Séminaire Point de Vue

Bayésien non-paramétrique, quelques propriétés fréquentistes

Ismaël Castillo

CNRS, LPMA Paris

Paris, 15 novembre 2010

PREMIÈRE PARTIE

Introduction

Cadre bayésien

- Observations $\mathbf{X}$
  $X = X^{(n)} = (X_1, \ldots, X_n)$
- Modèle pour les observations $\mathbf{X}|\,\theta$
  Famille de lois $p_\theta(\mathbf{X})$
- Modèle pour le paramètre θ loi a priori Π
  $\theta \sim \Pi$ probabilité

Cadre bayésien

- Observations $\mathbf{X}$
  $X = X^{(n)} = (X_1, \ldots, X_n)$
- Modèle pour les observations $\mathbf{X}|\,\theta$
  Famille de lois $p_\theta(\mathbf{X})$
- Modèle pour le paramètre θ loi a priori Π
  $\theta \sim \Pi$ probabilité

L'estimateur bayésien est la loi conditionnelle $\theta|\,\mathbf{X}$ loi a posteriori $\Pi(\cdot|\,\mathbf{X})$

L'objet principal d'intérêt sera ici $\Pi(\cdot|\,\mathbf{X})$

- On peut aussi s'intéresser à des aspects la loi a posteriori
  - Moyenne a posteriori $\int \theta d\Pi(\theta|\,\mathbf{X})$
  - Médiane a posteriori
  - Maximum a posteriori, etc.

Cadre bayésien, points de vue

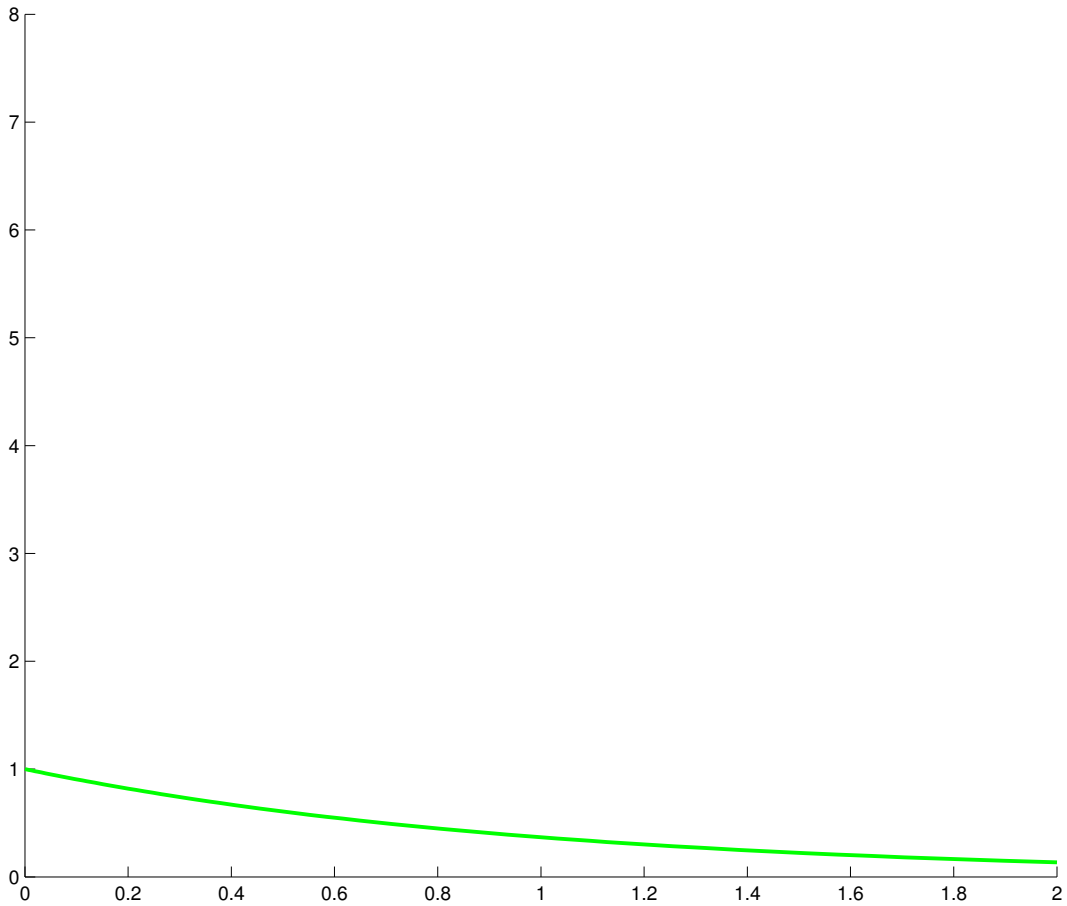
Différents points de vue

Fréquentiste	Bayésien I	Bayésien II
Modèle $p_\theta(\mathbf{X})$	Modèle $p_\theta(\mathbf{X})$	Modèle $p_\theta(\mathbf{X})$
Estimateur $\hat{\theta}(\mathbf{X})$	A priori $\theta \sim \Pi$ A posteriori $\theta \vert \mathbf{X}$	A priori $\theta \sim \Pi$ A posteriori $\theta \vert \mathbf{X}$
Il existe un "vrai" θ_0 Etude de $\hat{\theta}(\mathbf{X})$ sous $\mathbf{P}_{\theta_0}$	Il existe un "vrai" θ_0 Etude de $\theta \vert \mathbf{X}$ sous $\mathbf{P}_{\theta_0}$	Pas de "vrai" paramètre Tout est variable aléatoire

Nous étudions ici mathématiquement $\theta \vert \mathbf{X}$ sous $\mathbf{P}_{\theta_0}$

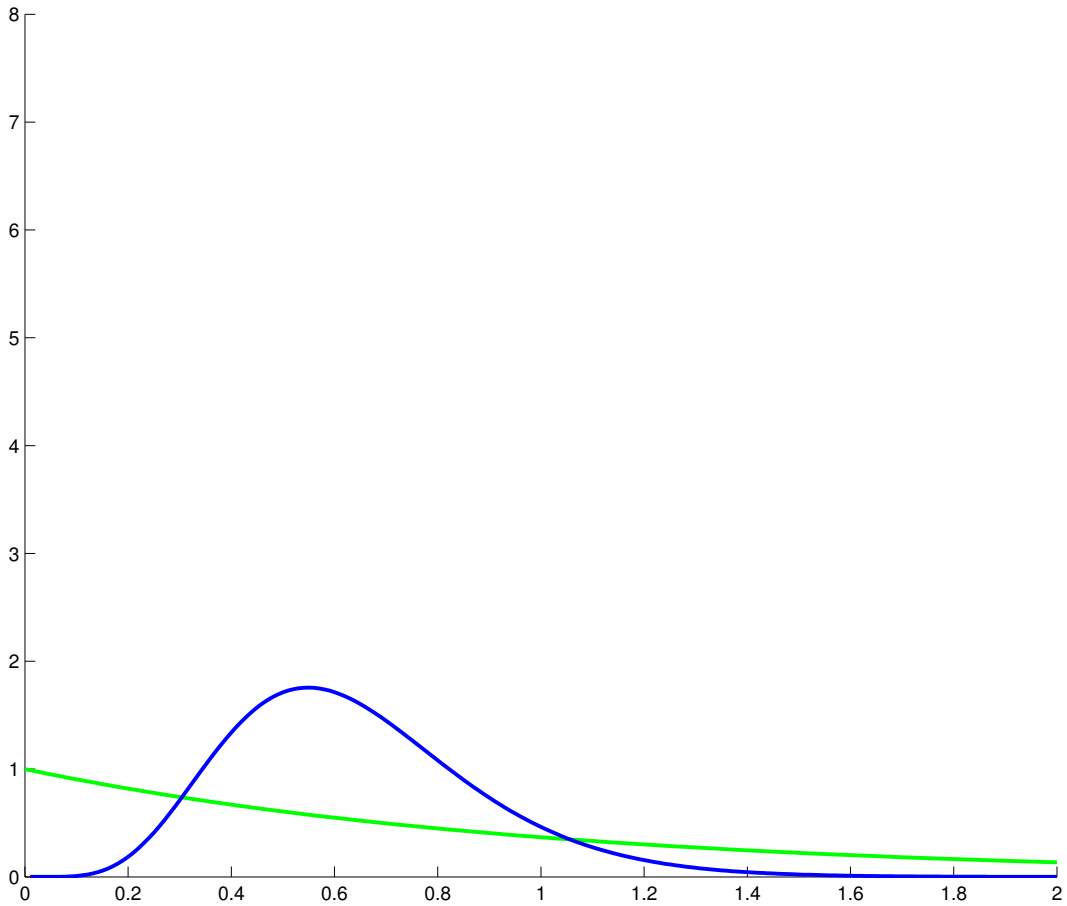
Si on suppose qu'il existe un vrai θ_0, peut-on dire que la loi a posteriori converge, en un certain sens, vers ce vrai paramètre ?

