Mélange de chaînes de Markov d'ordre variable pour l'analyse de séquences

Fabrice Rossi

CEREMADE, CNRS, UMR 7534, Université Paris-Dauphine, PSL University,
Fabrice.Rossi@dauphine.psl.eu

Résumé.

Nous proposons dans cette communication un modèle de mélange de chaînes de Markov d'ordre variable. Ces chaînes de Markov parcimonieuses permettent l'estimation de dépendances longues dans des séries temporelles à valeurs discrètes, sans pour autant nécessiter des observations longues (en comparaison). Elles sont donc particulièrement adaptées pour modéliser de trajectoires de vie et d'autres processus historiques décrits par un nombre relativement restreint de pas de temps, comme cela est fréquent en sciences humaines. Dans ces domaines, une tâche cruciale est l'identification de groupes de trajectoires présentant des comportements similaires. Nous proposons de réaliser cette tâche au moyen d'un modèle de mélange de chaînes d'ordre variable. L'estimation de ces modèles étant par nature non paramétrique, nous utilisons une vraisemblance pénalisée dont la version complète conduit de façon directe à un algorithme EM. Le nombre de composantes du mélange peut être choisi par l'utilisation de la même vraisemblance pénalisée.

Mots-clés. Séries temporelles discrètes, chaînes de Markov d'ordre variable, modèle de mélange, analyse de séquence

Abstract. In this talk, we propose a mixture model based variable length Markov chain. These sparse Markov chains allow for the estimation of long dependencies in discrete-valued time series without requiring long observations (in comparison). They are thus particularly suitable for modeling life trajectories and other historical processes described by a relatively small number of time steps, as is common in the social sciences. In these fields, a crucial task is the identification of groups of trajectories exhibiting similar behaviors. We propose to accomplish this task using a variable length Markov chain mixture model. Since the estimation of these models is inherently non-parametric, we employ a penalized likelihood, the complete version of which leads directly to an EM algorithm. The number of mixture components can be chosen using the same penalized likelihood.

Keywords. Discrete-valued time series, variable length markov chain, mixture model, sequence analysis

1 Introduction

Nous nous intéressons dans cette communication à l'analyse exploratoire de séries temporelles à valeurs discrètes. Cette tâche est, par exemple, particulièrement importante dans

l'analyse de séquences (Liao et al., 2022), une étape fondamentale des recherches sur les trajectoires de vie en sciences sociales. Dans ce contexte, un des objectifs de l'analyse est d'identifier des groupes de trajectoires (donc de séries temporelles à valeurs discrètes) « semblables », qui exhibent le même comportement général.

Historiquement, la première méthode proposée a été celle de l'appariement optimal (*optimal matching*), introduite par Abbott and Forrest (1986), qui s'appuie sur les distances d'édition (insertion, suppression et substitution d'observations pour passer d'une séquence à une autre). Plus généralement, la pratique en sciences sociales est dominée par le calcul de dissimilarités entre séries temporelles (Studer and Ritschard, 2015), mesures qui sont ensuite traitées par une classification hiérarchique ascendante ou par une variante adaptée des *k-means* (Kaufman and Rousseeuw, 1987).

L'utilisation de modèles Markoviens s'est développée plus récemment, notamment au moyen de modèles à ordre variable, les *Variable Length Markov Chain*, VLMC (Bühlmann and Wyner, 1999; Gabadinho and Ritschard, 2016) et des modèles latents, *Hidden Markov Model*, HMM (Bolano and Berchtold, 2016). Le développement de ces approches a naturellement conduit à formuler le problème de classification de trajectoires comme l'estimation d'un mélange de modèles markoviens, notamment dans (Helske and Helske, 2019; Helske et al., 2023).

À notre connaissance, les modèles à ordre variable n'ont jamais été utilisés dans le contexte de la classification. Nous nous proposons donc dans cet article d'étudier les mélanges de VLMC. Nous dérivons en particulier une stratégie d'estimation basée sur une vraisemblance pénalisée optimisée par un algorithme EM.

2 Rappel sur les chaînes de Markov d'ordre variable

Les VLMC ont été proposés par Rissanen (1983) et développés notamment par Bühlmann and Wyner (1999). On peut les voir comme des chaînes de Markov d'ordre supérieur parcimonieuses.

