Estimation adaptative non paramétrique pour les modèles à chaîne de Markov cachée

Introduction au domaine de recherche

Luc Lehéricy

11 octobre 2015

Introduction

Les modèles à chaîne de Markov cachée (ou HMM, Hidden Markov Models en anglais) ont été introduits dans les années 60, et sont utilisés depuis dans des domaines extrêmement variés tels que l'analyse du génome, la reconnaissance vocale, le traitement du signal, l'analyse d'image, l'économie, l'environnement...

Mathématiquement parlant, un HMM est un processus $(X_k, Y_k)_k$ tel que (X_k) est une chaîne de Markov et tel que conditionnellement à (X_k), les variables Y_k sont indépendantes de loi ne dépendant que de X_k.

Ces processus sont tout indiqués pour les modèles présentant une certaine cohérence spatiale ou temporelle. Typiquement, les bases du génome peuvent être modélisées comme les observations d'un HMM : les observations $(Y_k)_k$ sont les bases (A,C,G,T) et le processus caché $(X_k)_k$ est la catégorie à laquelle cette base appartient (par exemple gène codant / gène non codant), qui a de fortes chances d'être la même pour deux bases adjacentes. Ce modèle est dit paramétrique car il ne dépend que d'un nombre fini de paramètres : la loi initiale et la matrice de transition du processus caché $(X_k)_k$ et les probabilités d'émission, c'est-à-dire les probabilités conditionnelles $(\mathbb{P}(Y_k = y|X_k = x))_{x,y}$.

Ce n'est que récemment que des résultats théoriques ont été obtenus concernant l'estimation de HMM non paramétriques, autrement dit dont la loi dépend d'un nombre infini de paramètres. L'étude de ce cadre fait appel à des outils tels que la concentration de la mesure et la sélection de modèle.

Nous introduisons tout d'abord quelques notions d'estimation statistique et le principe de la sélection de modèle, puis nous les appliquons au cadre des modèles à chaîne de Markov cachée.

Table des matières

1 Fondements de l'estimation statistique **3**
1.1 Quelques définitions 3
1.2 Quels outils ? 3

2 Principe de la sélection de modèle **4**
2.1 Compromis biais-variance 5
2.2 Principe de la pénalisation 6
2.3 Notion de borne oracle 7

3 Application aux modèles à chaîne de Markov cachée **7**
3.1 Définition des modèles à chaîne de Markov cachée 8
3.2 Procédure d'estimation 9
3.3 Résultats 10

1 Fondements de l'estimation statistique

1.1 Quelques définitions

Considérons une suite de variables aléatoires $(Y_k)_{k\geq 1}$ à valeurs dans un espace $\mathcal{Y}$, de loi $\mathbb{P}$. L'objectif de l'estimation statistique est d'extraire des *observations* $(Y_k)_{k\geq 1}$ des informations sur la loi $\mathbb{P}$.

On appelle *estimateur* toute fonction mesurable sur $\mathcal{Y}^n$ pour un certain n. On appelle *famille d'estimateurs* une suite d'estimateurs $(T_n)_{n\geq 1}$ où chaque T_n est une fonction mesurable sur $\mathcal{Y}^n$, par exemple la famille des moyennes empiriques $(\frac{1}{n}\sum_{k=1}^n Y_k)_{n\geq 1}$.

On dit qu'un estimateur $\hat{g}$ *de la quantité* $g(\mathbb{P})$ est *sans biais* si $\mathbb{E}_{\mathbb{P}}(\hat{g}) = g(\mathbb{P})$. Généralement, on ne se contente pas du fait qu'un estimateur soit sans biais : on cherche à contrôler son écart par rapport à la quantité estimée.

On dit qu'un ensemble (aléatoire !) I est une *région de confiance de niveau* α de $g(\mathbb{P})$ si $\mathbb{P}(g(\mathbb{P}) \in I) \geq 1-\alpha$. Si $\mathcal{Y}$ est métrique, on choisit souvent I sous la forme d'une boule $B(\hat{g}, z)$, et la condition de niveau s'écrit alors : avec probabilité au moins $1-\alpha$, $\|\hat{g} - g(\mathbb{P})\| \leq z$.

