Un théorème de Bernstein-von Mises semi-paramétrique

Ismaël Castillo (VU Amsterdam)

Rennes,
le 29 août 2008

Introduction et cas paramétrique

Introduction : Cas i.i.d. Paramétrique

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

Θ intervalle ouvert de $\mathbb{R}$ (ou $\mathbb{R}^k$).

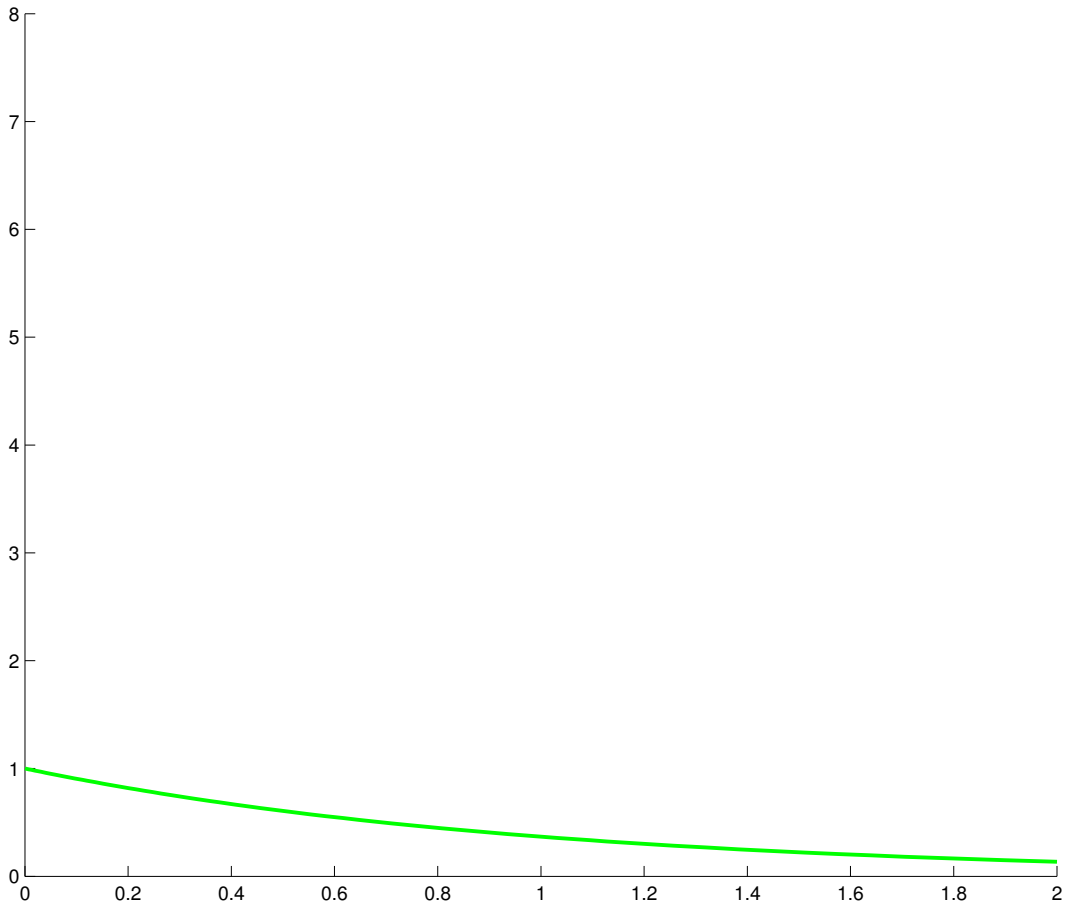
Cadre Bayésien. *A priori* sur θ, $d\pi(\theta) = \lambda(\theta)d\theta$.
Cette mesure est mise à jour avec les données $X^{(n)}$.

