Estimation de l'arbre de contexte dans les VLHMM

Colloque des jeunes probabilistes et statisticiens - CIRM

Thierry Dumont

ID Services - Université de Paris sud XI

April 19, 2012

1. VLMC : Variable Length Markov Chain.

2. VLHMM : Variable Length Hidden Markov Models
   - Notations et définitions
   - Résultats principaux
   - Algorithme
   - Simulations

Définition

Un processus stationnaire $(X_n)_{n\in\mathbb{Z}}$ d'espace d'états $\mathbb{X}$ est une chaîne de Markov de longueurs variables s'il existe deux fonctions l et c sur $\mathbb{X}^\infty$ telles que : pour tout "passé" $x_{-\infty:0}$,

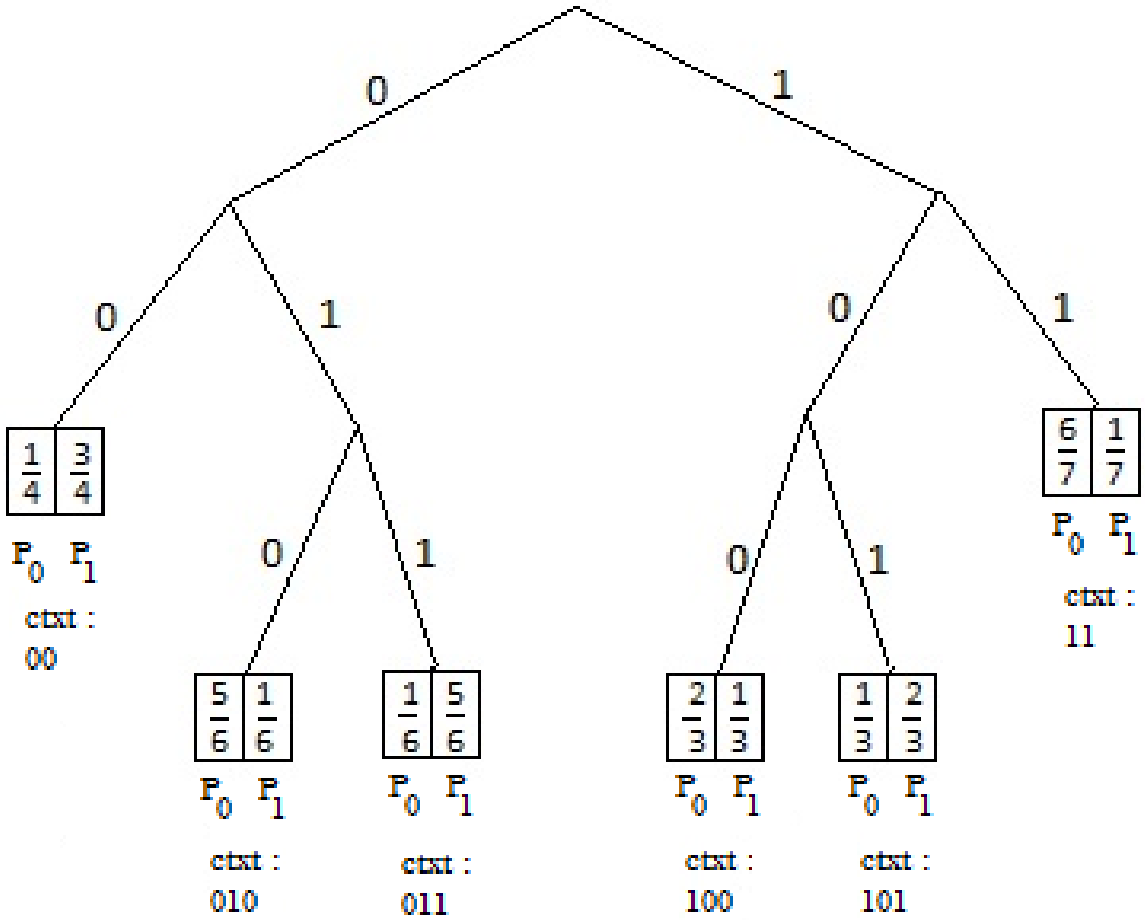
$$\ell(x_{-\infty:0}) = \min\{k \quad | \quad \forall x_1 \in E, \mathbb{P}(X_1 = x_1 | X_{-\infty:0} = x_{-\infty:0}) = \mathbb{P}(X_1 = x_1 | X_{-k+1:0} = x_{-k+1:0})\}$$

et,

$$c(x_{-\infty:0}) = x_{-\ell(x_{-\infty:0})+1:0}$$

Les éléments de l'image de c sont appelés contextes et possèdent une représentation d'arbre. La fonction c est appelée fonction contexte. $\longrightarrow$ Rissanen (1983), Csiszàr et Talata (2006).

Représentation en arbre de la fonction c :



Remarque

Connaitre l'arbre de contextes "minimal" d'un processus VLMC permet de connaitre le nombre minimal de paramètres à estimer nécessaires à l'analyse du processus.
Une des utilisations les plus répandues des modèles VLMC apparaît en théorie de l'information (Minimum Description Length), et en pratique pour la modélisation de séquences biologiques.

Définition

$(X_n, Y_n)_{n\in\mathbb{N}}$ est un modèle de Markov caché à longueurs variables (VLHMM pour Variable Length Hidden Markov Model) si $(X_n)_{n\in\mathbb{N}}$ est une chaîne de Markov à longueurs variables (irréductible dans notre étude) **non-observée** *et $(Y_n)_{n\in\mathbb{N}}$, appelées* **observations** *sont indépendantes conditionnellement à $(X_n)_{n\in\mathbb{N}}$:*

$$\mathbb{P}(Y_1 \in A_1, ..., Y_k \in A_k | X_1, ..., X_k) = \prod_{i=1}^{k} \mathbb{P}(Y_i \in A_i | X_i)$$

Remarque

- *Notre objectif dans cette présentation est l'étude d'un estimateur de l'arbre de contexte associé à la chaîne de Markov* **cachée** $(X_n)_{n\in\mathbb{Z}}$, *noté* $\tau^\star$, *et construit sur la base des observations* $(Y_k)_{k=1\ldots,n}$.