Exemple : Estimation de θ d'une loi $\Gamma(2, \theta^{-1})$



Densité a priori $\mathcal{E}(1)$

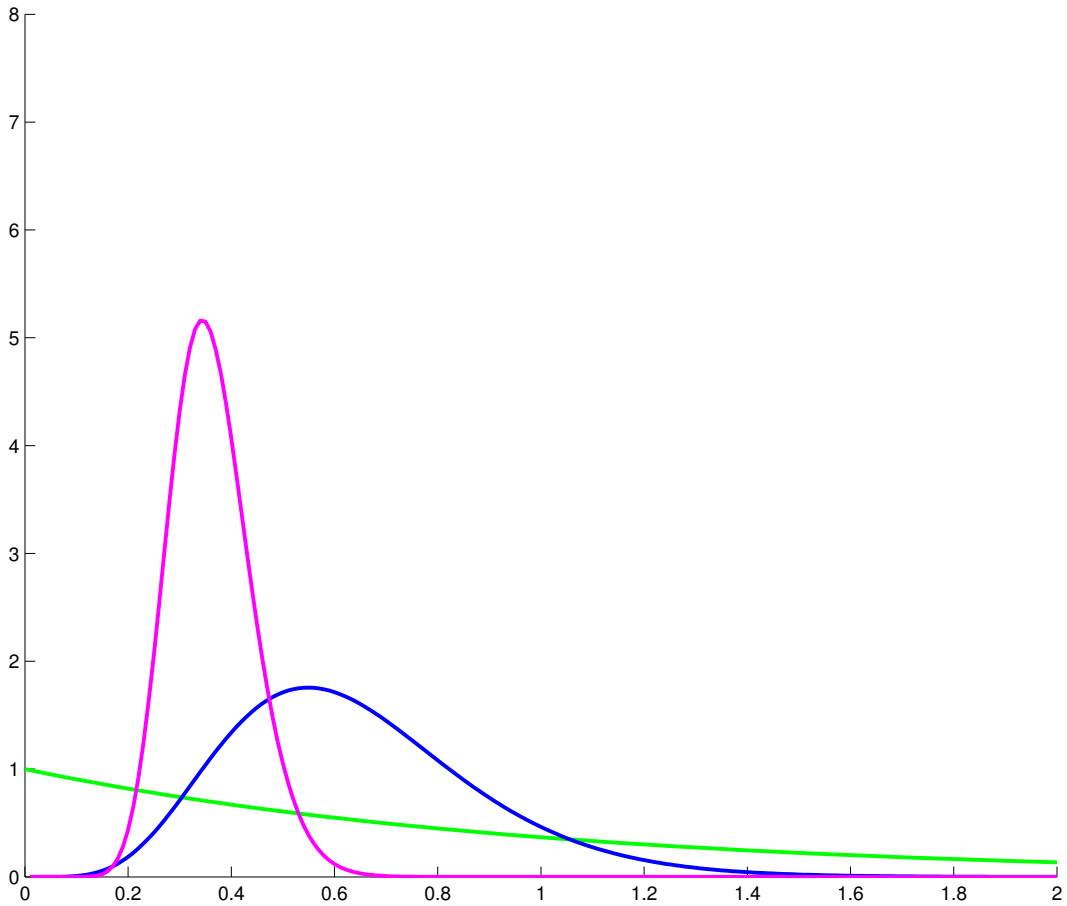
Exemple : Estimation de θ d'une loi $\Gamma(2, \theta^{-1})$



Données
X_1, X_2, X_3

Densité a posteriori pour $n = 3$

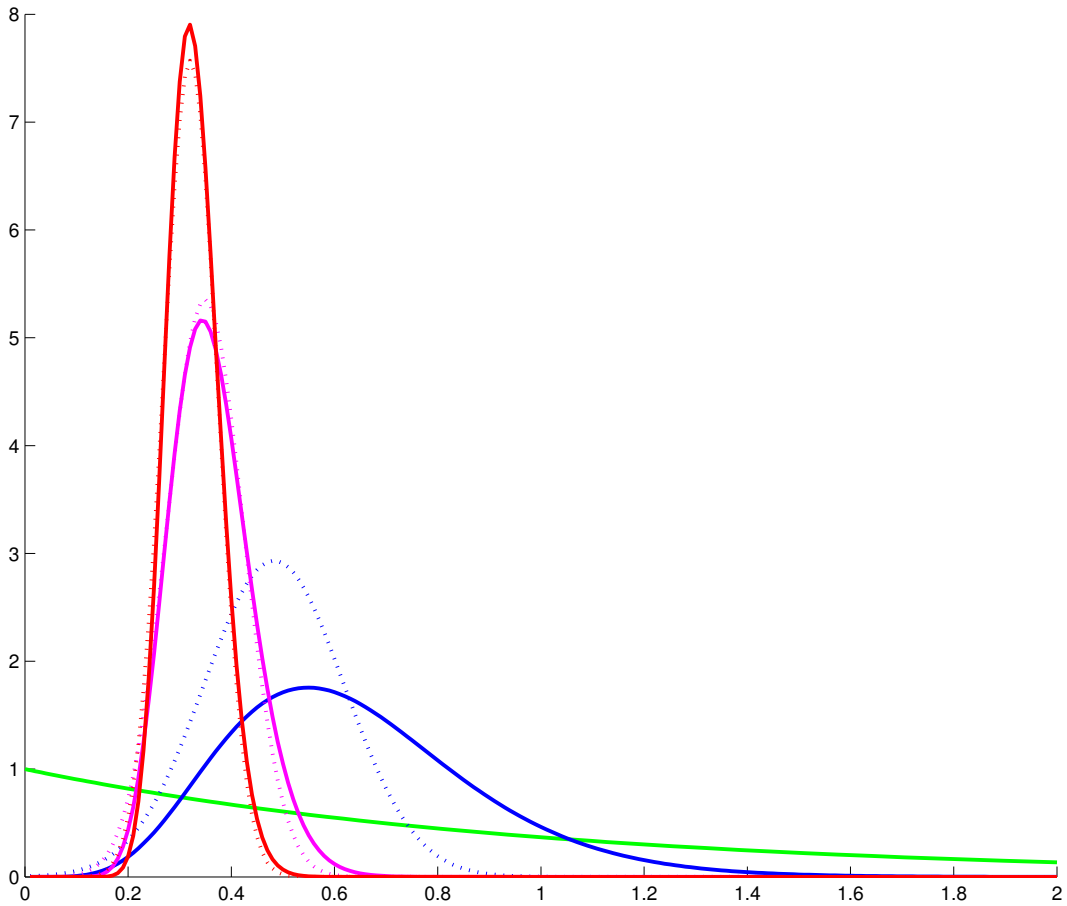
Exemple : Estimation de θ d'une loi $\Gamma(2, \theta^{-1})$



