EM variationnel pour les modèles de Markov cachés factorisés avec retour des données

Sebastian Le Coz [1] & Nathalie Peyrard[2]

[1] *MIAT - UR875 INRA Toulouse / sebastian.le-coz@inra.fr*
[2] *MIAT - UR875 INRA Toulouse / nathalie.peyrard@inra.fr*

Résumé. L'estimation dans les modèles de Markov cachés (Hidden Markov Model en anglais) est facile grâce à l'algorithme EM, lorsque la variable cachée est de dimension 1, ou faible. Dans le cas de problèmes spatiaux, la variable cachée peut être de très grande dimension et l'estimation exacte n'est plus possible, ne serait-ce que du fait de la taille des matrices de transition à représenter. Lorsque le vecteur des variables cachées est de grande dimension, et que la probabilité de transition se factorise (cadre des Factorial HMM ou des Coupled HMM) une estimation approchée par EM variationnel a été proposée avec succès. Nous considérons ici un autre cadre avec factorisation, celui des FHMM avec retour des données, et nous proposons plusieurs choix de distribution variationnelle pour construire un algorithme VEM, correspondant à différents compromis coût/qualité. Dans un FHMM avec retour des données, chaque variable cachée au temps t dépend non seulement de la même variable cachée au temps $t-1$, mais aussi de toutes les variables observées au temps t. Cette dépendance existe par exemple dans la dynamique spatio-temporelle d'espèces végétales avec survie de la banque de graines. Les différentes instanciations de VEM seront testées sur des exemples jouets inspirés des dynamiques d'espèces adventices dans les cultures.

Mots-clés. HMM multidimensionel, algorithme EM, méthodes variationnelles, dynamiques adventices

Abstract. Hidden Markov Model (HMM) estimation is easy with the EM algorithm, when the hidden variable is of dimension 1, or low. For spatial problems, the dimension of the hidden variable can be too large, and exact estimation is not possible anymore (in particular due to the size of the transition matrix that must be stored). When the dimension of the hidden variables vector is large, and the transition probability has factorisation properties (Factorial HMM or Coupled HMM) a VEM algorithm has been proposed for estimation, with success. We consider here another factored framework : FHMM with Data Feedback (DF-FHMM), and we propose several choices for the variational distribution when building the VEM algorithm, corresponding to different cost/quality trade-offs. In a DF-FHMM, each hidden variable at time t depends not only on the same variable at $t-1$ but also on all observed variables at t. This type of dependance occurs for instance in plant species spatio-temporal dynamics, for species with seed bank survival. The different VEM solutions will be tested on toy examples inspired from weeds dynamics in crop fields.

Keywords. multidimensional HMM, EM algorithm, variational methods, weeds dynamics

1 Introduction

Dans un paysage composé de plusieurs patchs, la dynamique des espèces végétales est pilotée par deux facteurs : la colonisation entre patchs et la capacité des graines à survivre dans la banque de graines. En pratique, l'état de la banque de graine n'est jamais observé car cela serait trop coûteux. Par contre, on peut disposer d'observations de la flore levée dans les différents patchs. La dynamique d'une espèce se modélise donc naturellement dans le cadre des modèles de Markov cachés [8]. Ce cadre a déjà été utilisé dans [4] et dans [2], pour modéliser la dynamique d'espèces adventices dans le cas d'une seule parcelle cultivée. Le cadre des Factorial Hidden Markov Model (FHMM, [6]) ou celui des Coupled-HMM (CHMM, [7]) pourrait permettre d'étendre la modélisation au cas multi patchs. Cependant, une spécificité des dynamiques des plantes à banque de graines est que la variable cachée au temps t (l'état de la banque de graine) dépend de la même variable au temps $t-1$ et de toutes les variables observées (la flore levée) au temps t puisque la flore levée produit des graines qui vont soit coloniser d'autres patchs, soit alimenter la banque de graines locale. On a donc un *retour des données* vers l'état caché.

