Consistance des modèles bayésiens non paramétriques de Markov cachés

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

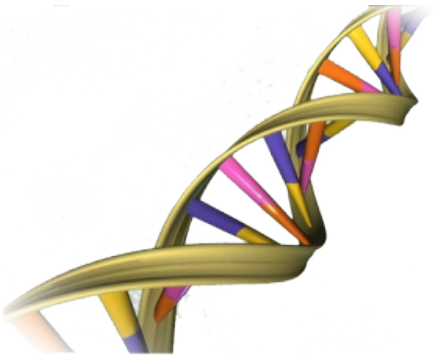
Directrices de thèse :

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

Colloque Jeunes Probabilistes et Statisticiens, jeudi 10 avril 2014

Introduction

Les modèles de Markov cachés



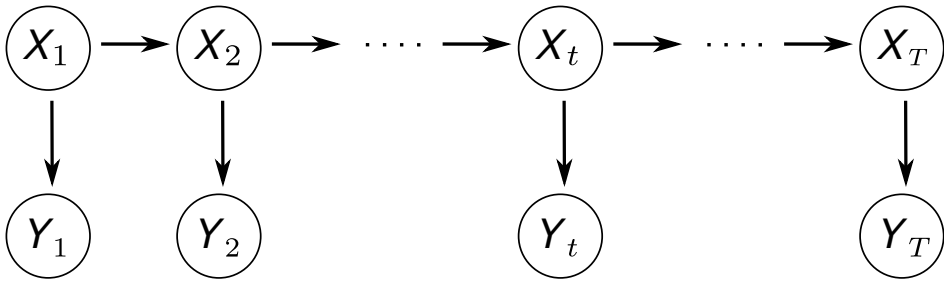
- permettent de traiter des données dépendantes
- sont très utilisés en pratique,
- leurs propriétés théoriques sont mal comprises.

But :

étudier les propriétés asymptotiques (lorsque le nombre d'observations augmente) de ces modèles dans le cadre bayésien.

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

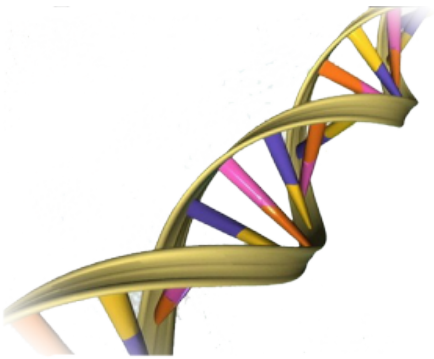
Alors $(X_t, Y_t)_{t\in\mathbb{R}_+}$ est une chaine de Markov cachée.

Exemples et utilisations

Les chaines de Markov cachés sont très utilisés en pratique en

- génomique,
- reconnaissance de parole,
- etc.

souvent pour classer des observations : suivant les états de la chaine.



t peut représenter un temps (reconnaissance de parole) ou une position (ADN).