Modèle

Supposons $(X_n, Y_n)_{n\in\mathbb{N}}$ VLHMM, X_n à valeur dans $\mathbb{X}$ fini et Y_n à valeurs dans $\mathbb{R}^B$.

- *notons $\tau^\star$ l'arbre de contextes du processus $(X_n)_n$,*
- *pour tout τ arbre de contextes notons*

$$\Theta_{t,\tau} = \left\{ (P_{s,x})_{s\in\tau,\ x\in\mathbb{X}},\ P_{s,x} \geq 0\ \forall (s,x) \in \tau\times\mathbb{X},\ \sum_{x\in\mathbb{X}} P_{s,x} = 1,\ \forall s\in\tau \right\}$$

Espace des paramètres de transition.
Remarque : $dim(\Theta_{t,\tau}) = |\tau|(|\mathbb{X}|-1)$.

Modèle

- $\mathbb{P}(Y_1 \in A_1, ..., Y_k \in A_k | X_1 = x_1, ..., X_k = x_k) = \prod_{i=1}^{k} \left[\int_{A_i} g_{\theta_{e,x_i},\eta}(y) d\mu_y \right]$
- $$\Theta_e = \left\{ ((\theta_{e,x})_{x \in \mathbb{X}}, \eta) \in \left(\mathbb{R}^{d_\theta} \right)^{|\mathbb{X}|} \times \mathbb{R}^{d_\eta} \right\}$$

  Espace des paramètres d'émission.
- *L'espace des paramètres du modèle est donc* $\Theta_\tau = \Theta_{t,\tau} \times \Theta_e$

Définition

La Vraisemblance

Pour tout τ arbre de contextes, pour tout $\theta = (\theta_t, \theta_e) \in \Theta_\tau$, nous définissons ce que nous appelons la vraisemblance par :

$$\forall y_{1:n} \in (\mathbb{R}^B)^n,\ g_\theta(y_{1:n}) = \sum_{x_{1:n} \in \mathbb{X}^n} \left[\prod_{i=1}^{n} g_{\theta_{e,x_i},\eta}(y_i) \right] g_{\theta_t}(x_{1:n})$$

où :

$$g_{\theta_t}(x_{1:n}) = \sum_{x_{-d(\tau)+1:0} \in \mathbb{X}^{d(\tau)}} \nu_{d(\tau)}(x_{-d(\tau)+1:0}) \prod_{i=1}^{n} P_{s_i,x_i}$$

avec

$$s_i = c_\tau(x_{-d(\tau)+1:i-1})$$

Définition

Si nous observons $Y_{1:n}$, nous définissons l'estimateur de l'arbre de contextes pour la chaîne de Markov cachée à longueur variable X par

$$\begin{aligned}\hat{\tau}_n &= \underset{\tau \text{ arbre complet}}{\operatorname{argmin}} \left[- \sup_{\theta \in \Theta_\tau} \log g_\theta(Y_{1:n}) + pen(n,\tau) \right] \\ &= \underset{\tau \text{ arbre complet}}{\operatorname{argmin}} \; sc(\tau)\end{aligned}$$

Où $pen(n,\tau) = pen_\alpha(n,\tau) = \left[\sum_{t=1}^{|\tau|} \frac{(|\mathbb{X}|-1)t+\alpha}{2} \right] \log n$

Theorem

Sous certaines conditions classiques sur le modèle, s'il existe un prior π_e sur Θ_e et une constante $b > 0$ tels que, si, pour tout $x_{1:n} \in \mathbb{X}^n$, on définit la loi mélange sur $\mathbb{R}^B$, conditionnellement à $x_{1:n}$ par,

$$\mathbb{KT}_e^n(y_{1:n}|x_{1:n}) = \int_{\Theta_e} \left[\prod_{i=1}^{n} g_{\theta_e,x_i,\eta}(y_i)\right] \pi_e(d\theta_e),$$

alors éventuellement p.s. :

$$\sup_{\theta_e \in \Theta_e} \sup_{x_{1:n}} \left[\log \prod_{i=1}^{n} g_{\theta_e,x_i,\eta}(Y_i) - \log \mathbb{KT}_e^n(Y_{1:n}|x_{1:n})\right] \leq b \log n$$

et si $\alpha > 2(b+1)$, **alors $\hat{\tau}_n \sim \tau^\star$ pour n assez grand p.s.**

Cas particulier important où les lois d'émission sont des gaussiennes de moyennes inconnues $(m_x)_{x\in\mathbb{X}}$ et de variance σ^2 commune à tous les $x \in \mathbb{X}$ mais inconnue.

Définition

Posons :

- $\eta = -\frac{1}{2\sigma^2}$
- $\theta_{e,x} = \frac{m_x}{\sigma^2}$ *pour tout* $x \in \mathbb{X}$

Alors

$$\Theta_e = \left\{ \left((\theta_{e,x})_{x \in \mathbb{X}}, \eta \right) \middle| \theta_{e,x} \in \mathbb{R},\ \eta < 0 \right\}$$

Et $\forall x_{1:n} \in \mathbb{X}^n$, $\forall y_{1:n} \in \mathbb{R}^n$,

$$\prod_{i=1}^{n} g_{\theta_{e,x_i},\eta}(y_i) = \frac{1}{\sqrt{2\pi}^n} \prod_{j=1}^{k} \exp\left[\eta \sum_{i \in I_j} y_i^2 + \theta_{e,j} \sum_{i \in I_j} y_i - n_j A(\eta, \theta_{e,j}) \right]$$

where $$A(\eta, \theta_{e,j}) = -\frac{\theta_{e,j}^2}{4\eta} - \frac{1}{2}\log(-2\eta)$$