Données
$X_1, \ldots, X_{10}$

Densité a posteriori pour $n = 10$

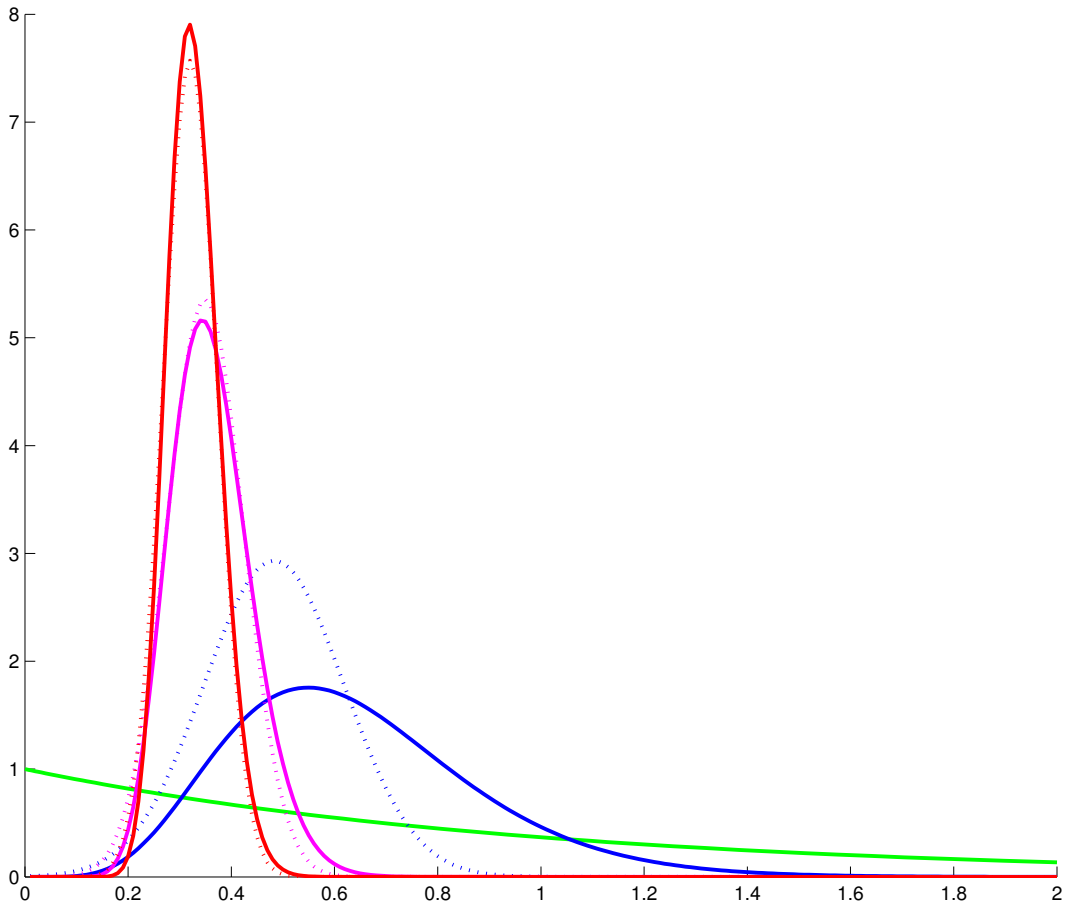
Exemple : Estimation de θ d'une loi $\Gamma(2, \theta^{-1})$



Données
$X_1, \ldots, X_{20}$

Densité a posteriori pour $n = 20$

Exemple : Estimation de θ d'une loi $\Gamma(2, \theta^{-1})$



Données
$X_1, \ldots, X_{20}$

Densité a posteriori pour $n = 20$

$\theta_0 = 1/3$

Modèles non-paramétriques

Exemple Bruit blanc Gaussien

Observation $dX^{(n)}(t) = f(t)dt + \frac{1}{\sqrt{n}}dB(t), \quad t \in [0,1],$
avec $f \in L^2[0,1]$ et B mouvement Brownien

Exemple Estimation de densité

Observations $X^{(n)} = (X_1, \ldots, X_n)$, i.i.d. densité f sur $[0,1]$

Sous les contraintes $f \geq 0$ et $\int_0^1 f(u)du = 1.$

Exemple Fonction de répartition, etc.

Question : A priori sur f ? "tirer au hasard une fonction"

Exemples d'a priori

- A priori sur les fonctions continues. Loi d'un processus X dont les trajectoires sont p.s. dans $\mathcal{C}^0[0,1]$

$$(B(s))_{0\leq s\leq 1} \quad \text{mvt Brownien}$$

- A priori sur les fonctions. Développement en série
  Soit $\{\varepsilon_k\}$ base orthonormale de $L^2[0,1]$

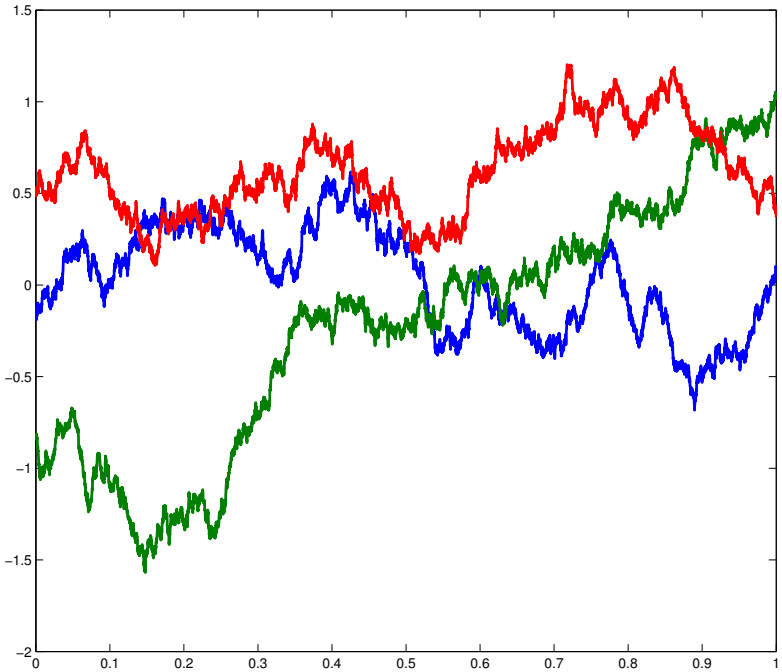
$$X(t) = \sum_{k\geq 1} \gamma_k \alpha_k \varepsilon_k(t),$$

  - $\{\gamma_k\}_{k\geq 1}$ suite dans ℓ^2,
  - $\alpha_k \sim \mathcal{N}(0,1)$ iid (ou $Laplace(1)$...)

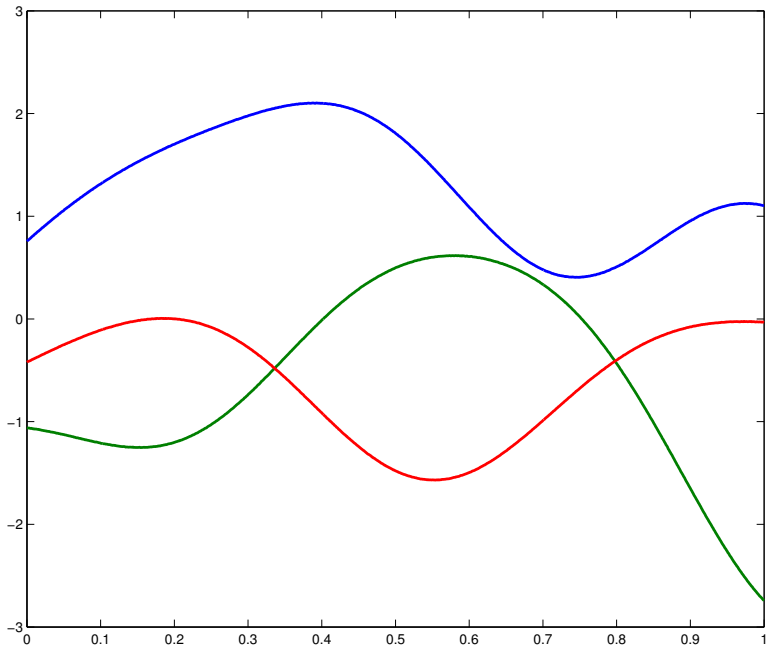
- A priori sur les densités via *normalisation*

$$p_X(u) = \frac{e^{X(u)}}{\int_0^1 e^{X(v)}dv}.$$

A priori gaussiens, exemples



$Z_t = B_t + N$, où $B_t =$ Mvt. Brownien et $N \sim \mathcal{N}(0,1)$.