Nous définissons ici le cadre des HMM factorisés avec retour des données (Data Feedback FHMM), puis nous nous intéressons à la question de l'estimation des paramètres d'un tel modèle. L'algorithme EM [3] ne peut plus être appliqué, à la fois du fait de la taille des matrices de transition à représenter, et de la complexité calculatoire, qui sont exponentielles en le nombre de variables cachées d'un pas de temps (i.e. le nombre de parcelles dans l'exemple adventices). L'algorithme Variational EM [1] est une version approchée du EM classique où, à l'étape E, la distribution conditionnelle des variables cachées est remplacée par une distribution plus simple (dite variationnelle). Il a déjà été appliqué avec succès au cadre FHMM [6]. Du fait du retour des données, l'algorithme VEM pour FHMM ne s'applique pas au DF-FHMM. Nous proposons donc différentes distributions variationnelles adaptées au cadre DF-FHMM, correspondant à différentes hypothèses simplificatrices et différents compromis temps/qualité de l'algorithme VEM associé.

2 Cadre des FHMM avec retour de données

Le modèle de DF-FHMM peut être vu comme un réseau bayésien dynamique [5] à $2n$ dimensions sur les variables ($X = \{X_{i,t}\}_{1\leq i\leq n, 1\leq t\leq T}, Y = \{Y_{i,t}\}_{1\leq i\leq n, 1\leq t\leq T}$), dont seules n dimensions sont observées à chaque pas de temps, ici les variables $Y_t = \{Y_{i,t}\}_{1\leq i\leq n}$. La variable observée $Y_{i,t+1}$ ne dépend que de $X_{i,t}$ (comme dans un HMM classique, après redéfinition des indices des pas de temps). Par contre, la variable $X_{i,t}$ dépend stochasti-

quement de $X_{i,t-1}$ et de $\{Y_{j,t}, \forall j = 1, \ldots, n\}$. Le graphe des indépendances conditionnelles associé à ce modèle est illustré sur la Figure 1.

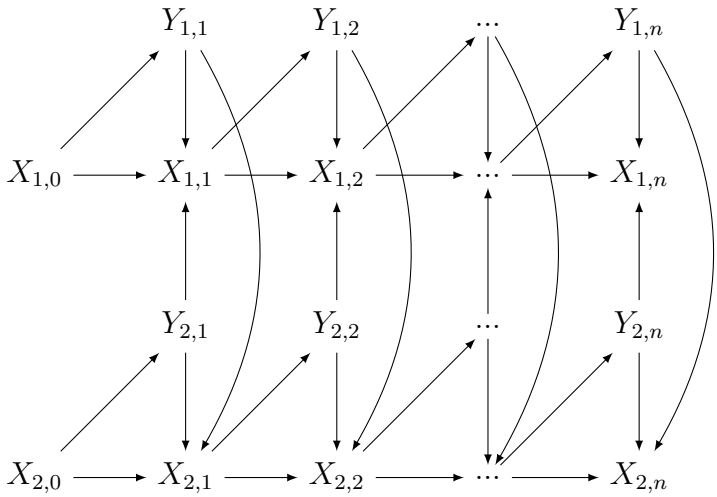

FIGURE 1 – Représentation graphique d'un FHMM avec retour de données, dans le cas de deux chaînes.

3 EM variationnel

Pour estimer les paramètres d'un DF-FHMM avec l'algorithme EM pour HMM, il faudrait dans un premier temps dupliquer les variables observées dans l'état caché, afin de retrouver la structure classique, "en peigne", d'un HMM. Ensuite, au cours de l'étape qui met en œuvre l'algorithme forward-backward, il faudrait stocker la matrice de transition des variables cachées. Dans le cas où toutes les variables sont binaires, cette matrice est de dimension 2^{4n}. Une réécriture des formules du forward-backward exploitant les indépendances conditionnelles du DF-FHMM peut permettre de réduire ce coût de représentation, mais le coût calculatoire reste prohibitif. Pour s'affranchir de cette complexité calculatoire, il est possible d'appliquer une approche variationnelle. L'étape E de l'algorithme EM est alors interprétée comme une étape de maximisation sur l'ensemble des lois des variables cachées conditionnellement aux observations. Le principe de l'algorithme VEM consiste à réduire l'espace de recherche dans l'étape E à un sous-ensemble $\mathcal{Q}$ de distributions plus simples du point de vue de l'inférence. Soit θ l'ensemble des paramètres du modèle et soit $\mathcal{F}$ la fonctionnelle définie comme suit (KL désigne la divergence de Kullback-Leibler) :

$$\mathcal{F}(q, \theta) = \log(P(y|\theta)) - KL(q(X) \mid p(X \mid y, \theta)).$$

