Estimation de l'ordre d'une chaîne de Markov cachée à émissions de la famille exponentielle

Cécile Low-Kam, André Mas

Estimation de l'ordre d'une chaîne de Markov cachée à émissions de la famille exponentielle

Cécile Low-Kam & André Mas

Institut de Mathématiques et de Modélisation de Montpellier
Université Montpellier 2 - CNRS
c.c. 051, Place Eugène Bataillon
34095 Montpellier Cedex, France

RESUME

Nous cherchons à estimer l'ordre, *i.e.* le nombre d'états cachés, d'un modèle de Markov caché (HMM), quand aucune borne supérieure sur cet ordre n'est connue. Nous nous intéressons aux HMM dont la distribution des états observables appartient à la famille exponentielle. Deux estimateurs pour l'ordre sont présentés : l'un est basé sur l'estimateur du maximum de vraisemblance, et l'autre sur un mélange bayésien. Tous deux sont pénalisés. Nous prouvons la consistance de ces estimateurs, et détaillons les pénalités pour des distributions de la famille exponentielle.

Mots-clé : modèle de Markov caché, famille exponentielle, maximum de vraisemblance, mélange bayésien, critère de choix de modèle.

ABSTRACT

We consider the problem of estimating the order, *i.e.* the number of hidden states, of a hidden markov model (HMM), when no prior bound on this order is known. We investigate the case of a HMM which observable states distribution lies in the exponential family. Two estimators for its order are provided: one is based on the maximum likelihood estimator, and the other on a bayesian mixture estimator. Both are penalized. A proof of the consistency of these estimators is presented, and penalties for distributions from the exponential family are detailed.

Key-words: hidden markov model, exponential family, maximum likelihood, bayesian mixture, model selection criterion.

1 Introduction

Les modèles de Markov cachés (HMM), introduits par Baum et Petrie (1966), ont d'abord été utilisés pour la reconnaissance de la parole, avant d'être appliqués à des domaines aussi variés que la bioinformatique ou l'économétrie.

Formellement, un HMM peut être décrit de la façon suivante : soit $\{X_n\}_{n\geq 1}$ une suite de variables aléatoires à valeurs dans un espace mesurable $(\mathcal{X}, \mathcal{A}, \mu)$. Le processus $\{X_n\}_{n\geq 1}$ est appelé *processus d'émission*, et notons que $\mathcal{X}$, l'alphabet d'émission, peut

être infini. Soit $\{Z_n\}_{n\geq 1}$ une chaîne de Markov à valeurs dans $\mathcal{Z} = \{1, ..., k\}$ telle que, conditionellement à $Z_1^n = (Z_1, ..., Z_n)$, les variables $X_1, ..., X_n$ sont indépendantes et la distribution de chaque X_i dépend seulement de Z_i pour $1 \leq i \leq n$. La cardinalité k de $\mathcal{Z}$ est appelée *l'ordre* du HMM.

Supposons maintenant que l'ordre k est inconnu, et que nous voulons l'estimer. Une procédure d'estimation de l'ordre consiste à définir une suite d'estimateurs $\hat{k}_1, ..., \hat{k}_n$, chacun associé à une séquence d'observations de longueur $1, ..., n$, tels que cette suite converge vers k presque sûrement. Nous nous intéressons à deux tels estimateurs pour k : Finesso (1991) et Kieffer (1993) proposent un estimateur de maximum de vraisemblance pénalisé, alors que Liu et Narayan (1994) introduisent un estimateur de mélange bayésien pénalisé. Lorsque k n'est pas borné *a priori*, la consistance de deux tels estimateurs est prouvée par Gassiat et Boucheron (2003) lorsque l'alphabet d'émission est fini, et par Chambaz *et al.* (2008) pour des cas particuliers d'alphabets d'émissions infinis (Poisson et Gaussien). Nous étendons cette dernière approche aux HMM à émissions de la famille exponentielle.

