

Estimation de chaînes de Markov cachées : méthodes et problèmes

Gilles Celeux, Jean Clairambault
INRIA Rocquencourt

Résumé

Nous présentons les principaux algorithmes d'identification des chaînes de Markov cachées : algorithmes de Baum, de Viterbi, estimation bayésienne, utilisation de l'échantillonnage de Gibbs. Nous mettons l'accent sur l'interprétation des chaînes de Markov cachées comme un modèle de mélange de lois de probabilité où les observations sont dépendantes, et sur le fait que le problème statistique posé est un problème à données incomplètes. Les caractéristiques de ces différentes méthodes sont comparées. Nous présentons une application ayant trait à l'analyse du rythme cardiaque de nouveau-nés réalisée à l'aide du langage synchrone SIGNAL.

Mots clés : Mélange de lois, dépendance markovienne, algorithmes de restauration-maximisation, rythme cardiaque.

Abstract

We present the main estimation procedures of Hidden Markov Chains: Baum forward-backward algorithm, Viterbi algorithm, Bayesian estimation and methods using Gibbs sampling. We interpret Hidden Markov Chains as mixtures with dependent data. We regard the statistical problem as an incomplete data model. The characteristics of the methods are discussed. An application, tackled with the synchronous language SIGNAL, concerning heart rate variability of neonates, is presented.

Keywords: Mixture, Markovian dependence, restoration-maximization algorithms, heart rate variability.

1 Le modèle

Soit un processus observé x_1, \dots, x_N à valeurs dans \mathbb{R}^s et un processus caché z_1, \dots, z_N à valeurs dans $\{1, \dots, k\}$ tels que

- les z_i forment une chaîne de Markov homogène et irréductible de matrice de transition $\mathbf{T} = (t_{j\ell} = P(z_{i+1} = \ell | z_i = j))$ et de distribution stationnaire $(p_\ell, \ell = 1, \dots, k) (0 < p_\ell < 1)$,
- les x_i sont indépendants conditionnellement aux z_i ,
- si $z_i = \ell$, x_i suit une loi de densité $\phi(x_i | \theta_\ell)$ appartenant à une famille paramétrée. (Par exemple, $\phi(\cdot | \theta)$ est la densité d'une loi normale et $\theta = (\mu, \sigma)$, μ étant la moyenne de la loi et σ sa matrice variance.)

Il s'ensuit que les x_i suivent la loi de densité

$$f(x) = \sum_{\ell=1}^k p_\ell \phi(x | \theta_\ell). \quad (1)$$

Ainsi le modèle des chaînes de Markov cachées est un modèle de mélanges pour lequel la traditionnelle hypothèse d'indépendance des données observées x_1, \dots, x_n a été remplacée par une hypothèse de dépendance markovienne.

Le problème est d'estimer les paramètres $(\theta_\ell, \ell = 1, \dots, k)$, la matrice de transition de la chaîne de Markov cachée $\mathbf{T} = (t_{j\ell}, j = 1, \dots, k; \ell = 1, \dots, k)$ et sa probabilité stationnaire $(p_\ell, \ell = 1, \dots, k)$. Avant de nous intéresser à ce problème, signalons que le modèle de chaînes de Markov cachées (CMC) dont la première étude significative est due à Baum, Petrie, Soulie et Weiss [BAU 70], est devenu un modèle des plus classiques en reconnaissance de la parole (cf., par exemple, [RAB 85], [GUC 90], [JUR 91] ou [AND 92], et, pour un état de l'art récent [GUE 92]). A l'heure actuelle, des applications dans de nombreux domaines se multiplient : données neurophysiologiques ([ALB 91], [LLP 92]), analyse de la structure de l'ADN ([CHU 89]),... Quant au présent article, il contient une application à l'analyse du rythme cardiaque du nouveau-né.

2 Algorithmes de restauration-maximisation

Le modèle des CMC rentre typiquement dans les modèles où les données complètes $\mathbf{y} = (y_1, \dots, y_n)$ de densité $g(\mathbf{y} | \eta)$ se décomposent en données observées $\mathbf{x} = (x_1, \dots, x_m)$ de densité $f(\mathbf{x} | \eta)$ et en données manquantes $\mathbf{z} = (z_1, \dots, z_p)$ (ici, on a $m = p = n$) de densité conditionnelle

$$k(\mathbf{z} | \mathbf{x}, \eta) = \frac{g(\mathbf{y} | \eta)}{f(\mathbf{x} | \eta)}, \quad (2)$$

avec $\mathbf{y} = (\mathbf{x}, \mathbf{z})$. Le fait que les données soient incomplètes rend souvent l'estimation par le maximum de vraisemblance du paramètre η difficile et a

conduit au développement d'algorithmes de *restauration-maximisation* (cf. [QIT 91]) qui fonctionnent comme suit. Partant de η^0 , ils utilisent deux étapes :

Restauration : “donner” des valeurs \mathbf{z}^m aux données manquantes sachant $\eta = \eta^m$.

Maximisation : trouver η^{m+1} qui maximise la vraisemblance complétée $g(\mathbf{y}^m|\eta)$ où $\mathbf{y}^m = (\mathbf{x}, \mathbf{z}^m)$.

Fondamentalement, les algorithmes diffèrent par leurs stratégies de restauration dont les principales sont décrites maintenant. Les valeurs manquantes sont remplacées par :

- leur espérance conditionnelle sachant \mathbf{x} et η^m . C'est la stratégie de l'algorithme EM [DLR 77], qui conduit, à chaque itération, à calculer η^{m+1} qui maximise $E[\log g(\mathbf{y}|\cdot)|\mathbf{x}, \eta^m]$.
- leur valeur la plus probable sachant \mathbf{x} et η^m . Cette stratégie est celle des algorithmes du MAP (maximum a posteriori).
- tirage(s) aléatoire(s) suivant la loi conditionnelle $k(\mathbf{z}|\mathbf{x}, \eta^m)$. C'est la stratégie de l'algorithme SEM [CED 85] où les z_i^m sont issus d'un seul tirage à chaque itération et celle de l'algorithme MCEM [WET 90] où les z_i^m sont obtenus comme moyenne de r tirages aléatoires (r augmentant en général avec le numéro d'itération).