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

Formule de Bayes. Pour tout B mesurable,

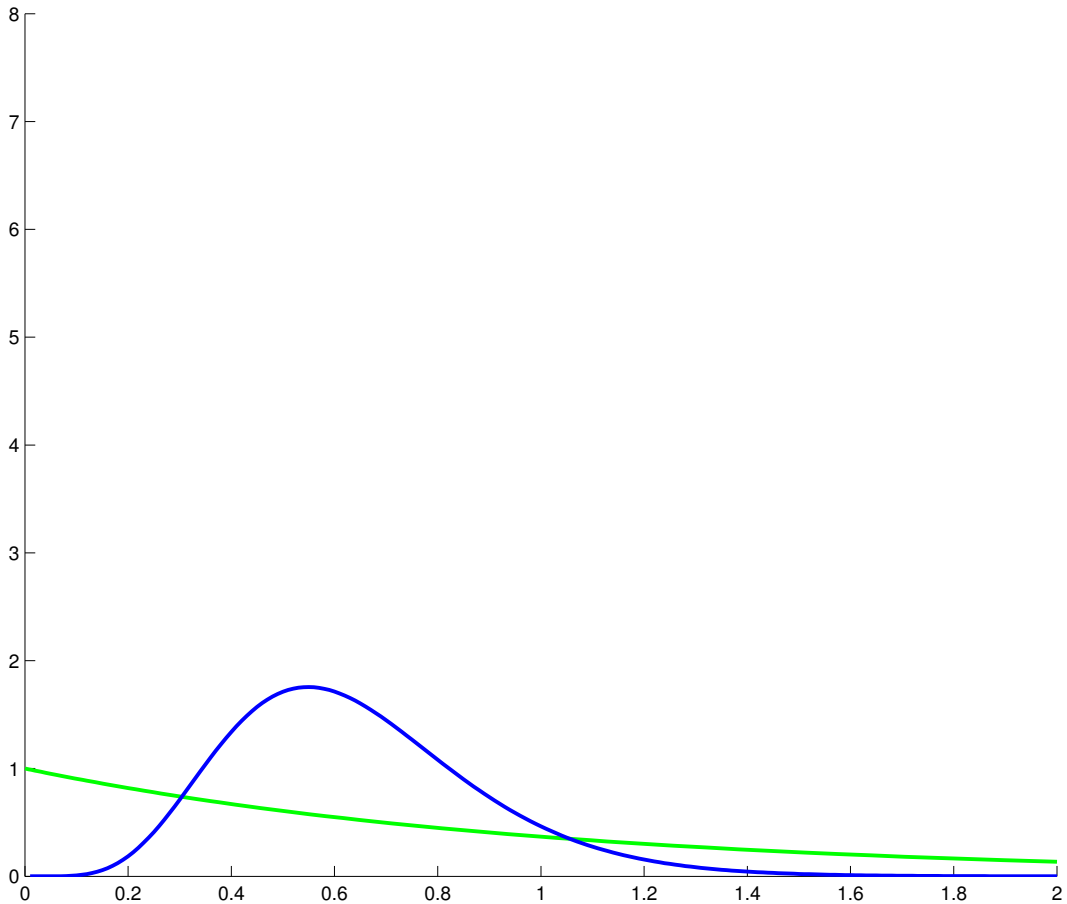
$$\pi(B|X^{(n)}) = \frac{\int_B \prod_{i=1}^n p_\theta(X_i) d\pi(\theta)}{\int \prod_{i=1}^n p_\theta(X_i) d\pi(\theta)}.$$

