# Variational EM pour Factorised Hidden Markov Models avec retour des données

Sebastian Le Coz

Séminaire des doctorants

Co-directeurs : Nathalie Peyrard, Pierre-Olivier Cheptou
Financement: Région + ANR AGROBIOSE

INRA SCIENCE & IMPACT

MIA TOULOUSE

CENTRE D'ECOLOGIE FONCTIONNELLE & EVOLUTIVE

# Motivation : modélisation de la dynamique des adventices

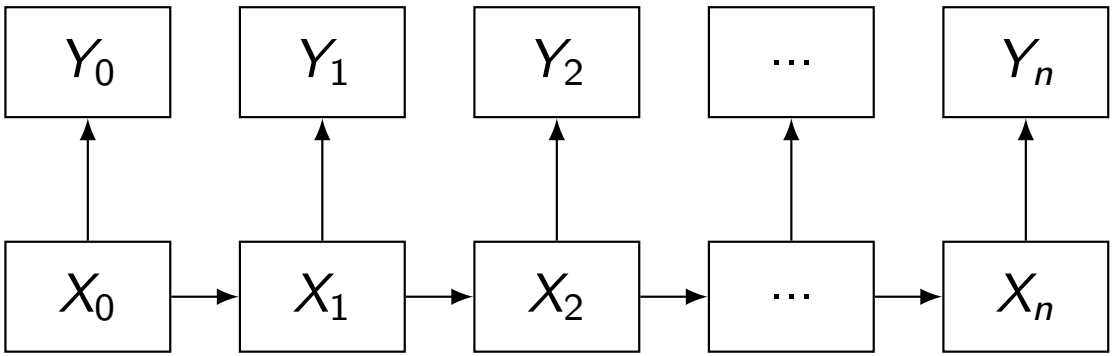
- En écologie, le rôle de la banque de graines est une boîte noire car ses effets sur la dynamique des adventices ne sont pas connus.
- La banque de graines est difficilement observable.

Certains modèles ne prennent pas en compte la banque de graines.
Exemple : Les modèles de métapopulation étudient la dispersion et la dynamique spatiale d'une espèce à l'aide de deux paramètres, la colonisation et l'extinction.

Problème : La dormance des graines peut engendrer une fausse déclaration d'extinction locale de l'espèce (Bullock et al 2006, Freckleton et al 2002).

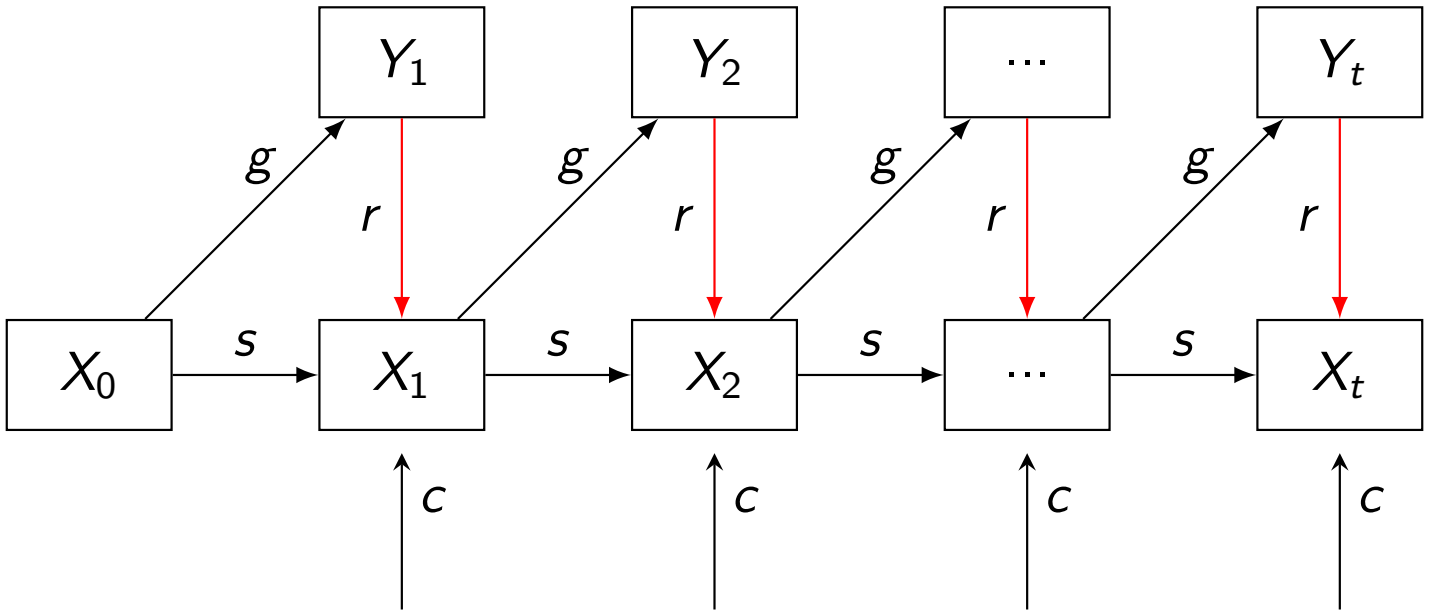
Par conséquent, en agronomie, la gestion des adventices repose en partie sur la compréhension du rôle de banque de graines dans la dynamique de l'adventice.

