Non parametric observation driven HMM

Hanna Bacave, Pierre-Olivier Cheptou, Nikolaos Limnios, Nathalie Peyrard

HMM non paramétrique piloté par les observations

Hanna Bacave[1] & Pierre-Olivier Cheptou[2] & Nikolaos Limnios[1] & Nathalie Peyrard[1]

[1] *INRAE, UR MIAT, Université de Toulouse, Castanet-Tolosan, France.* {*hanna.bacave, nathalie.peyrard*}*@inrae.fr*

[2] *CEFE-CNRS, Montpellier, France. pierre-olivier.cheptou@cefe.cnrs.fr*

[3] *Sorbonne University Alliance, Unversité de Technologie de Compiègne, LMAC, France. nikolaos.limnios@utc.fr*

Résumé. Les modèles de Markov cachés (HMM) sont largement utilisés dans de nombreux domaines pour étudier la dynamique d'un processus qui ne peut être observé directement. Cependant, dans certains cas, la structure des dépendances d'un HMM est trop simple pour décrire la dynamique du processus caché. En particulier dans certaines applications en finance et en écologie, les probabilités de transition de la chaîne de Markov cachée peuvent également dépendre de l'observation courante. Dans ce travail, nous nous intéressons à l'extension du HMM classique à cette situation. Nous définissons un nouveau modèle, le modèle de Markov caché piloté par l'observation (Observation-Driven HMM, OD-HMM). Nous présentons ensuite une étude complète du modèle dans le cadre non paramétrique avec des espaces d'états discrets et finis pour les variables cachées et observées. Nous étudions tout d'abord l'identifiabilité du modèle. Puis nous étudions la consistance de l'estimateur du maximum de vraisemblance. Nous établissons ensuite les équations forward-backward associées à l'étape E de l'algorithme EM. La qualité des estimations obtenues est testée sur des jeux de données simulées. Enfin, nous illustrons l'utilisation du modèle sur une application à l'étude de la dynamique des plantes annuelles. Ces travaux posent les bases théoriques et pratiques d'un nouveau cadre qui pourra ensuite être étendu au cas paramétrique, pour simplifier l'estimation, ou au cas semi-markovien pour aller vers plus de réalisme.

Mots-clés. HMM non homogène, identifiabilité, consistance, algorithme EM

Abstract. The hidden Markov models (HMM) are widely used in many different fields, to study the dynamics of a process that cannot be directly observed. However, in some cases, the structure of dependencies of a HMM is too simple to describe the dynamics of the hidden process. In particular, in some applications in finance or in ecology, the transition probabilities of the hidden Markov chain can also depend on the current observation. In this work we are interested in extending the classical HMM to this situation. We define a new model, referred to as the Observation Driven - Hidden Markov Model (OD-HMM). We present a complete study of the general non-parametric OD-HMM with discrete and finite state spaces (hidden and observed variables). We study its identifiability. Then we study the consistency of the maximum likelihood estimators. We derive the associated forward-backward equations for the E-step of the EM algorithm. The quality of the procedure is tested on simulated data sets. Finally, we illustrate the use of the model on an application on

the study of annual plants dynamics. This works sets theoretical and practical foundations for a new framework that could be further extended, on one hand to the non-parametric context to simplify estimation, and on the other hand to the hidden semi-Markov models for more realism.

Keywords. non homogeneous HMM, identifiability, consistency, EM algorithm

1 Introduction

Les modèles de Markov cachés (HMM, Cappé, Moulines et Rydén 2005) sont très utilisés dans différents domaines scientifiques comme, par exemple, en médecine (Le Strat et Carrat 1999) pour analyser des données d'épidémio-surveillance, en écologie (McClintock et al. 2020) pour étudier la dynamique d'un système écologique, ou encore en finance (Engel et Hamilton 1990) pour prédire le régime d'un système monétaire en fonction du taux de change. Leur intérêt vient du fait que les HMM permettent d'étudier la dynamique d'un système qui ne peut pas être directement observé.