Dans le cadre des CMC, l'algorithme EM, proposé par [BAU 70], est connu sous le nom de *Forward-Backward algorithm* et l'algorithme du MAP est connu sous le nom d'*algorithme de Viterbi*. Notons que l'algorithme du MAP est simplement un algorithme de classification de type nuées dynamiques dans le cadre de mélanges indépendants et l'algorithme ICM de Besag [BES 86] dans le cadre des champs de Markov.

Avant de présenter les algorithmes pour les CMC, il est utile de dresser un rapide panorama de leurs propriétés théoriques et pratiques pour l'identification de mélanges indépendants. (Pour plus de détails, on pourra se reporter à [CEL 92].) C'est en effet un domaine d'investigation très étudié des algorithmes de restauration-maximisation. Et tout laisse à penser que ces caractéristiques sont également valables pour l'identification des CMC et des champs de Markov puisque ces modèles sont également des modèles de mélange.

- L'algorithme du MAP présente l'intérêt de converger en un nombre fini d'itérations et toujours rapidement. Mais il peut fournir des estimations très biaisées, même avec des échantillons de grande taille, pour peu que les paramètres θ_ℓ intervenant dans l'équation (1) soient relativement proches et surtout que les proportions p_ℓ soient assez différentes.
- L'algorithme EM fournit souvent de très bons résultats. Un théorème [REW 84] garantit sa convergence vers la solution consistante des équations de la vraisemblance si sa position initiale est assez proche de cette

solution optimale. Cependant, lorsque les composants du mélange ne sont pas très bien séparés, il fournit des estimations très dépendantes de sa position initiale et converge souvent avec une lenteur insupportable.

- L'algorithme SEM fournit des estimations faiblement dépendantes de sa position initiale. Il parvient à identifier des mélanges assez imbriqués et permet de déceler le nombre de composants du mélange si on en connaît un majorant, pour peu que l'échantillon des données observées ne soit pas trop petit. Pour de très petits échantillons, l'algorithme MCEM est plus fiable que SEM. Sous des conditions techniques assez fortes, on a montré (cf. [DIC 91]) la convergence asymptotique (lorsque n tend vers l'infini) de la loi stationnaire de la suite des estimés fournis par SEM vers une loi normale de variance finie et centrée sur l'estimateur consistant du maximum de vraisemblance.

3 Les algorithmes de Baum et de Viterbi

Nous présentons maintenant les deux algorithmes les plus répandus pour l'identification des CMC. Avant de présenter les formules de l'algorithme EM pour les CMC, il convient de signaler le résultat de consistance de l'approche du maximum de vraisemblance [LER 92] qui conforte cette approche.

Théorème ([LER 92]) *Si le mélange est identifiable¹, si sa matrice de transition \mathbf{T} est irréductible, et sous des conditions de régularité classiques (à la Wald), la suite des estimateurs du maximum de vraisemblance tend p.s. vers les vraies valeurs des paramètres du modèle des CMC.*

3.1 L'algorithme de Baum

On peut décrire ainsi l'itération m de l'algorithme EM ou algorithme forward-backward de Baum :

Etape E : l'étape de restauration revient à calculer les quantités

$$D_{j\ell}^m(i) = P(z_{i+1} = \ell, z_i = j \mid \mathbf{x}, \eta^m) \quad (1 \leq i \leq n-1) \quad (1 \leq j, \ell \leq k)$$

$\eta^m = (\mathbf{T}^m, \theta_1^m, \dots, \theta_k^m)$ étant l'estimation courante des paramètres du modèle. Ces calculs reposent sur l'égalité ([BAU 70])

$$D_{j,\ell}^m(i) = \frac{F_i^m(j) t_{j\ell}^m \phi(x_{i+1} \mid \theta_\ell^m) B_{i+1}^m(\ell)}{f(\mathbf{x} \mid \eta^m)} \quad (3)$$

avec

$$F_i(j) = P(x_1, \dots, x_i, z_i = j \mid \eta),$$

$$B_i(\ell) = P(x_{i+1}, \dots, x_n \mid z_i = \ell, \eta)$$

¹Un mélange est identifiable ssi $\sum_{j=1}^k p_\ell \phi(x, \theta_\ell) = \sum_{j=1}^{k'} p'_\ell \phi(x, \theta'_\ell)$ implique $k = k'$, $p_\ell = p'_\ell$ et $\theta_\ell = \theta'_\ell$ pour tout $\ell = 1, \dots, k$.

et $f(\mathbf{x} | \eta) = \sum_{j=1}^k F_n(j)$.

Le calcul des F_i se fait par la procédure avant suivante :

- a) $F_1^m(j) = p_j^m \phi(x_1 | \theta_j^m)$ ($j = 1, \dots, k$)
- b) $F_{i+1}^m(j) = \phi(x_{i+1} | \theta_j^m) \sum_{\ell=1}^k F_i^m(\ell) t_{\ell j}^m$ ($j = 1, \dots, k$).

Le calcul des B_i se fait par la procédure arrière suivante :

- a) $B_n^m(\ell) = 1$ ($\ell = 1, \dots, k$).
- b) $B_i^m(\ell) = \sum_{j=1}^k t_{\ell j}^m \phi(x_{i+1} | \theta_j^m) B_{i+1}^m(j)$ ($i < n$) ($\ell = 1, \dots, k$).