Define now the conjugate exponential prior on Θ_e :

$$\pi_e^n(d\theta_e) = \exp\left[\alpha_1^n \eta + \sum_{j=1}^{k} \alpha_{2,j}^n \theta_{e,j} - \sum_{j=1}^{k} \beta_j^n A(\eta, \theta_{e,j}) - B\left(\alpha_1^n, \alpha_{2,1}^n, \ldots, \alpha_{2,k}^n, \beta_1^n, \ldots, \beta_k^n\right)\right] d\eta d\theta_{e,1} \cdots d\theta_{e,k}$$

where the parameters α_1^n, $(\alpha_{2,j}^n)_{j=1,\ldots,k}$ and $(\beta_j^n)_{j=1,\ldots,k}$ will be chosen later, and the normalizing constant may be computed as

$$\exp\left\{B\left(\alpha_1^n, \alpha_{2,1}^n, \ldots, \alpha_{2,k}^n, \beta_1^n, \ldots, \beta_k^n\right)\right\} = \frac{2^{k+\frac{\sum_{j=1}^k \beta_j^n}{2}} \pi^{\frac{k}{2}} \Gamma\left(\frac{\sum_{j=1}^k \beta_j^n + k + 2}{2}\right)}{\left(\prod_{j=1}^k \sqrt{\beta_j^n}\right)\left(\alpha_1^n - \sum_{j=1}^k \frac{(\alpha_{2,j}^n)^2}{\beta_j^n}\right)^{\frac{\sum_{j=1}^k \beta_j^n + k + 2}{2}}}$$

Proposition 4. *If* **(A1)** *holds, it is possible to choose the parameters* α_1^n, $(\alpha_{2,j}^n)_{j=1,\ldots,k}$ *and* $(\beta_j^n)_{j=1,\ldots,k}$ *such that for any* $\epsilon > 0$,

$$\max_{x_{1:n}} \left\{ \sup_{\theta_e \in \Theta_e} \log \prod_{i=1}^{n} g_{\theta_{e,x_i},\eta}(Y_i) - \log \mathbb{KT}_e^n(Y_{1:n}|x_{1:n}) \right\} \leq \frac{k+1+\epsilon}{2} \log n$$

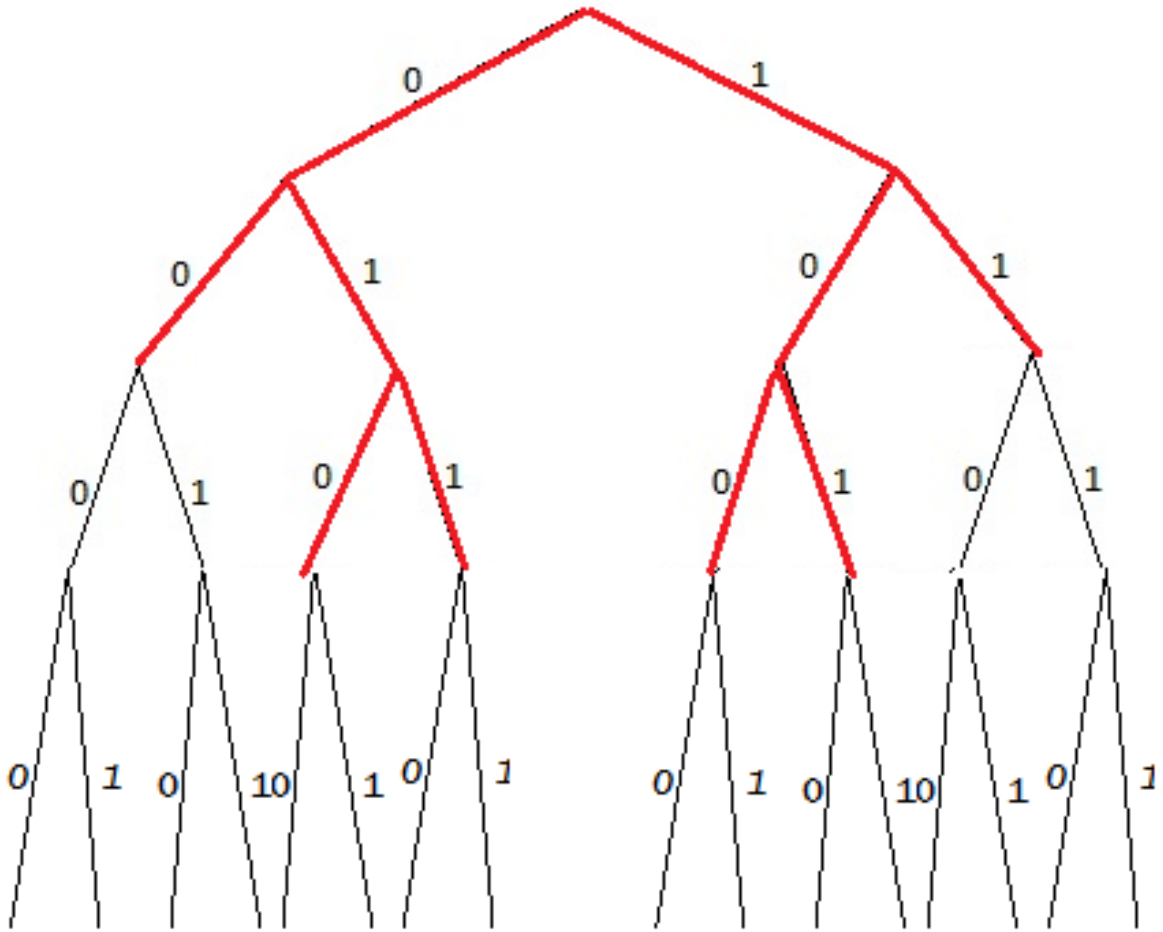
$\mathbb{P}_{\theta^\star}$ - *eventually a.s.*