## Le cadre Hidden Markov Model (HMM) pour les adventices




- Le modèle de David et al (2010) comporte trop de paramètres et de variables cachés, le rendant difficile à estimer.
- Le modèle de Fréville et al (2013) est un modèle qui limite la survie de la banque de graines à 1 an.
- Le modèle de Borgy et al (2015) ne prend pas en compte la colonisation extérieure.
- Le modèle de Pluntz et al (2015).

## Le Modèle de Pluntz et al : HMM avec retour des données




- Estimation sur HMM avec retour des données facile.
- Parametres de colonisation, de germination et de survie des graines.
- Résultat cohérant avec les données biologique.
- Article en cours.

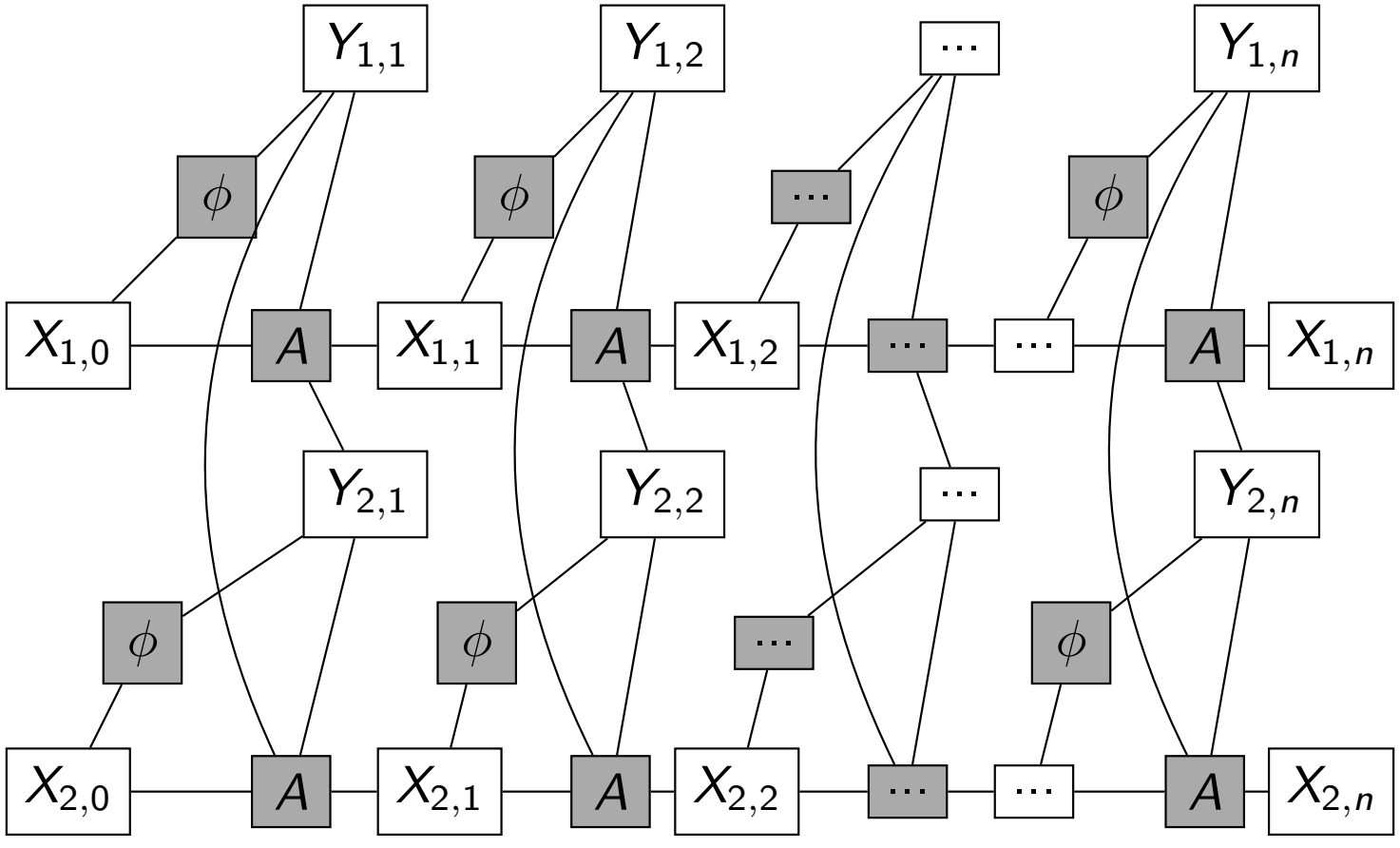
La dynamique spatiale n'est pas prise en compte !

## Objectifs :

Prendre en compte la dynamique spatiale des adventices à partir de la colonisation extérieure et interparcelle, la survie des graines, la germination des graines et la reproduction.

- Etendre le modèle HMM avec retour de données au cas multi-chaines.