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



Densité a priori $\mathcal{E}(1)$

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

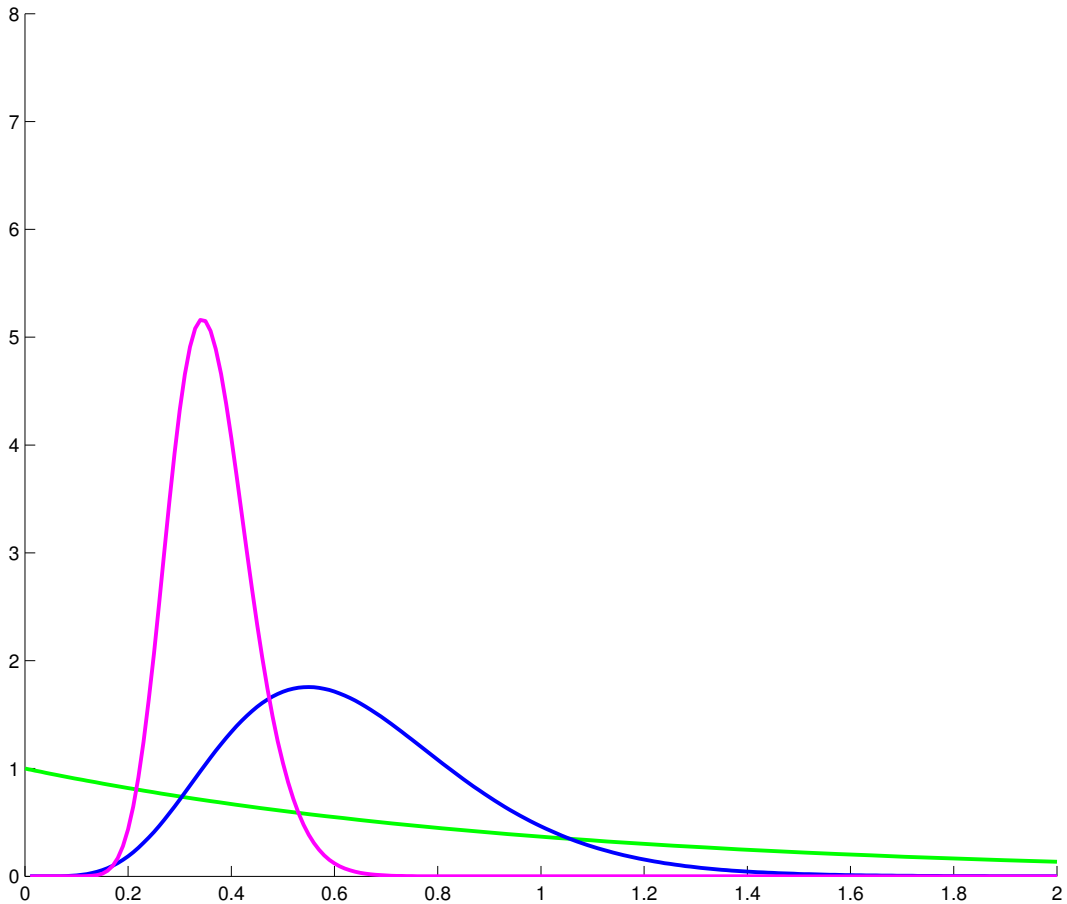


$\theta_0 = 1/3$

Données
X_1, X_2, X_3

Densité a posteriori pour $n = 3$

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

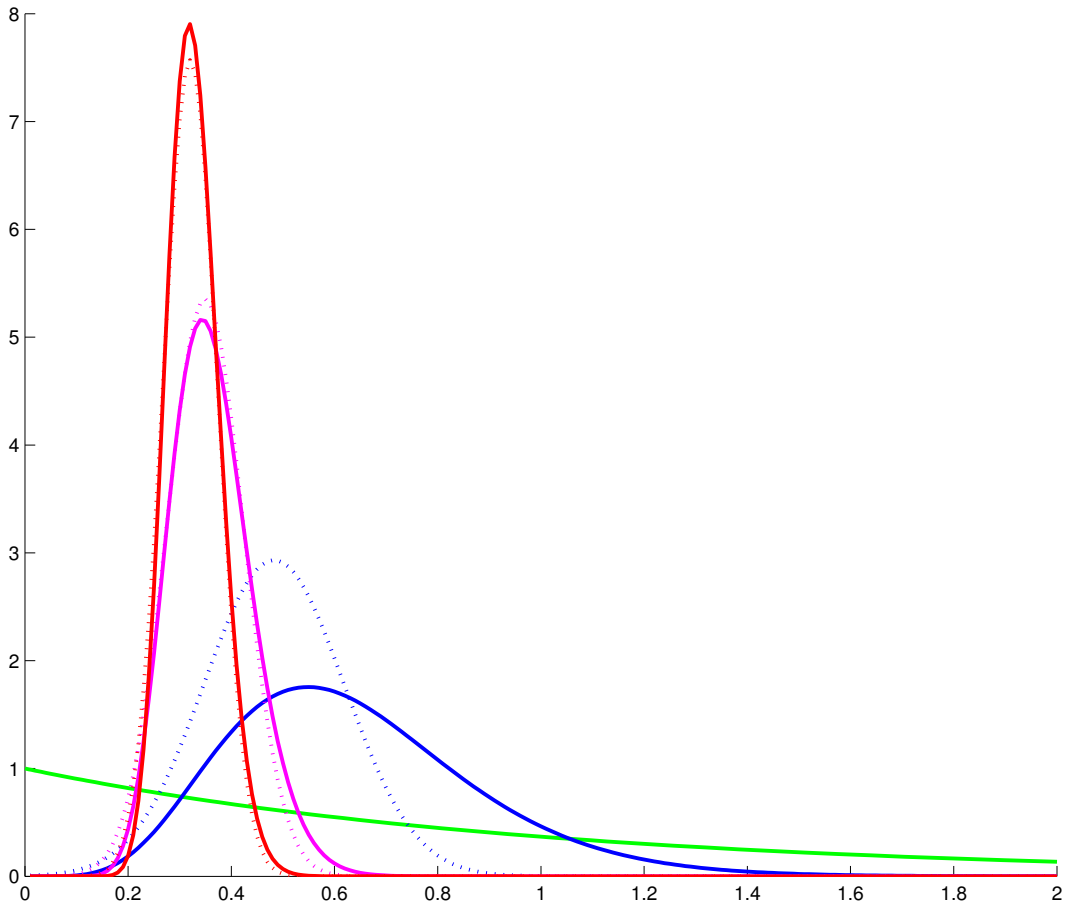


$\theta_0 = 1/3$

Données
$X_1, \ldots, X_{10}$

Densité a posteriori pour $n = 10$

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



$\theta_0 = 1/3$

Données
$X_1, \ldots, X_{20}$

Densité a posteriori pour $n = 20$

Notations et définitions

Vraisemblance $p_\theta^{(n)} \triangleq \prod_{i=1}^n p_\theta(X_i)$

Le modèle est lisse en θ_0 si $\theta \longrightarrow \ell_\theta(x) \triangleq \log p_\theta(x)$ est dérivable deux fois en θ_0 ["dérivable en moyenne quadratique" suffit]

Score $\dot{\ell}_\theta = \frac{\partial}{\partial\theta} \log p_\theta$.

Information de Fisher $\mathcal{I}_{\theta_0} \triangleq \mathbb{E}_{\theta_0} \dot{\ell}^2_{\theta_0}$.

Distance en variation totale entre deux mesures μ et ν sur une tribu $\mathcal{B}$

$$\|\mu - \nu\| \triangleq \sup_{B \in \mathcal{B}} |\mu(B) - \nu(B)|.$$

Théorème de Bernstein - von Mises paramétrique

Théorème ([Le Cam, van der Vaart])

Supposons que

- $\forall \varepsilon > 0$, *il existe une suite de tests* φ_n *tels que*

$$P_{\theta_0}^n \varphi_n \to 0, \qquad \sup_{|\theta - \theta_0| \geq \varepsilon} P_\theta^n (1 - \varphi_n) \to 0.$$

- *modèle lisse en* θ_0 *avec* $\mathcal{I}_{\theta_0} > 0$.

Si l'a priori π *admet une densité continue et positive en* θ_0, *alors*

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

où $\Delta_{n,\theta_0} \triangleq \frac{1}{\sqrt{n}} \sum_{i=1}^n \mathcal{I}_{\theta_0}^{-1} \dot{\ell}_{\theta_0}(X_i)$.

Idées de preuve

La preuve du théorème BVM paramétrique repose sur deux idées

- La "concentration" de l'a posteriori dans des boules de rayon $M/\sqrt{n}$ autour de θ_0 ="consistence" a vitesse $1/\sqrt{n}$.

  Obtenue, par exemple, par une hypothèse Tests

- La normalité asymptotique locale = "forme" du modèle.

