Consistance de l'a posteriori pour les modèles de Markov cachés à espace d'états fini

Elodie Vernet
elodie.vernet@math.u-psud.fr

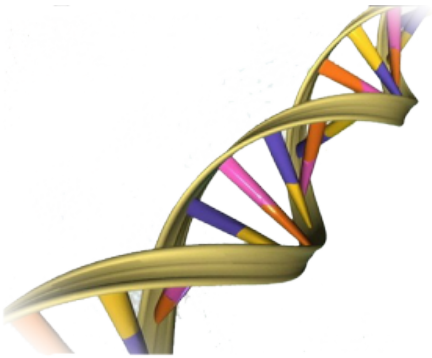
Directrices de thèse :

- Elisabeth Gassiat (Université Paris Sud)
- Judith Rousseau (CREST)

Journées MAS, Toulouse, mercredi 27 août 2014

Introduction

Les modèles de Markov cachés (HMMs)



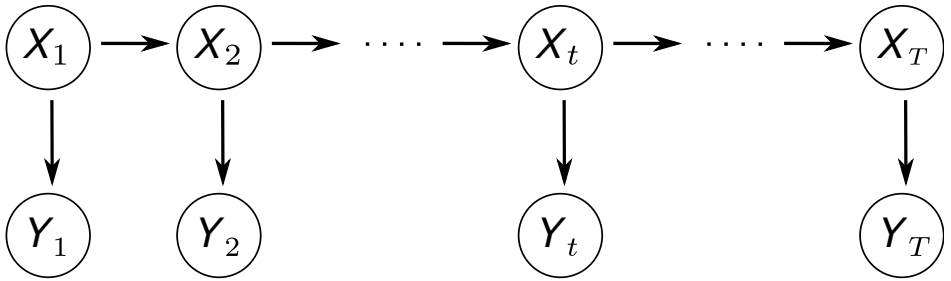
- permettent de traiter des données dépendantes
- sont très utilisés en pratique,
- leurs propriétés théoriques sont mal comprises.

Objectif :

étudier les propriétés asymptotiques (lorsque le nombre d'observations tend vers $+\infty$) de ces modèles dans le cadre bayésien non paramétrique.
Dans cet exposé je m'attarderai sur la consistance.

Plan

Définition



Si

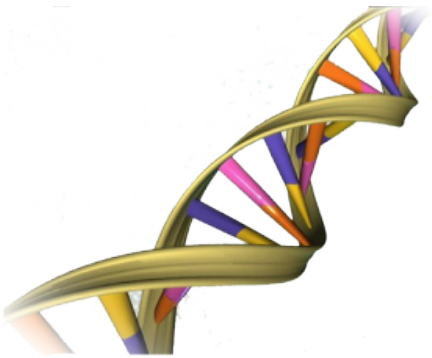
- $(X_t)_{t\in\mathbb{N}}$ est une chaine de Markov,
- Y_t est une pertubation de l'état X_t de la chaine : sachant la chaine de Markov $(X_t)_{t\in\mathbb{N}}$, les Y_t sont indépendants et ne dépendent que de X_t,
- les états X_t sont cachés,
- on observe les Y_t

Alors $(X_t, Y_t)_{t\in\mathbb{N}}$ est une chaine de Markov cachée.

Exemples et utilisations

Les chaines de Markov cachés sont très utilisés en pratique en

- génomique,
- reconnaissance de parole,
- etc.



souvent pour classer des observations : on essaie alors de retrouver les états cachés.

Modèle de Markov caché étudié

On supose qu'on a un nombre k fini et connu d'états :

