands

CpG 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

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

Consider : (Z, Y) hidden $SM1 - M0$ chain

1. Generate (Z, Y)

State space : $Z \to E = \{0, 1\}$, $Y \to A = \{0, 1, 2, 3\}$

kernel : $q_{0,1} = W_{0.7,\,0.9}$, $q_{1,0} = W_{0.5,\,0.7}$ – discrete time, first type, Weibull distributions

$$W_{q,b}(k) := \begin{cases} q^{(k-1)^b} - q^{k^b}, & k \geq 1 \\ 0, & k = 0 \end{cases}$$

conditional distribution of Y :

$$(R_{0;a})_{a \in A} = (0.25\ 0.25\ 0.25\ 0.25) \quad (R_{1;a})_{a \in A} = (0.1\ 0.4\ 0.4\ 0.1)$$

1. Generate (Z, Y)

State space : $Z \to E = \{0, 1\}$, $Y \to A = \{0, 1, 2, 3\}$

kernel : $q_{0,1} = W_{0.7,\,0.9}$, $q_{1,0} = W_{0.5,\,0.7}$ – discrete time, first type, Weibull distributions

$$W_{q,b}(k) := \begin{cases} q^{(k-1)^b} - q^{k^b}, & k \geq 1 \\ 0, & k = 0 \end{cases}$$

conditional distribution of Y :

$$(R_{0;a})_{a\in A} = (0.25\ 0.25\ 0.25\ 0.25) \quad (R_{1;a})_{a\in A} = (0.1\ 0.4\ 0.4\ 0.1)$$

1. Generate (Z, Y)

State space : $Z \to E = \{0, 1\}$, $Y \to A = \{0, 1, 2, 3\}$

kernel : $q_{0,1} = W_{0.7,\,0.9}$, $q_{1,0} = W_{0.5,\,0.7}-$ discrete time, first type, Weibull distributions

$$W_{q,b}(k) := \begin{cases} q^{(k-1)^b} - q^{k^b}, & k \geq 1 \\ 0, & k = 0 \end{cases}$$

conditional distribution of Y :

$(R_{0;a})_{a\in A} = (0.25\ 0.25\ 0.25\ 0.25)$ $(R_{1;a})_{a\in A} = (0.1\ 0.4\ 0.4\ 0.1)$

1. Generate (Z, Y)

State space : $Z \to E = \{0, 1\}$, $Y \to A = \{0, 1, 2, 3\}$

kernel : $q_{0,1} = W_{0.7,\,0.9}$, $q_{1,0} = W_{0.5,\,0.7}-$ discrete time, first type, Weibull distributions

$$W_{q,b}(k) := \begin{cases} q^{(k-1)^b} - q^{k^b}, & k \geq 1 \\ 0, & k = 0 \end{cases}$$

conditional distribution of Y :

$$(R_{0;a})_{a\in A} = (0.25\ 0.25\ 0.25\ 0.25)\ \ (R_{1;a})_{a\in A} = (0.1\ 0.4\ 0.4\ 0.1)$$

2. **Estimate** the characteristics of the model (EM algorithm)

Initial value

- $q_{0,1}^{(0)} = U(1, \ldots, 15)-$ uniform discrete distribution
- $q_{1,0}^{(0)} = U(1, \ldots, 10)-$ uniform discrete distribution

Initial value

- $q_{0,1}^{(0)} := W_{0.9,1}-$ discrete time Weibull distribution
- $q_{1,0}^{(0)} := W_{0.9,1.5}-$ discrete time Weibull distribution

2. **Estimate** the characteristics of the model (EM algorithm)

Initial value

- $q_{0,1}^{(0)} = U(1,\ldots,15)-$ uniform discrete distribution
- $q_{1,0}^{(0)} = U(1,\ldots,10)-$ uniform discrete distribution

Initial value

- $q_{0,1}^{(0)} := W_{0.9,1}-$ discrete time Weibull distribution
- $q_{1,0}^{(0)} := W_{0.9,1.5}-$ discrete time Weibull distribution