Theorem

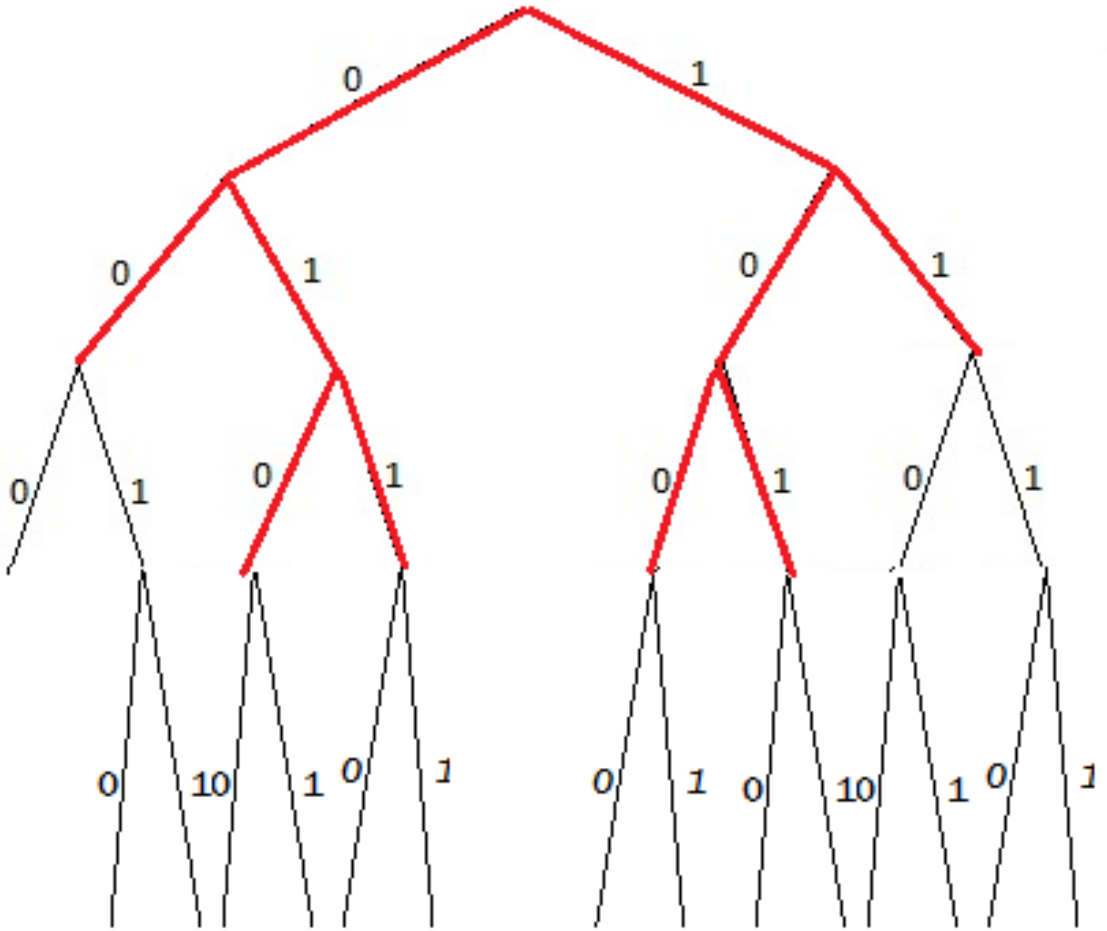
Si $\alpha > k + 3$, alors $\hat{\tau}_n \sim \tau^\star$ pour n assez grand p.s.

Nous proposons un algorithme d'élagage pour le cas général. Il permet, se donnant une suite d'observations $Y_{1:n}$ de calculer $\hat{\tau}_n$ en "élaguant" un arbre *couvrant* notre $\tau^\star$ dans l'esprit des premiers travaux de Rissanen dans le cadre des VLMC.

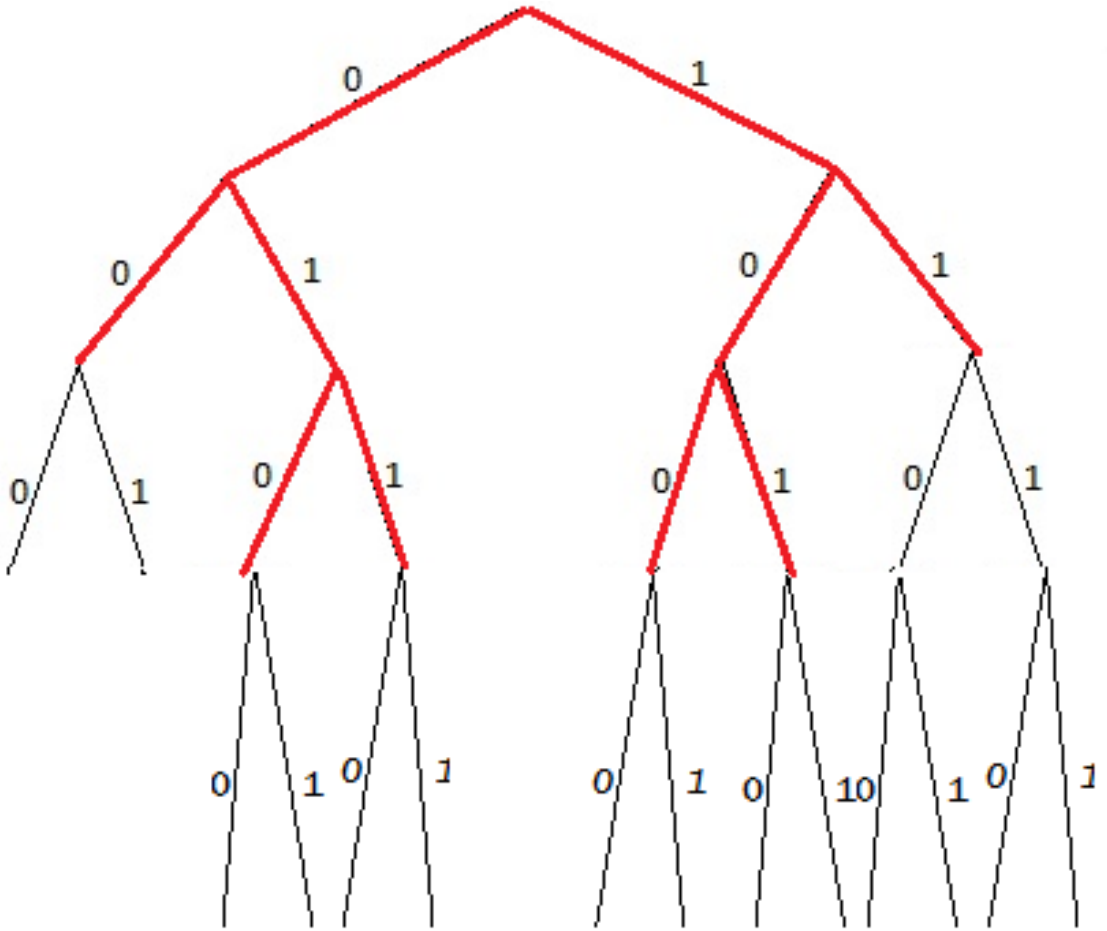
Arbre couvrant *maximal*.