  Si le modèle est lisse en θ_0, alors, pour h réel, quand $n \to \infty$,

$$\log \prod_{i=1}^{n} \frac{p_{\theta+h/\sqrt{n}}}{p_\theta}(X_i) = \frac{h}{\sqrt{n}} \sum_{i=1}^{n} \dot{\ell}_{\theta_0}(X_i) - \frac{1}{2}\mathcal{I}_{\theta_0} h^2 + o_{P_{\theta_0}^{(n)}}(1).$$

  Normalité asymptotique locale

Objectifs

Intérêts de ce type de résultat

- Généralité
- Obtention d'intervalles de confiance
- Concordance Bayésien/Fréquentiste

Objectif. Etendre le théorème BVM à un cadre semi-paramétrique d'inconnue (θ, f). Il faut notamment construire un a priori sur la partie non-paramétrique.

- Comprendre le rôle et l'influence de l'a priori non-paramétrique
- Lien avec la théorie Bayésienne Non-paramétrique

Bibliographie.
Bayes NP: [Ghosal, Ghosh, van der Vaart 2000], [Shen, Wasserman 2001]
BVM: [Kim, Lee 2002], [Kim 2004], [Shen 2002]

Cadre semi-paramétrique

Cadre semi-paramétrique

Modèle statistique $(\mathcal{X}^{(n)}, \mathcal{G}^{(n)}, \mathbf{P}^{(n)}_{\theta,f})$. Données $X^{(n)}$.
Les paramètres inconnus sont le couple $(\theta, f) = \eta$.

Espace des paramètres $A = \Theta \times \mathcal{F}$.
Θ intervalle ouvert de $\mathbb{R}$ et $\mathcal{F}$ inclus dans un espace fonctionnel (ex. : $L^2[0,1]$, $\mathcal{C}^0[0,1]$, ...)

Mesure dominante $d\mathbf{P}^{(n)}_{\theta,f} = p^{(n)}_{\theta,f} d\mu^{(n)}$.

A Priori $\Pi = \pi_\theta \otimes \pi_f$ sur A.

Formule de Bayes. La probabilité a posteriori de $C = C_1 \times C_2$ est

$$\Pi(C_1 \times C_2 | X^{(n)}) = \frac{\int_{C_1 \times C_2} p^{(n)}_{\theta,f}(X^{(n)}) d\Pi(\theta, f)}{\int p^{(n)}_{\theta,f}(X^{(n)}) d\Pi(\theta, f)}.$$

Cadre semi-paramétrique

Modèle statistique $(\mathcal{X}^{(n)}, \mathcal{G}^{(n)}, \mathbf{P}^{(n)}_{\theta,f})$. Données $X^{(n)}$.
Les paramètres inconnus sont le couple $(\theta, f) = \eta$.

Espace des paramètres $A = \Theta \times \mathcal{F}$.
Θ intervalle ouvert de $\mathbb{R}$ et $\mathcal{F}$ inclus dans un espace fonctionnel (ex. : $L^2[0,1]$, $\mathcal{C}^0[0,1]$, ...)

Mesure dominante $d\mathbf{P}^{(n)}_{\theta,f} = p^{(n)}_{\theta,f} d\mu^{(n)}$.

A Priori $\Pi = \pi_\theta \otimes \pi_f$ sur A.

Formule de Bayes. La marginale en θ a posteriori de $B \subset \Theta$ est

$$\Pi(B \times \mathcal{F} | X^{(n)}) = \frac{\int_B \int p^{(n)}_{\theta,f}(X^{(n)}) d\pi_f(f) d\pi_\theta(\theta)}{\int \int p^{(n)}_{\theta,f}(X^{(n)}) d\pi_f(f) d\pi_\theta(\theta)}.$$

Exemple : paramètre de translation en bruit blanc

$$dX^{(n)}(t) = f(t-\theta)dt + \frac{1}{\sqrt{n}}dB(t)\ , \quad t \in [-1/2, 1/2].$$

- Cadre semi-paramétrique
  - θ réel positif à estimer.
  - f fonction inconnue dans L^2 1-périodique symétrique.
  - $B(t)$ mouvement brownien standard.
- Cadre *Asymptotique* : $n \to +\infty$.

Propriété LAN et cadre SP simplifié

Le modèle suit une propriété *LAN linéaire* en $\eta_0 = (\theta_0, f_0)$, s'il existe un produit scalaire $\langle \cdot\,, \cdot \rangle_L$ sur $\mathbb{R} \times H \supset \Theta \times \mathcal{F}$ tel que

$$\log \frac{d\mathbf{P}^{(n)}_{\theta,f}}{d\mathbf{P}^{(n)}_{\theta_0,f_0}} = -n\|\theta - \theta_0, f - f_0\|_L^2/2 + \sqrt{n} W_n(\theta - \theta_0, f - f_0) + o_{P^{(n)}_{\eta_0}}(1),$$

où $W_n : \mathbb{R} \times H \to \mathbb{R}$ linéaire et pour $v_1, \ldots, v_d \in \mathbb{R} \times H$,

$$(W_n(v_1), \ldots, W_n(v_d)) \rightsquigarrow N(0, \{\langle v_i\,,\, v_j \rangle_L\}_{i,j}).$$

Paramètres locaux. $h = \sqrt{n}(\theta - \theta_0)$, $a = \sqrt{n}(f - f_0)$.

