Deux extensions du cadre HSMM : cas de plusieurs chaînes cachées factorisées et cas où la dynamique de la chaîne cachée dépend des observations

Nathalie Peyrard[1], avec les collègues du projet ANR HSMM-INCA

[1]MIAT INRAE Toulouse

RMR 2024 - Rouen

Modèle markovien

On se place dans le cas temps discret et variables discrètes.
$(Z_t)_{\{t=1,2,\ldots\}}$ est un modèle markovien si

$$\mathbb{P}(Z_t \mid Z_{t-1}, Z_{t-2}, \ldots Z_1) = \mathbb{P}(Z_t \mid Z_{t-1})$$

- Le modèle est défini par les probabilités de transition

$$R(i,j) = \mathbb{P}(Z_t = j \mid Z_{t-1} = i)$$

- Propriété : la durée de séjour dans un état suit une distribution géométrique

Modèle semi-markovien

C'est le couple (état, durée) = (J_n, X_{n+1}) qui est markovien

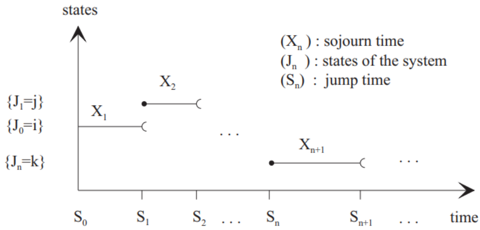


Figure: A sample path of a discrete time semi-Markov process

$$p_{(i,d),(i',d')} = \mathbb{P}(X_{n+1} = d', J_n = i' \mid X_n = d, J_{n-1} = i)$$

- Extension du modèle markovien à une distribution quelconque de la durée de séjour X_n.

Figure issue de V. S. Barbu et N. Limnios, Semi-Markov Chains and Hidden Semi-Markov Models toward Applications, Springer, 2008

Exemple : modèle semi-markovien 'Explicit Duration HMM'

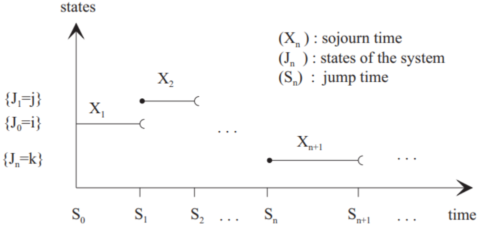


Figure: A sample path of a discrete time semi-Markov process

$$
\begin{aligned}
p_{(i,d),(i',d')} &= \mathbb{P}(X_{n+1} = d', J_n = i' \mid X_n = d, J_{n-1} = i) \\
&= \mathbb{P}(J_n = i' \mid J_{n-1} = i)\mathbb{P}(X_{n+1} = d' \mid J_n = i')
\end{aligned}
$$

avec $P(X_{n+1} = d' \mid J_n = i')$ probabilité d'une **durée de séjour** égale à d' sachant qu'on est dans l'état i'

Modèle semi-markovien caché (HSMM)

Le modèle

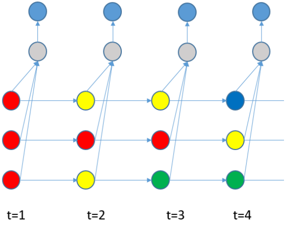
- l'état caché est un modèle de semi-Markov
- à chaque pas de temps on observe Y_t
- cas le plus classique Y_t : ne dépend que de Z_t (comme en HMM)

Estimation des lois du modèle

- classiquement : EM
- à l'étape E : un algorithme de forward-backward un peu plus tricky que pour un HMM
- à l'étape M : formules explicites dans le cas non paramétrique

Deux extensions des HSMM

Cas de plusieurs chaînes cachées qui émettent conjointement l'observation



Cas où l'observation influence l'état caché

Algorithme VBEM pour l'estimation d'un modèle HSMM multichaine factoriel, application à l'inférence de chemin migratoire des oiseaux