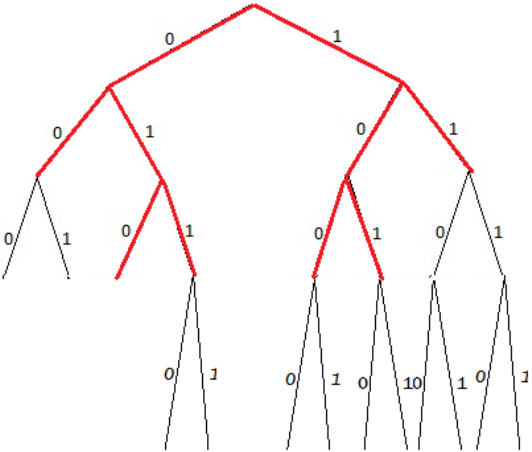
Processus d'élagage



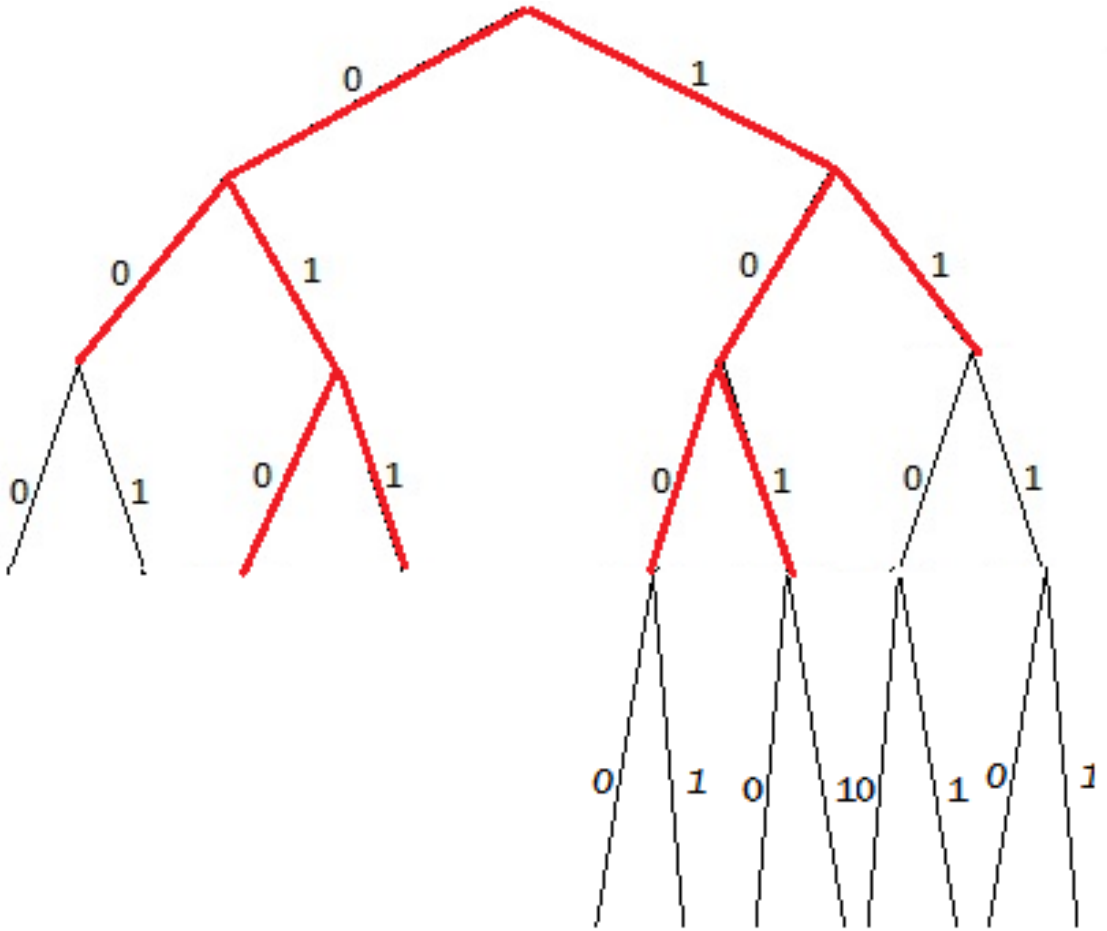
Processus d'élagage



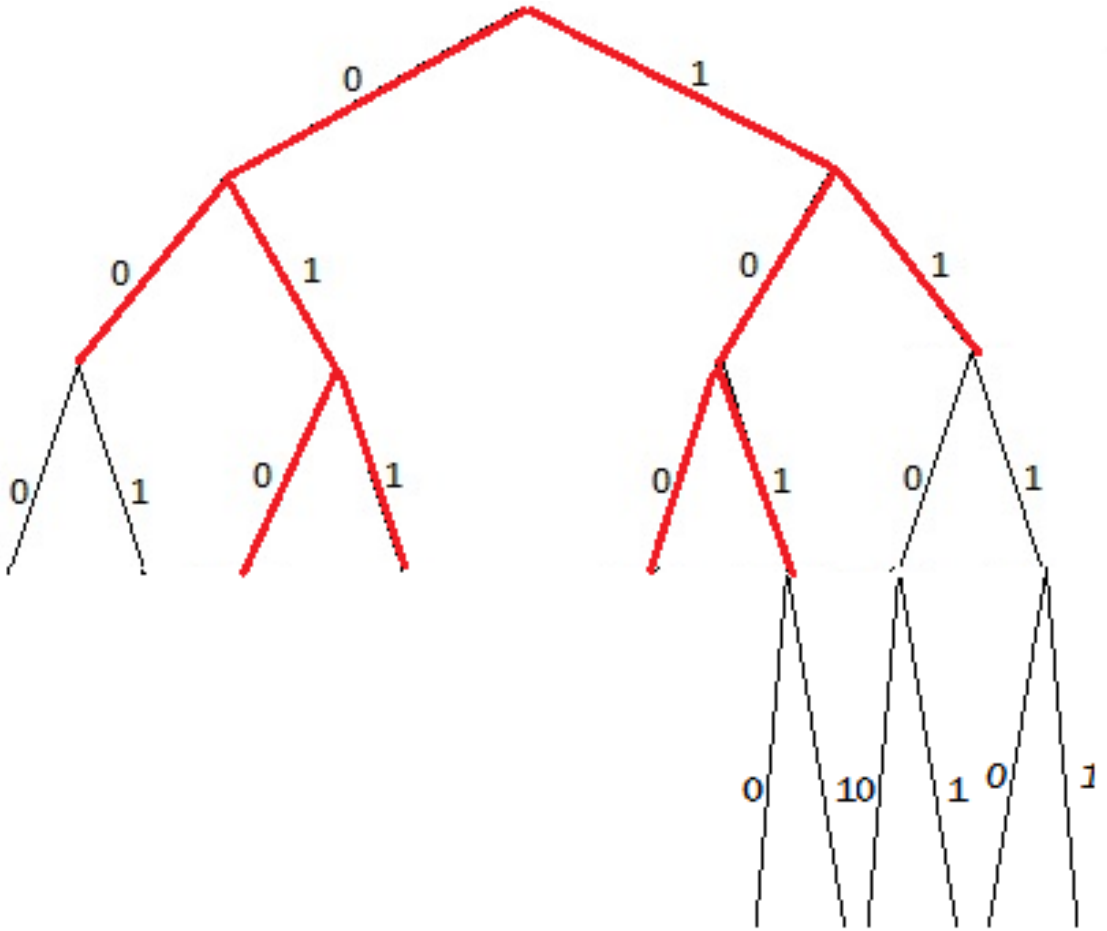
Processus d'élagage



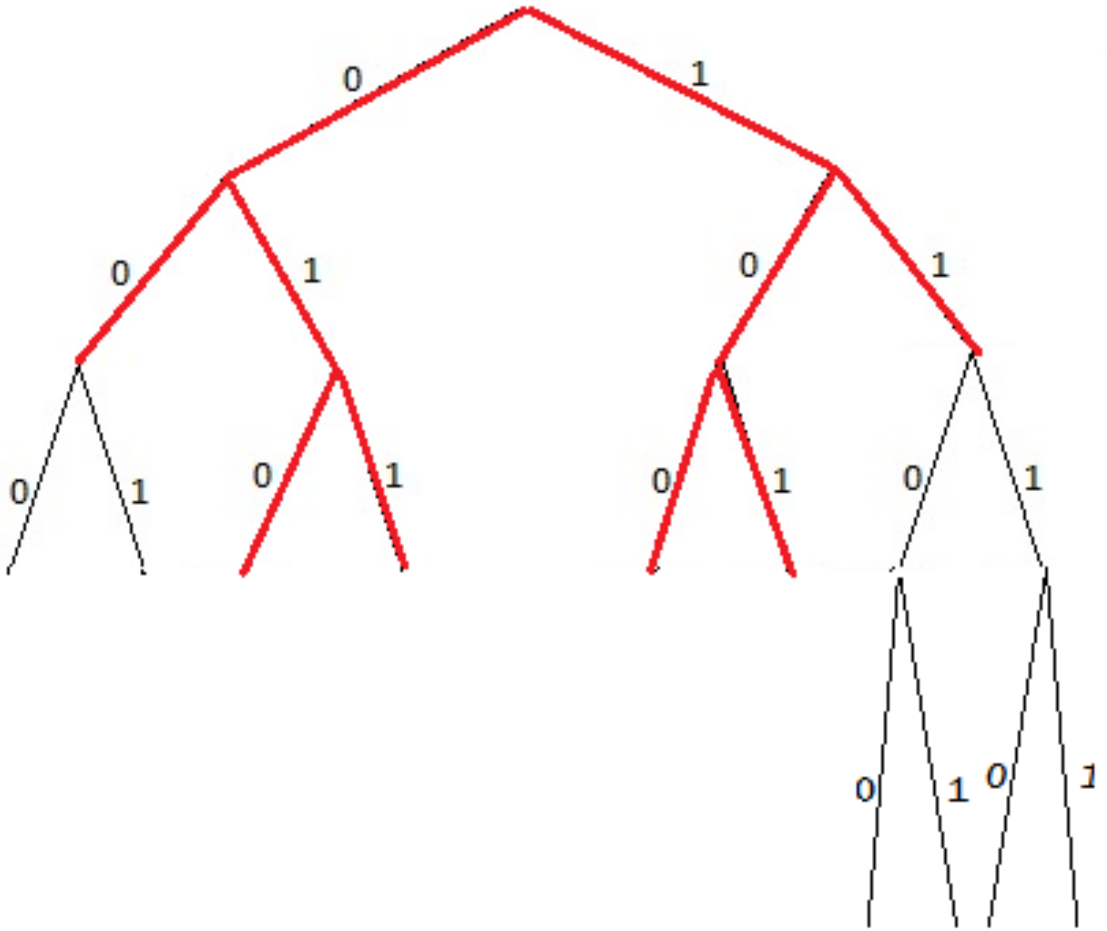
Processus d'élagage



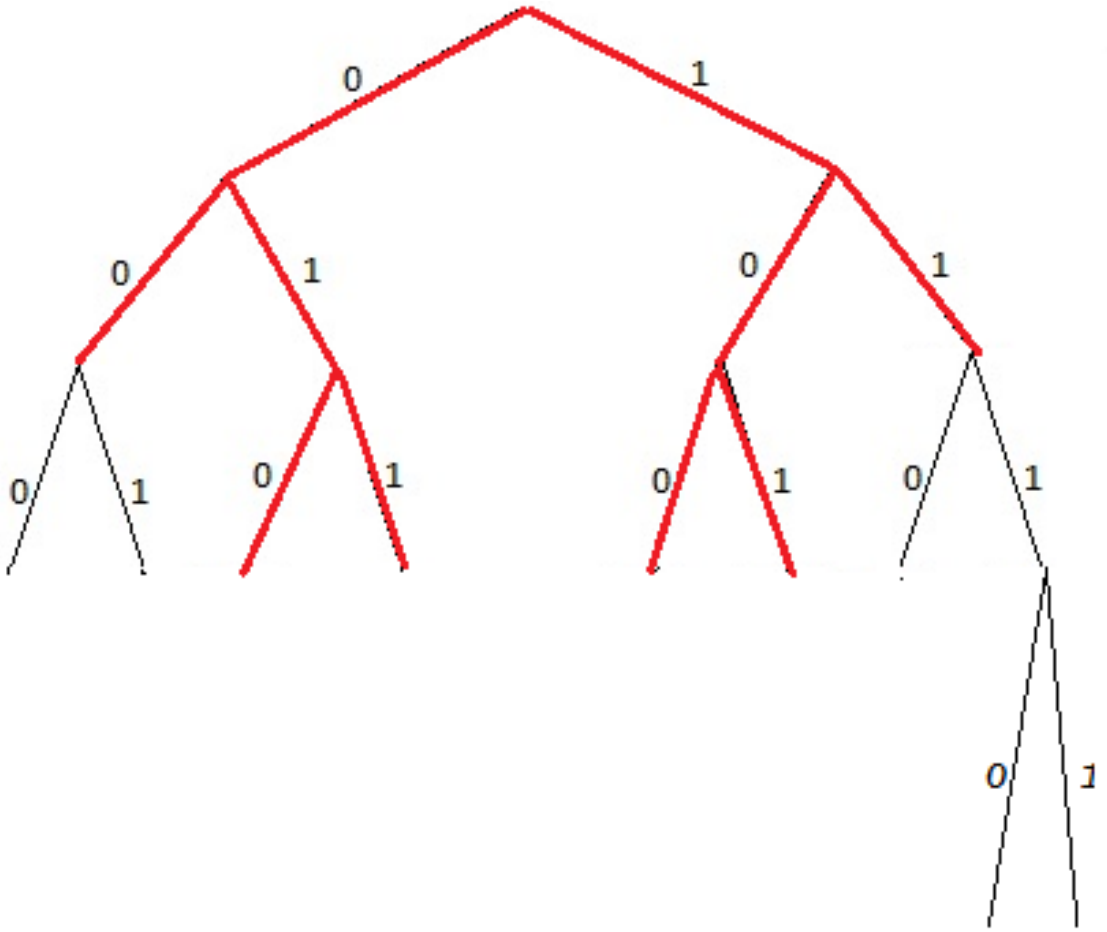
Processus d'élagage



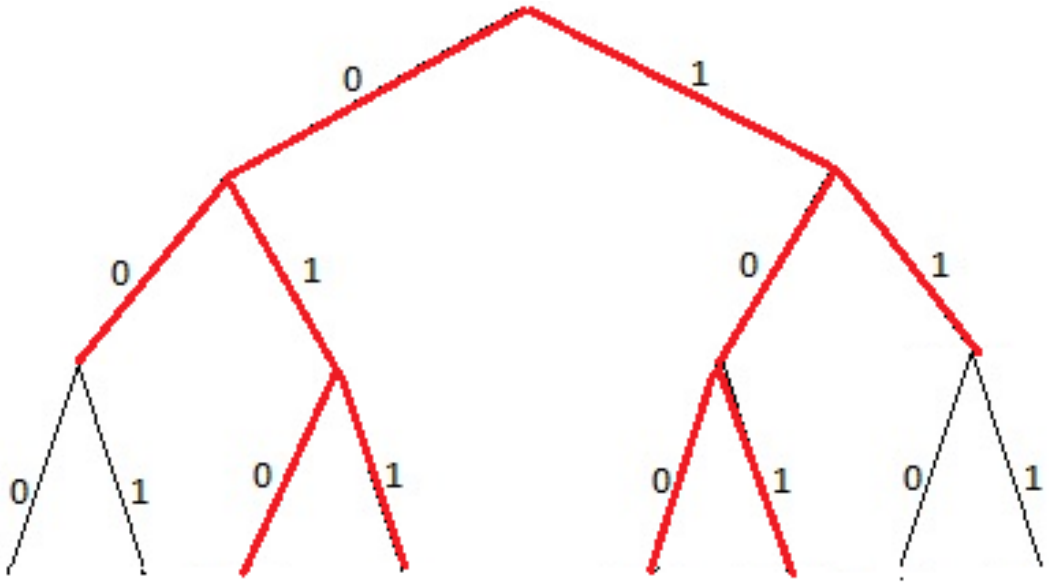
Processus d'élagage



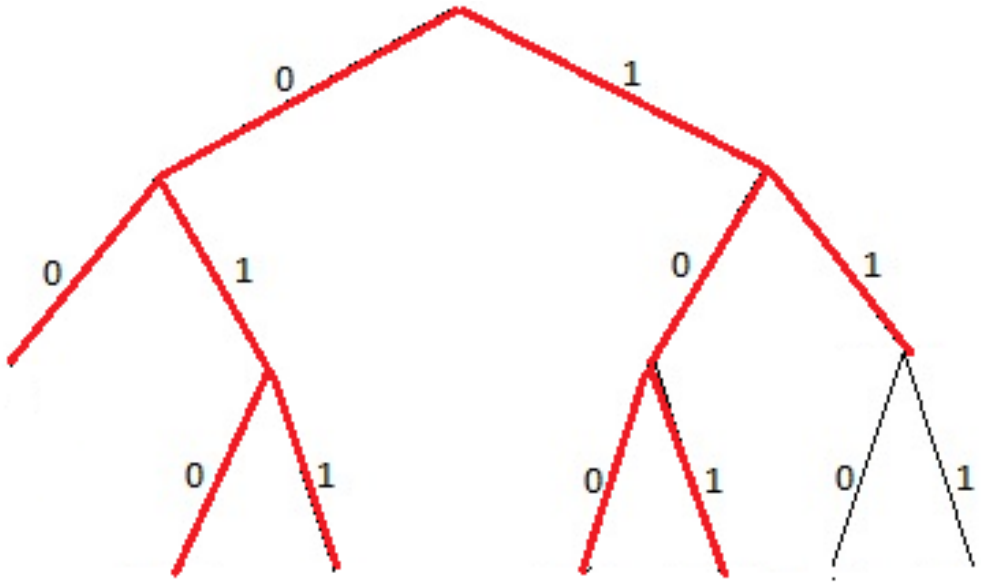
Processus d'élagage



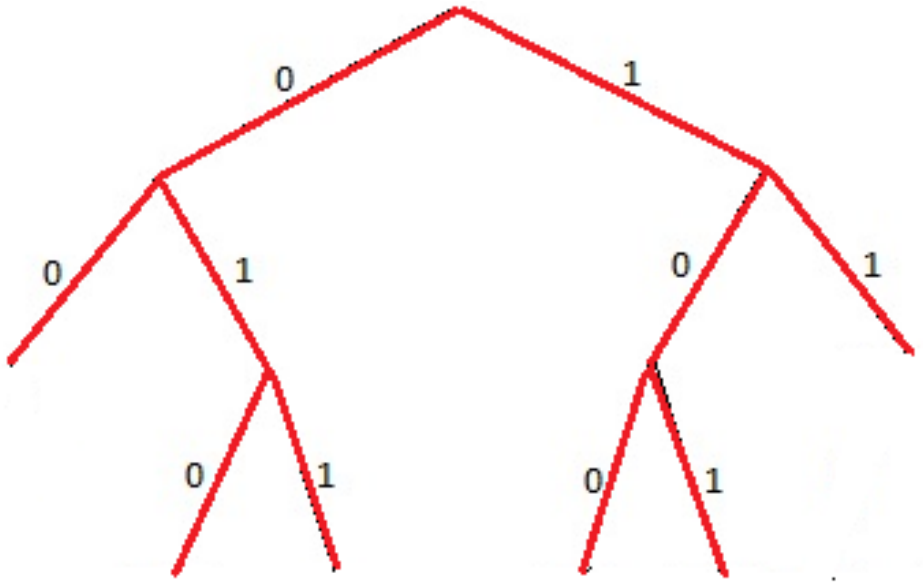
Processus d'élagage



Processus d'élagage



Processus d'élagage



Pour tester nos résultats, nous simulons une VLHMM $(X_k, Y_k)_{k=1..,n}$ où

- $\mathbb{X} = \{0, 1\}$
- l'arbre de contextes minimal est celui représenté précédemment.