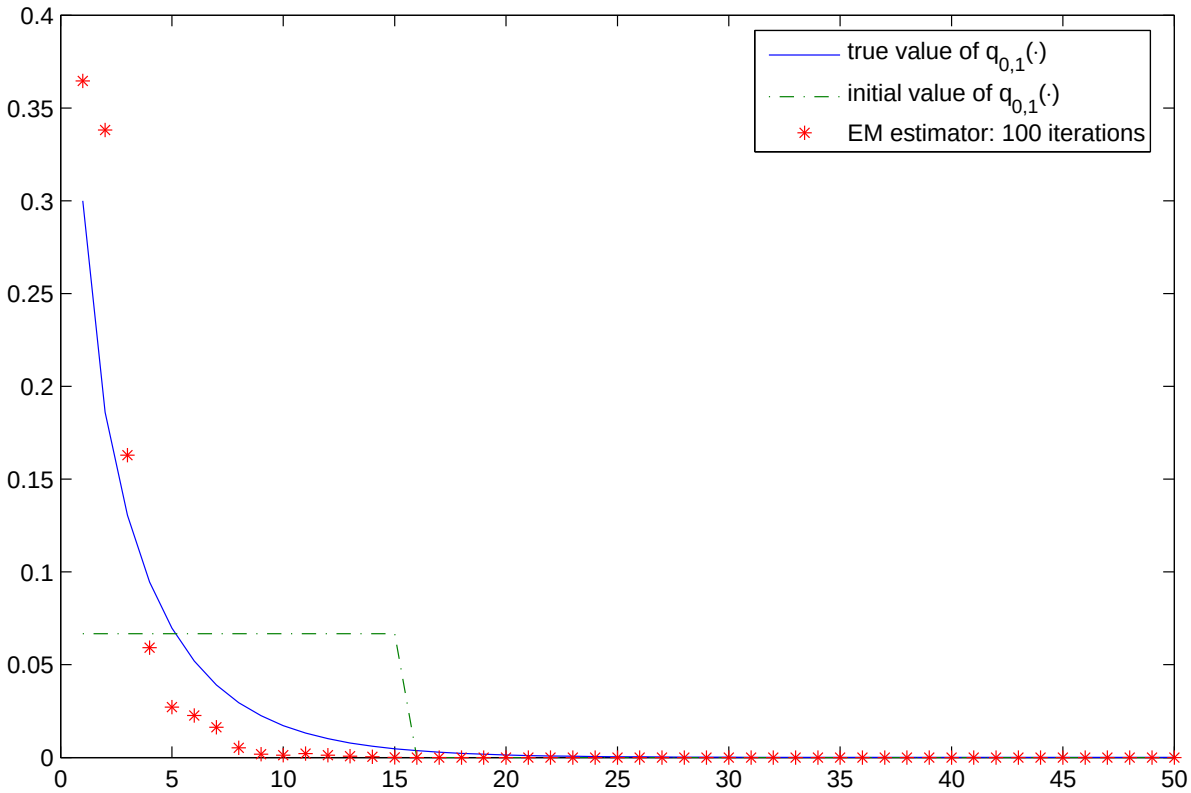

FIG.: Estimator of the semi-Markov kernel

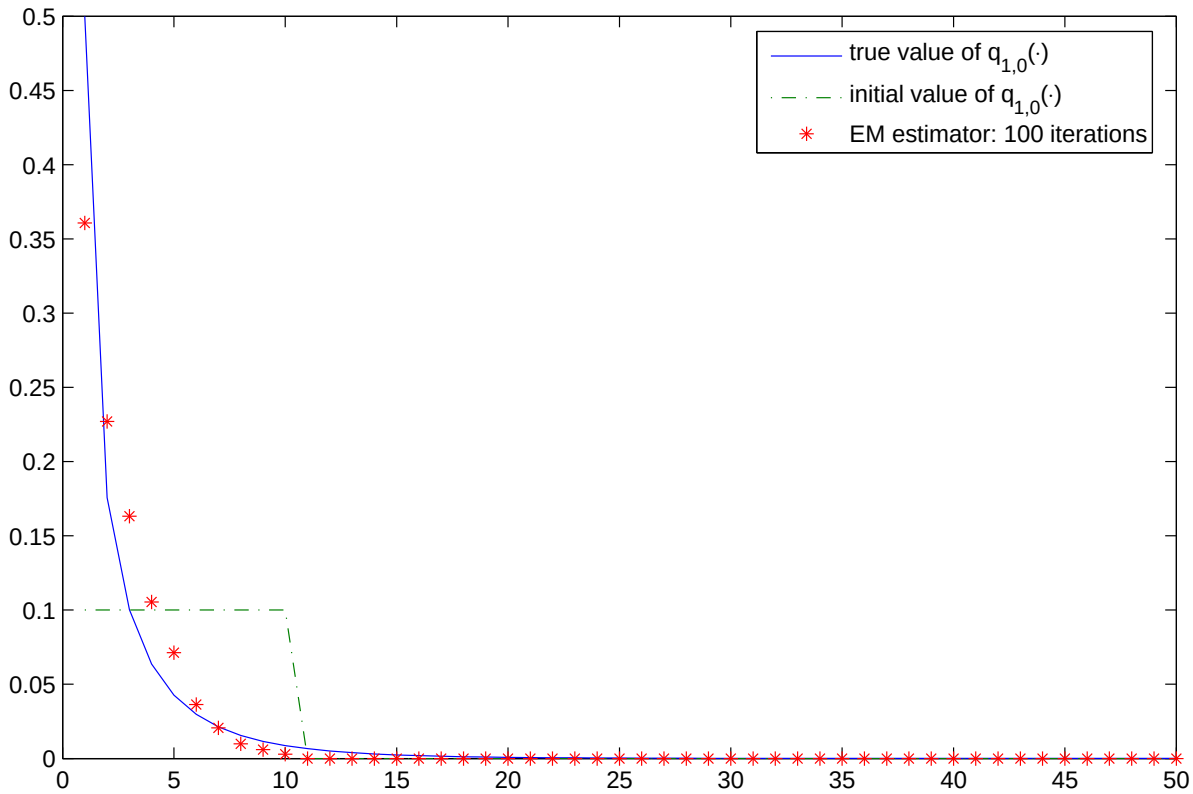

FIG.: Estimator of the semi-Markov kernel

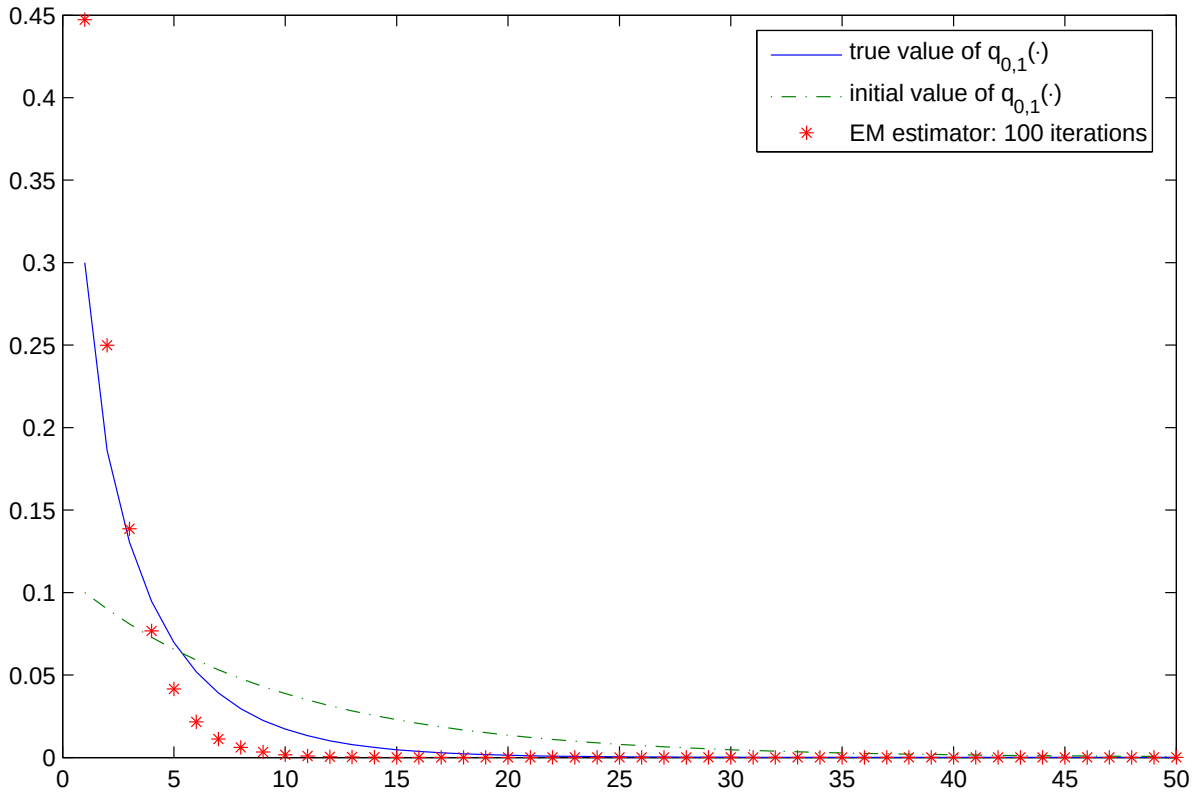

FIG.: Estimator of the semi-Markov kernel

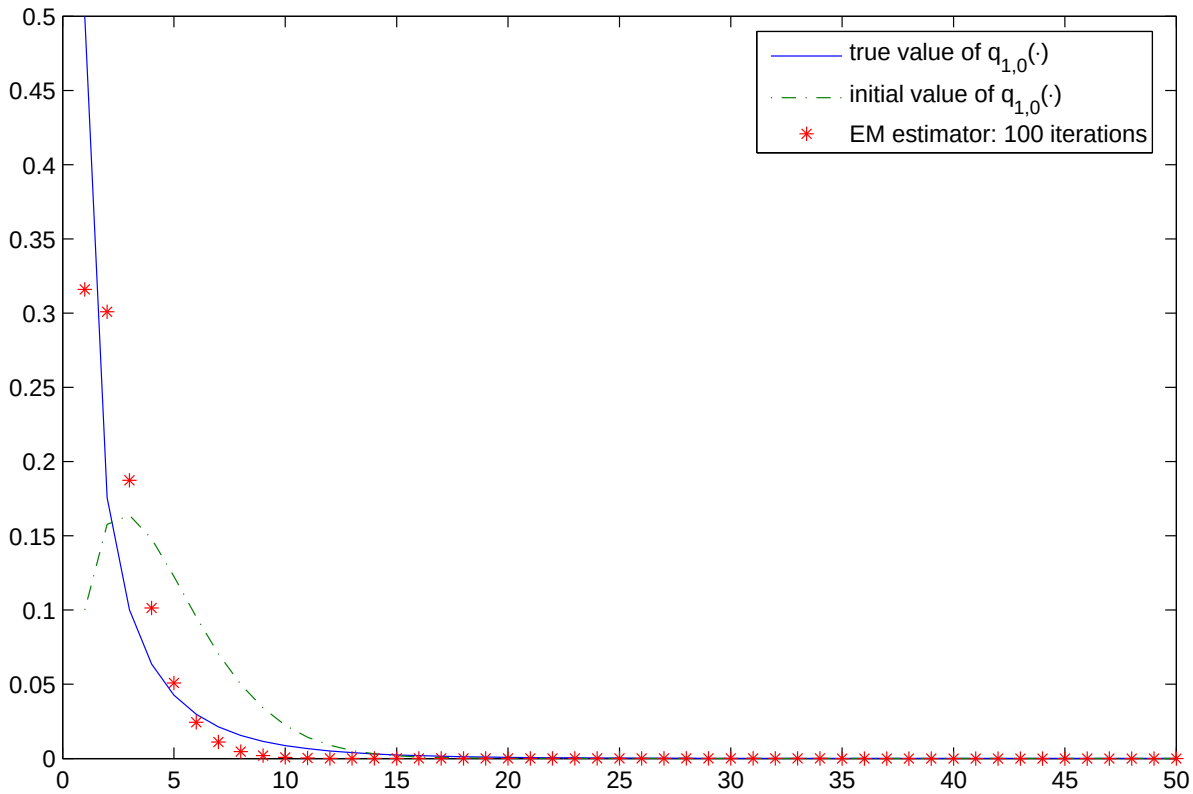

FIG.: Estimator of the semi-Markov kernel

Remarks

technique used for CpG islands research - generality
- patterns research
- correlation of the patterns with biological function

- takes into account the heterogeneity of DNA

  coding / non-coding zones, exons / introns, intron/exon borders, etc
- detection of homogeneous zones along DNA (Muri, 1997, use of HMM)

Plan

Future research

- statistical studies
- adapted algorithms
- applications :
  genetics, reliability, speech/image recognition, . . .

Future research

- statistical studies
- adapted algorithms
- applications :
  genetics, reliability, speech/image recognition, . . .