- Developper une méthode d'estimation pour FHMM avec retour des données.

# Factor graphe du FHMM avec retour des données




Les variables cachées :
$X^{C,N} = (X_{1,1}, X_{2,1}, ..., X_{c,1}, ..., X_{C,1}, X_{1,2}, ..., X_{1,n}, ..., X_{1,N}, ..., X_{C,N})$
Les variables observées :
$Y^{C,N} = (Y_{1,1}, Y_{2,1}, ..., Y_{c,1}, ..., Y_{C,1}, Y_{1,2}, ..., Y_{1,n}, ..., Y_{1,N}, ..., Y_{C,N})$

## Estimation d'un FHMM avec retour des données

MCMC : temps de calcul important et estimation variable.
EM : temps de calcul important sur FHMM et problème de mémoire.

Variational EM (VEM) développé par Beal (2003):

- approche la solution du EM à l'aide de simplifications imposées sur le modèle par son utilisateur.
- développé pour un FHMM sans retour des données par Ghahramani et Jordan (1997).

## Démarche

- Etendre le VEM à un FHMM avec retour des données.
- Développer plusieurs simplifications imposées pour le VEM.

EM

- Initialisation des paramètres $\lambda_0 = (\pi_0, A_0, \phi_0)$.
- Itèrer les étapes suivantes jusqu'à convergence de $\lambda_{it}$.