Un *modèle* est un ensemble $\mathcal{P}$ de lois de probabilités sur $\mathcal{Y}^{\mathbb{N}}$. Lorsque $\mathcal{P}$ peut être mis en correspondance avec un sous-ensemble de $\mathbb{R}^N$, on dit que le modèle est *paramétrique*. Pour estimer la loi, on fait appel à une hypothèse de modélisation de la forme $\mathbb{P} \in \mathcal{P}$.

Exemple 1.

- $\mathcal{P} := \{\mathcal{N}(\theta, 1) | \theta \in \mathbb{R}_+\}$ est un modèle paramétrique.
- $\mathcal{P} := \{f\mathrm{d}\mathcal{L} | f \in \mathbf{L}^2(\mathbb{R}, \mathbb{R}), f \geq 0, \int f\mathrm{d}\mathcal{L} = 1\}$, où $\mathcal{L}$ désigne la mesure de Lebesgue, est un modèle non paramétrique.
  Il est possible d'approcher $\mathcal{P}$ par la suite de modèles paramétriques $(\mathcal{P}_M)_M$ avec $\mathcal{P}_M := \{f\mathrm{d}\mathcal{L} | f \in E_M, f \geq 0, \int f\mathrm{d}\mathcal{L} = 1\}$, où $(E_M)_M$ est une suite croissante de sous-espaces de dimension finie de $\mathbf{L}^2(\mathbb{R}, \mathbb{R})$ d'union dense dans $\mathbf{L}^2(\mathbb{R}, \mathbb{R})$.

1.2 Quels outils ?

Deux approches permettent d'étudier la qualité d'une famille d'estimateurs.

La première étudie le comportement asymptotique de la famille. Un exemple élémentaire est le théorème central limite :

Théorème 1 (Théorème central limite)**.** *Notons* $\overline{Y}_n = \frac{1}{n}\sum_{k=1}^n Y_k$. *Supposons que les* Y_i *sont i.i.d. et de carré intégrable, alors :*

$$\sqrt{n}(\overline{Y}_n - \mathbb{E}[\overline{Y}_n]) \xrightarrow{(loi)} \mathcal{N}(0, \mathit{Var}(Y)).$$

Ce théorème met en évidence la normalité asymptotique de la moyenne empirique sous l'hypothèse que les variables aléatoires sont de carré intégrable. Ce résultat permet ensuite de construire des intervalles de confiance asymptotiques pour $\mathbb{E}[Y]$.

La seconde approche se concentre sur des résultats non asymptotiques. Elle repose essentiellement sur la théorie de la concentration de la mesure (voir par exemple [Mas07]), qui contrôle l'écart entre une quantité empirique et sa moyenne. Une inégalité de concentration simple est l'inégalité de Hoeffding :

Théorème 2 (Inégalité de Hoeffding)**.** *Soit $(a_i)_{i\geq 1}$ et $(b_i)_{i\geq 1}$ deux suites de réels. Supposons que les Y_i sont indépendants et que pour tout $i \geq 1$, p.s., $Y_i \in [a_i; b_i]$. Alors pour tout $n \geq 1$, pour tout $t > 0$*

$$\mathbb{P}\left(\left|\overline{Y}_n - \mathbb{E}[\overline{Y}_n]\right| \geq t\right) \leq 2\exp\left(\frac{-2t^2}{\sum_{i=1}^n (b_i - a_i)^2}\right).$$

Autrement dit, pour tout $n \geq 1$, pour tout $x > 0$, on a avec probabilité $1 - 2e^{-x}$:

$$\left|\overline{Y}_n - \mathbb{E}[\overline{Y}_n]\right| \leq \sqrt{\frac{\sum_{i=1}^n (b_i - a_i)^2}{2}}\sqrt{x}.$$

Cette inégalité permet de construire des intervalles de confiance très similaires à ceux issus du théorème central limite, mais sans avoir besoin de passer à la limite. Ce sont des résultats de ce type que nous utilisons dans les méthodes d'estimation non paramétrique par sélection de modèle.