Modèle de Markov caché étudié

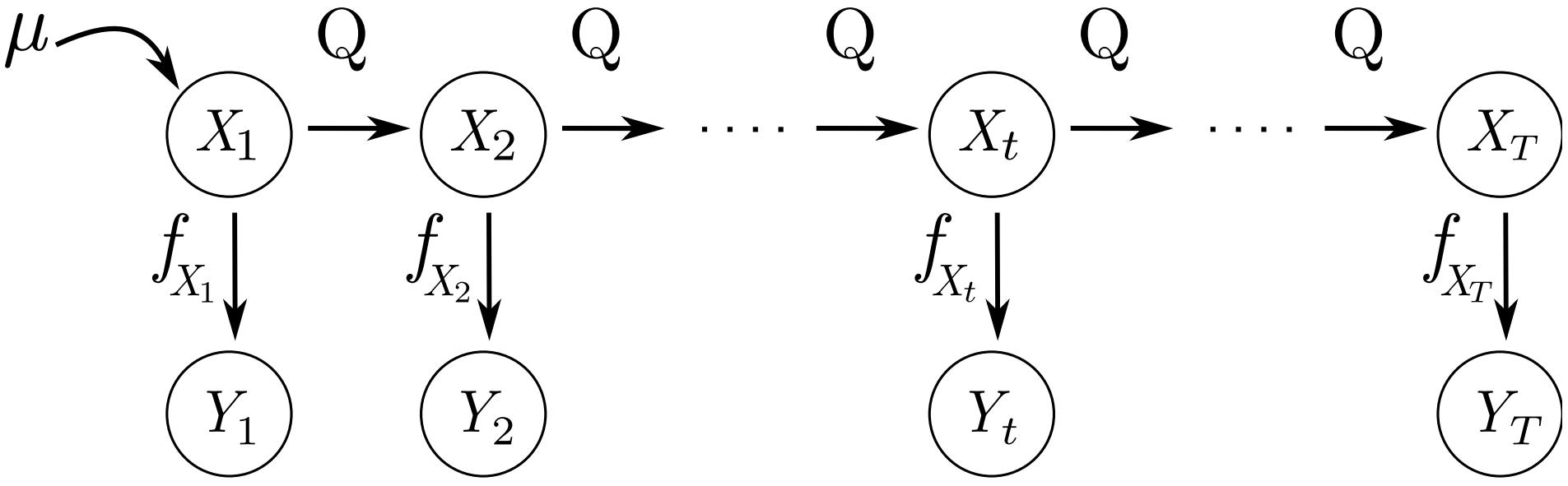
On supose qu'on a un nombre k fini et connu d'états :

$$X_t \in \{1, \ldots, k\}, \forall t$$

Paramètres du modèles :

- probabilité initiale μ, $\mu_i = P(X_1 = i)$, $i = 1, \ldots, k$,
- matrice de transition Q, $k \times k$, $Q_{i,j} = P(X_{t+1} = j | X_t = j)$,
- lois d'émission : $f = (f_1, \ldots, f_k)$, pour $1 \leq i \leq k$, f_i est la densité de Y_t sachant $X_t = i$

Schéma d'une chaine de Markov cachée



paramétrique – 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 vraissemblance (estimateur fréquentiste) pour les modèles paramétriques de Markov cachés (Douc, Matias, Moulines, Rydén).

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

Résumé

- en pratique modèles très utilisés
- en théorie des résultats en paramétrique
- mais pas de résultat théoriques en non paramétriques car l'identifiabilité n'a été montré que très récemment

Plan

Point de vue bayésien

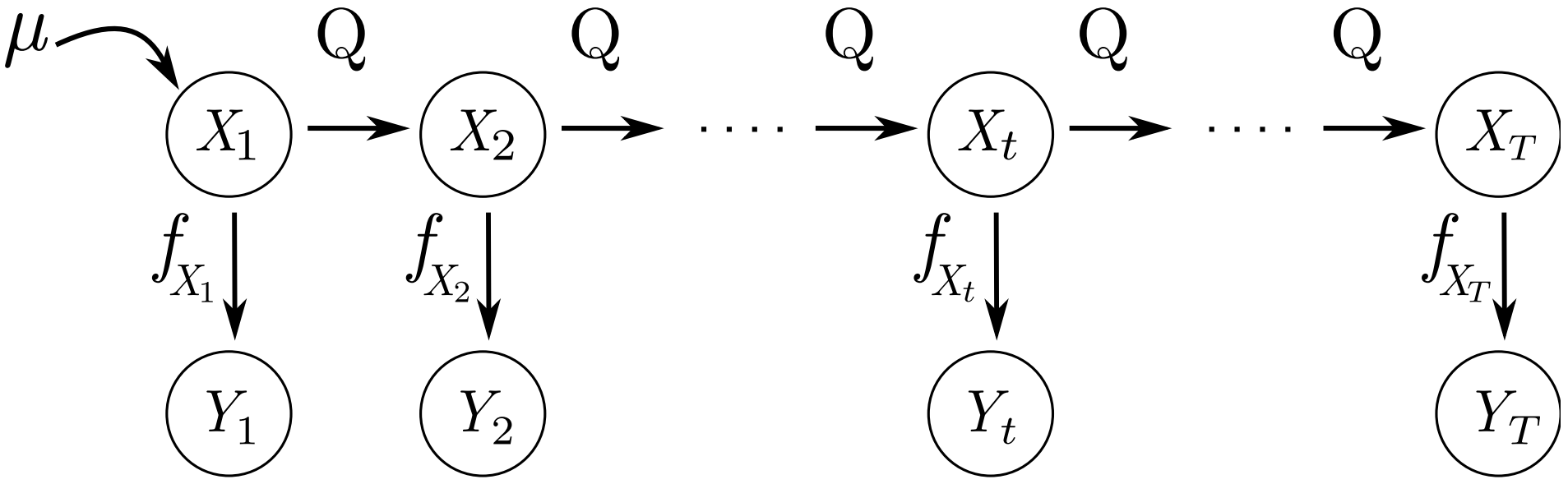
- on met une probabilité π sur l'ensemble des paramètres Θ : ce que l'on sait sur les paramètres avant de faire une expérience