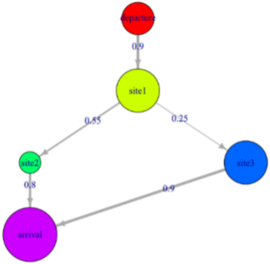
En collaboration avec

- Ronan Trépos, Régis Sabbadin, MIAT INRAE Toulouse
- Sam Nicol, CSIRO Brisbane

Motivation : inférence de chemins migratoires à partir de données ebird

Questions

- Quels sont les chemins utilisés préférentiellement par les oiseaux?
- Où faut-il acquérir plus de données pour améliorer l'inférence du réseau migratoire?



Les données ebird

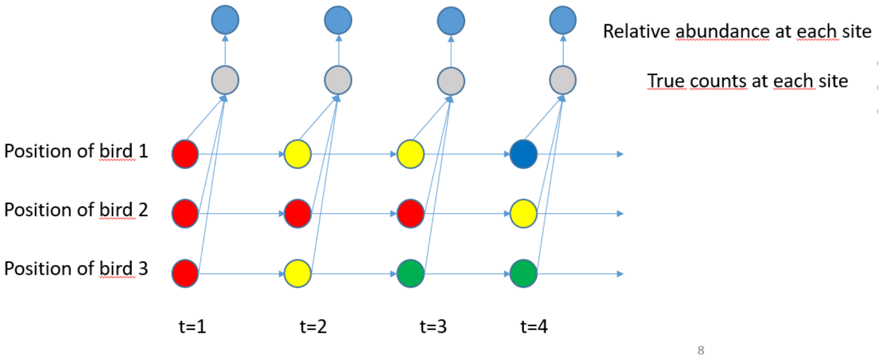
Idéalement : des séries temporelles de comptages à chaque site

En pratique : abondance relative hebdomadaire au niveau des pixels d'une grille (post-traitement ebird)

$\longrightarrow$ données à agréger à l'échelle d'un site

$\longrightarrow$ information imparfaite sur les comptages

BIRDNET, un modèle de type Factorial HSMM: (1) la structure



BIRDNET, un modèle de type Factorial HSMM: (2) les lois

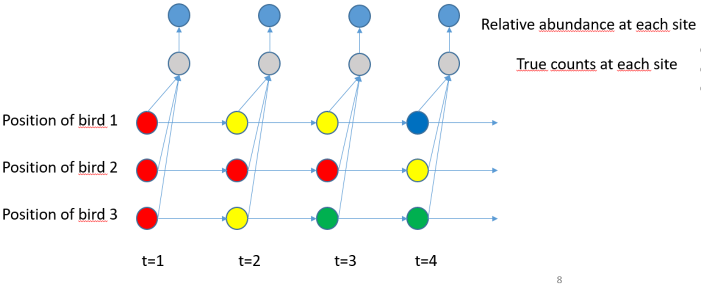
- n oiseaux/chaînes cachées; I sites
- Chaîne cachée
  - durée de séjour au site i : loi Poisson de paramètre λ_i
  - Proba du nouvel état : loi catégorielle, $(R(i, i+1), R(i, i+2), \ldots, R(i, I))$
- Emission commune
  - N_i^t : nombre d'oiseaux au site i à la date t
  - O_i^t : abondance relative au site i à la date t
  - $O_i^t = \alpha_i N_i^t$
  - $P(O_i^t \mid N_i^t)$: loi log-normale, i.e. $\ln(\alpha_i)$ suit une loi normale de paramètres (μ_i, σ_i)