Soit S un ensemble fini et S^∞ l'ensemble des séquences de longueurs arbitraires sur S. Une série temporelle $(X_i)_{i\in\mathbb{Z}}$ indexée par les entiers relatifs et à valeurs dans S est un VLMC s'il existe une fonction l de S^∞ dans $\{0,\dots,l_{\max}\}$ telle que pour tout t et toute séquence $x_{-\infty}^t$

$$\mathbb{P}(X_t = x_t \mid X_{-\infty}^{t-1} = x_{-\infty}^{t-1}) = \mathbb{P}\left(X_t = x_t \mid X_{t-l(x_{-\infty}^t)}^{t-1} = x_{t-l(x_{-\infty}^t)}^{t-1}\right). \tag{1}$$

Dans cette définition on a utilisé la notation suivante : pour toute séquence $(t_i)_i$ indexée sur $\mathbb{N}$ ou $\mathbb{Z}$, t_i^j désigne la séquence $(t_i, t_{i+1}, \dots, t_j)$, avec $i < j$. En particulier $t_{-\infty}^i$ désigne la séquence infinie à gauche.

On voit qu'un VLMC est donc une chaîne de Markov d'ordre $l_{\max}$ parcimonieuse. En effet, sa mémoire, et donc le besoin de définir une probabilité conditionnelle, dépend du *contexte*. La fonction contexte correspondante, c, de S^∞ dans lui-même, associe à un passé quelconque $x_{-\infty}^t$ le passé « important », le contexte, $x_{t-l(x_{-\infty}^t)}^{t-1}$. Au lieu de spécifier $S^{l_{\max}}$ lois

conditionnelles, on se contente d'autant de lois qu'il existe de contextes différents (l'image de S^∞ par c). On découple ainsi $l_{\max}$ du nombre de paramètres du modèle et on évite l'explosion de ce dernier avec la longueur de la mémoire. Il ainsi possible d'observer des dépendances temporelles longues conjointement à des dépendances courtes.

3 Mélange de VLMC

3.1 Mélange de chaînes de Markov

Un modèle VLMC peut être vu comme une chaîne de Markov d'ordre $l_{\max}$ dans laquelle on force les lois conditionnelles de même contexte à être identique. De ce fait, un mélange de VLMC est un mélange de chaînes de Markov (cf par exemple van de Pol and Langeheine (1990)), au moins d'un point de vue superficiel.

Concrètement, on suppose qu'on observe des séries temporelles à valeurs dans l'ensemble fini S, engendrées par un mélange de K VLMC, $\mathcal{M}^1, \ldots, \mathcal{M}^K$. Le mélange est caractérisé par des probabilités *a priori* $\pi_1, \ldots, \pi_K$. À chaque série temporelle $X^i = (X^i_{1 \leq t \leq m_i})$, on associe une variable latente Z^i distribuée selon π (i.e. $\mathbb{P}(Z^i = k \mid \pi) = \pi_k$). Les Z^i sont indépendantes, de même que les X^i. X^i est engendrée par $\mathcal{M}^k$ si $Z^i = k$. On suppose les m_i déterministes.

Étant données N séries temporelles, on cherche à estimer π et les K VLMC, c'est-à-dire pour ces derniers les fonctions de contexte $(c^k)_{1 \leq k \leq K}$ et les probabilités conditionnelles associées.

3.2 Estimation d'un VLMC

L'estimation d'un VLMC à partir d'une série temporelle se fait classiquement à partir de l'algorithme contexte de Rissanen (1983). Il identifie les sous-séquences qui apparaissent dans la série et dont l'interprétation comme un contexte apporte assez d'information au sens d'un test de rapport de vraisemblance (cf par exemple Bühlmann and Wyner (1999) pour des détails). Les probabilités conditionnelles sont estimées au sens du maximum de vraisemblance, mais ce n'est pas le cas de la fonction de contexte elle-même. Comme l'observe par exemple Mächler and Bühlmann (2004), la complexité du modèle (mesurée par son nombre de contextes) dépend fortement du niveau retenu pour le test.