  π est l'a priori

- on observe $Y_1, \ldots, Y_n$,

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

Le modèle étudié



L'a priori

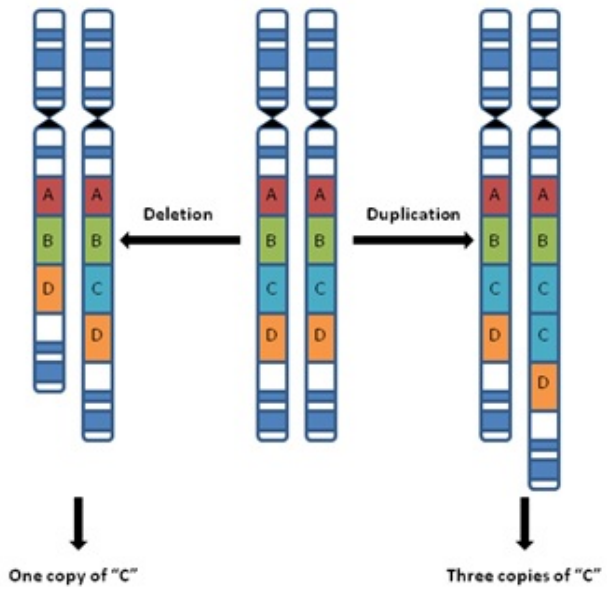
$$\pi = \mu \otimes \pi_Q \otimes \pi_f$$

- 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 $\{\text{les densités sur } \mathcal{Y}\}^k$

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).



- Il montre qu'en pratique utiliser un modèle non paramétrique bayésien fonctionne bien
- et en théorie ?

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 augmente.
- 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

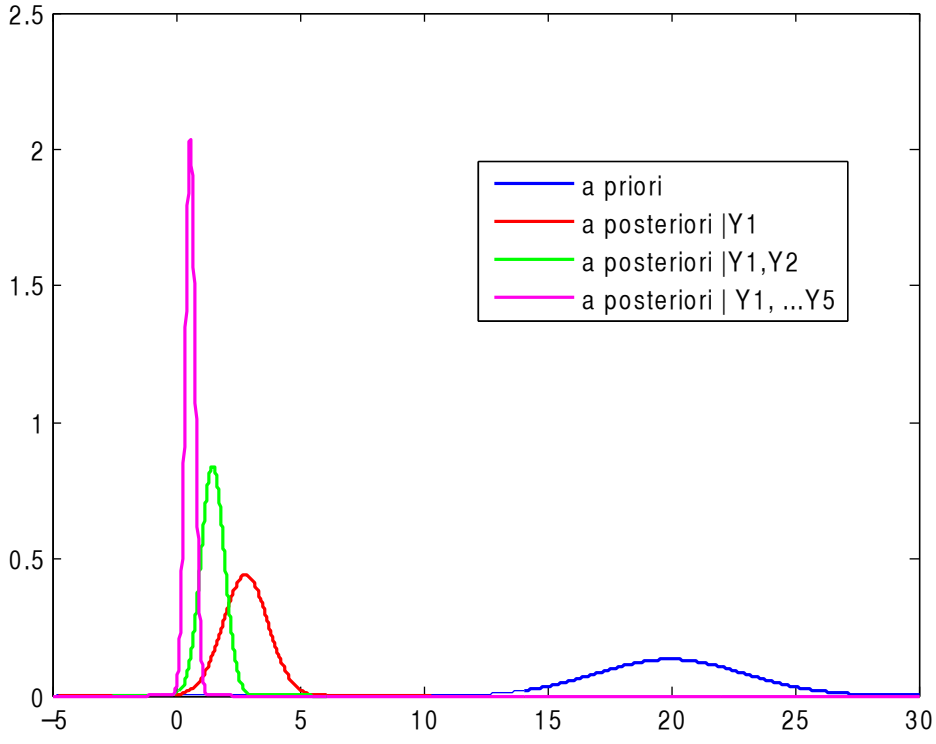
L'a posteriori est consistant en θ^* si pour tout voisinage U de θ^*, P^{θ^*}-presque sûrement