Processus Gaussien Z centré
$\mathbf{E}(Z_s Z_t) = \exp(-(s-t)^2/L)$

A priori à sauts, processus de Dirichlet

Processus de Dirichlet $DP(\alpha, G_0)$ sur $\mathbb{R}$

G_0 proba sur $\mathbb{R}$ loi moyenne et $\alpha > 0$ paramètre de concentration

Il existe une mesure de probabilité aléatoire sur $\mathbb{R}$ telle que pour toute partition finie $(B_1, \ldots, B_r)$ de $\mathbb{R}$,

$$(G(B_1), \ldots, G(B_r)) \sim Dir(\alpha G_0(B_1), \ldots, \alpha G_0(B_r))$$

[Ferguson 1973]

A priori à sauts, processus de Dirichlet

$G \sim DP(\cdot|\, G_0, \alpha)$ A quoi cela ressemble-t-il ?

G est p.s. une loi discrète

On a la représentation [Sethuraman 94]

$$G = \sum_{k=1}^{\infty} \pi_k \delta_{\theta_k}$$

- $\theta_k \sim G_0(\cdot)$ i.i.d.
- π_k poids donnés par stickbreaking
  - $V_1, V_2, \ldots \sim \text{Beta}(1, \alpha)$ i.i.d.
  - $\pi_k = V_k \prod_{i=1}^{k-1}(1 - V_i)$

A priori à sauts, processus de Dirichlet

$G \sim DP(\cdot|\, G_0, \alpha)$ A quoi cela ressemble-t-il ?

G est p.s. une loi discrète

On a la représentation [Sethuraman 94]

$$G = \sum_{k=1}^{\infty} \pi_k \delta_{\theta_k}$$

- $\theta_k \sim G_0(\cdot)$ i.i.d.
- π_k poids donnés par stickbreaking
  - $V_1, V_2, \ldots \sim \text{Beta}(1, \alpha)$ i.i.d.
  - $\pi_k = V_k \prod_{i=1}^{k-1}(1 - V_i)$

Processus de Pitman-Yor $\text{PY}(\alpha, d, G_0)$

Même représentation mais cette fois $V_k \sim \text{Beta}(1 - d, \alpha + kd)$

A priori à sauts, hiérarchies

Exemple d'a priori à sauts hiérarchiques

Hierarchical Dirichlet process (HDP) [Teh, Jordan, Beal, Blei (2005)]

Supposons que les données soient divisées en J groupes

$$G_0 | H, \gamma \sim DP(\cdot | H, \gamma)$$

G_0 sert ensuite lui-même de mesure de base : pour $j = 1, \dots, J$,

$$G_j | G_0, \alpha \sim DP(\cdot | G_0, \alpha)$$

Simulations : quelques mots

Le but en pratique est de pouvoir simuler une variable qui suit le loi a posteriori

- Dans certains cas (rares), c'est possible dire directement (de façon exacte)
- Les méthodes MCMC = Markov Chain Monte Carlo ont pour but de construire une chaîne de Markov $\theta_1, \theta_2, \ldots$ dont la loi stationnaire est la loi a posteriori

  Depuis une quinzaine d'année, de nombreux algorithmes de ce type ont été proposés

  C'est un champ très actif de recherches (rapidité de simulation, rapidité d'approximation de la loi stationnaire ...)

Parfois, on se contente de simuler suivant un aspect de l'a posteriori (moyenne et/ou variance par exemple)

Notions de convergence

Cadre bayésien dominé, cas paramétrique i.i.d.

Observations. $X^{(n)} = (X_1, \dots, X_n)$ i.i.d. de loi $dP_\theta = p_\theta d\mu$.

$\Theta \subset \mathbb{R}$ (ou $\mathbb{R}^k$).

Cadre Bayésien. *A priori* sur θ, $d\pi(\theta) = \lambda(\theta)d\theta$.
Cette mesure est mise à jour avec les données $X^{(n)}$.

L'*a posteriori* sachant $X^{(n)}$ est la loi conditionnelle $\pi(\cdot|X^{(n)})$.

Formule de Bayes. Pour tout B mesurable,

$$\pi(B|X^{(n)}) = \frac{\int_B \prod_{i=1}^n p_\theta(X_i) d\pi(\theta)}{\int \prod_{i=1}^n p_\theta(X_i) d\pi(\theta)}.$$

Cadre bayésien dominé

Observations. $X^{(n)}$ de loi $P_\eta^{(n)}$ indexée par η
η appartient à un espace de paramètres (potentiellement grand)

Densité. Mesure dominante $\mu^{(n)}$,

$$dP_\eta^{(n)} = p_\eta(X^{(n)})d\mu^{(n)}$$

Cadre bayésien. *A priori* Π sur η
Cette mesure est mise à jour avec les données $X^{(n)}$.

L' *a posteriori* sachant $X^{(n)}$ est la loi conditionnelle $\Pi(\cdot|X^{(n)})$.

Formule de Bayes. Pour tout ensemble mesurable B,

$$\Pi(B|X^{(n)}) = \frac{\int_B p_\eta^{(n)}(X^{(n)})d\Pi(\eta)}{\int p_\eta^{(n)}(X^{(n)})d\Pi(\eta)}.$$

Consistance

Consistance L'a posteriori est consistant en η_0 si, quand $n \longrightarrow \infty$,

$$\Pi(\cdot|X^{(n)}) \xrightarrow{w} \delta_{\eta_0}, \quad \text{en } \mathbf{P}_{\eta_0}^{(n)}\text{-proba.}$$

ou encore, dans un espace métrique séparable muni d'une distance d, pour tout $\varepsilon > 0$,

$$\Pi(\eta : \ d(\eta, \eta_0) < \varepsilon | \, X^{(n)}) \longrightarrow 1, \quad \text{en } \mathbf{P}_{\eta_0}^{(n)}\text{-proba.}$$

Résultats généraux

- [Doob (1949)]
- [Schwartz (1965)]
- [Barron, Schervish, Wasserman (1999)]

Vitesses de convergence

L'a posteriori converge à vitesse $\varepsilon_n \to 0$ pour la distance d en η_0 si

$$\mathbf{E}_{\eta_0}\Pi(\eta : d(\eta, \eta_0) \leq \varepsilon_n | X^{(n)}) \longrightarrow 1$$

C'est une borne supérieure : on cherche ε_n le plus petit possible

On dit que ζ_n est une borne inférieure pour la vitesse pour d en η_0 si

$$\mathbf{E}_{\eta_0}\Pi(\eta : d(\eta, \eta_0) \geq \zeta_n | X^{(n)}) \longrightarrow 1$$

Exemple Modèles paramétriques "réguliers" Pour tout $M_n \to +\infty$,

$$\mathbf{E}_{\theta_0}\pi(\frac{1}{M_n\sqrt{n}} \leq |\theta - \theta_0| \leq \frac{M_n}{\sqrt{n}} | X^{(n)}) \to 1$$

Que se passe-t-il dans un cadre non-paramétrique ?