Future research

- statistical studies
- adapted algorithms
- applications :
  genetics, reliability, speech/image recognition, . . .

References

- Barbu, V., Boussemart, M. and Limnios, N. (2004). Discrete time semi-Markov model for reliability and survival analysis, *Communications in Statistics-Theory and Methods*, **33** (11) 2833-2868.
- Barbu, V. and Limnios, N. (2006). Maximum likelihood estimation for hidden semi-Markov models, *C. R. Acad. Sci. Paris, Ser. I*, **342** 201-205.
- Baum, L.E. and Petrie,T. (1966).Statistical inference for probabilistic functions of finite state Markov chains, *The Annals of Mathematical Statistics* **37** 1554-1563.
- P.J. Bickel, Y. Ritov and T. Rydén (1998). Asymptotic normality of the maximum-likelihood estimator for general hidden Markov models *Ann. Statist.* **26** 1614-1635.
- Churchill, G.A. (1989). Stochastic Models for Heterogeneous DNA Sequences. *Bulletin of Mathematical Biology*, 51, 1, 79-94.
- Churchill, G. (1992). Hidden Markov chains and the analysis of genome structure. *Computers Chem.*, 16, 107-115.
- Csenki, A. (2002). Transition analysis of semi-Markov reliability models - A tutorial review with emphasis on discrete-parameter approaches. In : *Stochastic Models in Reliability and Maintenance* (ed. S. Osaki), Springer, Berlin, 219-251.
- Douc, R. (2005) Non singularity of the asymptotic Fisher information matrix in hidden Markov models, École Polytechnique, preprint.
- R. Durbin, S.R. Eddy, A. Krogh and G.J. Mitchison (1998). *Biological Sequence Analysis : Probabilistic Models of Proteins and Nucleic Acids*, Cambridge University Press, Cambridge UK.

- Ferguson, J. D. (1980). Variable duration models for speech. *Proc. of the symposium on the Application of Hidden Markov Models to Text and Speech* , Princeton, New Jersey, 143-179.
- Guédon, Y. (2003). Estimating hidden semi-Markov chains from discrete sequences. Journal of *Computational and Graphical Statistics* 12(3), 604–639.
- Guédon, Y. and Cocozza-Thivent, C. (1990). Explicit state occupancy modelling by hidden semi-Markov models : application of Derin's scheme. *Computer Speech and Language* **4**, 167–192.
- Howard, R. (1971). *Dynamic Probabilistic systems.* vol II, Wiley, New York.
- Levinson, S.E.(1986). Continuously variable duration hidden Markov models for automatic speech recognition. *Computer Speech and Language* **1** 29-45.
- Limnios, N. and Oprişan, G. (2001). *Semi-Markov Processes and Reliability*, Birkhäuser, Boston.
- Muri, F. (1997). *Comparaison d'algorithmes d'identification de Chaînes de Markov Cachées et application à la détection de régions homogènes dans les séquences d'ADN*, Thèse de doctorat, Université Paris 5.
- Mode, C.J. and Sleeman, C.K. (2000). *Stochastic processes in epidemiology*, World Scientific, New Jersey.