Dans certains cas, la structure de dépendance d'un HMM est trop simple pour décrire la dynamique du système caché. En particulier, les probabilités de transition de la chaîne de Markov cachée peuvent également dépendre de l'observation courante. Dans ce travail, nous nous intéressons à l'extension du HMM classique à cette situation. Nous appelons le nouveau modèle Observation Driven HMM (OD-HMM). Les résultats théoriques (Allman, Matias et Rhodes 2009 ; Cappé, Moulines et Rydén 2005) et les algorithmes d'inférence (Cappé, Moulines et Rydén 2005) proposés dans le cas du HMM ne s'appliquent pas directement à l'OD-HMM, principalement parce que dans l'OD-HMM les probabilités de transition ne sont pas constantes en fonction du temps car elles dépendent de l'observation courante. Pour cette nouvelle structure de dépendance, il est donc nécessaire d'étudier les conditions d'identifiabilité du modèle et les propriétés de l'Estimateur du Maximum de Vraisemblance (EMV). De plus, les équations forward - backward de l'algorithme de Baum-Welch (algorithme EM pour HMM) doivent être adaptées.

L'intérêt d'étudier le modèle OD-HMM vient directement de certaines applications, dans lesquelles les observations ont une influence sur les états cachés. Par exemple dans l'étude de la dynamique des plantes il est possible d'observer les plantes sur pied, tandis que les graines enfouies dans le sol ne sont pas visibles. Ainsi, le cadre des HMM semble adapté pour comprendre le rôle de ces graines dans la dynamique. Cependant, le HMM ne permet pas de modéliser l'étape de grenaison, lorsque la plante produit et disperse ses graines dans le sol. L'OD-HMM permet de la prendre en compte. En finance, l'OD-HMM peut être utile comme un extension du modèle Hamilton's Markov-switching (Hamilton 1989) utilisé pour les séries financières qui oscillent entre deux régimes cachés, afin de tenir compte d'une influence des données financières sur ces régimes (Engel et Hamilton 1990).

Il existe déjà des extensions du cadre HMM dans lesquelles les observations influencent la transition entre les états cachés. Ces dernières ont été étudiées soit pour traiter une application particulière (Pluntz et al. 2018 ; Le Coz, Cheptou et Peyrard 2019), soit, de

façon théorique, pour étudier les propriétés de l'EMV (Ailliot et Pène 2015). Dans Ailliot et Pène (2015) la consistance de l'EMV a été étudiée pour des structures plus complexes, incluant le cas de l'OD-HMM mais sans traiter le cas particulier correspondant à l'OD-HMM.

Dans cet article, nous considérons donc un OD-HMM général non paramétrique avec des espaces d'états discrets et finis (variables cachées et observées) et nous présentons une étude complète du modèle. Nous commençons dans la section 2 par définir ce modèle et étudier son identifiabilité. Ensuite, dans la section 3, nous étudions la consistance de l'estimateur du maximum de vraisemblance. Nous détaillons les modifications nécessaires, par rapport au cas HMM, pour obtenir l'algorithme EM associé au OD-HMM dans la section 4. Enfin, dans la section 5, nous validons la procédure d'estimation sur des données simulées selon le modèle OD-HMM, puis dans la section 6 nous l'illustrons sur des données décrivant la dynamique de plantes avec dormance, obtenues par simulation.

2 L'OD-HMM non paramétrique

2.1 Définition du modèle

Considérons deux ensembles de variables aléatoires, indexés par le temps (discret) : Y_t, le système observé au temps t, dont l'espace d'état est $\Omega_Y = \{1, ..., D\}$ et Z_t l'état caché au temps t, dont l'espace d'état est $\Omega_Z = \{1, ..., S\}$. On désigne le vecteur des observations entre $t = 0$ et $t = M$ par $Y_{0:M} = (Y_0, Y_1, Y_2, ..., Y_M)$. De la même manière, le vecteur des états cachés entre $t = 0$ et $t = M$ est noté $Z_{0:M} = (Z_0, Z_1, ..., Z_M)$. Dans la version la plus classique d'un HMM, dont les dépendances sont schématisées dans la Figure 1a, (Z_t) est une chaîne de Markov et $\mathbb{P}(Y_{0:M} = y_{0:M} \mid Z_{0:M} = z_{0:M}) = \prod_{t=0}^{M} \mathbb{P}(Y_t = y_t \mid Z_t = z_t)$ [1].

Nous considérons ici une extension du HMM classique, où l'observation Y_{t-1} a une influence sur la variable cachée suivante Z_t. Cela correspond à la représentation graphique des indépendances conditionnelles illustrée sur la Figure 1b.

La distribution jointe de $(Z_{0:M}, Y_{0:M})$ est entièrement déterminée par les distributions suivantes. La probabilité initiale $\mathbb{P}(Z_0 = z_0)$ est notée $\pi(z_0)$. La probabilité d'émission $\mathbb{P}(Y_t = y_t | Z_t = z_t)$ est notée $R(z_t, y_t)$. Enfin, la matrice de transition $\mathbb{P}(Z_t = z_t | Z_{t-1} = z_{t-1}, Y_{t-1} = y_{t-1})$ est notée $P_{y_{t-1}}(z_{t-1}, z_t)$. En raison de l'influence des observations sur les états cachés, la matrice de transition dépend de l'observation y_{t-1}. Par conséquent, une spécificité de ce modèle est qu'il y a autant de matrices de transition que d'états observés, contrairement au HMM classique.