BIRDNET, un modèle de type Factorial HSMM: (2) les lois

Modèle bayésien

- n oiseaux/chaînes cachées; I sites
- Chaîne cachée : SMM
  - durée de séjour au site i : loi Poisson de paramètre λ_i, prior Gamma
  - proba du nouvel état : loi catégorielle, $(R(i,i+1), R(i,i+2), \ldots, R(i,I))$, prior Dirichlet
- Emission commune
  - N_i^t : nombre d'oiseaux au site i à la date t
  - O_i^t : abondance relative au site i à la date t
  - $O_i^t = \alpha_i N_i^t$
  - $P(O_i^t \mid N_i^t)$: loi log-normale, i.e. $\ln(\alpha_i)$ suit une loi normale de paramètres (μ_i, σ_i), prior Gaussienne sur μ_i (σ_i pas estimé)

Remarque sur la structure de dépendance



- Les chaînes cachées sont indépendantes
- MAIS elles ne le sont plus conditionnellement aux observations : inférence difficile

Premières solutions mises en oeuvre pour l'inférence

- développement d'un algorithme MCEM et d'un ABC
- de bons résultats sur données simulées, et des résultats cohérents sur données ebird
- mais recherche d'une méthode plus rapide.

S. Nicol, M.-J. Cros, N. Peyrard, R. Sabbadin, R. Trépos, R. A. Fuller, B. K. Woodworth. FlywayNet: A hidden semi-Markov model for inferring the structure of migratory bird networks from count data, Methods in Ecology and Evolution, 2022

Principe de l'algorithme VBEM (1)

Une méthode déterministe pour faire de l'inférence approchée d'un modèle bayésien

- Le problème à résoudre
  - soit y observation, x variable cachée, θ paramètre du modèle
  - trouver $p(\theta \mid y)$ et $p(x \mid y)$ qui maximisent $p(y)$.
- L'approche VBEM (Beal, 2003)
  - maximiser une borne inf de $\ln p(y)$ parmi le sous-ensemble des distributions contraintes q telles que $q(x, \theta) = q(x)q(\theta)$

M.J. Beal, Variational algorithms for approximate Bayesian inference, Ph.D. Dissertation, UCL, 2003

Principe de l'algorithme VBEM (2)

Une borne inférieure de la vraisemblance :

$$\ln p(y) = \ln \int d\theta \; dx \; q(x,\theta) \frac{p(x,y,\theta)}{q(x,\theta)} \geqslant \int d\theta \; dx \; q(x,\theta) \; \ln \frac{p(x,y,\theta)}{q(x,\theta)}.$$

On note :

$$\mathscr{F}(q(x), q(\theta)) := \int d\theta \; dx \; q(x)q(\theta) \; \ln \frac{p(x,y,\theta)}{q(x)q(\theta)}.$$

Maximiser $\mathscr{F}$ revient à minimiser une divergence de Kullbac-Leibler :

$$\mathscr{F}(q(x), q(\theta)) = \ln p(y) - KL(q(x)q(\theta))|p(x,\theta \mid y))$$

Principe de l'algorithme VBEM (3)

VBEM : algorithme iteratif qui maximise de manière alternée $\mathscr{F}(q(x), q(\theta))$ selon $q(x)$ et $q(\theta)$

- **VBE step**: calculer
$$q(x \mid y) \propto e^{E_\theta^q[\ln p(x,y|\theta)]}$$
avec E_θ^q l'espérance pour la valeur courante de $q(\theta \mid y)$.
- **VBM step**: calculer
$$q(\theta \mid y) \propto p(\theta) e^{E_x^q[\ln p(x,y|\theta)]}$$
avec $p(\theta)$ la prior sur θ et E_x^q l'espérance pour la valeur courante de $q(x \mid y)$.

VBEM appliqué au modèle BIRDNET : la théorie

VBM step :

- les mises à jour sont simples car on a choisi des lois conjuguées
- cela demande de savoir calculer

$$\begin{aligned}
A(i,t) &= q(\text{bird is in patch i at time t} \mid O) \\
B(i) &= E_\pi[\text{sojourn time of bird at i} \mid O] \\
C(i) &= q(\text{bird stays at i} \mid O) \\
D(i,j) &= q(\text{bird stays consecutively at site i then j} \mid O)
\end{aligned}$$

VBEM appliqué au modèle BIRDNET : la théorie

Etape VBE

- Hypothèse supplémentaire 'mean field'
  - ici x est l'ensemble des n chaînes des n oiseaux : ça reste trop complexe.