Etape M : l'étape de maximisation conduit à calculer

$$t_{j\ell}^{m+1} = \frac{\sum_{i=1}^{n-1} D_{j\ell}^m(i)}{\sum_{i=1}^{n-1} \sum_{r=1}^k D_{jr}^m(i)}$$

Quant à l'estimation des θ_j^{m+1} , elle dépend bien sûr de la famille paramétrée $\phi(x|\theta)$ considérée. Par exemple, dans le cas gaussien réel, $\theta_j = (\mu_j, \sigma_j^2)$, μ_j étant la moyenne de la loi et σ_j^2 sa variance, on obtient

$$\begin{aligned} \mu_j^{m+1} &= \frac{\sum_{i=1}^n w_j^{m+1}(i) x_i}{\sum_{i=1}^n w_j^{m+1}(i)} \\ (\sigma_j^2)^{m+1} &= \frac{\sum_{i=1}^n w_j^{m+1}(i) (x_i - \mu_j^{m+1})^2}{\sum_{i=1}^n w_k^{m+1}(i)} \end{aligned}$$

où $w_j^{m+1}(i) = \sum_{\ell=1}^k D_{j\ell}^{m+1}(i)$.

Enfin, les proportions du mélange sont estimées par

$$p_j^{m+1} = \frac{1}{n} \sum_{i=1}^n w_j^{m+1}(i).$$

La procédure avant-arrière intervenant dans l'étape E est bien sûr très lourde. De plus, elle induit très rapidement un problème d'“underflow”. Aussi, en pratique, on utilise une procédure avant-arrière modifiée pour éviter ce problème. La solution la plus élégante et la plus simple consiste à travailler sur des probabilités conditionnelles (cf. [DEV 85]). On peut écrire (l'indice m d'itération est omis) :

$$D_{j\ell}(i) = \tilde{F}_i(j) t_{j\ell} \tilde{B}_i(\ell)$$

avec

$$\tilde{F}_i(j) = P(z_i = j | x_1, \dots, x_i, \eta)$$

et

$$\tilde{B}_i(\ell) = \frac{P(x_{i+1}, \dots, x_n | z_i, \eta)}{P(x_{i+1}, \dots, x_n | x_1, \dots, x_i, \eta)}.$$

Et les \tilde{F}_i et les \tilde{B}_i se calculent par des procédures avant et arrière analogues aux précédentes.

Procédure avant :

$$\begin{aligned}\tilde{F}_1(j) &= \frac{t_{j\ell}\phi(x_1|\theta_j)}{\sum_{\ell=1}^k p_\ell\phi(x_1|\theta_\ell)} \\ \tilde{F}_{i+1}(j) &= \frac{\sum_{\ell=1}^k \tilde{F}_i(\ell)t_{\ell j}\phi(x_{i+1}|\theta_j)}{\sum_{r=1}^k \sum_{\ell=1}^k \tilde{F}_i(\ell)t_{\ell r}\phi(x_{i+1}|\theta_r)}.\end{aligned}$$

Procédure arrière :

$$\begin{aligned}\tilde{B}_n(\ell) &= 1 \\ \tilde{B}_i(\ell) &= \frac{\sum_{j=1}^k t_{\ell j}\phi(x_{i+1}|\theta_j)\tilde{B}_{i+1}(j)}{\sum_{r=1}^k \sum_{j=1}^k \tilde{F}_i(j)t_{jr}\phi(x_{i+1}|\theta_r)} \quad (1 < i < n) \\ \tilde{B}_1(\ell) &= \frac{\sum_{j=1}^k \phi(x_2|\theta_j)\tilde{B}_2(j)}{\sum_{j=1}^k p_j\phi(x_2|\theta_j)}.\end{aligned}$$

3.2 L'algorithme de Viterbi

Cet algorithme est l'algorithme du MAP adapté au modèle des CMC. On décrit ainsi l'itération m de cet algorithme :

Etape C : l'étape de restauration des données manquantes est une étape de classification. On construit \mathbf{z}^m par le principe du MAP. Cela donne (cf., par exemple, [JUR 91]) :

$$z_i^m = \arg \max_{\ell=1,\dots,k} t_{z_{i-1}^m \ell}^m \phi(x_i|\theta_\ell^m) t_{z_{i+1}^m \ell}^m.$$

Etape M : les termes de la matrice D^{m+1} sont calculés par la formule

$$D_{j\ell}^{m+1} = \frac{1}{n} \sum_{i=2}^n \mathbf{I}_{\{z_{i-1}^m=j, z_i^m=\ell\}}$$

De là, on déduit aisément l'actualisation de la matrice de transition \mathbf{T} , des θ_ℓ et des proportions p_ℓ par les mêmes formules que celles de l'algorithme de Baum.

Ainsi les formules de l'algorithme de Viterbi sont particulièrement simples et, de plus, cet algorithme converge rapidement. Aussi il peut être recommandé pour l'identification de CMC dont les composants du mélange sont bien séparés ou pour initialiser un algorithme comme celui de Baum. Par contre, il peut s'avérer peu fiable lorsque les composants du mélange sont assez proches. Une simulation illustrant cette caractéristique est présentée dans la section 5.

4 Deux algorithmes stochastiques

L'algorithme SEM (version stochastique naturelle de EM) peut bien sûr s'appliquer sans difficulté pour estimer les CMC. Il a d'ailleurs été utilisé avec succès pour estimer des champs de Markov en analyse d'images (cf., par exemple, [QIT 91] et [MAP 91]). Cependant, cet algorithme n'évite pas le recours à la procédure avant-arrière puisque les tirages aléatoires pour compléter les données se font suivant la loi conditionnelle donnée dans la formule (2). Dans cette section, nous allons présenter deux algorithmes stochastiques qui évitent le recours à la procédure avant-arrière grâce à l'échantillonnage de Gibbs qui constitue un outil puissant pour l'approximation de calculs bayésiens. (Voir [GES 90] ou [ROB 92] chapitre 9 pour une présentation détaillée de l'échantillonnage de Gibbs.) Le premier algorithme propose ainsi une solution bayésienne pour estimer une CMC ; le second propose une utilisation de l'échantillonnage de Gibbs dans un cadre fréquentiste et permet de proposer une version stochastique de l'algorithme EM évitant la procédure avant-arrière de l'étape E.

