Vitesse de convergence de l'a posteriori pour les modèles non paramétriques de Markov cachés à espace d'état fini

Elodie Vernet [1]

[1] *Laboratoire de Mathématiques d'Orsay, Université Paris Sud, elodie.vernet@math.u-psud.fr*

Résumé. Les modèles de Markov cachés (HMMs) sont très utilisés en pratique, comme en génomique, reconnaissance de parole ou économétrie. Comme la modélisation paramétrique des densités d'émission peut conduire à de mauvais résultats en pratique, un récent intérêt pour les modèles de Markov cachés non paramétriques est apparu dans les applications. Or ces modèles ont peu été étudiés en théorie. Je présenterai des résultats asymptotiques sur les modèles bayésiens non paramétriques de Markov cachés à espace d'états fini. Je donnerai des hypothèses garantissant l'obtention de vitesses de convergence. Je finirai par exhiber des vitesses obtenues pour des a priori usuels.

Mots-clés. Statistique bayésienne non-paramétrique, chaine de Markov cachée, vitesse de convergence de l'a posteriori.

Abstract. Hidden Markov models (HMMs) have been widely used in diverse fields such as speech recognition, genomics, econometrics. Because parametric modeling of emission distributions may lead to poor results in practice, in particular for clustering purposes, recent interest in using non-parametric HMMs appeared in applications. Yet little thoughts have been given to theory in this framework. We present asymptotic results on Bayesian hidden Markov models with finite state space. Assumptions to obtain posterior rate of convergence will be given. Rates of convergence for usual priors will be exhibited.

Keywords. Bayesian non-parametrics, hidden Markov models, posterior rates of convergence.

1 Introduction

Les modèles de Markov cachés (HMMs) sont très utilisés en pratique, comme en génomique, reconnaissance de parole ou économétrie depuis leur introduction dans [1]. Le livre [2] fournit différents exemples d'application des HMMs. Les HMMs sont des processus stochastiques $(X_t, Y_t)_{t\in\mathbb{N}}$ tels que $(X_t)_{t\in\mathbb{N}}$ est une chaine de Markov et conditionnellement à $(X_t)_{t\in\mathbb{N}}$, les variables aléatoires Y_t, $t \in \mathbb{N}$, sont indépendantes, la loi de Y_t ne dépendant que de X_t. Le terme « modèle de Markov caché »(hidden Markov model en anglais) vient du fait que les observations sont seulement constitués des Y_t, les états X_t de la chaine de Markov sont eux cachés, on n'y a pas accès.

Les propriétés asymptotiques des estimateurs fréquentistes des paramètres des HMMs sont étudiées depuis les années 90. Les consistance et normalité asymptotique du maximum de vraisemblance ont été établis dans le cas paramétrique par, entre autres, [4], [5] et [6] pour le résultat de consistance le plus général jusqu'ici. Les résultats asymptotiques concernant les modèles bayésiens paramétriques sont plus récents. [3] se sont intéressés au cas où le nombre d'états cachés est connu et [10] au cas où celui-ci est inconnu. Comme la modélisation paramétrique des lois d'émission (c'est-à-dire de la loi de Y_t sachant X_t) peut aboutir à de mauvais résultats en pratique, en particulier dans le but de classifier ; un intérêt grandissant pour les HMMs non paramétriques est apparu dans les applications, comme dans [15]. Dans le cadre des HMMs non paramétriques, des résultats théoriques sur les procédures d'estimation ne sont apparus que très récemment comme dans [8] et [7] car l'indentifiabilité restait jusque très récemment [9] un problème ouvert.

Après avoir étudié la consistance de l'a posteriori dans le cadre des modèles non paramétriques de Markov cachés à espace d'état fini [14], nous nous intéressons, à la vitesse de convergence de cet a posteriori. Dans la Partie 2, nous effectuons quelques rappels sur les comportements asymptotiques de l'a posteriori. Nous précisons le modèle étudié dans la Partie 3. Et nous finissons par présenter des hypothèses qui permettent d'obtenir des vitesses a posteriori dans la Partie 4.