- le vrai paramètre d'émission est donné par $m_0^\star = 0$, $m_1^\star$ valant 2,3 ou 4 (cf tableau suivant), et $\sigma^{2,\star} = 1$.
- n varie entre 100 et 50000

Nous comparons notre estimateur $\hat{\tau}_n$ avec l'estimateur BIC, qui vérifie

$$pen_{BIC}(n, \tau) = \frac{|\mathbb{X}| - 1}{2} |\tau| \log n$$

qui est beaucoup moins "lourde" que $pen_\alpha(n, \tau)$.

Résultats de simulations :

$\tau^\star = \tau_1^\star$, $\lvert\tau^\star\rvert = 6$						
	Penalty (5)			BIC penalty		
$n/m_1^\star$	2	3	4	2	3	4
100	2	2	2	2	3	3
1000	2	2	2	7	**6**	**6**
2000	2	2	4	**6**	**6**	**6**
5000	2	4	4	7	**6**	**6**
10000	4	**6**	**6**	7	**6**	**6**
20000	5	**6**	**6**	**6**	**6**	**6**
30000	5	**6**	**6**	**6**	**6**	**6**
40000	**6**	**6**	**6**	7	**6**	**6**
50000	**6**	**6**	**6**	7	**6**	**6**

Table I: Case $\tau^\star = \tau_1^\star$. Comparison of $\lvert\hat{\tau}_n\rvert$ between our estimator and the BIC estimator for different values of n and $m_1^\star$.

$\tau^\star = \tau_1^\star$, $\lvert\tau^\star\rvert = 6$						
	Penalty (5)			BIC penalty		
$n/m_1^\star$	2	3	4	2	3	4
100	-202	-202	-190	-6	-6	2
1000	-235	-213	-155	4	-2	25
2000	-221	-129	-88	8	-4	4
5000	-144	-36	-20	5	-4	-5
10000	-75	-5	-4	4	-5	-4
20000	-6	-4	-4	10	-4	-4
30000	21	-5	-4	10	-5	-4
40000	12	-4	-3	10	-4	-3
50000	12	-7	-4	10	-4	-4

Table II: Case $\tau^\star = \tau_1^\star$. Score difference $sc(\hat{\tau}_n) - sc(\tau^\star)$.


J. Rissanen, "A universal data compression system," *IEEE Trans. Inform. Theory*, vol. 29, pp. 656 – 664, 1983.


I. Csiszár and Z. Talata, "Context tree estimation for not necessarily finite memory processes, via bic and mdl," *IEEE Trans. Inf. Theory*, vol. 52, pp. 1007–1016, 2006.


Z. L. W. J. Wang, Y. and Z. Liu, "Mining complex time-series by learning markovian models," in *Proceedings ICDM'06, sixth international conference on data mining*, China, 2005.


Y. Wang, "The variable-length hidden markov model and its applications on sequential data mining," Departement of computer science, Tech. Rep., 2005.


P. Collet, A. Galves, and F. Leonardi, "Random perturbations of stochastic processes with unbounded variable length memory," *Electron. J. Probab.*, vol. 13, pp. no. 48, 1345–1361, 2008.

Merci beaucoup pour votre attention.