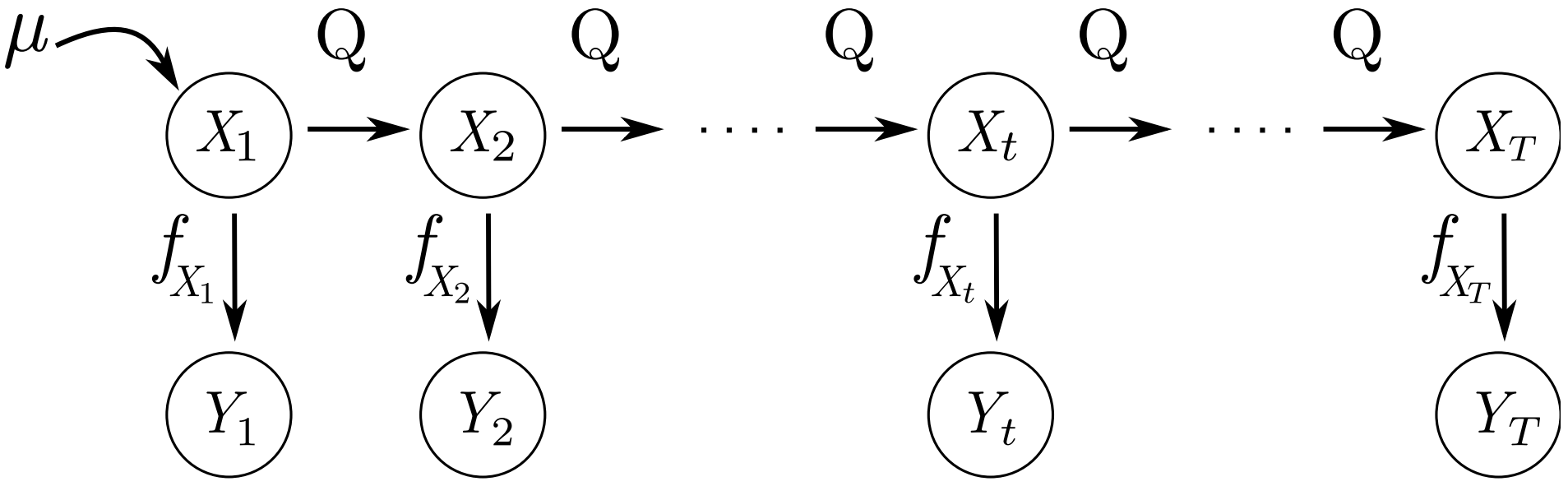
$$X_t \in \{1, \ldots, k\}, \quad Y_t \in \mathbb{R}^d, \forall t,$$

Paramètre du modèle : (μ, θ) avec $\theta = (Q, f)$

- probabilité initiale μ, $\mu_i = P(X_1 = i)$, $i = 1, \ldots, k$,
- matrice de transition Q, $k \times k$, $Q_{i,j} = P(X_{t+1} = j | X_t = j)$,
- densités d'émission : $f = (f_1, \ldots, f_k)$, pour $1 \leq i \leq k$, f_i est la densité de Y_t sachant $X_t = i$ par rapport à une mesure λ sur $\mathbb{R}^d$.

P^θ est la loi de probabilité de $(X_t, Y_t)_{t \in \mathbb{N}}$ sous θ et pour loi initiale "la" loi stationnaire.

Représentation du modèle étudié



grande dimension – non paramétrique

Petite histoire des chaines de Markov cachées

Années 60 : introduction des chaines de Markov cachées par Baum et Petrie. Depuis développement de nombreux algorithmes pour analyser les modèles de Markov cachés.

Années 90 : étude asymptotique de maximum de vraisemblance (estimateur fréquentiste) pour les modèles paramétriques de Markov cachés (Douc, Matias, Moulines, Rydén).

Années 2010 : étude asymptotique des modèles de Markov cachés paramétriques bayésiens, non paramétriques fréquentistes, identifiabilité des chaines de Markov cachées (Gassiat, Rousseau, Dumont, Le Corff...).

Résumé

- en pratique modèles très utilisés
- en théorie des résultats en paramétrique
- mais très peu de résultats théoriques en non paramétrique notamment car l'identifiabilité n'a été montrée que très récemment.