  $\longrightarrow$ on suppose l'indépendance $q(x) = \prod_{s=1}^{n} q(x_s)$.

- la solution de la maximisation, $q(x)$, est alors un HSMM monochaine

  $\longrightarrow$ on récupère tous les outils des HSMM (Fwd-Bwd) pour le calcul des quantités d'intérêt A, B, C et D.

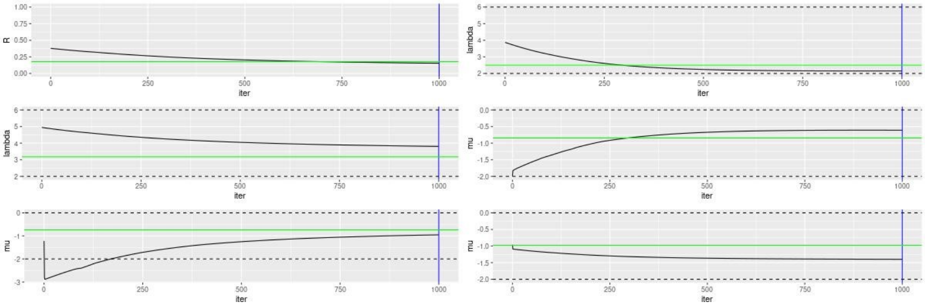
VBEM appliqué au modèle BIRDNET : la pratique

Quelques calibarges nécessaires

- problèmes d'underflow dans le Forward-Backward: réécriture des formules de récursion basée sur les lois a posteriori (Yu 2016) et passage au log
- résolution de l'équation du point fixe à l'étape VBE : une itération, plusieurs, combien?
- durées de séjour moyennes : si trop faibles, pas assez d'observations

Shun-Zheng Yu. Hidden Semi-Markov Models Theory, Algorithms and Applications. Elsevier, 2016

Exemple de résultat pour $I = 3$ sites



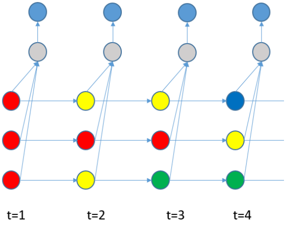
Conclusion

V(B)EM versus ABC pour F-HMM

- V(B)EM
  - \+ des formules simples de mise à jour
  - \+ exécution rapide (pas de simulation)
  - \- implémentation pas si simple
  - \- approximation champ moyen trop grossière?
- ABC
  - \+ implémentation simple (paquet easyABC)
  - \- exécution longue
  - \- choix des statistiques

Deux extensions des HSMM

Cas de plusieurs chaînes cachées qui émettent conjointement l'observation



Cas où l'observation influence l'état caché

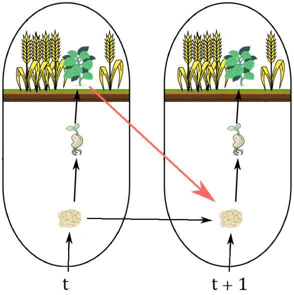
HSMM piloté par les observations pour l'estimation de la dynamique des adventices

Dans le cadre de la thèse d'Hanna Bacave, MIAT INRAE Toulouse
Co-encadrement avec Nikolaos Limnios, UTC Compiègne, et Pierre-Olivier Cheptou, CEFE Montpellier

Motivation : modélisation de la dynamique d'une population d'adventices dans une parcelle

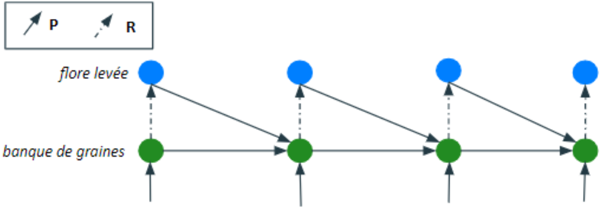
Question

Comment estimer les paramètres clés de la dynamique d'une espèce adventice à partir d'observation de flore levée uniquement?