2 Comportement asymptotique en statistiques bayésiennes

On se place par la suite dans le cadre des modèles dominés, on considère un ensemble de lois $\{p_n^\theta \lambda^{\otimes n}, \theta \in \Theta\}$ dominées par une mesure $\lambda^{\otimes n}$ et paramétrées par $\theta \in \Theta$. En statistiques bayésiennes, on considère un a priori π, c'est-à-dire une mesure de probabilité sur les paramètres $\theta \in \Theta$ et on s'intéresse à l'a posteriori $\pi(\cdot|Y_1, \ldots, Y_n)$, c'est-à-dire la loi des paramètres sachant les observations $Y_1, \ldots, Y_n$

$$\pi(A|Y_1, \ldots, Y_n) = \frac{\int_A p_n^\theta(Y_1, \ldots, Y_n)\pi(d\theta)}{\int_\Theta p_n^\theta(Y_1, \ldots, Y_n)\pi(d\theta)}, \quad A \subset \Theta.$$

Une question légitime est l'impact du choix de l'a priori sur l'a posteriori, notamment quand le nombre d'observations augmente : l'a priori joue-t-il un rôle important quand le nombre d'observations augmentent ou disparait-il au profit des observations et à quelle vitesse ? Étudier le comportement asymptotique de l'a posteriori implique de prendre un point de vue fréquentiste en supposant que les observations proviennent d'un vrai paramètre θ^*, i.e. $Y_1, \ldots, Y_n$ sont distribués selon $p_n^{\theta^*} \lambda^{\otimes n}$ puis on se demande si l'a posteriori se concentre autour du vrai paramètre θ^* et à quelle vitesse. Formellement, on dit que l' a posteriori $\pi(\cdot|Y_1, \ldots, Y_n)$ converge à vitesse ϵ_n, avec ϵ_n tendant vers 0, pour une distance d, si

$$\pi(\{\theta : d(\theta, \theta^*) > \epsilon_n\}|Y_1, \ldots, Y_n)$$

tend vers 0 en P^{θ^*}-probabilité.

Une méthode générale et usuelle pour déterminer la vitesse de convergence d'un a posteriori a été proposée par [11]. Elle repose sur deux étapes,

1. le contrôle de la masse a priori d'un certain voisinage de θ^*,
2. l'existence de tests.

Cette méthode permet par exemple d'obtenir des vitesses adaptatives pour l'estimation de densité avec des observations i.i.d., pour des classes de densités Höldériennes dans [13] et [12].

Dans le cadre des modèles de Markov cachés, on se propose d'utiliser les tests construits dans [10] pour satisfaire l'Étape 2. Notre but est de contrôler l'Étape 1 sous des hypothèses qui pourront être vérifiées en utilisant des techniques permettant l'obtention de vitesse dans le cadre de l'estimation de densité avec des observations i.i.d..

3 Le modèle étudié et quelques notations

Le modèle étudié est précisé ici et peut être visualisé sur la Figure 1. Les HMMs à espace d'état fini sont des processus stochastiques $(X_t, Y_t)_{t\in\mathbb{N}}$ tels que $(X_t)_{t\in\mathbb{N}}$ est une chaine de Markov prenant ses valeurs dans un espace d'état fini. On suppose, par la suite, que le nombre k d'états cachés est connu, ainsi l'espace d'états de la chaine de Markov est l'ensemble $\{1, \dots, k\}$.

On note $\Delta_k(\underline{q}) = \{(x_1, \dots, x_k) \ : \ x_i > \underline{q},\ i = 1, \dots k \ ; \ \sum_{i=1}^k x_i = 1\}$ une restriction du simplex de dimension $k-1$ et Q la matrice de transition $k \times k$ de la chaine de Markov ; en identifiant Q comme le k-uplet de lois de distributions (les lignes de la matrice), on écrit $Q \in \Delta_k^k(\underline{q})$. L'appartenance de Q à $\Delta_k^k(\underline{q})$ implique l'existence et l'unicité d'une mesure stationnaire associée. On suppose que l'ensemble des observations est $\mathbb{R}^d$ muni de la tribu Borélienne. On note $\mathcal{F}$ l'ensemble des densités par rapport à une mesure de référence λ sur $\mathbb{R}^d$. $\mathcal{F}^k$ est l'ensemble des densités d'émission possibles, c'est-à-dire pour $f = (f_1, \dots, f_k) \in \mathcal{F}^k$, la loi de Y_t sachant $X_t = i$ est $f_i\lambda$, $i = 1, \dots, k$. L'ensemble des