Note $\widetilde{\sigma} = \sigma^{-1}(\theta^0)$. Partition the matrix $\widetilde{\sigma}$:

$$\left[\begin{array}{ccccccc}
\overbrace{\widetilde{\sigma}_{1,1}}^{s-1} & \cdots & \overbrace{\widetilde{\sigma}_{1,s\widetilde{n}}}^{s-1} & \overbrace{\widetilde{\sigma}_{1,s\widetilde{n}+1}}^{\widetilde{n}-1} & \cdots & \overbrace{\widetilde{\sigma}_{1,s\widetilde{n}+s}}^{\widetilde{n}-1} & \\
\vdots & & \vdots & \vdots & & \vdots & \\
\widetilde{\sigma}_{s\widetilde{n},1} & \cdots & \widetilde{\sigma}_{s\widetilde{n},s\widetilde{n}} & \widetilde{\sigma}_{s\widetilde{n},s\widetilde{n}+1} & \cdots & \widetilde{\sigma}_{s\widetilde{n},s\widetilde{n}+s} & \widetilde{\sigma}_A \\
\widetilde{\sigma}_{s\widetilde{n}+1,1} & \cdots & \widetilde{\sigma}_{s\widetilde{n}+1,s\widetilde{n}} & \widetilde{\sigma}_{s\widetilde{n}+1,s\widetilde{n}+1} & \cdots & \widetilde{\sigma}_{s\widetilde{n}+1,s\widetilde{n}+s} & \\
\vdots & & \vdots & \vdots & & \vdots & \\
\widetilde{\sigma}_{s\widetilde{n}+s,1} & \cdots & \widetilde{\sigma}_{s\widetilde{n}+s,s\widetilde{n}} & \widetilde{\sigma}_{s\widetilde{n}+s,s\widetilde{n}+1} & \cdots & \widetilde{\sigma}_{s\widetilde{n}+s,s\widetilde{n}+s} & \\
 & & & \widetilde{\sigma}_B & & & \widetilde{\sigma}_C
\end{array}\right]
\begin{array}{ll}
\} & \\
\vdots & s-1 \\
\} & s-1 \\
\} & \widetilde{n}-1 \\
\vdots & \widetilde{n}-1 \\
\} &
\end{array}$$

For all $k = 2, \ldots, \widetilde{n}$ and $i \in E$, define :

- the scalars $\alpha_{i;k} = \overline{H_i^0}(k-1)$
- the matrices $B_{i;k} = (B_{i;k}(u,v))_{\substack{u=1,\ldots,s-1 \\ v=1,\ldots,\widetilde{n}-1}} \in \mathcal{M}_{(s-1)\times(\widetilde{n}-1)}$

$$B_{i;k}(u,v) = \begin{cases} q_{i,u}^0(k) \dfrac{\overline{H_i^0}(v-1)\overline{H_i^0}(v+1)\overline{H_i^0}(k-1)}{\overline{H_i^0}(v-2)\overline{H_i^0}(v)} & v \leq k-1, u < i \\ q_{i,u}^0(k+1) \dfrac{\overline{H_i^0}(v-1)\overline{H_i^0}(v+1)\overline{H_i^0}(k-1)}{\overline{H_i^0}(v-2)\overline{H_i^0}(v)} & v \leq k-1, u \geq i \\ 0 & \text{otherwise} \end{cases}$$

where $\overline{H_i}(\cdot)$ is *the survival function in state i* defined by

$$\overline{H_i}(n) := \mathbb{P}(X_1 > n \mid J_0 = i) = 1 - \sum_{j \in \mathbf{E}} \sum_{k=1}^{n} q_{ij}(k), \ n \in \mathbb{N}^*$$