Information efficace

Soit $\Pi_{\perp}^{\overline{\mathcal{F}}}$ la projection orthogonale sur $\overline{\mathcal{F}}$.
Soit $\gamma(\cdot) = \Pi_{\perp}^{\overline{\mathcal{F}}}(1,0)$. Alors $\Pi_{\perp}^{\overline{\mathcal{F}}}(h,a) = h\gamma(\cdot) + a(\cdot)$. Puis

$$\begin{aligned} \|h,a\|_L^2 &= \|h,-h\gamma(\cdot)\|_L^2 + \|0,h\gamma(\cdot)+a(\cdot)\|_L^2 \\ &= \{\underbrace{\|1,0\|_L^2 - \|0,\gamma\|_L^2}_{\triangleq \tilde{\mathcal{I}}_{\eta_0} \leq \mathcal{I}_{\eta_0}}\}h^2 + \|0,h\gamma+a\|_L^2. \end{aligned}$$

Information efficace $\tilde{\mathcal{I}}_{\eta_0} = \|1,0\|_L^2 - \|0,\gamma(\cdot)\|_L^2$.

Il n'y a pas de perte d'information si $\tilde{\mathcal{I}}_{\eta_0} = \mathcal{I}_{\theta_0} = \|1,0\|_L^2$ ou encore si $\gamma(\cdot) = 0 \Leftrightarrow \|h,a\|_L^2 = \|h,0\|_L^2 + \|0,a\|_L^2$.

LAN pour le modèle de translation

$$dX^{(n)}(t) = f(t-\theta)dt + \frac{1}{\sqrt{n}}dB(t)\ , \quad t \in [-1/2, 1/2].$$

Vraisemblance par rapport à la mesure dominante $B/\sqrt{n}$

$$d\mathbf{P}^{(n)}_{\theta,f}/d\mathbf{P}_{B/\sqrt{n}} = \exp\left\{ n\int_{-1/2}^{1/2} f(u-\theta)dX^{(n)}(u) - \frac{n}{2}\int_{-1/2}^{1/2} f(u-\theta)^2 du \right\}.$$

$$\log\left(d\mathbf{P}^{(n)}_{\theta,f}/d\mathbf{P}^{(n)}_{\theta_0,f_0}\right) = -\|(h,a)\|_L^2/2 + W(h,a) + R_n(h,a), \quad \text{où}$$

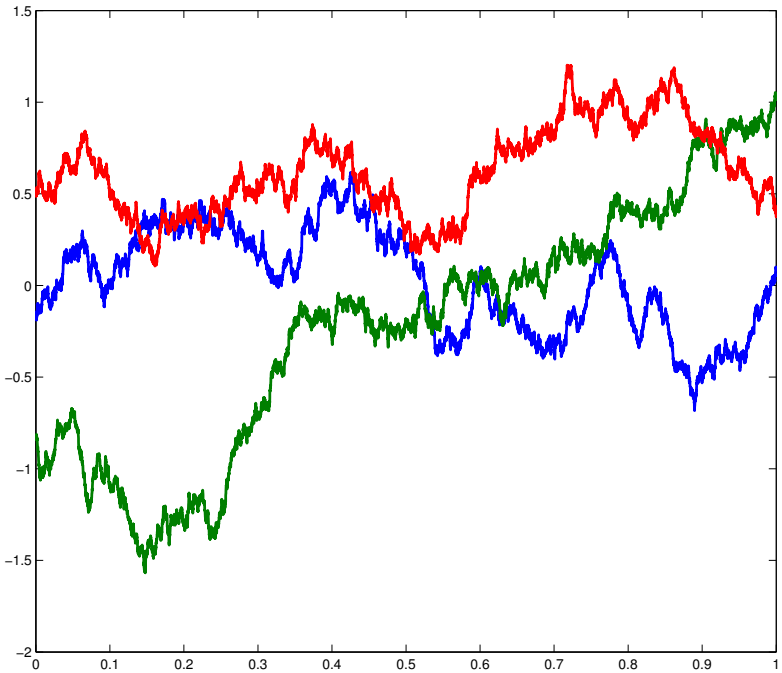
$$h = \sqrt{n}(\theta-\theta_0),\ a = \sqrt{n}(f-f_0)$$

$$\begin{aligned} \|(h,a)\|_L^2 &= h^2\int_{-1/2}^{1/2} f_0'(u)^2 du + \int_{-1/2}^{1/2} a(u)^2 du \\ W(h,a) &= \int_{-1/2}^{1/2} \left\{-hf_0'(t-\theta_0) + a(t-\theta_0)\right\} dB(t) \end{aligned}$$

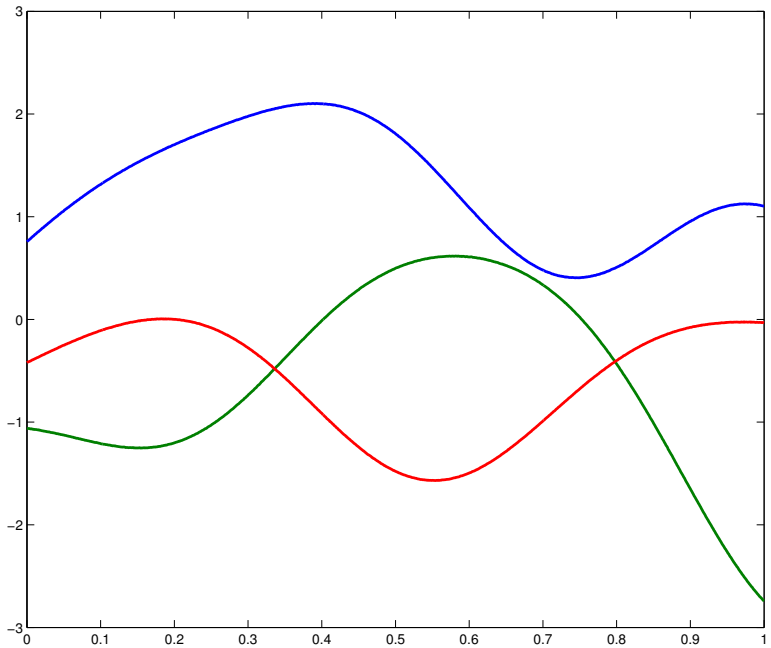
$(1,0)$ est orthogonal à $\{0\}\times\mathcal{F}$. Pas de perte d'information.