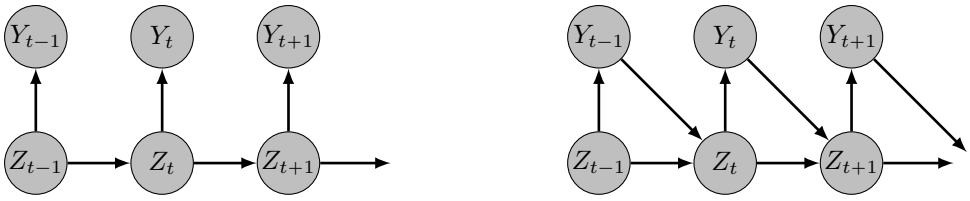
Bien que le modèle soit non-paramétrique, par souci de simplicité, nous appellerons ces distributions les paramètres du modèle. Puisqu'il s'agit de probabilités, avec une contrainte de somme égale à un, l'ensemble des paramètres est

$$\theta = (P_y(z', z), y \in \Omega_Y, z' \in \Omega_Z, z \in \Omega_Z \backslash \{S\}) \cup (R(y, z), y \in \Omega_Y \backslash \{D\}) .$$

Il prend ses valeurs dans $\Theta = [0, 1]^{|\Omega_Z|(|\Omega_Z|-1)|\Omega_Y| + (|\Omega_Y|-1)|\Omega_Z|}$.

1. Par convention, nous utilisons des lettres majuscules, Z_t ou Y_t, pour les variables aléatoires et des lettres minuscules, z_t ou y_t, pour les réalisations.

Definition 1 (OD-HMM). *On dit que (Z_t, Y_t) suit un modèle Observation-Driven HMM (OD-HMM) dès lors que l'indépendance conditionnelle entre les variables observées et les variables cachées sont telles que décrites dans la Figure 1b. Dans le cas non paramétrique, le modèle est défini par θ et il est noté $\mathcal{M}_\theta^{ODHMM}$.*



(a) HMM (b) OD-HMM

FIGURE 1 – Représentation graphique des dépendances conditionnelles dans la chaîne (Z_t, Y_t) : (a) cadre HMM, (b) cadre OD-HMM.

2.2 Identifiabilité du modèle

Afin d'obtenir des résultats sur l'identifiabilité d'un modèle $\mathcal{M}_\theta^{ODHMM}$, nous nous appuyons sur les travaux de Allman, Matias et Rhodes (2009) (théorème 6), dans lesquels les auteurs donnent des conditions suffisantes pour obtenir l'identifiabilité générique d'un HMM.

Definition 2 (Identifiabilité générique). *Soit $\mathcal{F}(\Theta) = \{\mathbb{P}_\theta, \theta \in \Theta\}$ une famille de distributions de probabilité. On dit que les paramètres du modèle θ sont génériquement identifiables si les éléments de Θ qui ne satisfont pas $\mathbb{P}_\theta = \mathbb{P}_{\theta'} \Rightarrow \theta = \theta'$ sont de mesure nulle dans l'espace des paramètres.*

Afin d'appliquer le théorème 6 de Allman, Matias et Rhodes (2009), nous reformulons le modèle $\mathcal{M}_\theta^{ODHMM}$ comme un HMM où les états cachés sont Z et Y et où l'état observé est une copie de Y et nous obtenons la proposition suivante :

Proposition 1 (Identifiabilité générique pour l'OD-HMM). *Les paramètres θ d'un ODHMM, $\mathcal{M}_\theta^{ODHMM}$, avec $|\Omega_Z|$ états cachés et $|\Omega_Y|$ états observables sont génériquement identifiables dès lors qu'on a $2L+1$ variables observées consécutives, pour lesquelles L satisfait la condition suivante :*

$$\binom{L + |\Omega_Y| - 1}{|\Omega_Y| - 1} \geqslant |\Omega_Z||\Omega_Y|.$$

3 Consistance de l'EMV

Dans Ailliot et Pène (2015), les auteurs donnent des conditions suffisantes pour la consistance de l'EMV pour une famille de modèles dits *non homogeneous Markov switching*

models. La structure de dépendance de l'OD-HMM en fait un cas particulier de ces modèles, même si les auteurs ne traitent pas du calcul de l'EMV en pratique pour cette structure. Ils obtiennent des résultats sur la consistance de l'EMV dans le cas où les espaces d'états sont continus et nous les adaptons ici au cas du OD-HMM pour lequel les espaces d'états considérés sont discrets finis.