- Adventices : plantes avec dormance
- Banque de graines dans le sol : non visible
- Dynamique : germination, grenaison, colonisation, dormance

Travaux précédents : modélisation par un HMM 'Observation Driven'



Pas de difficulté de mise en œuvre

- la probabilité de transition dépend aussi de l'observation précédente
- estimation par EM : en adaptant les formules de l'algorithme du Forward-Backward

M. Pluntz, S. Le Coz, N. Peyrard, R. Pradel, R. Choquet, P.-O. Cheptou. A general method for estimating seed dormancy and colonization in annual plants from the observation of existing flora, Ecology Letters 21(7), 2018

Comment modéliser l'influence de l'observation dans un cadre HSMM?

Rappel du modèle Explicit Duration HMM

$$\begin{aligned} p_{(i,d),(j,d')} &= \mathbb{P}(X_{n+1} = d', J_n = j \mid X_n = d, J_{n-1} = i) \\ &= \mathbb{P}(J_n = j \mid J_{n-1} = i)\mathbb{P}(X_{n+1} = d' \mid J_n = j) \end{aligned}$$

avec $P(X_{n+1} = d' \mid J_n = j)$ probabilité d'une durée de séjour égale à d' sachant qu'on est dans l'état j

Deux extensions possibles

- soit l'influence ne joue qu'au moment des changements d'état mais pas sur la durée de séjour : $\mathbb{P}(J_n = j \mid J_{n-1} = i, Y_{S(n)-1}) \rightarrow$ facile à modéliser mais pas réaliste
- soit elle joue aussi sur la durée de séjour : on ne peut plus fixer à l'entrée dans l'état la durée de séjour

Solution proposée : se ramener à un HMM

Autre façon de modéliser un HSMM

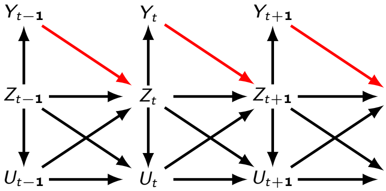
On considère les variables suivantes :

- $Z = (Z_t)_{t \in \mathbb{N}}$ la chaîne de semi-Markov cachée, **indexée sur le temps**
- $U = (U_t)_{t \in \mathbb{N}}$ le temps écoulé depuis l'entrée dans l'état courant de Z, lui aussi caché
- $Y = (Y_t)_{t \in \mathbb{N}}$ le processus observé

Le triplet $(Z, U, Y) = (Z_t, U_t, Y_t)_{t \in \mathbb{N}}$ est une **chaîne de Markov cachée**

On étend cette modélisation au cas où l'observation influence les états cachés
$\rightarrow$ modèle OD-HSMM

Structure de dépendance et lois définissant le OD-HSMM



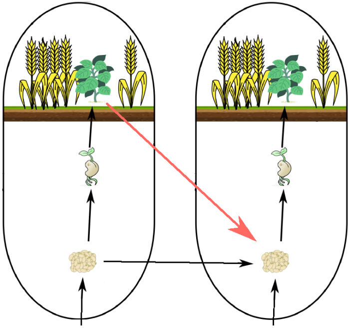
- Loi initiale :
  $\pi(z_0, u_0) = \mathbb{P}(Z_0 = z_0, U_0 = u_0)$.
- Loi d'émission :
  $R(z_t, y_t) = \mathbb{P}(Y_t = y_t | Z_t = z_t)$,
- Loi de transition :

$$P_{y_{t-1}}((z_{t-1}, u_{t-1}), (z_t, u_t)) = \mathbb{P}(Z_t = z_t, U_t = u_t | Z_{t-1} = z_{t-1}, U_{t-1} = u_{t-1}, Y_{t-1} = y_{t-1})$$

Un modèle OD-HSMM paramétrique pour la dynamique de l'adventice

Modèle en présence/absence. Les paramètres en jeu dans sa dynamique sont :

- la **colonisation**, de probabilité c ;
- la **survie** de la graine dans le sol entre t et $t+1$, de probabilité $s = s_0 e^{-\lambda(u_{n-1}-1)}$;
- la **germination** de la graine, de probabilité g ;
- la **grenaison** de la fleur, de probabilité d.