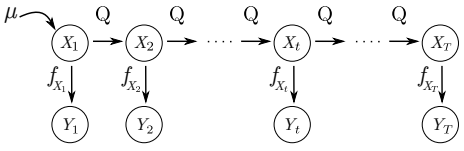

FIGURE 1 – The model

paramètres du modèle est $\Theta = \{\theta = (Q, f) \ : \ Q \in \Delta_k^k(\underline{q}), f \in \mathcal{F}^k\}$. Ainsi $\mathbb{P}^\theta$ est la loi stationnaire de $(X_t, Y_t)_{t\in\mathbb{N}}$ sous θ et p_l^θ est la densité de probabilité de $Y_1, \dots, Y_l$ par

rapport à $\lambda^{\otimes l}$ sous $\mathbb{P}^\theta$. Ainsi pour tout $\theta \in \Theta$ associé à la mesure stationnaire μ :

$$p_l^\theta(y_1, \ldots, y_l) = \sum_{x_1,\ldots,x_l=1}^{k} \mu_{x_1} Q_{x_1,x_2} \ldots Q_{x_{l-1},x_l} f_{x_1}(y_1) \ldots f_{x_l}(y_l).$$

Dans le cadre d'une approche bayésienne, on suppose que l'a priori π est une mesure de probabilité produit sur Θ,

$$\pi = \pi_Q \otimes \pi_f$$

telle que π_Q est une mesure de probabilité sur $\Delta_k^k(\underline{q})$ et π_f est une mesure de probabilité sur $\mathcal{F}^k$. Les observations sont maintenant distribuées selon la mesure stationnaire $\mathbb{P}^{\theta^*}$.

Pour déterminer la vitesse de convergence de l'a posteriori, on doit choisir une distance sur l'ensemble des paramètres Θ. On choisit d'étudier la vitesse de convergence pour le problème de l'estimation de la densité. On choisit donc la distance d_l qui est la distance L_1 sur les densités jointes p_l^θ :

$$d_l(\theta, \tilde{\theta}) = \|p_l^\theta - p_l^{\tilde{\theta}}\|_{L_1}.$$

On note $N(\epsilon, A, d_l)$ le nombre minimal de boules de rayon ϵ pour la distance d_l pour couvrir l'ensemble A.

4 Vitesse de convergence de l'a posteriori pour les HMMs

Le Théorème suivant permet d'obtenir des vitesses de convergence a posteriori dans le cadre des HMMs.

Theorem 4.1 *Soit une suite $\tilde{\epsilon_n} > 0$ tendant vers 0 telle que $n\tilde{\epsilon}_n^2 \to +\infty$ et $C_n > 0$,*

(a) s'il existe une suite B_n de sous-ensembles de $\Delta_k^k(\underline{q}) \times \mathcal{F}^k$ telle que

$$\pi(B_n) \gtrsim \exp(-C_n n\tilde{\epsilon}_n^2)$$

et telle que pour tout $\theta \in B_n$,

$$\|Q - Q^*\| \leq \tilde{\epsilon}_n^{2/(2-\alpha)}$$

et f appartient à un certain voisinage $V_{\tilde{\epsilon}_n^{4/(2-\alpha)}}(f^*)$ de f^* de rayon $\tilde{\epsilon}_n^{4/(2-\alpha)}$,

(b) s'il existe une suite $(\mathcal{F}_n)_{n\geq 1}$ de sous ensembles de $\Delta_k^k(\underline{q}) \times \mathcal{F}^k$ telle que

$$\pi(\mathcal{F}_n^c) = o(\exp(-n\tilde{\epsilon_n}^2(1 + C_n)))$$

(c) et s'il existe une suite $\epsilon_n \geq \tilde{\epsilon_n}$ *tendant vers* 0 *telle que* $n\tilde{\epsilon_n}^2(1+C_n)/(n\epsilon_n^2)$ *tend vers* 0 *et*

$$N\left(\frac{\epsilon_n}{12}, \mathcal{F}_n, d_l\right) \leq \exp\left(\frac{n\epsilon_n^2(k\min_{1\leq i;j\leq k} Q^*_{i,j})^2}{16l(2-k\min_{1\leq i,j\leq k} Q^*_{i,j})^2}\right)$$