Plan

Point de vue bayésien

- On met une probabilité π sur l'ensemble des paramètres Θ : ce que l'on sait sur les paramètres avant de faire une expérience

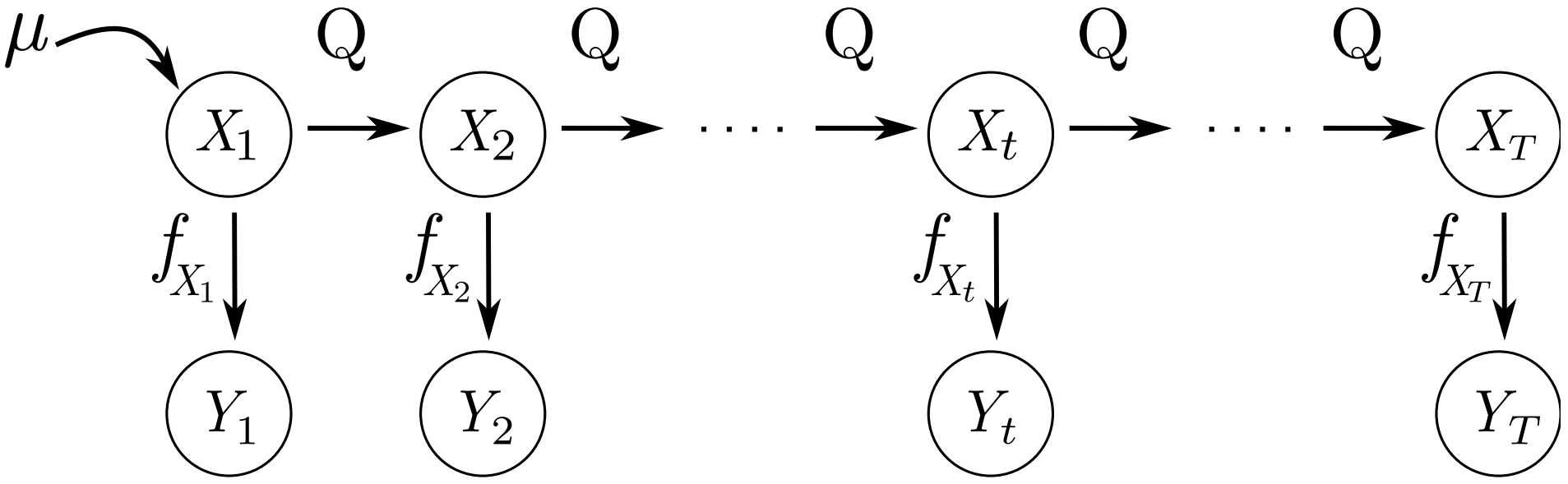
  π est l'a priori

- on observe $(Y_1, \ldots, Y_n)$,
  qui sachant θ est distribué selon P^θ,

- on étudie l'a posteriori : $\pi(\cdot | Y_1, \ldots, Y_n)$: probabilité sur Θ
  $\pi(\theta \in A | Y_1, \ldots, Y_n)$ probabilité que la paramètre θ appartienne à un sous-ensemble A de Θ, sachant qu'on a observé $Y_1, \ldots, Y_n$,

$$\pi(\theta \in A | Y_1, \ldots, Y_n) = \frac{\int_A p^\theta(Y_1, \ldots, Y_n)\pi(d\theta)}{\int_\Theta p^\theta(Y_1, \ldots, Y_n)\pi(d\theta)}.$$

A priori pour le modèle étudié



L'a priori

$$\pi = \mu \otimes \pi_Q \otimes \pi_f$$

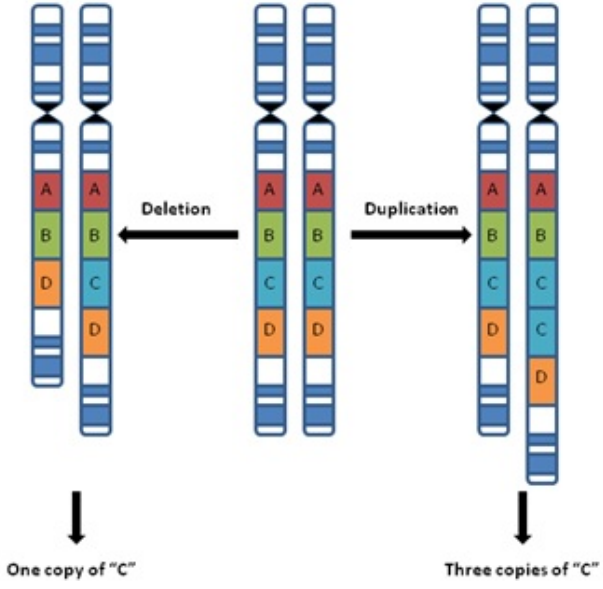
- μ est une probabilité initiale fixée (il n'est pas possible de retrouver la vraie probabilité initiale),
- π_Q est une probabilité sur les matrices de transition $k \times k$,
- π_f est une probabilité sur $\{\text{les densités sur } \mathbb{R}^d\}^k$.