Étape E :

Forward-Backward pour calculer $p(X_n^C | Y^{C,N}, \lambda_{it})$ et $p(X_n^C, X_{n-1}^C | Y^{C,N}, \lambda_{it})$ nécessaires pour évaluer

$$E_{\mathbb{P}}[ln(\mathbb{P}(X^{C,N}, Y^{C,N}|\lambda))|Y^{C,N} = y^{C,N}, \lambda_{it}]$$

Étape M :

Mise à jour du paramètre $\lambda$ avec :

$$\lambda_{it+1} = \arg\max_{\lambda} E_{\mathbb{P}}[ln(\mathbb{P}(X^{C,N}, Y^{C,N}|\lambda))|Y^{C,N} = y^{C,N}, \lambda_{it}]$$

Soit q une distribution sur $X^{C,N}$

$$Q(q,\lambda) = E_q[ln(\frac{\mathbb{P}(Y^{C,N} = y^{C,N}, X^{C,N}|\lambda)}{q(X^{C,N})})]$$

$$Q(q,\lambda) = E_q[ln(P(Y^{C,N}, X^{C,N}|\lambda))] - KL(q|\mathbb{P}(X^{C,N} = x^{C,N}|Y^{C,N} = y^{C,N}, \lambda))$$

Si $q(X^{C,N}) = \mathbb{P}(X^{C,N}|Y^{C,N}, \lambda)$ alors $KL = 0$.
Donc :

$$Q(q,\lambda) = E_{\mathbb{P}}[ln(P(Y^{C,N}, X^{C,N}|\lambda))|Y^{C,N} = y^{C,N}, \lambda))]$$

Ainsi on retrouver la quantilé utilisée dans l'étape M du EM.

## EM Variationnel

Impose des contraintes à la probabilité q.

Initialisation des paramètres $\lambda_0$

Itération sur 2 étapes.

- Étape E :

$$q_{it+1} = \arg\max_q Q(q, \lambda_{it})$$

qui est équivalent à

$$q_{it+1} = \arg\min_q KL(q|\mathbb{P}(X^{C,N} = x^{C,N}|Y^{C,N} = y^{C,N}, \lambda_{it}))$$

- Étape M :

$$\lambda_{it+1} = \arg\max_\lambda Q(q_{it+1}, \lambda)$$

**Deux type de simplification étudiées VEM non-stationnaire**

$q$ dépend du temps et peut s'écrire comme :

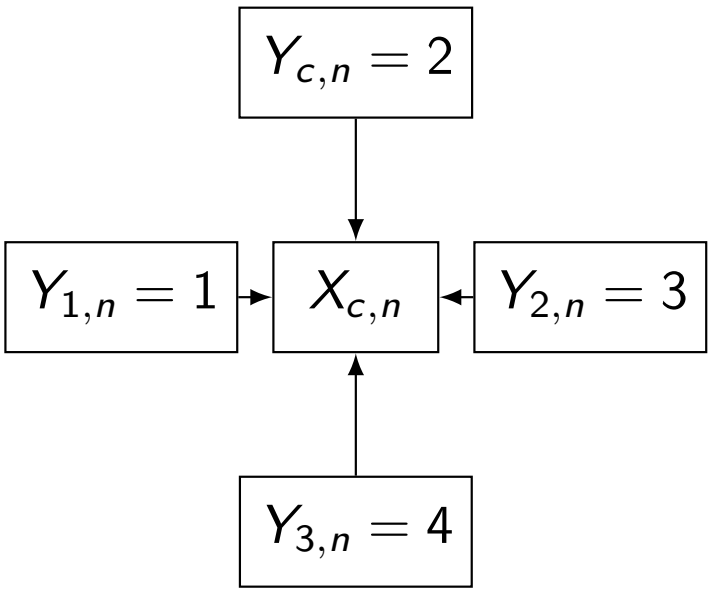
$$q_n(x^{C,N}) = \prod_{c=1}^{C} q_n(x_c^N) = \prod_{c=1}^{C} \prod_{n=0}^{N} q_n(x_{c,n})$$