Forme de l'a posteriori

Encore plus précisément, on peut s'intéresser à la forme de la loi a posteriori

Le théorème de Bernstein-von Mises (cadre paramétrique) est un exemple de tel résultat

Theorem ([Bernstein-von Mises])

Sous des hypothèses adéquates de régularité,

$$\left\| \pi(\cdot|X^{(n)}) - N\left(\hat{\theta}, \frac{\mathcal{I}_{\theta_0}^{-1}}{n}\right)(\cdot) \right\| \underset{n\to+\infty}{\longrightarrow} 0 \quad \text{sous } P_{\theta_0}^{(n)},$$

où $\mathcal{I}_{\theta_0}$ information de Fisher et $\hat{\theta}$ estimateur efficace de θ_0 et $\|\cdot\|$ distance en variation totale

On remarque que l'a priori est asymptotiquement "effacé" de la loi a posteriori

Objectifs et résultats

Objectifs

- Comprendre la comportement de l'a posteriori sous $\mathbf{P}_{\eta_0}$
- Trouver des classes d'a priori Π pour lesquels l'a posteriori se "comporte bien"

Résultats

Nous allons voir que, pour des problèmes non-paramétriques/de grande dimension, le choix de l'a priori est crucial

- l'a posteriori peut être inconsistant
- la vitesse de convergence dépend souvent de l'a priori
- pour l'estimation de fonctionnelles (cadre semi-paramétrique), le choix de l'a priori joue encore un rôle

Exemples

Exemple *Estimation de densité et a priori gaussiens*

[van der Vaart, van Zanten 2008, 2009]

Observations $X^{(n)} = (X_1, \dots, X_n)$, i.i.d. densité $f_0 > 0$ dans une classe de Hölder $\mathcal{C}^\beta[0,1]$

Soit $(W_t)_{t\in[0,1]}$ processus Gaussien centré de noyau de covariance $K(s,t) = \mathbf{E}(W_s W_t)$
Exemple Mouvement brownien $K(s,t) = s \wedge t$

A priori Π = loi engendrée par

$$\Pi \sim \frac{e^{W_t}}{\int_0^1 e^{W_u}\, du}$$

Distance : h distance de Hellinger

On veut montrer que, pour M assez grand,

$$\mathbf{E}_{f_0}\Pi(h(f, f_0) \leq M\varepsilon_n |\, X^{(n)}) \to 1 \qquad (n \to \infty)$$

C'est vrai pour les exemples suivants

Exemple *Estimation de densité et a priori gaussiens*

Mouvement brownien + gaussienne

$W_t = B_t + Z_0$, with $Z_0 \sim \mathcal{N}(0,1)$

$$\varepsilon_n = n^{-\frac{1}{4} \wedge \frac{\beta}{2}} = \begin{cases} n^{-1/4} & \text{si } \beta \geq 1/2 \\ n^{-\beta/2} & \text{si } \beta \leq 1/2 \end{cases}$$

Processus de Riemann-Liouville de paramètre $\alpha > 0$

$W_t = \int_0^t (t-s)^{\alpha - 1/2} dB_s + \sum_{k=0}^{\lceil \alpha \rceil} Z_k t^k$, avec $Z_k \sim \mathcal{N}(0,1)$ iid

$$\varepsilon_n \approx n^{-\frac{\alpha \wedge \beta}{2\alpha+1}} = \begin{cases} n^{-\frac{\alpha}{2\alpha+1}} & \text{si } \beta \geq \alpha \\ n^{-\frac{\beta}{2\alpha+1}} & \text{si } \beta \leq \alpha \end{cases}$$

De plus, ces vitesses sont optimales sur les classe de Hölder : bornes inférieures correspondantes [Castillo 2008]

Peut-on obtenir un résultat adaptatif ?

Exemple *Estimation de densité et a priori gaussiens*

Noyau de covariance à forme gaussienne

Processus gaussien Z_t centré de noyau de covariance

$$\mathbf{E}(Z_t Z_s) = e^{-(s-t)^2/L}$$

Dans ce cas, [van der Vaart, van Zanten (2010)] montrent qu'à L fixé, la vitesse de convergence associée est (au mieux) logarithmique

$$\varepsilon_n \approx (\log n)^{-\gamma(\beta)}$$

Exemple *Estimation de densité et a priori gaussiens*

Noyau de covariance à forme gaussienne

Processus gaussien Z_t centré de noyau de covariance

$$\mathbf{E}(Z_t Z_s) = e^{-(s-t)^2/L}$$

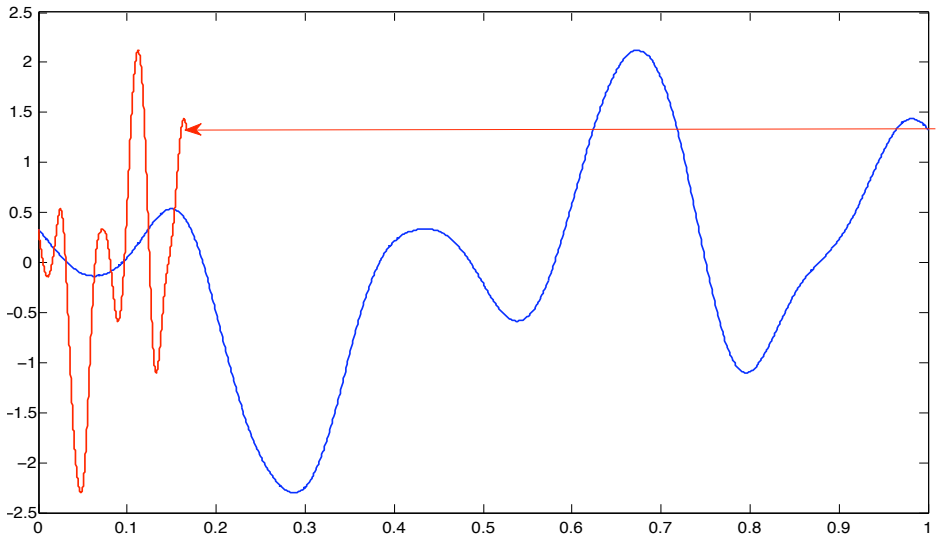
Dans ce cas, [van der Vaart, van Zanten (2010)] montrent qu'à L fixé, la vitesse de convergence associée est (au mieux) logarithmique

$$\varepsilon_n \approx (\log n)^{-\gamma(\beta)}$$

Cependant, ces a priori sont utilisés dans la communauté du 'machine learning' et semblent donner de bons résultats lorsque le paramètre L est "bien calibré" ...

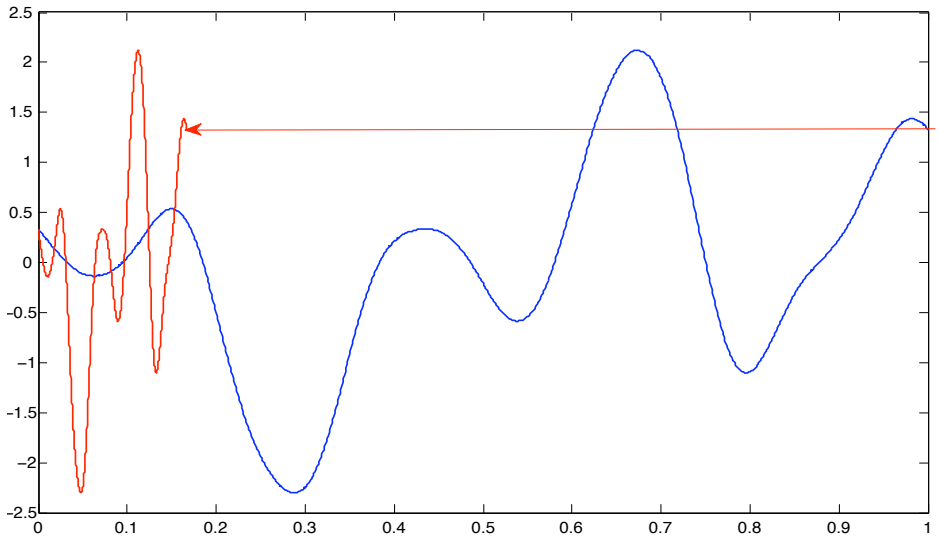
Exemple *Estimation de densité et a priori gaussiens, adaptation*

[van der Vaart, van Zanten (2009)]



Exemple *Estimation de densité et a priori gaussiens, adaptation*

[van der Vaart, van Zanten (2009)]



A priori Π : on considère Z_{At} et on pose $t \longrightarrow \dfrac{e^{Z_{At}}}{\int_0^1 e^{Z_{Au}}\,du}$, où

- A loi gamma
- $u \longrightarrow Z_u$ processus gaussien centré à noyau gaussien

Alors l'a posteriori converge à vitesse minimax (à un log près)

$$\mathbf{E}_{f_0}\Pi(h(f,f_0) \leq M(\log n)^{\gamma(\beta)} n^{-\frac{\beta}{2\beta+1}} \,|\, X^{(n)}) \longrightarrow 1$$