Theorem

Under assumptions **A1**, **A2** *and* **A3**, *the MLE of* $(q_{ij}^0(k))_{i,j\in E,k\in D}$ *is asymptotically normal :*

$$\sqrt{M}\left[(\widehat{q}_{ij}(k,M))_{i,j\in E,k\in D} - (q_{ij}^0(k))_{i,j\in E,k\in D}\right] \xrightarrow[M\to\infty]{\mathcal{D}} \mathcal{N}\left(\mathbf{0}, \sigma_q^2\right),$$

where

$$\sigma_q^2 = \mathit{Function}\left((\widetilde{\sigma}_{m,n})_{m,n=1,\ldots,s\tilde{n}+s}, (\alpha_{i;k})_{\substack{i\in E\\ k=2,\ldots,\tilde{n}}} (B_{i;k})_{\substack{i\in E\\ k=2,\ldots,\tilde{n}}}\right).$$

$$\sigma_q^2 = \begin{bmatrix} M_{1,1} & \dots & M_{1,\widetilde{n}} \\ \vdots & & \vdots \\ M_{\widetilde{n},1} & \dots & M_{\widetilde{n},\widetilde{n}} \end{bmatrix}, M_{u,v} \in \mathcal{M}_{s(s-1)\times s(s-1)}$$

$$M_{1,1} = \begin{bmatrix} \widetilde{\sigma}_{1,1} & \dots & \widetilde{\sigma}_{1,s} \\ \vdots & & \vdots \\ \widetilde{\sigma}_{s,1} & \dots & \widetilde{\sigma}_{s,s} \end{bmatrix}$$

For $2 \leq v \leq \widetilde{n}$,

$$M_{1,v} = \begin{bmatrix} \alpha_{1;v}\widetilde{\sigma}_{1,sv-s+1} + \widetilde{\sigma}_{1,s\widetilde{n}+1}B_{1;v}^T & \dots & \alpha_{s;v}\widetilde{\sigma}_{1,sv} + \widetilde{\sigma}_{1,s\widetilde{n}+s}B_{s;v}^T \\ \vdots & & \vdots \\ \alpha_{1;v}\widetilde{\sigma}_{s,sv-s+1} + \widetilde{\sigma}_{s,s\widetilde{n}+1}B_{1;v}^T & \dots & \alpha_{s;v}\widetilde{\sigma}_{s,sv} + \widetilde{\sigma}_{s,s\widetilde{n}+s}B_{s;v}^T \end{bmatrix}$$

For $2 \leq u \leq \widetilde{n}$,

$$M_{u,1} = \begin{bmatrix} \alpha_{1;u}\widetilde{\sigma}_{su-s+1,1} + B_{1;u}\widetilde{\sigma}_{s\widetilde{n}+1,1} & \cdots & \alpha_{1;u}\widetilde{\sigma}_{su-s+1,s} + B_{1;u}\widetilde{\sigma}_{s\widetilde{n}+1,s} \\ \vdots & & \vdots \\ \alpha_{s;u}\widetilde{\sigma}_{su,1} + B_{s;u}\widetilde{\sigma}_{s\widetilde{n}+s,1} & \cdots & \alpha_{s;u}\widetilde{\sigma}_{su,s} + B_{s;u}\widetilde{\sigma}_{s\widetilde{n}+s,s} \end{bmatrix}$$

For $2 \leq u, v \leq \widetilde{n}$, $M_{u,v} = (M_{u,v}^{l,r})_{l,r=1,\ldots,s}-$ block matrix

$$\begin{aligned} M_{u,v}^{l,r} &= (\alpha_{l;u}\widetilde{\sigma}_{su-s+l,sv-s+r} + B_{l;u}\widetilde{\sigma}_{s\widetilde{n}+l,sv-s+r})\alpha_{r;v} \\ &\quad +(\alpha_{l;u}\widetilde{\sigma}_{su-s+l,s\widetilde{n}+r} + B_{l;u}\widetilde{\sigma}_{s\widetilde{n}+l,s\widetilde{n}+r})B_{r;v}^{T} \end{aligned}$$