A priori non-paramétriques Gaussiens

Bayésien non-paramétrique : exemples d'a priori



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



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

Bayésien non-paramétrique : A priori Gaussiens

L'a priori π_f est la loi de W processus Gaussien centré à valeurs dans $(\mathbb{B}, \|\cdot\|)$ Banach séparable. Notons $K(s,t) = \mathbf{E}(W_s W_t)$ noyau de covariance.

Soit $(\mathbb{H}, \|\cdot\|_{\mathbb{H}})$ *noyau auto-reproduisant* (RKHS) de W dans $\mathbb{B}$

$\mathbb{H}$ est l'adhérence de l'ensemble des combinaisons linéaires $\sum_{i=1}^{p} a_i K(s_i, \cdot)$ pour le produit scalaire $\langle \sum_{i=1}^{p} a_i K(s_i, \cdot), \sum_{j=1}^{q} b_j K(t_j, \cdot)\rangle_{\mathbb{H}} = \sum_{i,j} a_i b_j K(s_i, t_j)$.

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

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

Bayésien non-paramétrique : A priori Gaussiens

[van der Vaart et van Zanten (2008)] :

Soit un problème non-paramétrique où l'inconnue est une fonction $f_0 \in \mathbb{B}$.
A priori π_f associé au processus Gaussien W sur $\mathbb{B}$, de RKHS $\mathbb{H}$.

Si f_0 est dans le support de l'a priori W et que ε_n vérifie

$$\varphi_{w_0}(\varepsilon_n) \leq n\varepsilon_n^2.$$

Alors, souvent, on est capable de montrer un résultat du type

$$\pi_f(d(f, f_0) > M\varepsilon_n | X^{(n)}) \to 0 \quad (n \to +\infty),$$

pour M assez grand, en $\mathbf{P}_{f_0}^{(n)}$-probabilité, pour une distance d sur $\mathcal{F}$.

Bayésien non-paramétrique : Exemples

Soit $\mathbb{B} = (\mathcal{C}^0[0,1], \|\cdot\|_\infty)$.
Supposons w_0 dans la classe de Hölder $\mathcal{C}^\beta[0,1]$.

Mouvement brownien + Gaussienne,
$W_t = B_t + Z_0$, où $Z_0 \sim \mathcal{N}(0,1)$. Alors

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

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

$$Y_t = \int_0^t (t-s)^{\alpha - 1/2} dB_s.$$

Si $W_t = Y_t + \sum_{k=0}^{\lceil \alpha \rceil} Z_k t^k$, où Z_k *i.i.d.* $\sim \mathcal{N}(0,1)$, alors

$$\varepsilon_n \approx n^{-\frac{\alpha \wedge \beta}{2\alpha + 1}}$$

Théorème de Bernstein - von Mises semi-paramétrique

Théorème BVM SP : Cas $\tilde{\mathcal{I}} = \mathcal{I}$, Hypothèses

Soit $\Pi = \pi_\theta \otimes \pi_f$ un a priori sur $\Theta \times \mathcal{F}$, supposons

$$(\mathbf{P}) \quad d\pi_\theta = \lambda(\theta)d\theta, \text{avec } \lambda \text{ continue et positive en } \theta_0.$$

Soient $\varepsilon_n \to 0$ et $\mathcal{F}_n \subset \mathcal{F}$ mesurables ("sieves")

$$\begin{aligned} A_n &= \{(\theta, f) \in \Theta \times \mathcal{F}_n, \quad \|\theta - \theta_0, f - f_0\|_L \leq \varepsilon_n\}. \\ A_n^{\theta=\theta_0} &= \{f \in \mathcal{F}_n, \quad \|0, f - f_0\|_L \leq \varepsilon_n/2\}. \end{aligned}$$

Supposons que

$$(\mathbf{C}) \quad \begin{aligned} &\Pi(A_n|X^{(n)}) \to 1 \qquad \text{en } \mathbf{P}_{\eta_0}^{(n)}\text{-probabilité,} \\ &\Pi^{\theta=\theta_0}(A_n^{\theta=\theta_0}|X^{(n)}) \to 1 \qquad \text{en } \mathbf{P}_{\eta_0}^{(n)}\text{-probabilité.} \end{aligned}$$

où

$$\Pi^{\theta=\theta_0}(D|X^{(n)}) = \frac{\int_D p_{\theta_0,f}(X^{(n)})d\pi_f(f)}{\int p_{\theta_0,f}(X^{(n)})d\pi_f(f)}$$

Théorème BVM SP : Cas $\tilde{\mathcal{I}} = \mathcal{I}$, Hypothèses

Supposons que le modèle vérifie la propriété LAN linéaire en η_0.