Exemple *Mélanges*

[Rousseau (2009)]

Observations $X_1, \dots, X_n$ i.i.d. densité $f_0 > 0$ et Hölder $\mathcal{C}^\beta[0,1]$

Densités Beta $\quad g(x|a,b) = \frac{x^{a-1}(1-x)^{b-1}}{B(a,b)} \qquad g_{\alpha,\varepsilon}(x) = g(x|\frac{\alpha}{1-\varepsilon}, \frac{\alpha}{\varepsilon})$

paramétrisation habituelle $\qquad$ reparamétrisation

A priori Π = hiérarchie mélange de densités Beta

$$g_{\alpha,k,\mathbf{p}^k,\epsilon k}(\cdot) = \sum_{j=1}^{k} p_j g_{\alpha,\varepsilon_j}(\cdot)$$

Exemple *Mélanges*

[Rousseau (2009)]

Observations $X_1, \dots, X_n$ i.i.d. densité $f_0 > 0$ et Hölder $\mathcal{C}^\beta[0,1]$

Densités Beta $\quad g(x|a,b) = \frac{x^{a-1}(1-x)^{b-1}}{B(a,b)} \qquad g_{\alpha,\varepsilon}(x) = g(x|\frac{\alpha}{1-\varepsilon}, \frac{\alpha}{\varepsilon})$

paramétrisation habituelle $\qquad$ reparamétrisation

A priori Π = hiérarchie mélange de densités Beta

$$g_{\alpha,k,\mathbf{p}^k,\epsilon^k}(\cdot) = \sum_{j=1}^{k} p_j g_{\alpha,\varepsilon_j}(\cdot)$$

- $\alpha \sim \pi_\alpha$
- $k \sim \pi_k$
- $\mathbf{p}^k | \, k \sim (p_1, \dots, p_k)$ loi sur le simplexe canonique de $\mathbb{R}^k$
- $\epsilon^k | \, k \sim (\varepsilon_1, \dots, \varepsilon_k)$ loi dans $(0,1)^k$

Alors l'a posteriori se concentre à vitesse

$$\mathbf{E}_{f_0} \Pi(h(f, f_0) \leq M(\log n)^{\rho(\beta)} n^{-\frac{\beta}{2\beta+1}} | \, X^{(n)}) \to 1$$

Exemple *Sparsité*

Modèle de suite gaussienne

$$X_i = \theta_i + \varepsilon_i, \qquad \varepsilon_i \overset{i.i.d}{\sim} \mathcal{N}(0,1), \qquad 1 \leq i \leq n$$

Signaux sparses $\qquad \ell_0[p_n] = \{\theta \in \mathbb{R}^n, \quad \#\{k,\ \theta_k \neq 0\} \leq p_n\} \qquad (p_n/n \to 0)$

But : estimation de θ pour la norme euclidienne $\|\cdot\|$

Π a priori sur θ : "Seuillage" bayésien au seuil α_n

α_n dans $[0,1]$

g densité sur $\mathbb{R}$ queues au moins exponentielles

$\Pi \sim \bigotimes_{i=1}^n (1-\alpha_n)\delta_0 + \alpha_n g$

Comment choisir α_n ?

Exemple *Sparsité, adaptation*

Π a priori sur θ : Seuillage bayésien hiérarchique

$$\begin{aligned} \alpha &\sim Beta(\kappa n, 1) \\ \Pi|\,\alpha &\sim \bigotimes_{i=1}^{n}(1-\alpha)\delta_0 + \alpha g \end{aligned}$$

[Castillo, van der Vaart (2010)] Pour M assez grand,

$$\sup_{\theta_0 \in \ell_0[p_n]} \mathbf{E}_{\theta_0}\Pi(\|\theta - \theta_0\| > M p_n \log(n/p_n)|\, X) \to 0$$

Exemple *Sparsité, distances ℓ^q*

Considérons l'estimation de $\theta \in \ell_0[p_n]$ pour la d_q-distance, avec $0 < q < 2$.

$$d_q(\theta, \psi) = \sum_{i=1}^{n} |\theta_i - \psi_i|^q.$$

Risque minimax $r^*_{n,q} = O(p_n \log^{q/2}(n/p_n))$

L'a posteriori converge à vitesse minimax pour tout $0 < q < 2$. Pour M assez grand,

$$\mathbf{E}_{\theta_0} \Pi(d_q(\theta, \theta_0) > M r^*_{n,q} | X) \to 0$$

Cela est vrai aussi pour l'a priori seuillage oracle Π^* avec $\alpha_n = p_n/n$

Exemple *Sparsité, distances ℓ^q*

Considérons l'estimation de $\theta \in \ell_0[p_n]$ pour la d_q-distance, avec $0 < q < 2$.

$$d_q(\theta, \psi) = \sum_{i=1}^{n} |\theta_i - \psi_i|^q.$$

Risque minimax $r^*_{n,q} = O(p_n \log^{q/2}(n/p_n))$

L'a posteriori converge à vitesse minimax pour tout $0 < q < 2$. Pour M assez grand,

$$\mathbf{E}_{\theta_0} \Pi(d_q(\theta, \theta_0) > M r^*_{n,q} | X) \to 0$$

Cela est vrai aussi pour l'a priori seuillage oracle Π^* avec $\alpha_n = p_n/n$

Conséquence Posons $\hat{\theta}^*_n = \int \theta \Pi^*(\theta)$ moyenne a posteriori pour Π^*

- [Johnstone, Silverman 2004] avaient constaté que si $q < 1$, la vitesse de convergence de $\hat{\theta}_n$ sous-optimale
- En revanche, la mesure a posteriori converge à vitesse optimale pour tout $0 < q < 2$

Mesure et moyenne a posteriori ont ici
des comportements différents pour $q < 1$

Un exemple inattendu ?

Modèle d'alignement de courbes. Observations : pour $t \in [0,1]$,

$$dY(t) = f(t)dt + dW_1(t),$$
$$dZ(t) = f(t-\theta)dt + dW_2(t).$$

A priori sur (θ, f) $\quad \Pi = \pi_\theta \otimes \pi_f^\sigma$

$$\pi_\theta \sim \text{Unif}([0,1])$$
$$\pi_f^\sigma \sim \sum_{k=1}^{+\infty} \sigma_k \nu_k \varepsilon_k(u) \qquad \text{[A priori gaussien]}$$

- $(\sigma_k)_{k\geq 1}$ suite décroissante de réels ℓ^2
- $(\nu_k)_{k\geq 1}$ i.i.d. $\mathcal{N}(0,1)$
- $\varepsilon_k(\cdot)$ base trigonométrique $\quad [\varepsilon_{2p}(\cdot) = \sqrt{2}\cos(2\pi p \cdot),\ \varepsilon_{2p+1}(\cdot) = \sqrt{2}\sin(2\pi p \cdot)]$

Un exemple inattendu ?

Deux a priori équivalents π_f^σ on f

$$\pi_f^\sigma \quad \sim \quad \sum_{k=1}^{+\infty} \sigma_k \nu_k \varepsilon_k(u)$$

Considérons les deux choix

$$\sigma_k^* = k^{-\frac{1}{2}-\alpha} \quad \leftrightarrow \quad \Pi^* \sim \pi_\theta \otimes {\pi_f^\sigma}^*$$

$$\sigma_k = \begin{cases} (2p)^{-\frac{1}{2}-\alpha} & \text{if } k = 2p \\ (2p)^{-\frac{1}{2}-\alpha} & \text{if } k = 2p+1. \end{cases} \quad \leftrightarrow \quad \Pi \sim \pi_\theta \otimes \pi_f^\sigma$$

Un exemple inattendu ?

Soit $f_0 = f_0^{[\beta]}$ avec $f_{0,k}^{[\beta]} = k^{-1/2-\beta}$.

Proposition

Soit $\alpha = 4$ et $\beta = 2$.