Nous nous plaçons dans l'espace probabilisé $(\Omega_Z \times \Omega_Y, \mathcal{P}(\Omega_Z \times \Omega_Y), \mathbb{P}_\theta, \theta \in \Theta)$ avec μ comme mesure de référence. Comme la chaîne (Z_t) est non homogène par rapport au temps, car P_y dépend de y, nous considérons la matrice de transition du couple (Z_t, Y_t) notée :

$$\forall (i,a),(j,b) \in \Omega_Z \times \Omega_Y, \tilde{P}_\theta(i,a;j,b) = P_a(i,j;\theta)R(j,b;\theta).$$

La chaîne de Markov du couple (Z_t, Y_t) est homogène, avec une matrice de transition $\tilde{P}_\theta$. Nous supposons que les éléments de $\tilde{P}_\theta$ sont tous strictement positifs. Cette hypothèse nous conduit à la Proposition 2.

Proposition 2. *Il existe une probabilité invariante $\tilde{\pi}_\theta$ de la chaîne de Markov (Z_t, Y_t) pour tout $\theta \in \Theta$.*

La proposition ci-dessus nous permet d'obtenir l'existence de la loi stationnaire. Nous nous plaçons dans le cas où la distribution initiale du processus au temps $t = 0$ est $\tilde{\pi}_\theta$. Ainsi, le processus est stationnaire. Notons $\overline{\mathbb{P}}_\theta^Y$ la marginale de $\tilde{\pi}_\theta$ pour la chaîne observée. La transposition des conditions suffisantes énoncées dans Ailliot et Pène (2015) au modèle OD-HMM nous permet d'obtenir la Proposition 3.

Proposition 3 (Consistance de l'EMV des OD-HMM)**.** *Sous l'hypothèse que les éléments de P_y et R sont continus en θ, pour tout y dans Ω_Y, alors, pour tout $z_0 \in \Omega_Z$, les valeurs limites de l'EMV $\hat{\theta}$ sont $\mathbb{P}_{\theta^*}$-p.s. contenues dans l'espace $\{\theta \in \Theta; \overline{\mathbb{P}}_\theta^Y = \overline{\mathbb{P}}_{\theta^*}^Y\}$, où θ^* est la vraie valeur des paramètres.*

Notons que puisque nous nous sommes placés dans un cas non-paramétrique, l'hypothèse est satisfaite.

4 Algorithme EM

Nous utilisons l'algorithme EM pour calculer l'EMV. Pour prendre en compte la dépendance entre les observations (Y_t) et les états cachés (Z_t), nous proposons une adaptation de l'étape E, dite de forward-backward pour HMM. Supposons que nous avons C réalisations $(y_{c,t})$ de C OD-HMM indépendants et identiquement distribués, où $c \in \{1, ..., C\}$. Dans la suite, on note le vecteur d'état caché au temps t pour la chaîne 1 à la chaîne C, $Z_{1:C,t} = (Z_{1,t}, Z_{2,t}, ..., Z_{C,t})$. De même, on note $Y_{1:C,t} = (Y_{1,t}, Y_{2,t}, ..., Y_{C,t})$ le vecteur des observations au temps t pour les chaînes 1 à C. Enfin, on écrit $\pi(z_{c,0};\theta)$, $P_{y_{c,t-1}}(z_{c,t-1}, z_{c,t};\theta)$ et $R(z_{c,t}, y_{c,t};\theta)$ les probabilités prises conditionnellement à θ.

L'algorithme EM repose sur la quantité intermédiaire suivante, où $\theta^{(m)}$ est la valeur courante du paramètre :