4.1 Estimation bayésienne

Le modèle des CMC nécessite l'estimation de nombreux paramètres. Aussi, l'initialisation des algorithmes est un point sensible pour une bonne estimation des CMC. En règle générale, il vaut mieux partir d'une position proche des bonnes valeurs des paramètres et faire si possible usage des informations a priori disponibles... Ce qui invite à une analyse bayésienne des CMC. Cette analyse a été réalisée par [RCD 92a] et nous la résumons ici.

On suppose que les densités $\phi(x|\theta)$ appartiennent à une famille exponentielle : $\phi(x|\theta) = h(x) \exp(\theta.t(x) - \psi(\theta))$. De la sorte, il existe une loi a priori conjuguée, $\pi(\theta)$, pour θ . On considère que les colonnes \mathbf{t}_j de la matrice de transition \mathbf{T} suivent des lois a priori de Dirichlet indépendantes $\mathcal{D}(\alpha_1^j, \dots, \alpha_k^j)$.

Dans ce cadre, la loi a posteriori de $\eta = (\theta_1, \dots, \theta_k, \mathbf{T})$ est

$$\pi(\eta|\mathbf{x}) \propto \left(\sum_{i_2=1}^k \dots \sum_{i_n=1}^k t_{1,i_2} \phi(x_2|\theta_{i_2}) \dots t_{i_{n-1},i_n} \phi(x_n|\theta_{i_n}) \right) \phi(x_1|\theta_1) \prod_{j=1}^k \left\{ \pi(\theta_j) \left(\prod_{\ell=1}^k t_{j,\ell}^{\alpha_\ell^j - 1} \right) \right\} \quad (4)$$

Elle est incalculable car elle comporte une somme de k^{n-1} termes. Mais la loi a posteriori conditionnelle $\pi(\eta|\mathbf{x}, \mathbf{z})$ est plus simple :

$$\pi(\eta|\mathbf{x}, \mathbf{z}) \propto \prod_{j=1}^k \left\{ \pi(\theta_j) \left(\prod_{\ell=1}^k t_{k,\ell}^{\alpha_\ell^j - 1} \right) \right\} \left\{ \prod_{i=2}^n t_{z_{i-1}, z_i} \phi(x_i|\theta_{z_i}) \right\} \phi(x_1|\theta_1).$$

Ainsi la loi conditionnelle de $\mathbf{t}_j = (t_{j,1}, \dots, t_{j,k})$:

$$\mathbf{t}_j \sim \mathcal{D} \left(\alpha_1^j + \sum_{i=2}^n \mathbf{I}_{\{z_{i-1}=j\}} \mathbf{I}_{\{z_i=1\}}, \dots, \alpha_k^j + \sum_{i=2}^n \mathbf{I}_{\{z_{i-1}=j\}} \mathbf{I}_{\{z_i=k\}} \right) \quad (5)$$

est facile à simuler. De même, par exemple, si la loi de $\phi(x|\theta_j)$ est une loi normale $\mathcal{N}(\theta_j, 1)$ et si la loi a priori conjuguée est une loi normale $\mathcal{N}(\mu_j, 1)$, on a

$$\theta_j | \mathbf{x}, \mathbf{z} \sim \mathcal{N} \left(\frac{\sum_{i=1}^n \mathbf{I}_{\{z_i=j\}} x_i + \mu_j}{\sum_{i=1}^n \mathbf{I}_{\{z_i=j\}} + 1}, \frac{\sum_{i=1}^n \mathbf{I}_{\{z_i=j\}}}{\sum_{i=1}^n \mathbf{I}_{\{z_i=j\}} + 1} \right). \quad (6)$$

Par ailleurs, pour éviter la procédure avant-arrière, on raisonne de même sur la loi conditionnelle de z_i sachant “le reste”

$$\begin{aligned} f(z_i | \mathbf{x}, \eta, z_{j \neq i}) &= f(z_i | x_i, \eta, z_{i-1}, z_{i+1}) \\ &= \frac{t_{z_{i-1}, z_i} \phi(x_i | \theta_{z_i}) t_{z_i, z_{i+1}}}{\sum_{j=1}^k t_{z_{i-1}, j} \phi(x_i | \theta_j) t_{j, z_{i+1}}} \end{aligned} \quad (7)$$

$$f(z_n | \mathbf{x}, \eta, z_{j < n}) \propto t_{z_{n-1}, z_n} \phi(x_n | \theta_{z_n}).$$

Partant de ces équations, on peut utiliser l'échantillonnage de Gibbs pour, à partir d'un vecteur d'état \mathbf{z}^0 , simuler η^m selon les équations (5) et (6), puis simuler \mathbf{z}^m selon (7). On construit ainsi une chaîne de Markov ergodique (η^m, \mathbf{z}^m) dont l'unique densité stationnaire est $\pi(\eta, \mathbf{z} | \mathbf{x})$. De plus, on a le théorème suivant :

Théorème [RCD 92a] (i) *Il existe des constantes $C > 0$ and $0 \leq \varrho < 1$ telles que, quel que soit η^0 ,*

$$\|\pi^m(\cdot) - \pi(\cdot | \mathbf{x})\|_1 \leq C \varrho^m.$$

(ii) *Pour toute fonction $h(\eta)$ telle que $\mathbf{E}^\pi[|h(\eta)| | \mathbf{x}] < +\infty$ et quel que soit $\pi^0(\eta)$, il existe une constante $C' > 0$ telles que*

$$|\mathbf{E}^{\pi^m}[h(\eta)] - \mathbf{E}^\pi[h(\eta) | \mathbf{x}]| \leq C' \varrho^m.$$