- Quand $n \to +\infty$,

$$\mathbf{E}_{\eta_0} \| \Pi(\cdot \times \mathcal{F} | X) - N(\theta_0 + \frac{\Delta_n}{\sqrt{n}}, \frac{\tilde{I}^{-1}}{n})(\cdot) \| \to 0$$

L'a posteriori $\Pi(\cdot \times \mathcal{F} | X)$ vérifie le théorème BVM

- Pour tout $\delta > 4/9$ et $M > 0$

$$\mathbf{E}_{\eta_0} \Pi^*(|\theta - \theta_0| \leq M n^{-\delta} | X) \to 0$$

L'a posteriori $\Pi^*(\cdot \times \mathcal{F} | X)$ ne converge pas à vitesse $\sqrt{n}$!

Théorie

Théorie : bayésien non-paramétrique

Consistance

- [Doob (1949)] Consistance à un ensemble de Π-mesure nulle près
- [Schwartz (1965)] Consistance sous des hypothèses de tests et de masse a priori suffisante dans des voisinages de Kullback-Leibler
- [Diaconis, Freedman (1986)] Exemple d'a priori "innocent" dans un cadre semi-paramétrique, pour lequel l'a posteriori n'est pas consistant

Vitesses de convergence ?

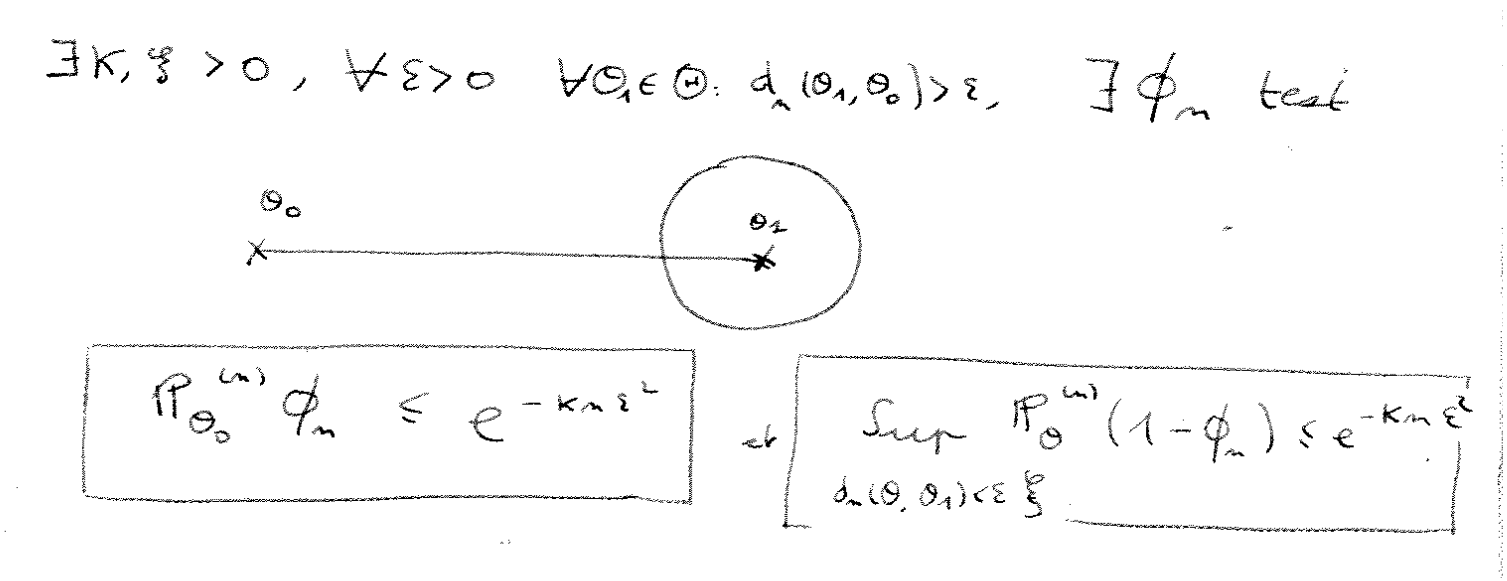
Théorie : bayésien non-paramétrique

Observations $\mathbf{X} = X^{(n)}$, famille de lois $\mathbf{P}_\theta$,
$\theta \in \Theta$ espace de paramètres (NP, SP ...) et Π a priori sur θ

On veut montrer que pour une distance d_n et une vitesse $\varepsilon_n \to 0$,

$$\mathbf{E}_{\theta_0} \Pi(\theta : \ d_n(\theta, \theta_0) > M\varepsilon_n | \mathbf{X}) \to 0$$

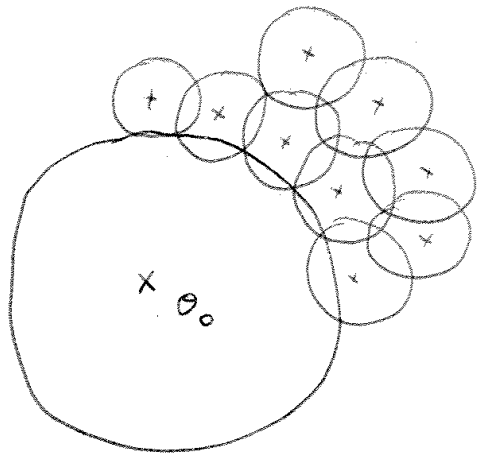
- Tests point/boule pour d_n **(T0)**



[Le Cam 73, 86] et [Birgé 83] : pour données $i.i.d.$, d_n =Hellinger convient

Théorie : bayésien non-paramétrique

- Test vraie valeur/Complémentaire d'une boule **(T1)**



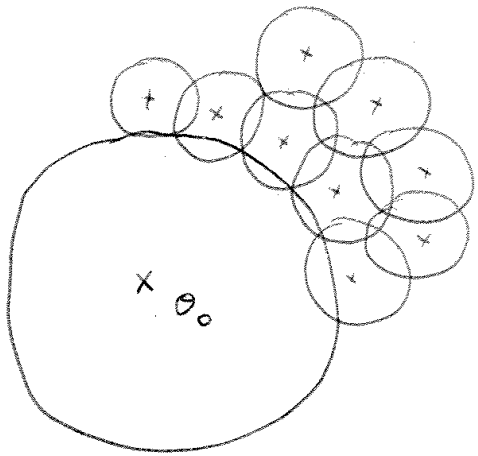
Il existe des tests ψ_n et des sieves Θ_n,

$$\mathbf{P}_{\theta_0}^{(n)}\psi_n \to 0$$

$$\sup_{\theta\in\Theta_n : d_n(\theta,\theta_0)>\varepsilon_n} \mathbf{P}_{\theta}^{(n)}(1-\psi_n) \lesssim e^{-Kn\varepsilon_n^2}$$

Théorie : bayésien non-paramétrique

- Test vraie valeur/Complémentaire d'une boule **(T1)**



Il existe des tests ψ_n et des sieves Θ_n,

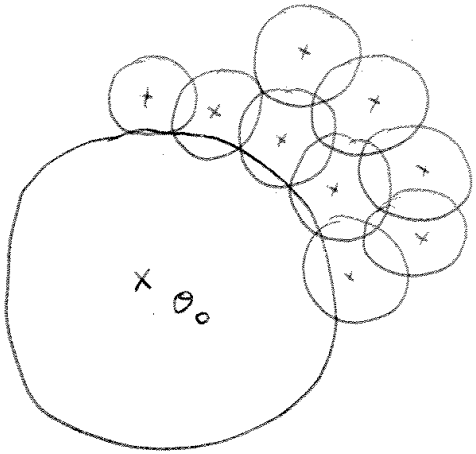
$$\mathbf{P}_{\theta_0}^{(n)}\psi_n \to 0$$

$$\sup_{\theta\in\Theta_n : d_n(\theta,\theta_0)>\varepsilon_n} \mathbf{P}_{\theta}^{(n)}(1-\psi_n) \lesssim e^{-Kn\varepsilon_n^2}$$

Une condition suffisante est un contrôle de l'entropie
$N(\varepsilon, \Theta_n, d_n)$ = nombre minimal de boules de rayon ε pour recouvrir Θ_n