$$\begin{aligned}
Q(\theta|\theta^{(m)}) &= \mathbb{E}\left[\ln \mathbb{P}(Y_{1:C,0:M}, Z_{1:C,0:M}|\theta)|Y_{1:C,0:M} = y_{1:C,0:M}, \theta^{(m)}\right] \\
&= \sum_{c=1}^{C} \sum_{z_0 \in \Omega_Z} \ln\left(\pi(z_0;\theta)\right) \mathbb{P}(Z_{c,0} = z_0|Y_{1:C,0:M} = y_{1:C,0:M}, \theta^{(m)}) \\
&+ \sum_{c=1}^{C} \sum_{t=1}^{M} \sum_{(z,z') \in \Omega_Z^2} \ln\left(P_{y_{c,t-1}}(z,z';\theta)\right) \mathbb{P}(Z_{c,t-1} = z, Z_{c,t} = z'|Y_{1:C,0:M} = y_{1:C,0:M}, \theta^{(m)}) \\
&+ \sum_{c=1}^{C} \sum_{t=0}^{M} \sum_{z' \in \Omega_Z} \ln\left(R(z', y_{c,t};\theta)\right) \mathbb{P}(Z_{c,t} = z'|Y_{1:C,0:M} = y_{1:C,0:M}, \theta^{(m)}).
\end{aligned}$$

Il s'agit d'un algorithme itératif et chaque itération est composée de deux étapes : à l'étape E on calcule les distributions marginales intervenant dans l'expression de la quantité intermédiaire et à l'étape M on met à jour l'ensemble des paramètres en résolvant $\theta^{(m+1)} = \arg\max_{\theta} Q(\theta|\theta^{(m)})$. Nous détaillons ces deux étapes dans les deux sections suivantes.

4.1 Etape E

L'étape E de l'algorithme EM pour OD-HMM est très similaire à celle pour HMM. Elle repose sur l'algorithme Forward-Backward. Cependant, la récursion backward a été modifiée pour prendre en compte le fait que la matrice de transition dépend de l'observation. L'étape E consiste à calculer les probabilités marginales d'intérêt apparues dans l'expression de $Q(\theta|\theta^{(m)})$, qui sont :

— $\forall 0 \leqslant t \leqslant M, \forall c \in \{1, ..., C\}, \forall z_t \in \Omega_Z$,

$$\rho_{c,t}^{(m)}(z_t) = \mathbb{P}(Z_{c,t} = z_t|Y_{1:C,0:M} = y_{1:C,0:M}, \theta^{(m)});$$

— $\forall 1 \leqslant t \leqslant M, \forall c \in \{1, ..., C\}, \forall z_{t-1}, z_t \in \Omega_Z^2$,

$$\xi_{c,t}^{(m)}(z_{t-1}, z_t) = \mathbb{P}(Z_{c,t-1} = z_{t-1}, Z_{c,t} = z_t|Y_{1:C,0:M} = y_{1:C,0:M}, \theta^{(m)}).$$

Pour obtenir $\rho_{c,t}^{(m)}(z_t)$ et $\xi_{c,t}^{(m)}(z_{t-1}, z_t)$, nous introduisons les variables suivantes :

— $\alpha_{c,t}^{(m)}(z_t)$, défini par $\forall 0 \leqslant t \leqslant M, \forall c \in \{1, ..., C\}, \forall z_t \in \Omega_Z$

$$\alpha_{c,t}^{(m)}(z_t) = \mathbb{P}(Y_{c,0:t} = y_{c,0:t}, Z_{c,t} = z_t|\theta^{(m)});$$

— $\beta_{c,t}^{(m)}(z_t)$, défini par $\forall 0 \leqslant t < M, \forall c \in \{1, ..., C\}, \forall z_t \in \Omega_Z$

$$\beta_{c,t}^{(m)}(z_t) = \mathbb{P}(Y_{c,t+1:M} = y_{c,t+1:M}|Z_{c,t} = z_t, Y_{c,t} = y_{c,t}, \theta^{(m)}).$$

La spécificité de l'algorithme Forward-Backward pour OD-HMM par rapport à celui pour HMM réside dans l'expression de $\beta_{c,t}^{(m)}(z_t)$: elle est calculée conditionnellement aux observations courantes. Dans l'algorithme Forward, $\alpha_{c,t}^{(m)}(z_t)$ est calculée grâce à la formule de récurrence suivante :

$$\forall 1 \leqslant t \leqslant M, \forall c \in \{1, ..., C\}, \forall z_t \in \Omega_Z,\ \alpha_{c,t}^{(m)}(z_t) = R^{(m)}(z_t, y_{c,t}) \sum_{z \in \Omega_Z} \alpha_{c,t-1}^{(m)}(z) P_{y_{c,t-1}}^{(m)}(z, z_t),$$

où $\alpha_{c,0}^{(m)}(z_0) = R^{(m)}(z_0, y_{c,0})\pi_(z_0)^{(m)}$. Dans l'algorithme Backward, $\beta_{c,t}^{(m)}(z_t)$ est calculée par la formule de récurrence suivante :