(iii) *Le processus (θ^m) est géométriquement φ -mélangeant.*

4.2 EM à la Gibbs

Afin de proposer une version stochastique de EM qui évite la procédure avant-arrière de l'étape E, un algorithme dénommé EM à la Gibbs, qui adapte l'échantillonnage de Gibbs au cadre fréquentiste pour simuler les données manquantes \mathbf{z} à chaque itération, a été conçu [RCD 92b]. Cet algorithme fonctionne ainsi : partant d'une position initiale η^0 , le vecteur \mathbf{z}^m est simulé selon la loi de probabilité définie par (7). L'actualisation de η^m dans l'étape de maximisation est identique à celle de l'algorithme de Viterbi (ou de l'algorithme SEM).

Toutefois on doit signaler une précaution numérique. Du fait que \mathbf{z}^m est tiré au hasard, il se peut que des éléments de la matrice \mathbf{T}^{m+1} soient nuls. Et dans ce cas ils resteront nuls et la chaîne \mathbf{z}^m ne sera plus irréductible. Pour éviter cela, on stabilise l'estimation de la matrice \mathbf{T} en utilisant la formule

$$p_{j,\ell}^{(m+1)} = \frac{\alpha + \sum_{i=2}^n \mathbf{I}_{\{z_{i-1}^m=j\}} \mathbf{I}_{\{z_i^m=\ell\}}}{\sum_{i=2}^n \mathbf{I}_{\{z_{i-1}^m=j\}} + n\alpha}$$

où α est un nombre positif à choisir petit devant n . De plus, cette modification dans le calcul de \mathbf{T}^m permet d’assurer le théorème suivant :

Théorème [RCD 92b] *La chaîne de Markov (\mathbf{z}^m) engendrée par l’algorithme EM à la Gibbs est uniformément géométriquement ergodique et possède une distribution stationnaire unique.*

Ce théorème ne renseigne pas sur la nature de cette distribution stationnaire. On peut conjecturer, et cela est confirmé par la pratique (cf. section 5), que cette distribution se concentre autour de l’estimateur du maximum de vraisemblance. Ainsi l’algorithme EM à la Gibbs apparaît comme une version de SEM spécialement adaptée au cadre des CMC.

5 Illustrations

Nous avons programmé deux des algorithmes étudiés plus haut : l’algorithme du MAP (ou de Viterbi) et l’algorithme “EM à la Gibbs” en langage SIGNAL. En effet, SIGNAL est un langage flot-de-données, parallèle et synchrone, conçu pour des applications en temps réel, et notre but à terme est d’identifier sur des fenêtres de taille fixe, en temps réel, des états cachés gouvernant une variable observée.

5.1 Simulations

Dans cette première étude, nous avons testé ces programmes sur une seule fenêtre de 4096 points. La figure 1 illustre les performances de ces deux algorithmes sur un signal simulé, mélange de deux gaussiennes $\mathcal{N}(10, 4)$ et $\mathcal{N}(12, 4)$ suivant une variable d’état simple ne prenant que deux valeurs.

L’algorithme de Viterbi atteint très rapidement (en 20 itérations) un état stationnaire, mais l’identification des paramètres et de la variable d’état est peu satisfaisante : 908 erreurs sur 4096 pour la variable d’état, les lois estimées sont $\mathcal{N}(9.82, 3.06)$ au lieu de $\mathcal{N}(10, 4)$, $\mathcal{N}(12.75, 2.43)$ au lieu de $\mathcal{N}(12, 4)$, et les probabilités de persistance dans un état sont 0.90 et 0.83, au lieu de 0.999... pour les deux états. On retrouve ici la tendance des algorithmes du MAP à surestimer les différences entre moyennes et à sous-estimer les variances (cf. [CEL 92]). En revanche, EM à la Gibbs converge plus lentement, mais très sûrement : on peut voir sur la figure 1 que l’identification est déjà bonne dès 50 itérations, et qu’elle est presque parfaite après 1600 itérations (il subsiste des erreurs d’identification, 38 valeurs erronées sur 4096, aux changements d’état). Les estimations des paramètres¹ : $\mathcal{N}(10.01, 3.89)$ et $\mathcal{N}(11.92, 3.94)$ pour les lois, 0.9991 et 0.9995 pour les probabilités de rester dans un état sont également très satisfaisantes.

¹Dans cette version, les estimations sont obtenues par un passage d’une itération de Viterbi à l’issue des itérations de EM à la Gibbs.

Signalons que les paramètres ont été initialisés ici à des valeurs ($\mathcal{N}(8, 3)$, $\mathcal{N}(14, 2)$, 0.98 et 0.99 pour les probabilités de non-transition) pas exagérément éloignées des vraies valeurs, mais que d'autres simulations ont montré qu'EM à la Gibbs supporte des perturbations importantes de ces valeurs initiales, tandis que l'algorithme de Viterbi ne donne plus que des performances très médiocres quand ces valeurs initiales s'éloignent trop des valeurs à identifier.

5.2 Application à l'étude du rythme cardiaque

L'application de modèles de Markov cachés à l'étude de signaux d'origine biologique n'est pas une nouveauté, mais, à notre connaissance, rien n'avait été tenté jusqu'ici sur le rythme cardiaque.