2 Principe de la sélection de modèle

Nous allons voir le principe de la sélection de modèle dans un cadre particulier qui se prête très bien aux calculs considérés : celui de la régression.

On part d'un vecteur d'observations $Y \in \mathbb{R}^N$, et on considère le modèle

$$Y = f^* + \epsilon$$

où $\epsilon \sim \mathcal{N}(0, \sigma^2 I_N)$ est un vecteur normal centré réduit, et $f^* \in \mathcal{F}$ pour un certain espace vectoriel $\mathcal{F}$.

Exemple 2. On cherche à estimer une fonction continue f^* à partir d'observations bruitées faites en N points $x_1, \ldots, x_N$. On pose Y_i l'observation faite en x_i, f_i^* la valeur $f^*(x_i)$, et $\mathcal{F}$ l'ensemble des valeurs prises par les fonctions continues sur $[0; 1]$ en les points d'observation.

L'estimateur usuel dans ce cadre est l'estimateur du maximum de vraisemblance, défini par

$$\hat{f} \in \arg\min_{f \in S} \left(-\log L(f)\right)$$

où S est un sous-espace de $\mathcal{F}$ et $L(f)$ est la vraisemblance du modèle.

Dans le cadre gaussien, la vraisemblance se réécrit directement sous la forme $-\log L(f) = \|Y - f\|_2^2/(2\sigma^2) + (N/2)\log(2\pi\sigma^2)$, et donc, en notant Π_S la projection orthogonale sur S :

$$\hat{f} = \Pi_S Y.$$

2.1 Compromis biais-variance

Définition 1. *Le risque de l'estimateur $\hat{f}$ est la quantité*

$$R(\hat{f}) = \mathbb{E}\left[\|\hat{f} - f^*\|_2^2\right].$$

Cette quantité sert de mesure de qualité : plus elle est faible, plus l'estimateur est bon. Lorsque $\hat{f} = \Pi_S Y$, le risque se décompose en deux termes :

$$\begin{aligned} R(\hat{f}) &= \mathbb{E}\left[\|f^* - \Pi_S f^* - \Pi_S \epsilon\|_2^2\right] \\ &= \|f^* - \Pi_S f^*\|_2^2 + \mathbb{E}\left[\|\Pi_S \epsilon\|_2^2\right] \\ &= \|f^* - \Pi_S f^*\|_2^2 + \sigma^2 \dim(S) \end{aligned}$$

et plus généralement

$$R(\hat{f}) = \|f^* - \Pi_S f^*\|_2^2 + \mathbb{E}\left[\|\hat{f} - \Pi_S f^*\|_2^2\right].$$

Le premier terme est appelé le biais, ou erreur d'approximation : c'est l'erreur commise quand on se restreint au modèle S. Le second est appelé la variance, ou erreur d'estimation : c'est l'erreur liée à l'estimation au sein du modèle S.

Ces deux termes ont un comportement opposé lorsqu'on modifie la taille du modèle S. En effet, prendre un plus grand modèle permet de réduire l'erreur d'approximation $\|f^* - \Pi_S f^*\|_2^2$, mais cela augmente l'erreur d'estimation, et inversement. Le risque zéro n'existe pas : un "bon" modèle doit donc réaliser un compromis entre biais et variance.

2.2 Principe de la pénalisation

Dans l'exemple 2, prendre $S = \mathcal{F}$ est peu judicieux : l'espace choisi est beaucoup trop gros. Cela n'apporte aucune contrainte sur les valeurs de f^*, ce qui amène à prendre $\hat{f} = Y$. C'est du sur-apprentissage : on colle très bien au données observées, mais on est incapable d'en prédire de nouvelles. Dans ce cas, le biais sera nul, mais la variance (qui est proportionnelle à la dimension du modèle) sera énorme, donc le risque sera grand.