$$\forall 0 \leqslant t < M, \forall c \in \{1, ..., C\}, \forall z_t \in \Omega_Z,\ \beta_{c,t}^{(m)}(z_t) = \sum_{z' \in \Omega_Z} R^{(m)}(z', y_{c,t+1}) \beta_{c,t+1}^{(m)}(z') P_{y_{c,t}}^{(m)}(z_t, z'),$$

où $\beta_{c,M}^{(m)}(z_M) = 1$.
A l'issue de l'algorithme Forward - Backward, on utilise $\alpha_{c,t}^{(m)}(z_t)$ et $\beta_{c,t}^{(m)}(z_t)$ pour calculer $\rho_{c,t}^{(m)}(z_t)$ et $\xi_{c,t}^{(m)}(z_{t-1}, z_t)$, comme suit : $\forall c \in \{1, ..., C\}$,

$$\forall 0 \leqslant t \leqslant M, \forall z_t \in \Omega_Z,\ \rho_{c,t}^{(m)}(z_t) = \frac{\alpha_{c,t}^{(m)}(z_t)\beta_{c,t}^{(m)}(z_t)}{\sum_{z_t \in \Omega_Z} \alpha_{c,t}^{(m)}(z_t)\beta_{c,t}^{(m)}(z_t)};$$

$$\forall 1 \leqslant t \leqslant M, \forall z_{t-1}, z_t \in \Omega_Z^2,\ \xi_{c,t}^{(m)}(z_{t-1}, z_t) = \frac{\alpha_{c,t-1}^{(m)}(z_{t-1}) P_{y_{c,t-1}}^{(m)}(z_{t-1}, z_t) R^{(m)}(z_t, y_{c,t}) \beta_{c,t}^{(m)}(z_t)}{\sum_{z_t \in \Omega_Z} \alpha_{c,t}^{(m)}(z_t)\beta_{c,t}^{(m)}(z_t)}.$$

4.2 Etape M

À l'étape M, on résout le problème de maximisation pour obtenir l'expression des paramètres actualisés en fonction de $\rho_{c,t}^{(m)}(z_t)$ et $\xi_{c,t}^{(m)}(z_{t-1}, z_t)$ comme suit :

$$\forall y \in \Omega_Y, \forall z_t \in \Omega_Z, \forall z_{t-1} \in \Omega_Z, P_y^{(m+1)}(z_{t-1}, z_t) = \frac{\sum_{c=1}^{C} \sum_{t=1}^{M} \xi_{c,t}^{(m)}(z_{t-1}, z_t) \mathbb{1}_{(y_{c,t-1}=y)}}{\sum_{c=1}^{C} \sum_{t=1}^{M} \sum_{z'_t \in \Omega_Z} \xi_{c,t}^{(m)}(z_{t-1}, z'_t) \mathbb{1}_{(y_{c,t-1}=y)}};$$

$$\forall z_t \in \Omega_Z, \forall y \in A, R^{(m+1)}(z_t, y) = \frac{\sum_{c=1}^{C} \sum_{t=0}^{M} \rho_{c,t}^{(m)}(z_t) \mathbb{1}_{(y_{c,t}=y)}}{\sum_{c=1}^{C} \sum_{t=0}^{M} \rho_{c,t}^{(m)}(z_t)}.$$

5 Validation sur des données simulées

Pour évaluer la qualité de la procédure d'estimation, nous avons effectué trois tests avec différentes vraies valeurs des paramètres, en prenant les espaces d'états Ω_Z et Ω_Y de taille 2. Pour un test donné, nous simulons C chaînes $(Z_{c,t})$ et $(Y_{c,t})$ avec une longueur $M = 500$ à partir de la vraie valeur P_y^* et R^* des matrices (voir Table 1), avec pour distribution initiale $\pi = (1, 0)$ et cela pour $C = 10, 50, 100$. L'estimateur θ^{EM} est celui pour lequel la vraisemblance est la plus grande parmi celle obtenu, à l'issue de l'algorithme EM, pour chacune des 10 initialisations aléatoires. L'algorithme est stoppé si l'un des deux critères est vérifié : convergence ou nombre maximal d'itérations (500) atteint. Pour évaluer la convergence de l'algorithme, nous utilisons le critère suivant :

$$dist(\theta_i^{(m)}, \theta_i^{(m+1)}) = \frac{1}{K} \sum_{k=1}^{K} \frac{|\theta_{i,k}^{(m)} - \theta_{i,k}^{(m+1)}|}{\theta_{i,k}^{(m)}}$$

où i est une ligne d'une des matrices de transition ou d'émission, K est la longueur de la ligne θ_i et $\theta_{i,k}$ est le k-ième élément de la i-ième ligne de θ. L'algorithme est considéré comme ayant convergé si la moyenne de ces distances est inférieure à $0,001$. Ensuite la distance entre θ^* et θ^{EM} est calculée de la même manière. Pour un nombre de chaînes C fixé, le test est répété 30 fois.