2 Définitions

Les notations présentées dans cette section sont analogues à celles utilisées par Chambaz *et al.* (2008). Pour $k \geq 1$, soit $\{p_j^0 : 1 \leq j \leq k\} \in \mathbb{R}_+^k$ un ensemble de réels tels que $\sum_{j=1}^k p_j^0 = 1$. Soit $\mathcal{S}_k$ un ensemble de matrices $p = (p_{jj'})_{1\leq j,j'\leq k} \in \mathbb{R}_+^{k^2}$ telles que, pour tout $j \leq k$, $\sum_{j'=1}^k p_{jj'} = 1$. Soit $\{Z_n\}_{n\geq 1}$ une chaîne de Markov à valeurs dans $\{1, ..., k\}$, de distribution initiale $P_\delta\{Z_0 = j\} = p_j^0$ pour $1 \leq j \leq k$, et de probabilités de transition

$$P_\delta\{Z_{i+1} = j' | Z_i = j\} \quad = \quad p_{jj'}, \text{pour tout } j, j' \leq k. \tag{1}$$

Soit $\Delta_k = \left\{\delta = (p, \theta) : p \in \mathcal{S}_k, \theta = (\theta_1, ..., \theta_k) \in \left(\mathbb{R}^d\right)^k\right\}$ un espace de paramètres. Pour $k \geq 1$, pour $\delta \in \Delta_k$, soit $\{(X_n, Z_n)\}_{n\geq 1}$ un HMM à valeurs dans $\mathcal{X}$, tel que la distribution de chaque X_i est de la famille exponentielle de densité

$$\phi_{\theta_{Z_i}}(x) \quad = \quad h(x) \exp\left(\theta'_{Z_i} T(x) - A(\theta_{Z_i})\right), \tag{2}$$

où θ' est la transposée du vecteur θ, $T : \mathcal{X} \to \mathbb{R}^d$, et $A : \mathbb{R}^d \to \mathbb{R}$. L'espace naturel des paramètres est $\Theta = \{\theta : A(\theta) < \infty\} \subset \mathbb{R}^d$. La statistique $T(x) = (T_1(x), ..., T_d(x))$ est une statistique exhaustive de la distribution. On note g_δ la densité de $X_1^n = (X_1, ..., X_n)$ sous δ.

Pour tout $k \geq 1$, soit ν_k le prior sur Δ_k tel que, pour $\alpha = (\alpha_1, ..., \alpha_d) \in \mathbb{R}^d$, et $\beta \in \mathbb{R}$, sous ν_k:

- p et θ sont indépendants,

- $p_j^0 = 1/k$ pour tout $1 \leq j \leq k$,
- les vecteurs $\{p_{jj'} : j' \leq k\}$ (pour $1 \leq j \leq k$) dont indépendants, de distribution de Dirichlet $(1/2, ..., 1/2)$,
- $\theta_1, ..., \theta_k$ sont indépendants, identiquement distribués de densité $\phi_{\alpha,\beta}$:

$$\phi_{\alpha,\beta}(\theta) = \exp(\alpha'\theta - \beta A(\theta) - B(\alpha,\beta)), \quad (3)$$

où $B(\alpha,\beta) = \log \int_\Theta \exp(\alpha'\theta - \beta A(\theta))\, d\theta$ est une constante de normalisation.

Nous définissons le mélange suivant :

$$q_k(X_1^n) = \int_{\Delta_k} g_\delta(X_1^n)\, d\nu_k(\delta). \quad (4)$$

3 Résultats

Nous obtenons une inégalité entre le maximum de vraisemblance et le mélange défini ci-dessus. Pour $k, n \geq 1$, soit

$$c_{kn} = \log k - k \log \frac{\Gamma(k/2)}{\Gamma(1/2)} + k\frac{3k(k+1)+1}{12n}, \quad (5)$$

et

$$d_{kn} = kB(\alpha,\beta) - \frac{kd}{2}\log k - \frac{kd}{2}\log 2\pi. \quad (6)$$

Soit $z_0^n = (z_0, ..., z_n) \in \{1, ..., k\}^{n+1}$ des valeurs non-observables, et $x_1^n = (x_1, ..., x_n) \in \mathcal{X}^n$ une suite d'observations correpondantes. Soit aussi

$$n_j = \sum_{i=1}^n \mathbb{I}_{\{z_i=j\}},\ I_j = \{i \leq n : z_i = j\},\ \overline{T}_j = n_j^{-1} \sum_{i \in I_j} T(x_i), \text{et } \tilde{T}_j = \frac{n_j \overline{T}_j + \alpha}{n_j + \beta}. \quad (7)$$