Pour éviter ce problème et réaliser un bon compromis entre biais et variance, on introduit une famille $(S_M)_M$ de sous-espaces de $\mathcal{F}$, chacun correspondant à une structure possible pour f^*, et on définit pour chacun l'estimateur du maximum de vraisemblance $\hat{f}_M = \Pi_{S_M} Y$.

Exemple 3. Reprenons l'exemple 2. On peut choisir

$$S_M = \{(f(x_1), \dots, f(x_N)) | f \text{ polynôme de degré } M\}.$$

L'objectif est alors de minimiser le risque $R(\hat{f}_M)$ en M. Malheureusement, ce risque est inconnu : il faut l'estimer ou le contrôler à partir des données. Un critère permettant de le faire est le critère d'information d'Akaike (AIC).

Critère AIC. L'idée est de corriger la quantité $\|Y - \Pi_{S_M} Y\|_2^2$ pour en faire un estimateur sans biais du risque.

$$\begin{aligned}
\mathbb{E}\left[\|Y - \Pi_{S_M} Y\|_2^2\right] &= \mathbb{E}\left[\|f^* - \Pi_{S_M} f^* + \epsilon - \Pi_{S_M}\epsilon\|_2^2\right] \\
&= \|f^* - \Pi_{S_M} f^*\|_2^2 + \mathbb{E}\left[\|\epsilon - \Pi_{S_M}\epsilon\|_2^2\right] \\
&= \|f^* - \Pi_{S_M} f^*\|_2^2 + \sigma^2(N - \dim(S_M)) \\
&= R(\hat{f}_M) - 2\sigma^2 \dim(S_M) + N\sigma^2
\end{aligned}$$

donc $R(\hat{f}_M) = \mathbb{E}\left[\|Y - \Pi_{S_M} Y\|_2^2\right] + 2\sigma^2 \dim(S_M) - N\sigma^2$. On sélectionne alors l'espace d'approximation par

$$\hat{M}_{\text{AIC}} \in \arg\min_M \left\{ \|Y - \hat{f}_M\|_2^2 + 2\sigma^2 \dim(S_M) \right\}$$

et on définit l'estimateur final comme l'estimateur de l'espace sélectionné :

$$\hat{f} = \hat{f}_{\hat{M}_{\text{AIC}}}.$$

On dit qu'on a *pénalisé* la fonction de perte $\|Y - \hat{f}_M\|_2^2$ par la pénalité $2\sigma^2 \dim(S_M)$.

Bien que simple, ce critère ne marche pas toujours, en particulier quand le nombre de modèles devient grand, d'où l'intérêt de construire d'autres critères.

Cas général. Dans le cas général, on se donne une fonction de perte γ et une pénalité pen, et on considère

$$\begin{aligned}
&\hat{f}_M \in \arg\min_{t \in S_M} \gamma(t), \\
&\hat{M} \in \arg\min_M \left\{ \gamma(\hat{f}_M) + \mathrm{pen}(S_M) \right\}, \\
&\hat{f} = \hat{f}_{\hat{M}}.
\end{aligned}$$

2.3 Notion de borne oracle

Une fois une pénalité pen choisie, il reste à voir si l'estimateur qu'elle définit est de bonne qualité. Cela s'exprime la plupart du temps par une *borne oracle*, autrement dit une inégalité de la forme

$$R(\hat{f}) \leq C \inf_M \{R(P_{S_M} f^*) + \mathrm{pen}(S_M)\} + A$$

c'est-à-dire ici

$$\mathbb{E}\left[\|\hat{f} - f^*\|^2\right] \leq C \inf_M \left\{\|f^* - P_{S_M} f^*\|^2 + \mathrm{pen}(S_M)\right\} + A.$$

Obtenir une borne oracle signifie que l'on est capable de majorer le risque de l'estimateur par l'infimum de tous les biais pénalisés, à constante multiplicative et additive près. Autrement dit, on contrôle le risque de l'estimateur par le meilleur risque qu'il est possible d'obtenir, ce qui garantit que cet estimateur est toujours raisonnablement bon.