TABLE 1 – Valeurs des matrices de transition et d'émission pour les 3 tests.

	Vrais paramètres
Test 1	$P_0^* = \begin{pmatrix} 0.2 & 0.8 \\ 0.8 & 0.2 \end{pmatrix}, P_1^* = \begin{pmatrix} 0.8 & 0.2 \\ 0.2 & 0.8 \end{pmatrix}, R^* = \begin{pmatrix} 0.8 & 0.2 \\ 0.2 & 0.8 \end{pmatrix}$
Test 2	$P_0^* = \begin{pmatrix} 0.6 & 0.4 \\ 0.4 & 0.6 \end{pmatrix}, P_1^* = \begin{pmatrix} 0.4 & 0.6 \\ 0.6 & 0.4 \end{pmatrix}, R^* = \begin{pmatrix} 0.2 & 0.8 \\ 0.8 & 0.2 \end{pmatrix}$
Test 3	$P_0^* = \begin{pmatrix} 0.2 & 0.8 \\ 0.8 & 0.2 \end{pmatrix}, P_1^* = \begin{pmatrix} 0.8 & 0.2 \\ 0.2 & 0.8 \end{pmatrix}, R^* = \begin{pmatrix} 0.4 & 0.6 \\ 0.6 & 0.4 \end{pmatrix}$

Dans le premier test, le plus facile, les matrices définissant θ^* sont toutes très contrastées. Dans ce cas, l'estimation est de très bonne qualité. En effet, la distance moyenne entre les paramètres estimés et les vrais est de $0,098$ pour $C = 10$, $0,071$ pour $C = 50$ et enfin $0,068$ pour $C = 100$.

Le second test est un peu plus difficile, car nous avons choisi de prendre des matrices de transitions moins contrastées, c'est-à-dire que les lignes de ces matrices sont proches de $(\frac{1}{2}, \frac{1}{2})$. Ici, la qualité de l'estimation s'est légèrement dégradée. En effet, la distance moyenne entre les paramètres estimés et les vrais est de $0,143$ pour $C = 10$, $0,138$ pour $C = 50$ et enfin $0,127$ pour $C = 100$. On note que la distance moyenne pour les matrices P est supérieure à celle obtenue pour R. Par exemple, dans le meilleur des cas, pour $C = 100$, on a une distance moyenne pour P égale à $0,143$ et une distance moyenne pour R égale à $0,087$.

Enfin, dans le dernier cas, où la matrice R est moins contrastée, dans le sens où ses lignes sont similaires entre elles, on observe une nette dégradation de l'estimation et l'apparition de label-switching (distances moyennes entre les paramètres estimés et les vrais de $0,212$ pour $C = 10$, $0,177$ pour $C = 50$ et enfin $0,150$ pour $C = 100$). Cette dégradation est assez marquée pour l'estimation des matrices de transition, où les distances moyennes sont de $0,270$ pour $C = 10$, $0,235$ pour $C = 50$ et enfin $0,202$ pour $C = 100$. En revanche, elle ne l'est pas pour l'estimation de R dont les distances moyennes sont de $0,097$ pour $C = 10$, $0,06$ pour $C = 50$ et enfin $0,047$ pour $C = 100$.

Ainsi, bien qu'au cours des expérimentations la qualité de l'estimation se dégrade en fonction du contraste présent dans les matrices de vrais paramètres, on observe que l'estimation de R est toujours de très bonne qualité. Pour ce qui est de l'estimation de P, sa qualité se dégrade au cours du temps, mais elle reste toujours satisfaisante.