Théorème 1 *Pour $k, n \geq 1$,*

$$0 \leq \sup_{\delta \in \Delta_k} \log g_\delta(X_1^n) - \log q_k(X_1^n) \leq \frac{k}{2}(k+d-1)\log n + c_{kn} + d_{kn} + f_k(X_1, ..., X_n) + O(n^{-1}), \quad (8)$$

où

$$f_k(X_1, ..., X_n) = \sup_{\delta \in \Delta_k} \max_{z_0^n \in \{1,...k\}^{n+1}} \sum_{j=1}^k \left\{ \frac{-1}{2} \log\det \frac{n_j}{n_j+\beta} \ddot{A}^{-1}(\dot{A}^{-1}(\tilde{T}_j)) - \alpha' \dot{A}^{-1}(\tilde{T}_j) + \beta A\left(\dot{A}^{-1}(\tilde{T}_j)\right)\right\} \quad (9)$$

4 Application

Nous considérons les deux estimateurs suivants pour l'ordre k:

$$\hat{k}_{ML} \quad = \quad \arg\min_{k\geq 1}\{-\sup_{\delta\in\Delta_k}\log g_\delta(X_1^n) + pen(n,k)\} \tag{10}$$

et

$$\hat{k}_{mix} \quad = \quad \arg\min_{k\geq 1}\{-\log q_k(X_1^n) + pen(n,k)\} \tag{11}$$

où $pen(n,k) = o(n)$ est une fonction positive croissante de n et k. Nous montrons que pour un choix adéquat de la pénalité $pen(n,k)$, ces estimateurs sont consistants.

Soit $b > 2$. Supposons que la famille de mélanges d'au plus k élements de $\{\phi_\theta(x), \theta \in \Theta\}$ est identifiable, et qu'il existe $a > 1$ tel que :

$$P(f_k(X_1, ..., X_n) \geq k\varphi(n)) \quad \leq \quad n^{-a}, \tag{12}$$

où $\varphi(n) = o(n)$ est une fonction positive croissante de n, qui dépend de la forme des queues de distribution des états observables. Nous avons alors le résultat suivant :

Théorème 2 *$\hat{k}_{ML}$ est consistant si*

$$pen(n,k) \quad = \quad \sum_{\ell=1}^{k}\left(\frac{\ell(\ell+d-1)+b}{2}\log n + c_{\ell n} + d_{\ell n}\right) + k(k+1)\varphi(n), \tag{13}$$

et $\hat{k}_{mix}$ est consistant si

$$pen(n,k) \quad = \quad \sum_{\ell=1}^{k-1}\frac{\ell(\ell+d-1))+b}{2}\log n + k(k+1)\varphi(n). \tag{14}$$

Lors de l'exposé, nous comparerons ces estimateurs avec ceux obtenus par Chambaz *et al.* (2008) pour les émissions gaussiennes et poissoniennes, et nous détaillerons également nos résultats pour d'autres distributions de la famille exponentielle.

Bibliographie

[1] Baum, L. et Petrie, T. (1966) Statistical Inference for Probabilistic Functions of Finite State Markov Chains, *The Annals of Mathematical Statistics*, 37, 1554–1563.

[2] Chambaz, A., Garivier, A. et Gassiat, E. (2008) A MDL approach to HMM with Poisson and Gaussian emissions. Applications to order identification, *Journal of Statistical Planning and Inference*, 139, 962–977.

[3] Finesso, L. (1991) Consistent estimation of the order for Markov and hidden Markov chains, *Thèse de doctorat*, University of Maryland at College Park.

[4] Gassiat, E. et Boucheron, S. (2003) Optimal error exponents in hidden Markov models order estimation, *IEEE Transactions on Information Theory*, 49, 964–980.
[5] Kieffer, J. (1993) Strongly consistent code-based identification and order estimation for constrained finite-state model classes, *IEEE Transactions on Information Theory*, 39, 893–902.
[6] Liu, C.C. et Narayan, P. (1994) Order estimation and sequential universal data compression of a hidden Markov source by the method of mixtures, *IEEE Transactions on Information Theory*, 40, 1167–1180.