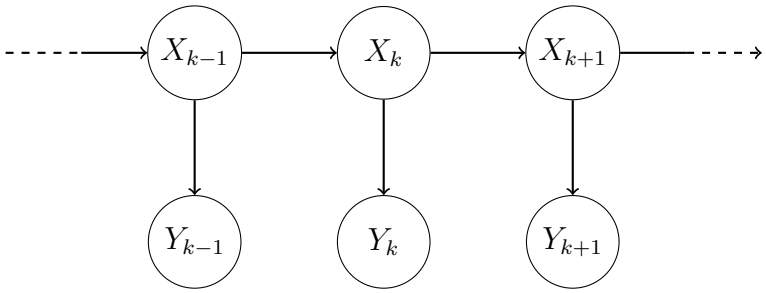
3 Application aux modèles à chaîne de Markov cachée

Nous appliquons la méthode vue à la section précédente aux modèles à chaîne de Markov cachée. Dans ce cadre, il est possible de construire par pénalisation un estimateur qui satisfait une borne oracle, pour une pénalité que nous précisons. La plupart des idées de cette partie proviennent de [dCGL15].

3.1 Définition des modèles à chaîne de Markov cachée

Définition 2. *Un modèle à chaîne de Markov cachée est un processus* $(X_k, Y_k)_{k\geq 1} \in (\mathcal{X} \times \mathcal{Y})^{\mathbb{N}}$ *tel que*

- $(X_k)_{k\geq 1}$ *est une chaîne de Markov à valeurs dans un espace* $\mathcal{X}$ *fini, autrement dit la loi de* X_k *conditionnellement à* $(X_1, \ldots, X_{k-1})$ *est égale à la loi de* X_k *conditionnellement à* X_{k-1}*,*
- *conditionnellement à* $(X_k)_{k\geq 1}$*,* $(Y_k)_{k\geq 1}$ *sont des variables indépendantes ;*
- *pour tout* $k \geq 1$*, la loi de* Y_k *conditionnellement à* $(X_{k'})_{k'\geq 1}$ *est égale à la loi de* Y_k *conditionnellement à* X_k*.*



On peut exprimer la loi de Y sachant X à l'aide d'un noyau markovien Q. La famille $(Q(x, \cdot))_{x\in\mathcal{X}}$ est la famille des lois d'émission du HMM :

$$\mathbb{P}(Y_k \in A | (X_{k'})_{k'\geq 1}) = Q(X_k, A).$$

Rappelons qu'un noyau markovien Q sur $(E, \mathcal{E}) \times (F, \mathcal{F})$ est une application $(E, \mathcal{F}) \to \mathbb{R}$ telle que

- $\forall A \in \mathcal{F}, \quad Q(\cdot, A)$ est une fonction mesurable $E \to \mathbb{R}$,
- $\forall x \in E, \quad Q(x, \cdot)$ est une probabilité sur F.

Paramètres du modèle. Les paramètres du HMM $(X_k, Y_k)_{k\geq 1}$ sont :

- la loi de transition de la chaîne de Markov $(X_k)_{k\geq 1}$, c'est-à-dire la matrice $\mathbf{Q}^*$ définie par $\mathbf{Q}^*(i, j) = \mathbb{P}(X_{k+1} = j | X_k = i)$
- sa loi initiale π^* définie par $\pi^*(i) = \mathbb{P}(X_1 = i)$
- ses lois d'émission.

On cherche à les estimer en n'ayant accès qu'aux observations $(Y_k)_k$.

Nous supposons toujours dans la suite que les lois d'émission sont absolument continues par rapport à la mesure de Lebesgue, et on note $f_i^* = \dfrac{dQ(i, \cdot)}{d\mathrm{Leb}}$ leurs densités, qu'on appellera densités d'émission.