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

$\hookrightarrow$ oubli de l'a priori au profit des observations.

Exemple facile de consistance

- Modèles i.i.d. de gaussiennes à variance connue 1 et de moyenne inconnue θ,
- $\pi = \mathcal{N}(20, 9)$: probabilité sur le paramètre θ,
- $Y_i \overset{i.i.d.}{\sim} \mathcal{N}(0, 1)$, $\theta^* = 0$,



↪ L'a priori s'efface devant les observations, concentration de l'a posteriori autour du vrai paramètre $\theta^* = 0$.

Résultats de consistance bayésienne

- Doob 1949 : l'a posteriori est consistant en π-presque toute valeur de θ^*
  -> OK pour modèles paramétriques mais
- Freedman 1963 : contre-exemples en non-paramétrique.
- 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 ...).

Ces résultats ne s'appliquent pas aux modèles de Markov cachés à cause la dépendance de Y_n avec toutes les précédentes observations.

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^*,

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

- Voisinage -> nécessité de choisir une topologie sur Θ
- Résultat sur l'identifiabilité de Gassiat, Cleynen et Robin (2013) -> on peut comparer deux paramètres par la loi jointe de trois observations consécutives.
- Par exemple on peut considérer la topologie associée à la convergence étroite sur ces lois jointes, un voisinage de $P_n^{\theta^*}$ contient un ensemble de la forme :

$$\left\{ P \ : \ \left| \int h_j dP - \int h_j dP_n^{\theta^*} \right| < \epsilon_j, j = 1,\ldots,N \right\},$$

où pour tout j, $\epsilon_j > 0$ et $h_j \ : \ \mathcal{Y}^n \to \mathbb{R}$ sont continues et bornées.

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 transisiton 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) pour tout y, $\inf_{\theta \in \Theta_\epsilon} \frac{1}{k} \sum_{i=1}^{k} f_i(y) > 0$,
- (H2b) $\sup_{\theta \in \Theta_\epsilon} \sup_{y} \max_{1 \leq i \leq k} f_i(y) < \infty$
- (H2c) $\sum_{i=1}^{k} \mu_i^* \int f_i^*(y) \left| \log \left(\inf_{\theta \in \Theta_\epsilon} \frac{1}{k} \sum_{j=1}^{k} f_j(y) \right) \right| \lambda(dy) < \infty$

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.

Consistance non paramétrique des chaines de Markov cachées

J'ai montré la consistance pour différentes topologies :

- la topologie associée à la convergence étroite sur le loi jointe de trois observations consécutives
- la distance L_1 sur les densités jointes de trois observations consécutives
- une topologie produit correspondant à la topologie sur les matrices de transitions Q $\times$ la topologie faible sur les probabilités d'émission f.

Mais avec restriction sur les matrices de transistion :

relaxation possible ??

En pratique ça fonctionne dans des cas moins restrictifs mais en théorie on n'y est pas encore !!!

Et par la suite ??

- La vitesse de concentration et adaptivité : travail en cours !!!
- Et si on ne connait pas le nombre d'états de la chaine de Markov ??...

Merci pour votre attention.