Dans le cadre un mélange de VLMC, il est donc préférable de recourir à une estimation par vraisemblance pénalisée, comme proposé (pour un seul modèle) dans Csiszar and Talata (2006); Garivier (2006); Garivier and Leonardi (2011). Si on note $|\mathcal{M}^k|$ le nombre de contextes du VLMC $\mathcal{M}^k$, une pénalité naturelle est $|\mathcal{M}^k| f(M)$ où M est le nombre total d'observations (ici $M = \sum_{i=1}^N m_i$) et f une fonction positive choisie comme dans Garivier and Leonardi (2011) : $f(m) \to \infty$ et $\frac{f(m)}{m} \to 0$ quand $m \to \infty$. Un exemple classique est la pénalité du BIC, avec $f(m) = \frac{|S|-1}{2} \log m$ (où $|S|$ est le cardinal de l'espace d'états, i.e. le nombre de valeurs discrètes possibles pour les séries observées).

3.3 Algorithme EM

Notons $p(\mathcal{M}^k \mid x^i)$ la vraisemblance du VLMC $\mathcal{M}^k$ pour la série x^i. La log-vraisemblance complète pénalisée du mélange de K VLMC s'écrit alors

$$L_p(\pi,\ \mathcal{M}^1, \ldots, \mathcal{M}^K \mid \mathbf{z}, \mathbf{x}) = \sum_{i=1}^{N} \sum_{k=1}^{K} z_k^i \left(\log \pi_k + \log p(\mathcal{M}^k \mid x^i) \right) - f(M) \sum_{k=1}^{K} |\mathcal{M}^k|,$$

où $\mathbf{z} = (z^1, \ldots, z^N)$, $\mathbf{x} = (x^1, \ldots, x^N)$ et où les z_k^i sont les indicatrices $z_k^i = \mathbb{I}_{z^i = k}$.

Cette formulation conduit de façon directe à un algorithme EM. Dans la phase E de l'algorithme, on calcule

$$\mathbb{E}_{\mathbf{Z} \sim q} \left\{ L_p(\pi,\ \mathcal{M}^1, \ldots, \mathcal{M}^K \mid \mathbf{Z}, \mathbf{x}) \right\} = \mathbb{E}_{\mathbf{Z} \sim q} \left\{ \sum_{i=1}^{N} \sum_{k=1}^{K} z_k^i \left(\log \pi_k + \log p(\mathcal{M}^k \mid x^i) \right) \right\} - f(M) \sum_{k=1}^{K} |\mathcal{M}^k|$$

où q désigne une distribution sur les variables latentes $\mathbf{Z} = (Z^1, \ldots, Z^N)$. Si l'estimation courante des paramètres est $\pi^{(t)}$, $\mathcal{M}^{1(t)}, \ldots, \mathcal{M}^{K(t)}$, la distribution optimale pour les variables latentes est donnée par

$$q(Z^i = k)^{(t)} = \tau_k^{i(t)} = \mathbb{P}(Z^i = k \mid x^i,\ \pi^{(t)},\ \mathcal{M}^{1(t)}, \ldots, \mathcal{M}^{K(t)}),$$

c'est-à-dire

$$\tau_k^{i(t)} = \frac{\pi_k^{(t)} \mathbb{P}(x^i \mid \mathcal{M}^{k(t)})}{\sum_{l=1}^{K} \pi_l^{(t)} \mathbb{P}(x^i \mid \mathcal{M}^{l(t)})}.$$

On a donc

$$\mathbb{E}_{\mathbf{Z} \sim q^{(t)}} \left\{ L_p(\pi,\ \mathcal{M}^1, \ldots, \mathcal{M}^K \mid \mathbf{Z}, \mathbf{x}) \right\} = \sum_{i=1}^{N} \sum_{k=1}^{K} \tau_k^{i(t)} \left(\log \pi_k + \log p(\mathcal{M}^k \mid x^i) \right) - f(M) \sum_{k=1}^{K} |\mathcal{M}^k|$$

Dans la phase M on maximise cette quantité par rapport à π et aux VLMC $\mathcal{M}^1, \ldots, \mathcal{M}^K$. Or, il est clair que les différents VLMC peuvent être optimisés séparément en maximisant pour chaque $\mathcal{M}^k$

$$\sum_{i=1}^{N} \tau_k^{i(t)} \log p(\mathcal{M}^k \mid x^i) - f(M) |\mathcal{M}^k|.$$

En pratique, l'implémentation de la phase M se base donc sur l'algorithme proposé dans Garivier (2006) qui peut être lui-même vu comme une variante de l'algorithme contexte. Deux adaptations sont nécessaires : la prise en compte de plusieurs séries temporelles et l'intégration d'une pondération pour chaque série.