Par exemple, on a $P_1((1,3),(0,1)) = (1 - s_0 e^{-\lambda\times 2})(1-c)(1-d)$. et $R(1,0) = 1-g$.

Estimation du OD-HSMM adventice par algorithme ABC

Principe (méthode par rejet)

Répéter un grand nombre de fois

1. générer θ^c selon une prior $P(\theta^c)$
2. générer $(Y^c_{0:M}, Z^c_{0:M}, U^c_{0:M})$
3. calculer des statistiques à partir de $Y^c_{0:M}$: $S^c = (S^c_1, S^c_2, \dots)$
4. si $Dist(S^c, S(Y_{0:M})) < \epsilon$ on accepte θ^c

Choix des statistiques

Les statistiques de base sur les données observées

1. nombre de 1
2. nombre de transitions de 0 vers 1
3. informations sur les longueurs de 0 et de 1 consécutifs (minimum, maximum, moyenne, quantile à 25%, médiane, quantile à 75%)
4. mode et valeur du mode dans les longueurs de 0 et de 1 consécutifs

Plusieurs groupes de statistiques testés

- Groupe 1 : statistiques 1, 2 et 3
- Groupe 2 : statistiques 1, 2 et 4
- Groupe 3 : toutes les statistiques

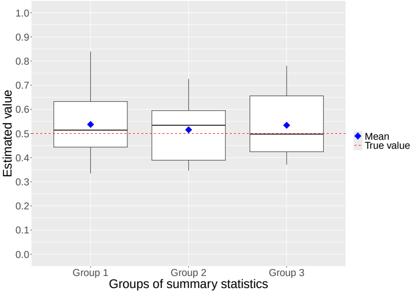
Comportement sur données simulées

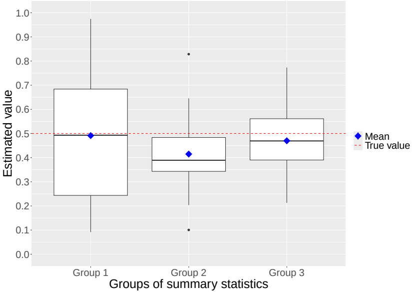
Protocole (répété 30 fois)

- tous les paramètres égaux à 0.5,
- simulation des données $(Z^*_{0:M}, Y^*_{0:M})$ pour $M = 500$
- pour chaque groupe de statistique, estimation avec paquet easyABC avec méthode basée sur du SMC de Lenormand et al (2012)

Lenormand, M., Jabot, F. & Deffuant, G. (2012) Adaptive approximate Bayesian computation for complex models.