Théorie : bayésien non-paramétrique

- Test vraie valeur/Complémentaire d'une boule **(T1)**



Il existe des tests ψ_n et des sieves Θ_n,

$$\mathbf{P}_{\theta_0}^{(n)}\psi_n \to 0$$

$$\sup_{\theta\in\Theta_n : d_n(\theta,\theta_0)>\varepsilon_n} \mathbf{P}_{\theta}^{(n)}(1-\psi_n) \lesssim e^{-Kn\varepsilon_n^2}$$

- Les sieves Θ_n capturent l'essentiel de la masse a priori

$$\Pi(\Theta\backslash\Theta_n) \leq e^{-(c+4)n\varepsilon_n^2}$$

- L'a priori charge suffisamment des voisinages de θ_0

$$\Pi(B_{KL}(\theta_0,\varepsilon_n)) \geq e^{-cn\varepsilon_n^2}$$

$$B_{KL}(\theta_0,\varepsilon_n) = \left\{ \int p_{\theta_0}^{(n)} \log \frac{p_{\theta_0}^{(n)}}{p_\theta^{(n)}} \leq n\varepsilon_n^2, \quad \int p_{\theta_0}^{(n)} \log^2 \frac{p_{\theta_0}^{(n)}}{p_\theta^{(n)}} \leq n\varepsilon_n^2 \right\}$$

Théorie : bayésien non-paramétrique

[Ghosal, Ghosh, van der Vaart (2000)]

S'il existe des ensembles $\Theta_n \subset \Theta$ et $c > 0$ tels que, pour d_n telle que **(T0)** soit vérifiée, il existe des tests ψ_n avec

$$\mathbf{P}_{\theta_0}^{(n)}\psi_n \to 0, \quad \sup_{\theta \in \Theta_n : d_n(\theta,\theta_0) > \varepsilon_n} \mathbf{P}_{\theta}^{(n)}(1-\psi_n) \lesssim e^{-Kn\varepsilon_n^2} \qquad \textit{tests}$$

$$\Pi(\Theta \backslash \Theta_n) \leq e^{-(c+4)n\varepsilon_n^2} \qquad \textit{masse restante}$$

$$\Pi(B_{KL}(\theta_0, \varepsilon_n)) \geq e^{-cn\varepsilon_n^2} \qquad \textit{masse a priori}$$

Alors pour $M > 0$ assez grand,

$$\mathbb{E}_{\theta_0}\Pi(\theta : d_n(\theta, \theta_0) \leq M\varepsilon_n | X^{(n)}) \to 1$$

Théorie : bayésien non-paramétrique

[Ghosal, Ghosh, van der Vaart (2000)]

S'il existe des ensembles $\Theta_n \subset \Theta$ et $c > 0$ tels que, pour d_n telle que **(T0)** soit vérifiée,

$$\log N(\varepsilon_n, \Theta_n, d_n) \leq n\varepsilon_n^2 \qquad \textit{entropie}$$

$$\Pi(\Theta \backslash \Theta_n) \leq e^{-(c+4)n\varepsilon_n^2} \qquad \textit{masse restante}$$

$$\Pi(B_{KL}(\theta_0, \varepsilon_n)) \geq e^{-cn\varepsilon_n^2} \qquad \textit{masse a priori}$$

Alors pour $M > 0$ assez grand,

$$\mathbb{E}_{\theta_0} \Pi(\theta : d_n(\theta, \theta_0) \leq M\varepsilon_n | X^{(n)}) \to 1$$

Théorie : a priori gaussiens

$W = (W_t : t \in T)$ processus Gaussien centré à valeurs dans $\mathcal{C}^0[0,1]$

Noyau de covariance $K(s,t) = \mathbf{E}(W_s W_t)$

Théorie : a priori gaussiens

$W = (W_t : t \in T)$ processus Gaussien centré à valeurs dans $\mathcal{C}^0[0,1]$

Noyau de covariance $K(s,t) = \mathbf{E}(W_s W_t)$

Noyau auto-reproduisant $\mathbb{H}$ (RKHS) associé à W

On définit une norme $\|\cdot\|_{\mathbb{H}}$ par

$$\langle \sum_{i=1}^{p} a_i K(s_i,\cdot), \sum_{j=1}^{q} b_j K(t_j,\cdot)\rangle_{\mathbb{H}} = \sum_{i,j} a_i b_j K(s_i,t_j)$$

Puis on pose

$$\mathbb{H} = \overline{\mathrm{Vect}\{K(s,\cdot), \quad s \in T\}}^{\mathbb{H}}$$

Exemple Mouvement brownien (B_t) et $\mathbb{B} = (\mathcal{C}^0[0,1], \|\cdot\|_\infty)$,

$$\mathbb{H} = \{ \int_0^{\cdot} f(u)du, \quad f \in L^2[0,1] \}$$

Théorie : a priori gaussiens

Fait : pour tout w dans le support de W dans $\mathbb{B}$, et tout $\varepsilon > 0$,

$$e^{-\varphi_{\boldsymbol{w}}(\varepsilon/2)} \leq \mathbf{P}(\|W - w\| < \varepsilon) \leq e^{-\varphi_{\boldsymbol{w}}(\varepsilon)}$$

Fonction de concentration . Soit $w_0 \in \mathbb{B}$, pour tout $\varepsilon > 0$,

$$\varphi_{w_0}(\varepsilon) = \inf_{h \in \mathbb{H} : \|h - w_0\| < \varepsilon} \frac{\|h\|_{\mathbb{H}}^2}{2} - \log \mathbb{P}(\|W\| < \varepsilon)$$

Terme d'approximation — Probabilité de petite boule

Exemple Mouvement brownien (B_t)

$$-\log \mathbb{P}(\|B\|_\infty < \varepsilon) \approx \varepsilon^{-2} \quad (\varepsilon \to 0)$$

Théorie : a priori gaussiens

[van der Vaart, van Zanten (2008)]

Soit un problème non-paramétrique, fonction inconnue $f_0 \in \mathbb{B}$.
Prior π_f = loi de W processus gaussien sur $\mathbb{B}$, de RKHS $\mathbb{H}$.

Supposons que

- f_0 est dans le support dans $\mathbb{B}$ de l'a priori
- $\|\cdot\|$=norme sur $\mathbb{B}$ *peut se relier* à la distance d du problème statistique considéré

Soit ε_n solution de l'équation

$$\varphi_{f_0}(\varepsilon_n) \leq n\varepsilon_n^2$$

Alors l'a posteriori se concentre à vitesse ε_n : pour M assez grand,

$$\mathbf{E}_0 \pi_f(d(f, f_0) > M\varepsilon_n | X^{(n)}) \to 0$$

Borne inférieure [C. 2008]

Théorie : a priori gaussiens

Ingrédients de preuve :

- Lien entre $\mathbf{P}(\|W - w\| < \varepsilon)$ et fonction de concentration
  masse a priori
- Inégalité de [Borell 75]
  Soient $\mathbb{B}_1$ et $\mathbb{H}_1$ les boules unité de $\mathbb{B}$ et $\mathbb{H}$ associées à W

$$\mathbf{P}(W \notin M\mathbb{H}_1 + \varepsilon\mathbb{B}_1) \leq 1 - \Phi(\Phi^{-1}(e^{-\phi_{\mathbf{0}}(\varepsilon)}) + M)$$

  sieves $\Theta_n = \sqrt{n}\varepsilon_n\mathbb{H}_1 + \varepsilon_n\mathbb{B}_1$
- Lien entropie de $\mathbb{H}_1$ et probabilité de petite boule
  entropie

Conclusion

Conclusion

- L'approche bayésienne est utile pour suggérer des estimateurs
- Permet naturellement d'intégrer des hyperparamètres via des hiérarchies
  cf. adaptation
- Nous avons présenté quelques outils permettant de garantir des propriétés de convergence sous $\mathbf{P}_{\theta_0}$

Perspectives

- Machine learning : de nombreux a priori considérés, pas ou peu de résultats garantissant la convergence de l'a posteriori
- Construction et convergence d'a priori dans des problèmes de grande dimension
  cadres semi-paramétriques, problèmes inverses, sparsité etc.