Utilisation de ce modèle

Exemple de génomique : variablilité du nombre de copies d'un gène

Cet exemple est issue d'un papier de Yau, Papaspilopoulos, Roberts et Holmes (2011).

- Ils obtiennent de meilleurs résultats avec leur modèle non paramétrique avec un a priori particulier
- et en théorie, comment déterminer si un a priori est "bon" ?



Consistance de l'a posteriori

But : Déterminer si un a priori est "bon".

- On s'intéresse au comportement asymptotique de l'a posteriori, i.e. quand le nombre d'observations tend vers $+\infty$.
- On veut savoir si l'a posteriori se concentre autour du vrai paramètre θ^* lorsque le nombre d'observations distribuées selon θ^* augmente.
- On s'intéresse à la consistance de l'a posteriori en θ^*.

Définition

L'a posteriori est consistant en θ^* si pour tout voisinage U de θ^*, P^{θ^*}-presque sûrement

$$\pi(U|Y_1, \ldots, Y_n) \to 1.$$

Résultats asymptotiques bayésiens

- Doob 1949 : l'a posteriori est consistant en π-presque toute valeur de θ^*
  -> OK pour les modèles paramétriques mais
- Freedman 1963 : donne un contre-exemple en non-paramétrique
  -> nécessité de déterminer des conditions sur l'a priori et θ^* assurant la consistance de l'a posteriori.
- Schwartz 1965 : Dans le cas i.i.d., si l'a priori met assez de poids dans un certain voisinage de θ^*, alors l'a posteriori est consistant en θ^* pour la topologie étroite.
- récemment : étude de la vitesse de concentration, adaptivité de modèles bayésiens non-paramétriques (Ghosh, Ghosal, Ramamoorthi, Rousseau, Tokdar, van der Vaart ...).

Plan

Oui mais consistant par rapport à quelle topologie ??

Rappel de la définition de consistance bayésienne

On dit que l'a posteriori est consistant en θ^* si pour tout voisinage U de θ^*,

$$\pi(U|Y_1, \ldots, Y_n) \to 1, P^\theta - p.s.$$

Voisinage -> nécessité de choisir une topologie sur Θ. On va comparer deux paramètres en comparant :

1. la loi jointe de l observations consécutives P_l^θ :
   - on va tout d'abord utilisé la topologie associée à la convergence étroite sur ces lois jointes,
   - puis la distance L_1 sur ces densités jointes p_l^θ,
2. une topologie produit sur Q et f (pour montrer la consistance pour chaque composante du paramètre Q et f).

Théorème de consistance

Si (H1) : il existe $\underline{q} > 0$ tel que

- $\mu_i \geq \underline{q}$ pour tout $1 \leq i \leq k$,
- π_Q ne met du poids que sur les matrices de transition Q telles que $Q_{i,j} \geq \underline{q}$, pour tout $1 \leq i, j \leq k$,

et si (H2) : pour tout $\epsilon > 0$, il existe $\Theta_\epsilon \subset \Theta$ tel que $\pi(\Theta_\epsilon) > 0$,

- (H2a) $\sum_{i=1}^{k} f_i(y) > 0$, pour tout y tel que $\sum_{i=1}^{k} f_i^*(y) > 0$,
- (H2b) $\sup_{y \,:\, \sum_{i=1}^{k} f_i^*(y) > 0} \max_{1 \leq i \leq k} f_i(y) < \infty$
- (H2c) $\sum_{i=1}^{k} \mu_i^* \int f_i^*(y) \left| \log \left(\sum_{j=1}^{k} f_j(y) \right) \right| \lambda(dy) < \infty$

Conditions sous lesquelles la chaine se mélange bien.

et pour tout $\theta = (Q, f) \in \Theta_\epsilon$,

- (H2d) $\|Q - Q^*\| < \epsilon$,
- (H2e) $\max_{1 \leq i \leq k} \int f_i^*(y) \max_{1 \leq i,j \leq k} \log \left(\frac{f_j^*(y)}{f_j(y)} \right) \lambda(dy) < \epsilon$,

L'a priori met du poids au "voisinage" de θ^*.

alors l'a posteriori est consistant en θ^* pour la topologie associée à la convergence étroite sur les lois de probabilité de $Y_1, \ldots, Y_l$.