Définissons le *terme de reste* du développement LAN par

$$R_n(\theta, f) \triangleq \log \frac{d\mathbf{P}^{(n)}_{\theta,f}}{d\mathbf{P}^{(n)}_{\theta_0,f_0}} - \left\{ n\|\theta - \theta_0, f - f_0\|_L^2/2 - \sqrt{n} W_n(\theta - \theta_0, f - f_0) \right\}.$$

Soit $A_n^{(2)} = \{(\theta, f) \in \Theta \times \mathcal{F}_n, \quad \|\theta - \theta_0, f - f_0\|_L \leq 2\varepsilon_n\}$, supposons

$$(\mathbf{N}) \qquad \sup_{(\theta,f)\in A_n^{(2)}} \frac{|R_n(\theta, f) - R_n(\theta_0, f)|}{1 + n(\theta - \theta_0)^2} = o_{P^{(n)}_{\eta_0}}(1).$$

Théorème BVM SP : Cas $\tilde{\mathcal{I}} = \mathcal{I}$, Enoncé

Dans le cas sans perte d'information, $\tilde{\mathcal{I}}_{\eta_0} = \|1, 0\|_L^2$.
Si $\tilde{\mathcal{I}}_{\eta_0} > 0$, notons $\Delta_{n,\eta_0} \triangleq \tilde{\mathcal{I}}_{\eta_0}^{-1} W_n(1, 0)$.

Théorème

Supposons qu'il n'y ait pas perte d'information et que $\tilde{\mathcal{I}}_{\eta_0} > 0$. *Si* **(P)**, **(C)**, **(N)** *sont vérifiées, alors pour tout* $B \subset \Theta$ *mesurable,*

$$\sup_B \left| \Pi(B \times \mathcal{F} | X^{(n)}) - N\left(\theta_0 + \frac{\Delta_{n,\eta_0}}{\sqrt{n}}, \frac{\tilde{\mathcal{I}}_{\eta_0}^{-1}}{n}\right)(B) \right| \to 0,$$

quand $n \to +\infty$, *en* $\mathbf{P}_{\eta_0}^{(n)}$*-probabilité.*

Remarque : la partie π_f de l'a priori Π n'est pas supposée gaussienne.

Conditions (C)-(N), discussion

- Hypothèse **(C)**
  - On utilise les techniques de [Ghosal, Ghosh, van der Vaart 00] et [van der Vaart, van Zanten 08]
  - On obtient des résultats en termes d'une distance d_T pour laquelle certains tests existent
    L'a posteriori se concentre alors à vitesse ε_n pour d_T.
  - Ensuite il faut relier d_T et $\|\cdot\|_L$.
- Hypothèse **(N)**.
  - Techniques de suprema de processus (empiriques)
  - Le supremum porte sur le sieve $\mathcal{F}_n$
    Pour les a priori Gaussiens, souvent on prend
    $\mathcal{F}_n = \alpha_n \mathbb{B}_1 + \sqrt{n}\alpha_n \mathbb{H}_1$.

Application : modèle de bruit blanc

$$dX^{(n)}(t) = f(t-\theta)dt + \frac{1}{\sqrt{n}}dB(t)\ , \quad t \in [-1/2, 1/2].$$

- $\mathcal{F} = \{f \in L^2[0,1],\ f$ symétrique, 1-périodique, $\int_0^1 f = 0\}$.

  $f_k \triangleq \sqrt{2}\int_0^1 f(u)\cos(2\pi k u)du$ pour $k \geq 1$.

  $\Theta = [-\tau_0, \tau_0] \subset]-1/4, 1/4[$.

- Conditions de regularité sur f_0

  Il existe ρ, L tels que

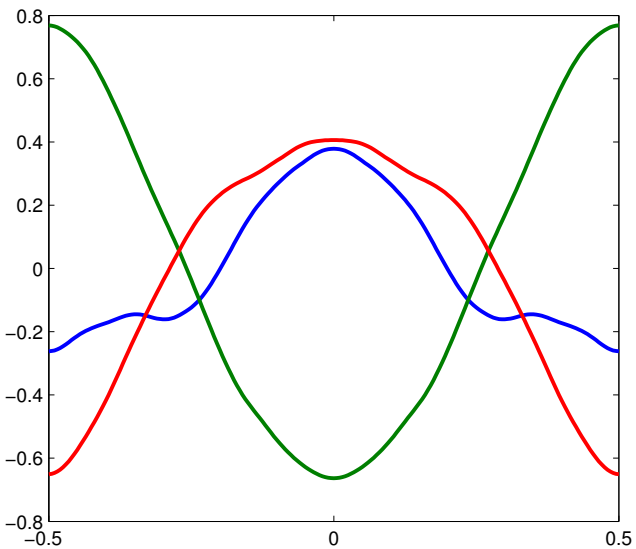
  - $|f_1| \geq \rho > 0$.
  - $\sum_{k\geq 1} k^{2\beta} f_k^2 \leq L^2$ pour $\beta > 1$.

Application : modèle de bruit blanc

A priori $\Pi = \pi_\theta \otimes \pi_f$ où π_θ vérifie **(P)**.