Et pourtant il est depuis longtemps connu que la variabilité du rythme cardiaque est un indicateur de l'état du système nerveux autonome, utilisé notamment en gynécologie-obstétrique (surveillance du bien-être foetal), en diabétologie (recherche de neuropathie diabétique), et bien sûr en cardiologie (surveillance des sujets à risque de mort subite, après un infarctus du myocarde par exemple). Le système nerveux autonome est subdivisé en deux branches : une branche dite sympathique, cardioaccélétratrice, responsable des fluctuations de basse fréquence du rythme cardiaque, et une branche parasympathique, cardiomodératrice, à l'origine de la variabilité de haute fréquence qui se manifeste par un couplage de la respiration et du rythme cardiaque (arythmie sinusale respiratoire) (cf. [CLA 92]).

Il est donc tentant d'aller identifier sur le rythme cardiaque une variable cachée "état du système nerveux autonome", qu'on peut schématiquement classer en deux catégories : à prédominance sympathique ou à prédominance parasympathique.

Mais nous n'avons pas accès, à l'inverse du cas d'un signal simulé, aux "vraies" valeurs de la variable d'état. En revanche, il est connu qu'en sommeil calme, l'état du système nerveux autonome est à prédominance parasympathique, et à prédominance sympathique en sommeil agité (il s'agit du sommeil de nouveau-nés, codé en stades par des cliniciens, au vu du tracé de l'électroencéphalogramme et des mouvements oculaires rapides, donc indépendamment de l'électrocardiogramme). On trouvera sur la figure 2 un extrait de tracé du rythme cardiaque d'un nouveau-né à terme et de sa variabilité à haute fréquence (obtenue par transformation de Fourier à court terme; bande étudiée : oscillations de période allant de 3 à 8 battements). Ce tracé présente une période de forte activité à haute fréquence, encadrée par deux périodes où l'activité à basse fréquence l'emporte. Ces périodes correspondent à une période de sommeil calme encadrée par deux périodes de sommeil agité. Mais bien sûr il subsiste de l'activité à basse fréquence en sommeil calme, et de l'activité à haute fréquence en sommeil agité.

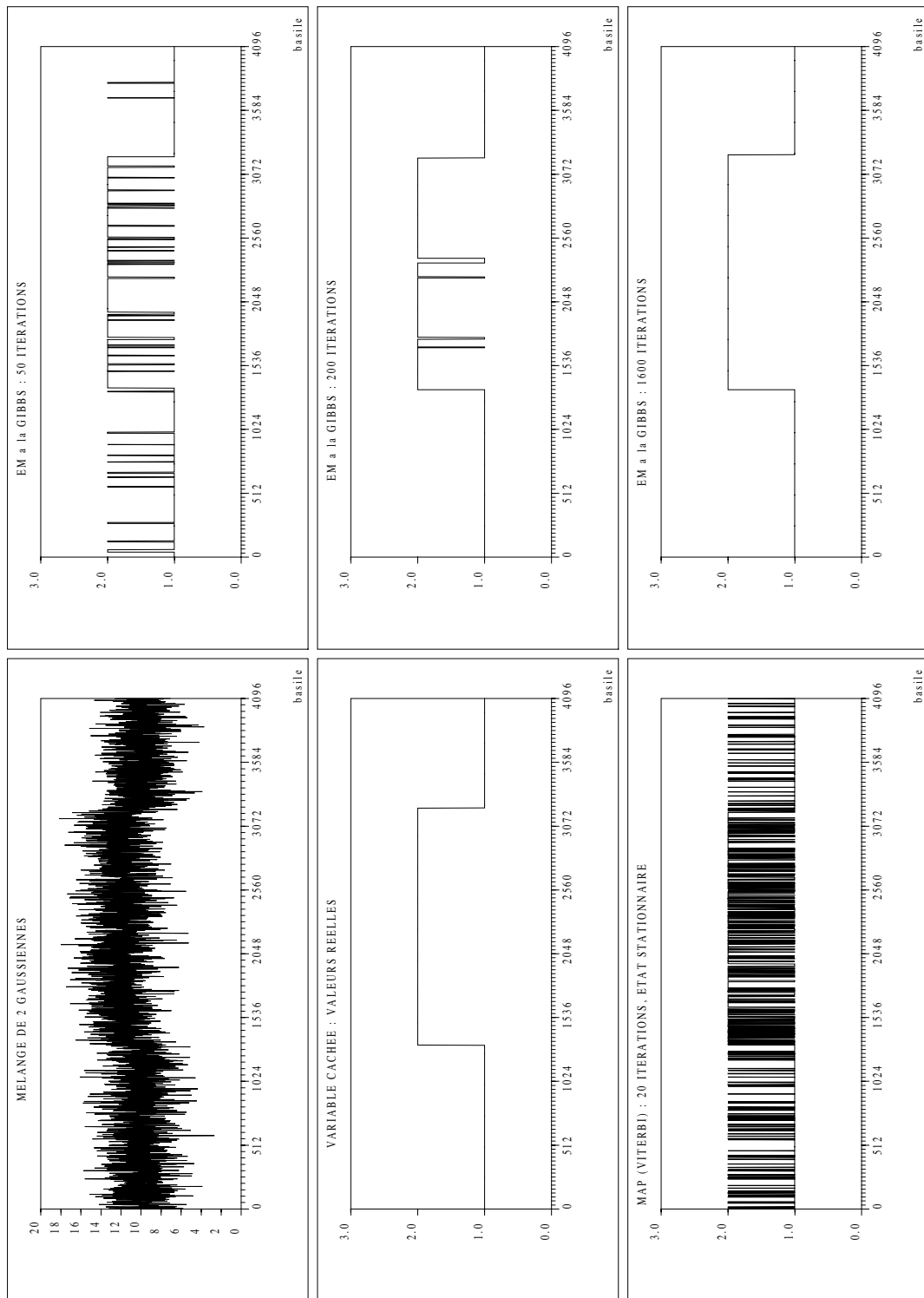


Figure 1. Performances comparées de l'algorithme du maximum a posteriori (MAP, ou algorithme de Viterbi) et de l'algorithme EM à la Gibbs sur l'identification d'une chaîne de Markov cachée simple à deux états gouvernant un mélange de deux signaux gaussiens.