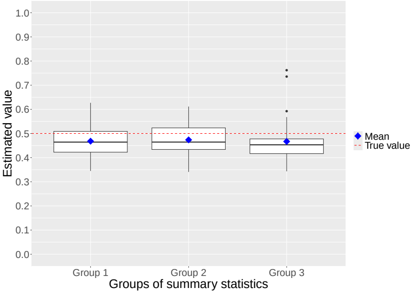
Résultats

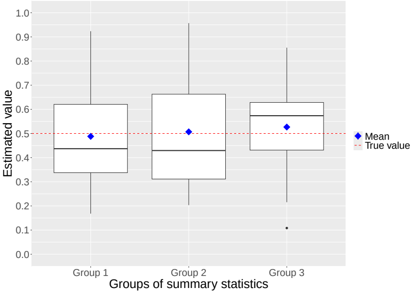




Estimation des paramètres *c* de colonisation et *d* de grenaison

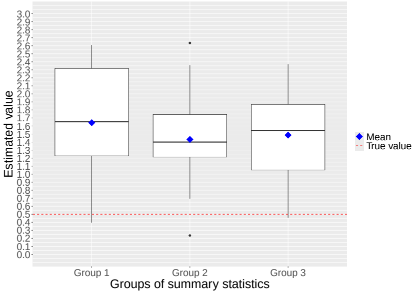
Résultats





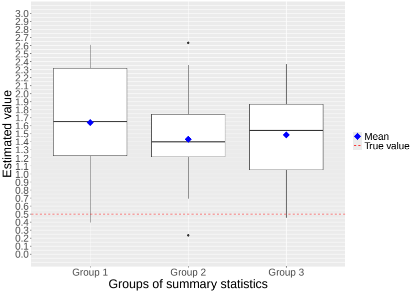
Estimation des paramètres g de germination et $s0$ impliqué dans la survie $s_0 e^{-\lambda(u_{n-1}-1)}$

Résultats

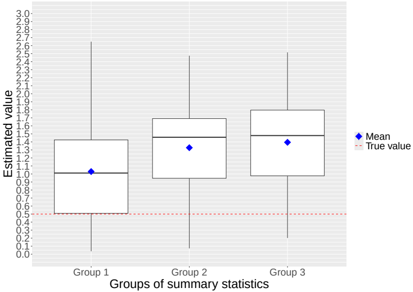


Estimation du paramètre λ impliqué dans la survie $s_0 e^{-\lambda(u_{n-\mathbf{1}}-\mathbf{1})}$

Influence de g?



$g = 0.5$



$g = 1$

Estimation du paramètre λ impliqué dans la survie $s_0 e^{-\lambda(u_{n-1}-1)}$

Conclusion

OD-HSMM pour dynamique adventice

- Pas vraiment de groupe de statistiques qui se détache
- Très bonne estimation de chaque paramètre sauf λ : quelle statistique pour mieux capter son rôle?
- La suite : exploration d'une approche alternative, avec EM

Intérêt de l'approche au delà du OD-HSMM

- Modélisation de l'influence de l'observation sur l'état caché : le même problème se pose pour
    - l'influence de covariables sur l'état caché
    - l'influence de l'état caché d'une autre chaîne cachée, en contexte spatio-temporel

Discussion générale

- **Aspects computationnels**. 'il n'y a plus qu'à faire tourner un (V)(B)EM' : ce n'est pas si trivial que ça!

- **Communauté HSMM** : deux visions
  - Les informaticiens.
    S.-Z. Yu. Hidden Semi-Markov Models Theory, Algorithms and Applications. Elsevier, 2016
  - Les statisticiens.
    V. S. Barbu et N. Limnios, Semi-Markov Chains and Hidden Semi-Markov Models toward Applications, Springer, 2008

  Des travaux qui se complètent mais pas exactement le même formalisme HSMM
  → pas toujours facile de faire les ponts, on y travaille!

Estimation du OD-HSMM adventice par algorithme EM

Principe

- étape E, on calcule $Q(\theta|\theta^{(m)}) = \mathbb{E}_{\theta^{(m)}}\left[\ln \mathbb{P}_\theta(Y_{0:M}, Z_{0:M}, U_{0:M}) | Y_{0:M} = y_{0:M}\right]$
- étape M, on maximise $Q(\theta|\theta^{(m)})$ pour mettre à jour les paramètres

Difficultés

- étape E : repose sur un forward-backward couteux en temps et espace
- étape M : non explicite -> optim num et la vraisemblance complète était plate (testé dans le cas SEM).

Une solution? (en cours)

- étape E : forward-backward
- étape M : mise à jour en mode 'non paramétrique', avec formules explicites
- à convergence : les probabilités estimées s'expriment en fonction des paramètres -> on identifie les paramètres par moindres carrés.