Pour π_f nous prenons des a priori Gaussiens de la forme suivante. Soient $\alpha_k \sim N(0,1)$ i.i.d. et $\alpha > 0$. On pose

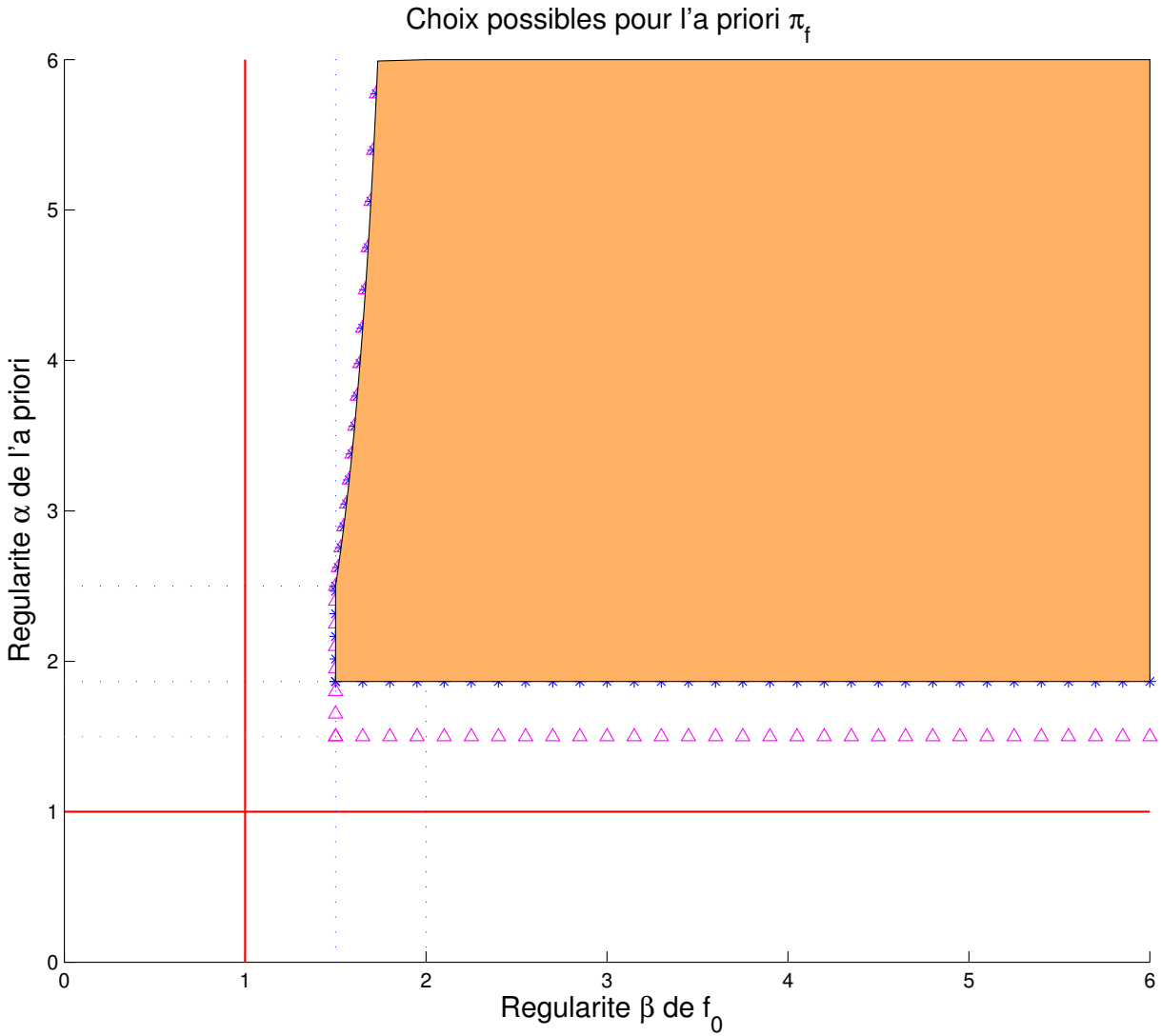
$$Z_{t,\alpha} = \sum_{k=1}^{+\infty} \frac{\alpha_k}{k^{\frac{1}{2}+\alpha}} \cos(2\pi k t)$$



α s'interprète comme la "regularité" de l'a priori.

Application : modèle de bruit blanc

Couples α, β pour lesquels les conditions du Théorème sont vérifiées :



Cas $\tilde{\mathcal{I}} < \mathcal{I}$ avec a priori Gaussiens

Lorsqu'il y a perte d'information,

$$n\|\theta - \theta_0, f - f_0\|_L^2 = n\tilde{\mathcal{I}}_{\eta_0}(\theta - \theta_0)^2 + n\|0, f - f_0 + (\theta - \theta_0)\gamma\|_L^2$$

et γ est non nulle.

Idée : si W processus Gaussien de RKHS $\mathbb{H}$, et si $h \in \mathbb{H}$ alors la mesure dP^{W+h} est absolument continue par rapport à dP^W, donc on peut envisager un *changement de variables*.

On pose alors $g = f + (\theta - \theta_0)\gamma$ (ou éventuellement $g = f + (\theta - \theta_0)\gamma_n$ si $\gamma \notin \mathbb{H}$).

On obtient alors un théorème dans le même esprit que le précédent, mais limité au cas d'a priori Gaussiens

Application: modèle de Cox avec a priori Gaussiens.

Conclusion

- Notre approche met en évidence les hypothèses de concentration de l'a posteriori à vitesse ε_n et de forme gaussienne locale = généralisation naturelle des hypothèses paramétriques.

- Pour l'application à des modèles particuliers, les limitations :
  - Choix du sieve parfois délicat pour vérifier **(N)**.
  - La vitesse ε_n doit parfois être suffisamment rapide (par ex. au moins $n^{-1/4}$). Cela impose parfois des conditions de régularité sur la fonction f.

- Perspectives futures
  - BVM : cas non Gaussien pour $0 < \tilde{\mathcal{I}}_{\eta_0} < \mathcal{I}_{\eta_0}$.
  - BVM : cadre plus géneral que le LAN linéaire avec "courbure".
  - BVM : estimation de fonctionnelles
  - Bayes NP : compréhension des vitesses de cv. pour a priori non Gaussiens