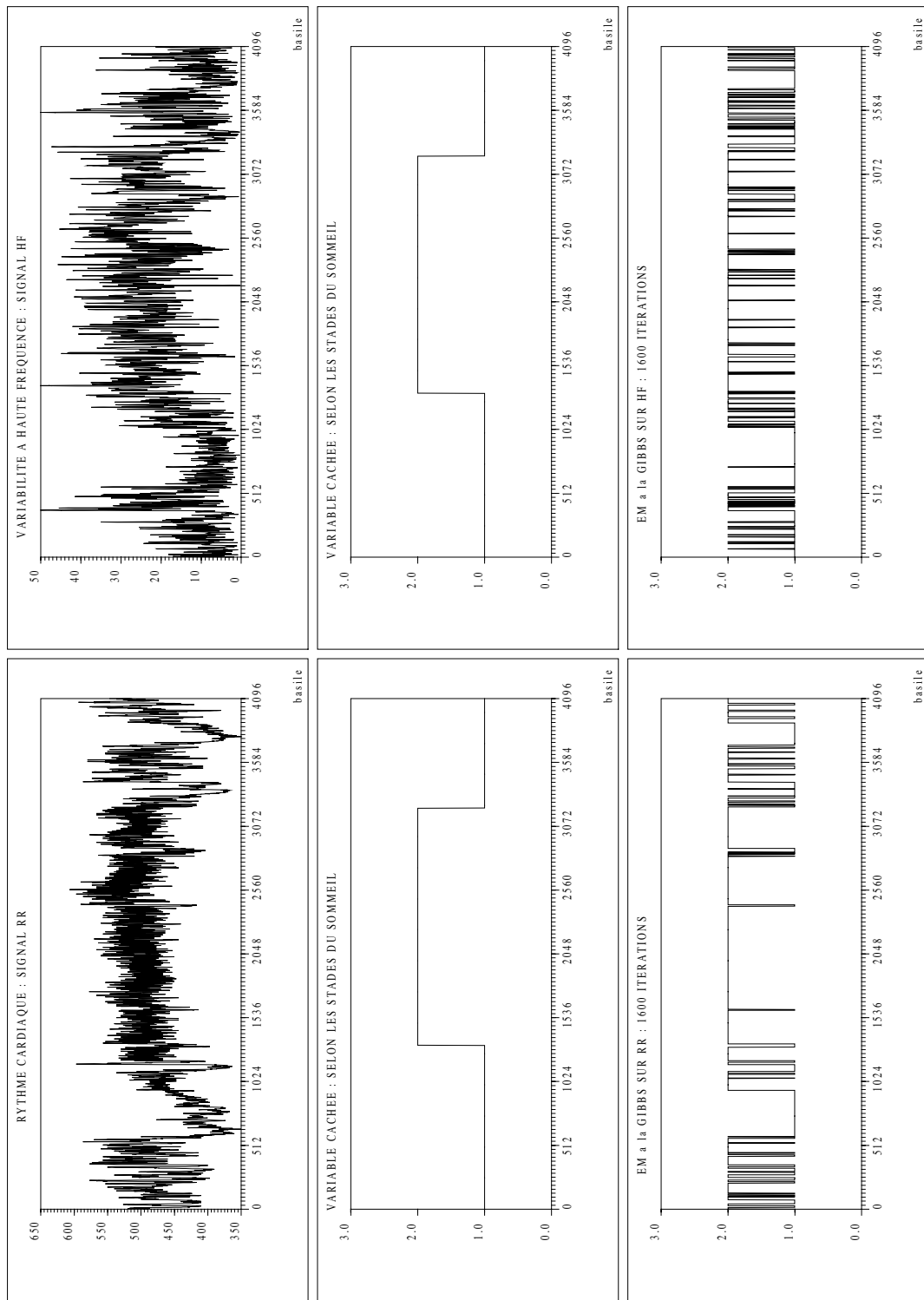


Figure 2. Tentative d'identification par l'algorithme EM à la Gibbs d'une variable cachée contrôlant le rythme cardiaque (état sympathique vs. parasympathique du système nerveux autonome), d'une part sur le signal RR brut (à gauche), d'autre part sur le signal HF, variabilité à haute fréquence extraite par transformation de Fourier à court terme du signal RR (à droite).

Il en résulte que l'identification sur ces tracés par l'algorithme EM à la Gibbs d'une hypothétique variable cachée est nécessairement imparfaite. Pourtant, EM à la Gibbs propose une classification du tracé en deux états qui n'est pas si éloignée du codage en stades du sommeil, et qu'une procédure de lissage prenant en compte un temps de séjour minimal dans chaque état, par exemple, permettrait sans doute d'améliorer. Après 1600 itérations de l'algorithme, il y a 2901 valeurs exactes (71%) pour l'identification sur le signal RR, et 3242 valeurs exactes (79%) pour l'identification sur le signal HF (variabilité à haute fréquence); les probabilités estimées de persistance dans un état sont de 0.963 et 0.985 sur RR, et de 0.950 et 0.962 sur HF. Ces résultats donnent à penser qu'un modèle multidimensionnel, qui prendrait en compte les signaux RR, HF, BF (variabilité à basse fréquence), pourrait améliorer encore ces résultats.

Une limitation de ce modèle, dans son état actuel, tient à sa trop grande simplicité : autant les caractéristiques statistiques du rythme cardiaque en sommeil calme sont adéquates au modèle (faible autocorrélation autorisant l'hypothèse d'indépendance, normalité), autant le sommeil agité semble polymorphe et devrait être subdivisé en plusieurs états. Une autre possibilité d'affinement de la méthode serait de remplacer (dans le cas du signal HF) l'estimation d'un modèle gaussien par un modèle de type χ^2 .