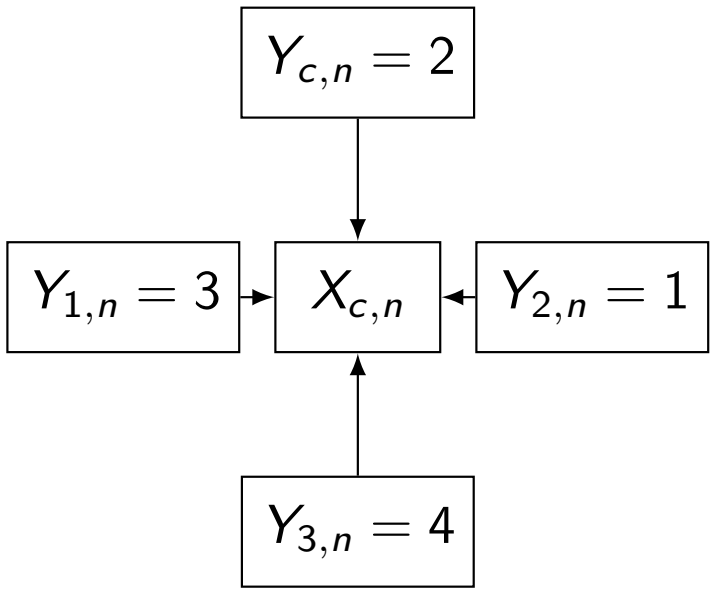
**VEM à observation interchangeables**

$q$ dépend des observations d'une année et peut s'écrire comme :

$$q_{c,n}(x^{C,N}) = \prod_{c=1}^{C} q_{c,n}(x_c^N) = \prod_{c=1}^{C} \pi_q(x_{c,0}) \prod_{n=1}^{N} q_{c,n}(x_{c,n}|y_n^C))$$

## VEM à observation interchangeables

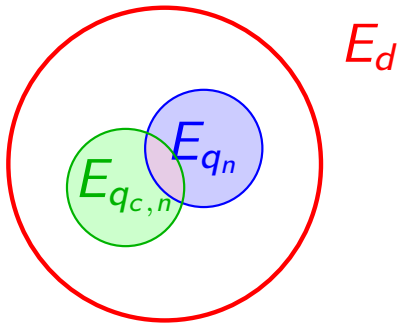



C'est-à-dire que pour toute permutation $\sigma$ l'égalité suivante est vérifié:

$$q_{c,n}(x_{c,n}|y_{c,n}, y_{c',n} \forall c' \neq c) = q_{c,n}(x_{c,n}|y_{c,n}, y_{(\sigma(c'),n)} \forall c' \neq c)$$

$E_d$ l'ensemble des distributions possibles
$E_{q_n}$ l'ensemble des distributions non-stationnaire
$E_{q_{c,n}}$ l'ensemble des distributions pour observation interchangeables




Nombre de probabilités à calculer pour l'étape E d'un FHMM avec retour des données:

| méthode d'estimation | nombre de probabilités |
|---|---|
| EM exact | $\|\Omega_X\|^{2C}$ x $N + \|\Omega_X\|^C$ x $N$ |
| VEM à observation interchangeables | $\|\Omega_X\|\ \|\Omega_Y\|\ \binom{\|\Omega_Y\| + (C-1) - 1}{C-1}$ |
| VEM non-stationnaire | N x $\|\Omega_X\|$ |

## Conclusion

Les méthodes variationnelles permettent de faire un compromis entre le temps de calcul et la qualité de l'estimation.

| temps de calcul | méthode d'estimation | qualité |
|---|---|---|
| $\wedge$ | EM exact | $\wedge$ |
| \| | VEM à observation interchangeables | \| |
| \| | VEM non-stationnaire | \| |

**En cours**

- Coder les différents VEM et tester les méthodes d'estimation avec données simulées.

**Perspectives**

- Coder différent VEM dans le cas paramétriques.
- Tester sur données Epoisse et analyser les résultats.
- Prédire l'état de la banque de graines et de la flore levée.

# Merci de votre attention !



Freckleton RP and Watkinson AR. *Large-scale spatial dynamics of plants: metapopulation regional ensembles and patchy populations*. Journal of Ecology, 2002, Vol. 90, pages: 419-434.