6 Illustration

Les processus impliqués dans la dynamique d'une plante sont les suivants. La germination, dont la probabilité est notée g, correspond au fait que la graine se développe en une plante. Ensuite, la production de graines, de probabilité d, décrit la capacité de la plante à produire puis à disperser ses graines dans le sol. Au lieu de germer, les graines peuvent survivre dans le sol d'une année sur l'autre, avec une probabilité s. Finalement, la colonisation, dont la probabilité est notée c, représente l'arrivée de graines exogènes, par exemple par le vent. A partir de ces quantités, il est possible de construire les matrices R^*, P_0^* et P_1^* comme suggéré dans Pluntz et al. (2018) :

$$R^* = \begin{pmatrix} 1 & 0 \\ 1-g & g \end{pmatrix} ; \; P_0^* = \begin{pmatrix} 1-c & c \\ (1-c)(1-s) & 1-(1-c)(1-s) \end{pmatrix}$$

$$P_1^* = \begin{pmatrix} (1-c)(1-d) & 1-(1-c)(1-d) \\ (1-c)(1-d)(1-s) & 1-(1-c)(1-d)(1-s) \end{pmatrix}.$$

Une quantité d'intérêt est la durée moyenne de présence de graines en continu dans le sol. Nous illustrons comment la calculer à partir de données pseudo-réelles obtenues par simulation pour des paramètres égaux à ceux estimés pour l'espèce *Alopecurus Myosoroides* par Pluntz et al. (2018) : $g = 0,59$, $s = 0,51$, $c = 0,09$. Nous choisissons $d = 0,5$, car cette quantité n'est pas estimée dans Pluntz et al. (2018). Pour ces données et en utilisant l'algorithme EM pour OD-HMM nous estimons θ^{EM} puis nous générons 100 chaînes Z_t de longueur 500. Le nombre moyen de pas de temps (ici le pas de temps est l'année) de présence consécutive de graines dans le sol est estimé par sa fréquence empirique. Ainsi, il y a des graines en permanence dans le sol pendant 3.26 pas de temps. Toutes les durées calculées oscillent entre 2.6 et 3.9.

7 Conclusion

Dans cet article, nous avons défini l'OD-HMM dans le cas non paramétrique à espaces d'états discrets et finis et nous avons montré comment il permet de modéliser de manière

plus précise qu'avec le HMM classique certaines dynamiques complexes. Pour le calcul de l'estimateur du maximum de vraisemblance par l'algorithme EM, nous avons dû nous restreindre, à des tailles d'espaces d'états de 2 et nous avons étudié des chaînes de longueur $M = 500$, avec un nombre maximal de chaînes égal à 100 afin d'éviter les problèmes numériques. Une première perspective à ces travaux serait de se placer dans un cadre paramétrique pour étudier des systèmes plus importants. Pour aller vers plus de réalisme, nous envisageons également d'étendre l'OD-HMM au cas où la chaîne cachée $(Z_t)_t$ est une chaîne de semi-Markov (Barbu et Limnios 2008). La loi de temps de séjour sera alors définie par une loi quelconque et non par une loi géométrique comme dans le cas particulier du HMM.

Références

Ailliot, P., et F. Pène. 2015. « Consistency of the maximum likelihood estimate for non-homogeneous Markov-switching models ». *ESAIM : PS* 19 : 268-292.

Allman, E., C. Matias et J. Rhodes. 2009. « Identifiability of parameters in latent structure models with many observed variables ». *Annals of Statistics* 37 (6A) : 3099-3132.

Barbu, V-S., et N. Limnios. 2008. *Semi-Markov Chains and Hidden Semi-Markov Models towards Applications.* Springer.

Cappé, O., E. Moulines et T. Rydén. 2005. *Inference in Hidden Markov Models.* Springer.

Engel, C., et J.-D. Hamilton. 1990. « Long Swings in the Dollar : Are They in the Data and Do Markets Know It ? » *The American Economic Review* 80 (4) : 689-713.

Hamilton, J.-D. 1989. « A New Approach to the Economic Analysis of Nonstationary Time Series and the Business Cycle ». *Econometrica* 57 (2) : 357-384.

Le Coz, S., P. O. Cheptou et N. Peyrard. 2019. « A spatial Markovian framework for estimating regional and local dynamics of annual plants with dormancy ». *Theoretical Population Biology* 127 : 120-132.

Le Strat, Y., et F. Carrat. 1999. « Monitoring epidemiologic surveillance data using hidden Markov models ». *Statistics in Medicine* 18 (24) : 3377-3513.

McClintock, B.-T., R. Langrock, O. Gimenez, E. Cam, D.-L. Borchers, R. Glennie et T.-A. Patterson. 2020. « Uncovering ecological state dynamics with hidden Markov models ». *Ecology Letters* 23 (12) : 1878-1903.

Pluntz, M., S. Le Coz, N. Peyrard, R. Pradel, R. Coquet et P. O. Cheptou. 2018. « A general method for estimating seed dormancy and colonisation in annual plants from the observation of existing flora ». *Ecology Letters* 21 : 1311-1318.