Il reste aussi à évaluer, concurremment au modèle des chaînes de Markov cachées qui a été étudié ici, des méthodes autorégressives classiques de détection de rupture (cf. [BAB 86]).

Enfin une validation de la méthode par des expérimentations pharmacologiques animales (il existe des médicaments qui bloquent spécifiquement une branche ou l'autre du système nerveux autonome) semble nécessaire et reste à faire.

References

- [ALB 91] ALBERT, P. S. : A two-state Markov mixture model for a time series of epileptic seizure counts. *Biometrics*, **47**, 1371-1381, 1991.
- [AND 92] ANDRE-OBRECHT, R. : Utilisation des chaînes de Markov cachées en traitement de la parole. Ce volume, 1992.
- [BAB 86] BASSEVILLE, M. et BENVENISTE, A. (Eds.) : Detection of abrupt changes in signals and dynamical systems. *Lecture Notes in Control and Information Sciences*, **77**. Springer Verlag, New-York, 1986.
- [BAU 70] BAUM L. E., PETRIE T., SOULES G. et WEISS N. : A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains. *Ann. Math. Statistic.*, **41**, 164-171, 1970.
- [BES 86] BESAG, J. : On the statistical analysis of dirty pictures (with discussion). *J.R.S.S. B*, **48**, 192-236, 1986.

- [CEL 92] CELEUX, G. : Modèles probabilistes en classification. In *Modèles pour l'analyse des données multidimensionnelles*, (Droesbeke, J.-J., Fichet, B. et Tassi. P. éditeurs). Economica, Paris, pp. 165-214, 1992.
- [CED 85] CELEUX, G. et DIEBOLT, J. : The SEM Algorithm: a probabilistic teacher algorithm derived from the EM algorithm for the mixture problem. *Computational Statistics Quarterly*, **2**, 73-82, 1985.
- [CHU 89] CHURCHILL, G. : Stochastic Models for heterogeneous DNA sequences. *Bull. Math. Biol.* **51**, 79-94, 1989.
- [CLA 92] CLAIRAMBAULT, J., CURZI-DASCALOVA, L., KAUFFMANN, F., MEDIGUE, C. et LEFFLER, L. : Heart rate variability in normal full-term and preterm neonates. *Early Human Development*, **28**, 169-183. Elsevier Scientific Publishers Ireland Ltd., 1992.
- [DLR 77] DEMPSTER, A., LAIRD, N. et RUBIN, D. : Maximum likelihood from incomplete data via the EM algorithm (with discussion). *J.R.S.S. B*, **39**, 1-38, 1977.
- [DEV 85] DEVIJVER, P.A. : Baum's forward-backward algorithm revisited. *Pattern Recognition Letters* **3**, 369-373, 1985.
- [DIC 91] DIEBOLT, J. et CELEUX G. : Asymptotic properties of a stochastic EM algorithm for estimating mixing proportions. Technical Report 228, Department of Statistics, University of Washington, Seattle, USA, 1991.
- [GES 90] GELFAND, A. et SMITH, A.F. : Sampling based approaches to calculating marginal densities. *J. American Statistical Society* **85**, 398-409, 1990.
- [GUE 92] GUEDON, Y. : Review of several stochastic speech unit models. Alcatel Alsthom Recherche (soumis à publication), 1992.
- [GUC 90] GUEDON, Y. et COCCOZA-THIVENT, C. : Explicit state occupancy modeling by hidden semi-Markov models: application of Derin's scheme. *Computer Speech Language* **4**, 167-192, 1990.
- [JUR 91] JUANG, B.H. et RABINER, L.R. : Hidden Markov Models for Speech Recognition. *Technometrics* **33**, 251-272, 1991.
- [LLP 92] LE, N. D., LEROUX B. G. et PUTERMAN M. L. : Exact likelihood in a Markov mixture model for time series of seizure counts (reader reaction). *Biometrics* **48**, 317-323, 1992.
- [LER 92] LEROUX B. G. : Maximum likelihood estimation for hidden Markov models. *Stochas. Proc. and their Applications* **40**, 127-143.
- [MAP 91] MASSON, P. et PIECZYNSKI, W. : Segmentation conceptuelle non supervisée des images SPOT. *Actes de RFIA 91*, pp. 997-1002, Lyon, 1991.
- [QIT 91] QIAN, W. et TITTERINGTON, D.M. : Estimation of parameters in hidden Markov models. *Phil. Trans. Roy. Soc. London A* **337**, 407-428, 1991.
- [RAB 85] RABINER, L. R., JUANG, B. H., LEVINSON, S. E. et SONDHI, M.M. : Recognition of isolated digits using hidden Markov models with continuous mixture densities. *ATT Technical Journal* **64**, 1211-1234, 1985.

- [REW 84] REDNER, R. A. et WALKER, H. F. : Mixture densities, maximum likelihood and the EM algorithm. *SIAM Review*, **26**, 195-239, 1984.
- [ROB 92] ROBERT, C.P. : *L'analyse statistique bayésienne*. Economica, Paris, 1992.
- [RCD 92a] ROBERT, C. P., CELEUX, G. et DIEBOLT J. : Bayesian estimation of hidden Markov chains: A stochastic implementation. A paraître dans *Statistics and Probability Letters*, 1992.
- [RCD 92b] ROBERT, C. P., CELEUX, G. et DIEBOLT J. : Hidden Markov chains: A stochastic implementation of two estimation schemes. Rapport technique LSTA 161, Université Paris 6, 1992.
- [WET 90] WEI, G.C.G. et TANNER, M. : A Monte Carlo implementation of the EM algorithm and the poor man's data augmentation algorithms. *J. American Statist. Assoc.* **85**, 699-704, 1990.