L'algorithme VEM itère sur les deux étapes suivantes (jusqu'à ce qu'un critère d'arrêt soit satisfait) :

Etape E : $q^{(k+1)} = \underset{q \in \mathcal{Q}}{\operatorname{argmax}}\ \mathcal{F}(q, \theta^{(k)})$
Etape M : $\theta^{(k+1)} = \underset{\theta}{\operatorname{argmax}}\ \mathcal{F}(q^{(k+1)}, \theta)$

4 Propositions pour la distribution variationnelle

Toute la difficulté de mise en œuvre du VEM réside dans le choix de la famille $\mathcal{Q}$. Nous proposons ici trois solutions correspondant à différents niveaux de simplification de la loi conditionnelles des variables cachées. Les trois familles font l'hypothèse que conditionnellement aux observations la variable cachée au temps t est indépendante de la variable cachée au temps $t-1$.

Les trois familles sont les suivantes :

- $\mathcal{Q}_1$ *famille non stationnaire.* La famille $\mathcal{Q}_1$ inclut toutes les distributions de la forme suivante
$$q(X) = \prod_{t=1}^{T} \prod_{i=1}^{n} q_t(X_{i,t}). \tag{1}$$
- $\mathcal{Q}_2$ *famille stationnaire.* La famille $\mathcal{Q}_2$ correspond aux distributions de la forme (1) pour lesquelles q_t est indépendant de t. Elle est incluse dans $\mathcal{Q}_1$.
- $\mathcal{Q}_3$ *famille non stationnaire à observations interchangeables.* Dans la famille $\mathcal{Q}_3$, on impose que la variable cachée $X_{i,t}$ dépende de manière spécifique de $Y_{i,t}$ et des autres variables observées au temps t : $q_t(X_{i,t} \mid Y) = q(X_{i,t} \mid Y_{i,t}, Y_{j,t}, \forall j \neq i)$. Cela permet d'introduire de la non stationarité, guidée par les données. Pour limiter la taille des distributions variationnelles à calculer, nous rajoutons l'hypothèse que les $n-1$ variables $\{Y_{j,t}\}_{j \neq i}$ (t fixé) sont interchangeables. La famille $\mathcal{Q}_3$ n'est pas incluse dans $\mathcal{Q}_1$, car dans ce cas, q_t dépend également de i.

5 Perspectives

Nous avons établi les équations définissant les algorithmes VEM associés aux trois choix de distribution variationnelle. La suite de ce travail consiste maintenant à mettre en oeuvre ces algorithmes et à comparer leurs comportements sur des données simulées, soit dans un cadre non paramétré des matrices de transition et d'émission, soit dans le cadre d'un modèle paramétrique inspiré de la modélisation de la dynamique d'espèces adventices dans un parcellaire.

Références

[1] Beal M. *Variational algorithm for approximate Bayesian inference*. M.A., M.Sci., Physics, University of Cambridge, UK, 2003.

[2] Borgy B., Reboud X., Peyrard N., Sabbadin R. and Gaba S. *Dynamics of Weeds in the Soil Seed Bank : A Hidden markov Model to Estimate Life History Traits from Standing Plant Time Series*. PLOS ONE, Oct 2015.

[3] Dempster A.P., Laird N.M., Rubin D.B. *Maximum Likelihood from Incomplete Data via the EM Algorithm*. Journal of the Royal Statistical Society, pages 1-38, Vol. 39, No. 1, 1977.

[4] Fréville H., Choquet R., Pradel R. and Cheptou P.-O. *Inferring seed bank from hidden Markov models : new insights into metapopulation dynamics in plants*. Journal of Ecology, pages 1572-1580, 2013.

[5] Ghahramani, Z. *Learning dynamic bayesian networks*. Lecture Notes in Computer Science, 1387 :168–197, 1997.

[6] Ghahramani Z., Jordan M. *Factorial Hidden Markov Models*. Kluwer Academic Publishers, pages 245-273, 1997.

[7] Wainwright M. J., Jordan, M. I. *Graphical Models, Exponential Families, and Variational Inference*. Foundations and Trends in Machine Learning,Vol. 1, No. 1, pages 1-305, 2008.

[8] Rabiner L. *A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition*. Proceedings of the IEEE, pages 257-286, Vol. 77, Issue : 2, 1989.