3.2 Procédure d'estimation

Identifiabilité du modèle Nous supposons les observations générées par un HMM $(X_k, Y_k)_k$ à K^* états cachés de paramètres $(\pi^*, \mathbf{Q}^*, \mathbf{f}^*)$ vérifiant les points suivants :

- la matrice de transition $\mathbf{Q}^*$ est inversible,
- la chaîne de Markov $(X_k)_k$ est irréductible apériodique,
- la famille des lois d'émission $\mathbf{f}^* := (f_1^*, \ldots, f_{K^*}^*)$ est libre.

Alors la donnée de la loi des trois premières observations (Y_1, Y_2, Y_3) permet de retrouver les paramètres du HMM à permutation des états cachés près (voir par exemple [GCR13]). Nous allons donc estimer la densité g^* des trois premières observations par rapport à la mesure de Lebesgue.

Construction de l'estimateur des moindres carrés pénalisés.

L'objectif de l'estimation est de trouver une fonction t qui minimise la distance $\mathbf{L}^2$ par rapport à la vraie densité, autrement dit qui minimise $\|t - g^*\|_2^2 = \|g^*\|_2^2 + \|t\|_2^2 - 2\mathbb{E}(Z)$ où le triplet $Z = (Y_1, Y_2, Y_3)$ est généré suivant la densité g^*.

Pour cela, nous introduisons la fonction de perte

$$\gamma_N : t \longmapsto \|t\|_2^2 - \frac{2}{N}\sum_{s=1}^{N} t(Z^{(s)}).$$

Notons que $\mathbb{E}[\gamma_N(t)] = \|t - g^*\|_2^2 - \|g^*\|_2^2$, ce qui justifie son usage comme critère empirique pour sélectionner l'estimateur.

Enfin, nous décomposons l'espace $\mathbf{L}^2([0,1])$ (de dimension infinie) dans lequel nous estimons les densités en une suite croissante d'espaces de dimension finie $(S_M)_M$ d'union dense dans $\mathbf{L}^2([0,1])$.

Notre hypothèse de simulation sera $\mathbf{f}^* \in \mathcal{F}^{K^*}$ où $\mathcal{F}$ vérifie :

- $\mathcal{F}$ est un sous-ensemble fermé borné de $\mathbf{L}^2([0,1]^D, \mathbb{R}_+)$ tel qu'il existe une constante $C_{\mathcal{F},\infty}$ telle que pour tout $f \in \mathcal{F}$, $\int f \mathrm{d}\mathcal{L}^D = 1$ et $\|f\|_\infty \leq C_{\mathcal{F},\infty}$, où $\mathcal{L}^D$ désigne la mesure de Lebesgue sur $[0,1]^D$.
- $\mathbf{f}^* \in \mathcal{F}^{K^*}$ et $\mathcal{F}$ est stable par projection sur $\mathfrak{P}_M$ pour tout $M \in \mathcal{M}$

Nous pouvons alors construire une suite d'estimateurs $\hat{g}_M$ et de projetés g^*_M définis pour chaque espace par

$$\begin{cases} g^*_M \in \underset{t \in \mathrm{HMM}(S_M)}{\arg\min} \|t - g^*\|_2^2 \\ \hat{g}_M \in \underset{t \in \mathrm{HMM}(S_M)}{\arg\min} \gamma_N(t) \end{cases}$$

où on note $\mathrm{HMM}(S_M)$ l'ensemble des densités des trois premières observations d'un HMM à K états cachés de densités d'émission dans $S_M \cap \mathcal{F}$.

Enfin, on pose $\hat{g} = \hat{g}_{\hat{M}}$ l'estimateur des moindres carrés pénalisés, avec

$$\hat{M} \in \underset{M}{\arg\min} \{\gamma_N(\hat{g}_M) + \mathrm{pen}(N, M)\}$$

pour une fonction de pénalisation pen judicieusement choisie.

3.3 Résultats

La question que l'on est alors amené à se poser est : comment choisir la pénalité pour obtenir des garanties théoriques sur l'erreur commise ?