Remarques sur le théorème

- Volontairement la condition (H2e) est très similaires à la condition souvent vérifiée pour montrer la consistance dans le cas de l'estimation de densité pour des observations i.i.d..
  - On peut reprendre les a priori pour lesquels l'a posteriori est consistant dans le cas i.i.d. sur f
  - et montrer la consistance pour les chaines de Markov cachées des a posteriori correspondants.
  - Exemples : mélanges de gaussiennes indépendants ou translatés sur f_i, processus de Dirichlet.

- Si de plus l'a priori met principalement du poids sur des espaces qui ne sont pas trop grands
  alors on obtient la consistance de l'a posteriori pour la distance L_1 sur les densités marginales de l observations.

Consistance pour Q et f

Identifiabilité : par Gassiat, Cleynen et Robin (2013)

- Si les lois d'émission $f_1 d\lambda, \ldots, f_k d\lambda$ sont des probabilités linéairement indépendantes
- et si la matrice de transition Q est de rang plein

alors on retrouve les paramètres du modèle de Markov caché à partir de la loi jointe de trois observations consécutives à permutation des états près.

Consistance de l'a posteriori pour Q et pour f

- La consistance de l'a posteriori pour la convergence L_1 sur la loi jointe de 3 ou plus observations,
- et les conditions d'identifiabilité sur θ^*,

implique la consistance de l'a posteriori pour Q et pour f (pour la topologie étroite) à permutation près.

Plan

Processus de Dirichlet

La loi de Dirichlet de paramètre $(\alpha_1, \ldots, \alpha_k)$ est une généralisation de la loi béta sur le simplex de dimension $k-1$,
elle a une densité par rapport à Lebesgue proportionnelle à

$$\prod_{i=1}^{k} x_i^{\alpha_i - 1}$$

sur le simplex de dimension $k-1$.

Processus de Dirichlet

Soit $\alpha > 0$, G_0 une loi de probabilité sur $\mathbb{R}^d$,
on dit que P suit un processus de Dirichlet si pour toute partition $A_1, \ldots, A_m$ de $\mathbb{R}^d$,
$(P(A_1), \ldots P(A_m))$ est distribué selon la loi de Dirichlet de paramètre $(\alpha G_0(A_1), \ldots, \alpha G_0(A_m))$.

Cas discret et processus de Dirichlet

Dans le cas discret, i.e. $Y_t \in \mathbb{N}$,

- si $\pi_f = DP(\alpha G_0)^{\otimes k}$, i.e. $f_i \overset{\text{i.i.d.}}{\sim} DP(\alpha G_0)$
- si pour tout $1 \leq i \leq k$,

$$\sum_{l \in \mathbb{N}} f_i^*(l)/G_0(l) < \infty$$

alors (H1a), (H1b), (H1c) et (H1e) sont vérifiées.

Ainsi si de plus il existe $\underline{q} > 0$ tel que

- $Q^*_{i,j} \geq \underline{q}$ pour tout $1 \leq i, j \leq k$,
- $mu_i \geq \underline{q}$ pour tout $1 \leq i \leq k$,
- le support de π_Q est le k-simplex restreint aux matrices telles que $Q_{i,j} \geq \underline{q}$ pour tout $1 \leq i, j \leq k$,

alors il y a consistance de l'a posteriori pour la distance l_1 sur les densités marginales de l observations.

Et par la suite ??

- La vitesse de concentration et adaptivité : travail en cours !
- Vitesse semi-paramétrique (sur Q et sur f).
- Consistance et vitesse pour les probabilités de smoothing $P^{\theta}(X_i = j \mid Y_1, \dots Y_n)$ (consistance déjà montrée dans le cas discret, i.e. $Y_t \in \mathbb{N}$).
- Et si on ne connait pas le nombre d'états de la chaine de Markov ??...

Merci pour votre attention.