Alors il existe une constante C *dépendant de* k *et* $\underline{q}$ *telle que*

$$\pi\left(\theta : d_l(\theta,\theta^*) \geq C\epsilon_n | Y_{1:n}\right) = o_{\mathbb{P}^{\theta^*}}(1) \tag{1}$$

L'hypothèse (a) permet de contrôler l'Étape 1. et les hypothèses (b) et (c) permettent de contrôler l'Étape 2. On a volontairement traduit le voisinage de θ^* à contrôler dans l'Étape 1. en un produit d'ensembles constitué d'un voisinage de Q^* et d'un voisinage de $(f_i^*)_{1,\dots k}$ qui ressemble aux voisinages qu'il est suffisant de contrôler pour obtenir des vitesses de convergence dans le cadre de l'estimation de densité avec des observations i.i.d..

On peut ainsi appliquer ce Théorème a des a priori dont on connait la vitesse dans le cadre de l'estimation de densité avec des observations i.i.d.. Ce théorème permet notamment d'obtenir une vitesse $\epsilon_n = n^{-\frac{\beta(2-\delta)/2}{(2-\delta)\beta+1}}$ pour tout $\delta > 0$ pour les classes de densité d'émission β-Höldérienne pour un a priori sur les densités d'émission qui est un produit de mélange de gaussiennes par processus de Dirichlet. Cette vitesse est optimale à δ près.

Références

[1] Leonard E. Baum and Ted Petrie. Statistical inference for probabilistic functions of finite state markov chains. *The Annals of Mathematical Statistics*, 37(6) :1554–1563, 1966.

[2] O. Cappé, E. Moulines, and T. Rydén. *Inference in Hidden Markov Models*. Springer, 2005.

[3] M. C. de Gunst and O. Shcherbakova. Asymptotic behavior of bayes estimators for hidden markov models with application to ion channels. *Mathematical Methods of Statistics*, 17(4) :342–356, 2008.

[4] R. Douc and C. Matias. Asymptotics of the maximum likelihood estimator for general hidden markov models. *Bernoulli*, 7 :381–420, 2001.

[5] R. Douc, E. Moulines, and T. Rydén. Asymptotic properties of the maximum likelihood estimator in autoregressive models with markov regime. *The Annals of statistics*, 32(5) :2254–2304, 2004.

[6] Randal Douc, Eric Moulines, Jimmy Olsson, and Ramon van Handel. Consistency of the maximum likelihood estimator for general hidden Markov models. *The Annals of Statistics*, 39(1) :474–513, 2011.

[7] Thierry Dumont and Sylvain Le Corff. Nonparametric regression on hidden phi-mixing variables : identifiability and consistency of a pseudo-likelihood based estimation procedure. *arxiv preprint arXiv :1209.0633*, 2012.

[8] E. Gassiat and J. Rousseau. Non parametric finite translation hidden markov models and extensions. *Bernoulli*, 2013. to appear.

[9] Elisabeth Gassiat, Alice Cleynen, and Stéphane Robin. Finite state space non parametric hidden markov models are in general identifiable. *arXiv preprint arXiv :1306.4657*, 2013.

[10] Elisabeth Gassiat and Judith Rousseau. About the posterior distribution in hidden Markov models with unknown number of states. *Bernoulli*, 20(4) :2039–2075, 2014.

[11] Subhashis Ghosal, Jayanta K Ghosh, and Aad W Van Der Vaart. Convergence rates of posterior distributions. *Annals of Statistics*, 28(2) :500–531, 2000.

[12] Willem Kruijer, Judith Rousseau, Aad Van Der Vaart, et al. Adaptive bayesian density estimation with location-scale mixtures. *Electronic Journal of Statistics*, 4 :1225–1257, 2010.

[13] Weining Shen, Surya T Tokdar, and Subhashis Ghosal. Adaptive bayesian multivariate density estimation with dirichlet mixtures. *Biometrika*, 100(3) :623–640, 2013.

[14] Elodie Vernet. Posterior consistency for nonparametric hidden markov models with finite state space. *arXiv preprint arXiv :1311.3092*, 2013.

[15] C. Yau, O. Papaspiliopoulos, G.O. Roberts, and C. Holmes. Bayesian non-parametric hidden markov models with applications in genomics. *Journal of the Royal Statistical Society*, 73 :37–57, 2011.