Théorème 3 (Borne oracle)**.** *Il existe des constantes N_0, ρ^* et A telles que si*

$$\forall N, M, \quad pen(N, M) \geq \rho^* M \frac{\log(N)}{N}$$

alors pour tout $N \geq N_0$, pour tout $x > 0$, on a avec probabilité $1 - (e-1)^{-2} e^{-x}$:

$$\|\hat{g}_{\hat{M}} - g^*\|_2^2 \leq 4 \inf_M \left\{ \|g^*_M - g^*\|_2^2 + pen(N, M) \right\} + 4A \frac{x}{N}.$$

Cette méthode d'estimation a l'avantage de pouvoir être aisément mise en pratique. La figure 1 représente les estimateurs des densités de HMM à 3 états cachés, obtenues à partir de 3 000 observations.

Il reste encore un certain nombre de questions ouvertes à résoudre. Par exemple, [dCGL15] fournit un lemme algébrique permettant de contrôler l'erreur commise sur les paramètres du HMM (et non plus sur la densité de 3 observations), mais celui-ci n'est démontré que pour les HMM à 2 états cachés. Dans une autre direction, sélectionner le nombre d'états cachés du HMM en plus des autres paramètres fait surgir des questions telles que l'impact de la séparation des densités sur la qualité de l'estimation, qui affecte également une autre méthode reposant quant à elle uniquement sur des considérations algébriques : l'estimation spectrale (voir par exemple [HKZ12]).

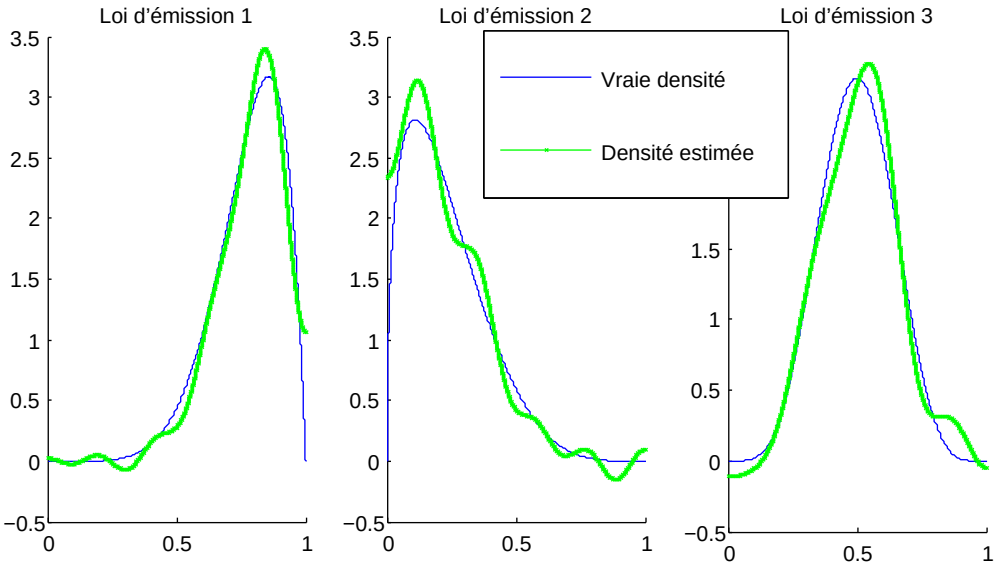


FIGURE 1 – Densités d'émission (vraies et estimées) pour un HMM à trois états cachés. On est parti de $N = 3\,000$ observations.

Références

[dCGL15] Yohann de Castro, Elisabeth Gassiat, and Claire Lacour. Minimax adaptative estimation of non-parametric hidden markov models. Soumis, 2015.

[GCR13] Elisabeth Gassiat, Alice Cleynen, and Stéphane Robin. Finite state space non parametric hidden markov models are in general identifiable. *arXiv preprint arXiv :1306.4657*, 2013.

[HKZ12] Daniel Hsu, Sham M Kakade, and Tong Zhang. A spectral algorithm for learning hidden markov models. *Journal of Computer and System Sciences*, 78(5) :1460–1480, 2012.

[Mas07] Pascal Massart. Concentration inequalities and model selection. In *Lecture Notes in Mathematics*, volume 1896. Springer, Berlin, 2007.