Notons pour conclure que le nombre de composantes sur mélange peut être lui-même choisi à partir de la vraisemblance pénalisée.

Bibliographie

A. Abbott and J. Forrest. Optimal matching methods for historical sequences. *The Journal of Interdisciplinary History*, 16(3) :471–494, 1986. doi : 10.2307/204500.

D. Bolano and A. Berchtold. General framework and model building in the class of hidden mixture transition distribution models. *Computational Statistics & Data Analysis*, 93 : 131–145, 2016. ISSN 0167-9473. doi : https://doi.org/10.1016/j.csda.2014.09.011. URL https://www.sciencedirect.com/science/article/pii/S0167947314002722.

P. Bühlmann and A. J. Wyner. Variable length markov chains. *The Annals of Statistics*, 27 (2) :480–513, April 1999. doi : 10.1214/aos/1018031204.

I. Csiszar and Z. Talata. Context tree estimation for not necessarily finite memory processes, via bic and mdl. *IEEE Transactions on Information Theory*, 52(3) :1007–1016, March 2006. ISSN 1557-9654. doi : 10.1109/TIT.2005.864431.

A. Gabadinho and G. Ritschard. Analyzing state sequences with probabilistic suffix trees : The pst r package. *Journal of Statistical Software*, 72(3) :1–39, 2016. doi : 10.18637/jss.v072.i03. URL https://www.jstatsoft.org/index.php/jss/article/view/v072i03.

A. Garivier. Consistency of the unlimited bic context tree estimator. *IEEE Transactions on Information Theory*, 52(10) :4630–4635, Oct 2006. ISSN 1557-9654. doi : 10.1109/TIT.2006.881742.

A. Garivier and F. Leonardi. Context tree selection : A unifying view. *Stochastic Processes and their Applications*, 121(11) :2488–2506, 2011. ISSN 0304-4149. doi : https://doi.org/10.1016/j.spa.2011.06.012.

S. Helske and J. Helske. Mixture hidden markov models for sequence data : The seqhmm package in r. *Journal of Statistical Software*, 88(3) :1–32, 2019. doi : 10.18637/jss.v088.i03. URL https://www.jstatsoft.org/index.php/jss/article/view/v088i03.

S. Helske, M. Keski-Säntti, J. Kivelä, A. Juutinen, A. Kääriälä, M. Gissler, M. Merikukka, and T. Lallukka. Predicting the stability of early employment with its timing and childhood social and health-related predictors : a mixture markov model approach. *Longitudinal and Life Course Studies*, 14(1) :73 – 104, 2023. doi : 10.1332/175795921X16609201864155.

L. Kaufman and P. J. Rousseeuw. Clustering by means of medoids. In Y. Dodge, editor, *Statistical Data Analysis Based on the L1-Norm and Related Methods*, pages 405–416. North-Holland, 1987.

T. F. Liao, D. Bolano, C. Brzinsky-Fay, B. Cornwell, A. E. Fasang, S. Helske, R. Piccarreta, M. Raab, G. Ritschard, E. Struffolino, and M. Studer. Sequence analysis : Its past, present, and future. *Social Science Research*, 107 :102772, 2022. doi : 10.1016/j.ssresearch.2022.102772.

M. Mächler and P. Bühlmann. Variable length markov chains : Methodology, computing, and software. *Journal of Computational and Graphical Statistics*, 13(2) :435–455, 2004. doi : 10.1198/1061860043524.

J. Rissanen. A universal data compression system. *IEEE Transactions on Information Theory*, 29(5) :656–664, Sep. 1983. ISSN 1557-9654. doi : 10.1109/TIT.1983.1056741.

M. Studer and G. Ritschard. What Matters in Differences Between Life Trajectories : A Comparative Review of Sequence Dissimilarity Measures. *Journal of the Royal Statistical Society Series A : Statistics in Society*, 179(2) :481–511, 07 2015. ISSN 0964-1998. doi : 10.1111/rssa.12125.

F. van de Pol and R. Langeheine. Mixed markov latent class models. *Sociological Methodology*, 20 :213–247, 1990.