Bullock JM, Shea K and Skarpaas O. *Measuring plant dispersal: an introduction to field methods and experimental design*. Plant Ecology, 2006, Vol. 186, pages : 217-234.


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, 2013, pages 1572-1580.


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.


David O, Garnier A, Larédo C and Lecomte J. *Estimation of Plant Demographic Parameters from Stage-Structured Censuses*. Biometrics, pages 875-882, 2010.


*http://www2.ufz.de/biolflor/index.jsp*


Matthew J. Beal. *Variational algorithm for approximate Bayesian inference*. M.A., M.Sci., Physics, University of Cambridge, UK, 2003.


Zoubin Ghahramani and Michael I Jordan. *Factorial Hidden Markov Models*. Kluwer Academic Publishers, pages 245-273, 1997.

$$
\begin{aligned}
ln(\mathbb{P}(Y^{C,N} = y^{C,N}|\lambda)) &= ln\left[\sum_{x^{C,N}\in|\Omega_X|^{CN}} \mathbb{P}(Y^{C,N} = y^{C,N}, X^{C,N} = x^{C,N}|\lambda)\right] \\
&= ln\left[\sum_{x^{C,N}\in|\Omega_X|^{CN}} q(x^{C,N})\frac{\mathbb{P}(Y^{C,N}=y^{C,N},X^{C,N}=x^{C,N}|\lambda)}{q(x^{C,N})}\right] \\
&\geq \sum_{x^{C,N}\in|\Omega_X|^{CN}} q(x^{C,N}) ln\left[\frac{\mathbb{P}(Y^{C,N}=y^{C,N},X^{C,N}=x^{C,N}|\lambda)}{q(x^{C,N})}\right] \\
&\geq Q(q(x^{C,N}),\lambda)
\end{aligned}
$$

## Réecriture de $Q$ avec la divergence de Kullback-Leibler

$$
\begin{aligned}
Q(q(x^{C,N}),\lambda) &= \sum_{x^{C,N}\in|\Omega_X|^N} q(x^{C,N}) ln\left[\frac{\mathbb{P}(Y^{C,N}=y^{C,N},X^{C,N}=x^{C,N}|\lambda)}{q(x^{C,N})}\right] \\
&= \sum_{x^{C,N}\in|\Omega_X|^{CN}} q(x^{C,N}) ln\left[\frac{\mathbb{P}(X^{C,N}=x^{C,N}|Y^{C,N}=y^{C,N},\lambda)\mathbb{P}(Y^{C,N}=y^{C,N}|\lambda)}{q(x^{C,N})}\right] \\
&= \sum_{x^{C,N}\in|\Omega_X|^{CN}} q(x^{C,N}) ln(\mathbb{P}(Y^{C,N}=y^{C,N}|\lambda)) \\
&+ \sum_{x^{C,N}\in|\Omega_X|^{CN}} q(x^{C,N}) ln\left[\frac{\mathbb{P}(X^{C,N}=x^{C,N}|Y^{C,N}=y^{C,N},\lambda)}{q(x^{C,N})}\right] \\
&= ln(\mathbb{P}(Y^{C,N}=y^{C,N}|\lambda)) \\
&- \sum_{x^{C,N}\in|\Omega_X|^{CN}} q(x^{C,N}) ln\left[\frac{q(x^{C,N})}{\mathbb{P}(X^{C,N}=x^{C,N}|Y^{C,N}=y^{C,N},\lambda)}\right] \\
&= ln(\mathbb{P}(Y^{C,N}=y^{C,N}|\lambda)) \\
&- KL(q(x^{C,N})|\mathbb{P}(X^{C,N}=x^{C,N}|Y^{C,N}=y^{C,N},\lambda))
\end{aligned}
$$

Ainsi, on remarquer que $Q(q(x^{C,N}),\lambda) = ln(\mathbb{P}(Y^{C,N}=y^{C,N}|\lambda))$ quand $q(x^{C,N}) = \mathbb{P}(X^{C,N}=x^{C,N}|Y^{C,N}=y^{C,N},\lambda)$.